

Université de Strasbourg

Amélioration de la classification supervisée par agrégation d'arbres de décision



Iris Versicolor



Iris Setosa



Iris Virginica

L'équipe des Fleurs :

Auteurs : Darren Delgado, Hassan El-Sahily, Marina Galstyan

Encadrante : Anne Jeannin-Girardon

Date : 21 décembre 2025

Projet réalisé dans le cadre de l'UE Intelligence Artificielle – Année universitaire 2024-2025

Table des matières

1	Introduction	2
2	Méthode	2
2.1	Données et Construction d'un arbre de décision	2
2.1.1	Préparation des données	2
2.1.2	Construction de l'arbre	2
2.1.3	Prédiction d'une instance : exemple détaillé	3
2.2	Méthode d'échantillonnage (Bootstrap)	3
2.3	Construction des ensembles ($t = 1, 4, 8$)	4
2.3.1	Différences entre les arbres	4
2.4	Agrégation des prédictions	5
2.5	Évaluation	5
2.5.1	Métriques utilisées	5
3	Résultats et discussions	5
3.1	Évolution des performances sur les ensembles selon le nombre d'arbres	5
3.1.1	Arbre unique($t = 1$)	5
3.1.2	Ensemble de 4 arbres ($t = 4$)	6
3.1.3	Ensemble de 8 arbres ($t = 8$)	6
3.1.4	Récapitulatif en terme de performance	7
3.2	Études de cas	7
3.2.1	Exemple avec PRÉDICTIONS CORRECTES	7
3.2.2	Exemple avec PRÉDICTIONS INCORRECTES	8
3.2.3	Interpretation	8
3.3	Diversité des arbres et impact sur les performances	9
4	Conclusion	10

1 Introduction

Les arbres de décision sont un type de technique d'apprentissage supervisé qui est fréquemment utilisé pour des tâches de classification et parfois de régression. Leur popularité découle de leur facilité d'utilisation et de leur interprétabilité; ils créent une hiérarchie de règles de décision basées sur les attributs des données, offrant ainsi une transparence dans le processus de prédiction. Contrairement à des modèles plus complexes comme les réseaux de neurones, les arbres de décision sont facilement observables et explicables, ce qui les rend particulièrement utiles dans des domaines où la justification des prédictions est essentielle (comme la médecine, la finance ou le diagnostic industriel).

Cependant, cette simplicité a ses limites. Les arbres de décision sont sensibles aux changements dans les données sous-jacentes : même une petite modification peut entraîner une structure radicalement différente. Leur robustesse peut être compromise par cette instabilité, qui résulte de leur nature gourmande (gloutonne) lors de la construction (en optimisant les critères de séparation locaux à chaque nœud). Par conséquent, bien qu'ils soient efficaces pour des jeux de données simples et bien structurés, ils ont tendance à surajuster (overfit) des données bruyantes ou peu représentatives.

Dans ce contexte, notre objectif est d'explorer dans quelle mesure la combinaison de plusieurs arbres de décision, construits indépendamment via des échantillonnages aléatoires, peut améliorer la performance d'un modèle de classification.

Pour répondre à ces questions, ce rapport est structuré comme suit : la première partie décrit la méthodologie mise en œuvre (de la construction des arbres aux méthodes d'échantillonnage et d'agrégation). La seconde partie présente et discute les résultats obtenus selon différents critères d'évaluation. Enfin, nous concluons par une synthèse des apports de l'agrégation, ainsi que ses limites et perspectives d'amélioration.

2 Méthode

Pour débiter cette exploration, nous présentons ci-dessous la méthodologie adoptée, en commençant par la préparation des données, suivie de la mise en œuvre d'un arbre de décision de référence. Cette première étape constitue une base solide pour les expérimentations ultérieures impliquant des ensembles d'arbres.

