Random forest

Prérequis

Pour comprendre la théorie du random forest il faut bien comprendre :

- L'ensemble learning
- Les arbres de décision

Le problèmes des arbres

- Les arbres de décision sont facile à entraîner, facile à utilisé et très interprétable, on peut savoir assez rapidement les règles qui permettent