2.1 Données et Construction d'un arbre de décision

2.1.1 Préparation des données

L'étude utilise le dataset Iris étendu, contenant des mesures morphologiques de fleurs d'iris (longueur/largeur des pétales et sépales) ainsi que des informations supplémentaires comme le type de sol. Les données sont prétraitées pour encoder les variables catégorielles (ex : soiltype) en variables numériques via one-hot encoding. L'ensemble est divisé en :

- 80% pour l'entraînement (construction des arbres)
- 20% pour le test (évaluation finale)

Cette séparation aléatoire garantit une bonne représentativité des classes dans les deux sous-ensembles.

2.1.2 Construction de l'arbre

Un arbre de décision binaire, d'une profondeur maximale de 10, a été construit à partir des données d'entraînement. À chaque nœud de l'arbre, la variable choisie pour effectuer la

séparation est celle qui maximise la réduction de l'entropie, en utilisant comme seuil de split la médiane de cette variable. Cette architecture permet un compromis entre complexité du modèle et risque de sur-apprentissage. La classe `DecisionTree` gère la structure (nœuds, branches, feuilles).

Fonctions clés :

- `entropy()` : Calcule l'impureté des données.
- `attribute_gain_median()` : Évalue le gain d'un attribut via split médian.
- `best_attribute()` : Identifie l'attribut optimal pour séparer les données.

2.1.3 Prédiction d'une instance : exemple détaillé

Pour illustrer le fonctionnement d'un arbre unique, nous avons implémenté une fonction `item1(tree, test_data)` permettant de sélectionner une instance aléatoire dans le jeu de test, de prédire sa classe, d'afficher la confiance du modèle, ainsi que le chemin complet parcouru dans l'arbre.

Voici un exemple de sortie réelle générée par la fonction : Dans ce cas, l'instance testée

```
----- DEBUT: ITEM 1 -----

=== Nouvelle prédiction ===
Instance réelle : versicolor

At petal_area (split value 5.48105): 6.347999999999999 > 5.48105
> Going right
  At petal_curvature_mm (split value 10.72): 8.61 <= 10.72
  > Going left
    At petal_texture_trichomes_per_mm2 (split value 13.65): 15.72 > 13.65
    > Going right
      At petal_length (split value 4.505): 5.52 > 4.505
      > Going right
        At elevation (split value 124.95): 130.9 > 124.95
        > Going right
          At sepal_width (split value 2.785): 2.69 <= 2.785
          > Going left
            Reached leaf node. Predicted class: versicolor (Proportion: 1.00)

Classe prédite : versicolor
Confiance : 1.00

----- FIN: ITEM 1 -----
```

FIGURE 1 – Illustration de l'arbre de décision utilisé.

appartient réellement à la classe `versicolor`. L'arbre a suivi 7 niveaux de décisions basées sur les features pour arriver à une prédiction correcte, avec une confiance de 100% (la feuille terminale contient exclusivement des exemples `versicolor` dans l'ensemble d'entraînement).

Cette prédiction/item illustre déjà premièrement le rôle de chaque features dans la prise de décision, celle effectivement faite et la capacité explicative du modèle (i.e chaque décision est justifiable).

2.2 Méthode d'échantillonnage (Bootstrap)

Notre méthode repose sur un échantillonnage bootstrap avec remise, implémenté via la fonction `sample_with_replacement()`. Cette approche suit trois étapes clés :

1. Tirage avec remise :

Pour chaque arbre de l'ensemble, nous tirons aléatoirement N instances avec remise depuis le jeu d'entraînement original (où N est la taille du jeu d'entraînement). Ce processus garantit que chaque arbre est entraîné sur un sous-ensemble légèrement différent des données.

2. Création des jeux de validation :

Lors de l'échantillonnage avec remise effectué pour construire chaque modèle de l'ensemble, certaines observations ne sont pas sélectionnées. Ces données peuvent alors être utilisées pour évaluer la performance du modèle correspondant sur des exemples qu'il n'a pas vus à l'entraînement. Cette méthode d'évaluation permet une validation croisée sans avoir besoin de réserver un jeu de données distinct.

3. Avons-nous des cas de sur-apprentissage ?

Sur un seul arbre, l'exactitude d'entraînement est de 1.00, tandis que celle de test est de 0.975. La faible différence (0.025) entre ces deux mesures montre que le modèle s'adapte bien aux données sans signe évident de sur-apprentissage.

Sur un ensemble de 8 arbres, les exactitudes d'entraînement sont :

[0.991, 0.995, 0.99, 0.986, 0.991, 0.988, 0.988, 0.989]

et celles de test :

[0.964, 0.967, 0.98, 0.963, 0.979, 0.967, 0.971, 0.963]

Les écarts entre entraînement et test :

[0.026, 0.028, 0.01, 0.024, 0.012, 0.021, 0.017, 0.026]

montrent que la généralisation reste bonne pour chaque modèle de l'ensemble. Cela illustre que l'agrégation de plusieurs arbres permet d'obtenir des résultats fiables et consistants, sans sur-ajustement visible aux données d'entraînement.

2.3 Construction des ensembles ($t = 1, 4, 8$)

Nous construisons trois configurations distinctes comprenant respectivement $t = 1$, $t = 4$ et $t = 8$ arbres. Chaque configuration repose sur t échantillonnages avec remise, afin que chaque arbre soit entraîné sur un sous-ensemble différent des données.

2.3.1 Différences entre les arbres

Cette méthode permet à chaque arbre d'un ensemble de voir des exemples légèrement différents. Les variations observées entre les arbres viennent des éléments suivants :

- Chaque arbre est entraîné sur un échantillon contenant environ 63% des données originales, obtenu par tirage avec remise.
- Environ 37% des données ne sont pas présentes dans un échantillon donné, ce qui implique que deux arbres entraînés séparément n'utilisent pas les mêmes exemples et peuvent produire des résultats différents.

2.4 Agrégation des prédictions

Lorsque plusieurs arbres sont construits, il est possible de combiner leurs résultats pour obtenir une prédiction finale. Cette combinaison peut se faire de différentes manières :

1. **Vote majoritaire** : Chaque arbre fournit une prédiction indépendante, et la classe la plus fréquemment choisie par l'ensemble des arbres est retenue comme classification finale. Cette approche permet de limiter les erreurs individuelles et d'obtenir une décision plus robuste.
2. **Vote pondéré par la confiance** : Chaque arbre attribue un poids à sa prédiction en fonction de la confiance qu'il accorde à son résultat (calculée à partir des proportions de classes dans ses feuilles terminales). La classe ayant le score le plus élevé après agrégation des poids est choisie. Cette méthode permet d'accorder une importance plus grande aux arbres produisant des prédictions fiables.

Cette approche duale permet d'adapter la stratégie d'agrégation aux spécificités des données tout en garantissant des performances optimales. Les résultats détaillés de l'évaluation comparative sont présentés dans la section 4.

En combinant ces approches, on renforce la précision du modèle et on réduit les risques liés aux classifications erronées d'arbres individuels.

2.5 Évaluation

L'évaluation permet de mesurer la performance du modèle en vérifiant sa capacité à bien classer les instances. Elle se déroule en deux étapes :

1. **Évaluation sur le jeu de validation** : Chaque ensemble d'arbres est testé sur des données mises de côté lors de l'entraînement afin d'obtenir une estimation préliminaire de sa précision.
2. **Évaluation finale sur le jeu de test** : Une fois le modèle validé, il est évalué sur un ensemble de test totalement indépendant pour mesurer sa capacité de généralisation.

2.5.1 Métriques utilisées

Plusieurs indicateurs permettent d'analyser la performance du modèle :

- **Exactitude** : Pourcentage d'instances correctement classifiées.
- **Matrice de confusion** : Tableau comparant les prédictions aux vraies classes, permettant d'identifier les classes difficiles à prédire.
- **Comparaison test/train (sur-apprentissage)** : Analyse de la différence entre la précision sur les données d'entraînement et de test pour détecter un éventuel sur-ajustement du modèle.

3 Résultats et discussions

3.1 Évolution des performances sur les ensembles selon le nombre d'arbres

3.1.1 Arbre unique($t = 1$)

Rapport de classification : Arbre unique ($t=1$)

precision	recall	f1-score	support
-----------	--------	----------	---------

setosa	1.00	1.00	1.00	114
versicolor	0.94	0.93	0.93	126
virginica	0.93	0.94	0.93	127
accuracy			0.95	367
macro avg	0.96	0.96	0.96	367
weighted avg	0.95	0.95	0.95	367

Même avec un arbre unique, les résultats sont bons : toutes les classes sont bien identifiées, et le modèle atteint une précision globale de 95%. Cependant, on observe quelques confusions entre *versicolor* et *virginica*, deux classes naturellement proches. Ce modèle reste sensible aux erreurs d'apprentissage spécifiques à un seul arbre.

3.1.2 Ensemble de 4 arbres ($t = 4$)

Rapport de classification : Ensemble d'arbres (t=4)				
	precision	recall	f1-score	support
setosa	1.00	1.00	1.00	274
versicolor	0.95	0.96	0.96	263
virginica	0.97	0.95	0.96	263
accuracy			0.97	800
macro avg	0.97	0.97	0.97	800
weighted avg	0.97	0.97	0.97	800

L'ajout d'arbres améliore nettement les performances : l'accuracy passe à 97%. L'ensemble devient plus fiable, avec une nette réduction des faux positifs, particulièrement entre *versicolor* et *virginica*. L'amélioration simultanée des scores de précision, de rappel et de F1-score reflète une capacité accrue à identifier correctement les échantillons, sans augmenter le taux d'erreur sur une classe particulière.

3.1.3 Ensemble de 8 arbres ($t = 8$)

Rapport de classification : Ensemble d'arbres (t=8)				
	precision	recall	f1-score	support
setosa	1.00	0.99	1.00	310
versicolor	0.96	0.98	0.97	315
virginica	0.98	0.97	0.97	309
accuracy			0.98	934
macro avg	0.98	0.98	0.98	934
weighted avg	0.98	0.98	0.98	934

À $t = 8$, les performances atteignent un excellent niveau avec 98% d'accuracy. L'ensemble capte encore mieux les motifs discriminants entre les classes. L'augmentation du recall de *versicolor* (0.98) indique qu'il réussit à détecter quasiment tous les échantillons de cette classe. Le score F1 reste élevé et équilibré, preuve que le modèle ne fait ni trop d'erreurs positives ni

d'omissions. Les moyennes pondérées, proches des moyennes macro, montrent que la performance est homogène, même en tenant compte des déséquilibres éventuels dans la distribution des classes.

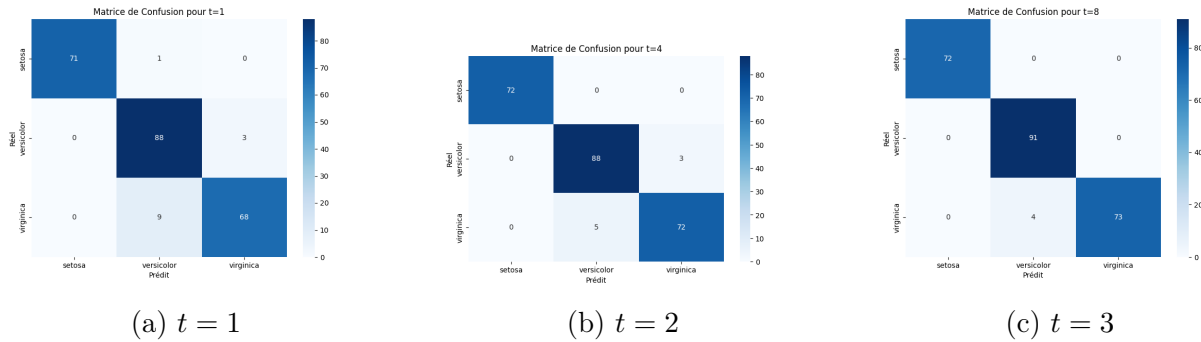


FIGURE 2 – Matrices de confusion obtenues selon la valeur de t

Les matrices de confusion permettent d'analyser les erreurs commises par le modèle pour différentes valeurs de t , et montrent qu'à mesure que t augmente, la fréquence des erreurs diminue, confirmant ainsi que l'agrégation d'arbres améliore la stabilité et la précision du modèle.

Les matrices de confusion pour différentes valeurs de t illustrent clairement l'amélioration des performances du modèle grâce à l'agrégation d'arbres. Pour $t = 1$, la classe *virginica* présente 9 erreurs de classification vers *versicolor* (soit 68 prédictions correctes sur 77), tandis que *versicolor* est confondue 3 fois avec *virginica* (88/91 correctes). À $t = 2$, les erreurs diminuent : *virginica* est mal classée 5 fois et *versicolor* conserve le même niveau d'erreur (3). Pour $t = 3$, la performance s'améliore encore, avec seulement 4 er

3.1.4 Récapitulatif en terme de performance

L'ensemble $t = 8$, considéré comme le plus performant, est testé sur un jeu de test final composé de données jamais utilisées lors de l'entraînement ou de la validation. Les résultats révèlent une meilleure robustesse, une amélioration générale des performances et des prédictions plus fiables.

L'agrégation progressive d'arbres améliore significativement la robustesse et la stabilité du modèle. Les hausses de précision réduisent les erreurs de classification, les rappels plus élevés montrent une meilleure couverture des cas pertinents, et les scores F1 confirment un bon compromis entre sensibilité et fiabilité. Ces résultats démontrent l'intérêt concret des ensembles dans les tâches de classification multiclassée comme le fait que les arbres ont capturé des motifs récurrents présents dans les nouvelles données.

3.2 Études de cas

3.2.1 Exemple avec PRÉDICTIONS CORRECTES

Instance 1

- Vrai label : **virginica**
- **Prédictions individuelles (classe, confiance) :**
 - Arbre 1-8 : virginica (confiance : 1.00)

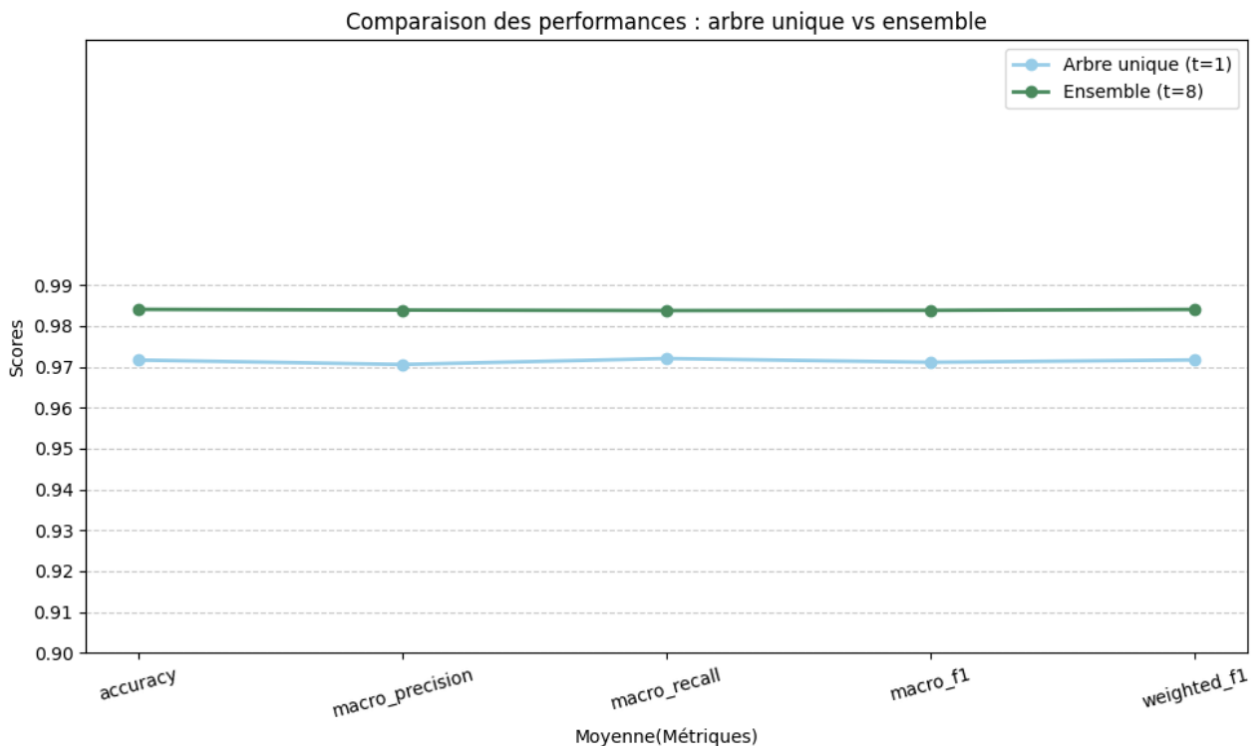


FIGURE 3 – Illustration de la comparaison des différentes métriques.

- Poids totaux par classe : virginica (8.00)
- **Prédiction finale :**
 - Vote pondéré : virginica
 - Vote majoritaire : virginica

3.2.2 Exemple avec PRÉDICTIONS INCORRECTES

Instance 112

- Vrai label : **virginica**
- **Prédictions individuelles (classe, confiance) :**
 - Arbre 1, 2, 4, 6, 7, 8 : versicolor (confiance : 1.00)
 - Arbre 3, 5 : virginica (confiance : 1.00)
- Poids totaux par classe : versicolor (6.00), virginica (2.00)
- **Prédiction finale :**
 - Vote pondéré : versicolor
 - Vote majoritaire : versicolor

3.2.3 Interpretation

L'objectif est d'identifier les cas où le modèle de classification fonctionne efficacement et ceux où il échoue, afin de mieux comprendre ses forces et faiblesses.

- **Instance bien classée :**

Ce cas illustre une forte cohérence entre les arbres, ce qui reflète un apprentissage stable.

L'unanimité dans les prédictions (avec confiance élevée) montre que l'instance est facilement classifiable : probablement située au centre du cluster virginica. Le système d'agrégation (moyenne ou majorité) ne change rien ici, car il n'y a pas de désaccord

— **Instance mal classée :**

Correspondent aux erreurs du modèle, où la classe prédite diffère de la classe réelle. Un examen des prédictions individuelles des arbres peut montrer un désaccord entre eux, révélant un manque de consensus dans le modèle. Et puisque la majorité des votes était pour versicolor l'arbre a supposé que c'était versicolor or que c'est virginica.

Le vote majoritaire consiste à retenir la classe prédite par le plus grand nombre d'arbres, sans prendre en compte leur niveau de certitude. Cette méthode, simple et souvent efficace lorsque les arbres sont bien entraînés, peut toutefois être biaisée si certains arbres commettent des erreurs récurrentes.

À l'inverse, le vote pondéré attribue à chaque prédiction un poids proportionnel à la confiance associée. Cette stratégie permet une meilleure gestion des cas ambigus, notamment lorsque plusieurs classes sont en compétition, et peut contribuer à réduire les erreurs systématiques.

Dans notre cas, la confiance attribuée à chaque prédiction est de 100%, ce qui s'explique par le fait que chaque feuille terminale de nos arbres ne contient que des instances d'une seule classe. Ainsi, dans cette configuration, le vote pondéré revient à un simple vote majoritaire, car tous les poids sont identiques.

Cependant, si l'on réduit la profondeur des arbres, les feuilles pourraient contenir des instances de classes différentes. Les prédictions deviendraient alors moins certaines (confiance < 100%), ce qui rendrait le vote pondéré réellement pertinent. Il permettrait alors de mieux refléter l'incertitude de chaque arbre et d'améliorer la robustesse globale du modèle.

3.3 Diversité des arbres et impact sur les performances

L'attribut racine d'un arbre de décision est celui qui offre la meilleure séparation des classes. Voici les attributs les plus discriminants identifiés pour chaque arbre :

- **Petal_length** : utilisé comme attribut racine pour 4 arbres (1, 4, 7, 8).
- **Petal_area** : identifié comme racine dans 3 arbres (2, 5, 6).
- **Petal_curvature_mm** : utilisé une seule fois (arbre 3).

L'analyse des variables met en évidence que **petal_length** et **petal_area** sont les caractéristiques les plus influentes dans le processus de classification. Cela suggère que la longueur et la surface des pétales jouent un rôle déterminant dans la différenciation entre les classes. L'utilisation d'attributs racines variés dans les arbres de décision contribue à la diversité structurelle au sein de l'ensemble, ce qui renforce la robustesse du modèle et sa capacité à généraliser sur des données nouvelles.

Dans notre approche, plusieurs sous-ensembles de données sont créés aléatoirement à partir du jeu d'apprentissage initial, en sélectionnant des exemples avec remise. Chaque arbre de décision est ensuite construit à partir d'un de ces sous-ensembles. Comme chaque arbre voit des données différentes, cela produit des structures différentes, notamment au niveau des premières décisions prises dans l'arbre.

Cette variation entre les arbres est un point important de la méthode : comme les arbres ne font pas tous les mêmes choix ni ne voient les mêmes exemples, les erreurs commises par certains

peuvent être compensées par les décisions des autres lors de la combinaison des résultats. Cela permet d'obtenir des prédictions finales plus stables et plus fiables.

- **Un arbre unique** peut produire des résultats différents selon les données qu'il a reçues pour son apprentissage.
- **Un ensemble d'arbres**, en combinant plusieurs points de vue, permet d'obtenir une décision plus stable et plus fiable.
- **Le fait que les arbres soient différents les uns des autres** garantit que les décisions prises ne dépendent pas d'un seul modèle.
- **La combinaison des décisions** (par vote majoritaire ou pondéré) permet de corriger les erreurs possibles de certains arbres grâce aux autres.

En résumé, utiliser plusieurs arbres construits à partir de données différentes est une stratégie efficace pour améliorer la qualité des décisions prises par le modèle.

4 Conclusion

Les arbres de décision, bien qu'efficaces et interprétables, souffrent de limitations notables lorsqu'ils sont utilisés seuls : instabilité et possible dépendance aux données d'entraînement. Notre étude avait pour objectif d'examiner dans quelle mesure la combinaison de plusieurs arbres, construits à partir d'échantillons aléatoires des données, pouvait pallier ces faiblesses et améliorer la performance globale d'un modèle de classification.

Les résultats expérimentaux ont mis en évidence que l'agrégation d'arbres, en particulier via l'échantillonnage bootstrap, permet de réduire la variance du modèle. Cette réduction de variance contribue à rendre les prédictions plus robustes et mieux adaptées aux nouvelles données non observées, améliorant ainsi la capacité de généralisation du modèle. Par conséquent, l'utilisation d'un ensemble d'arbres de décision se révèle être une approche efficace pour renforcer la précision, la stabilité et la fiabilité des classifieurs.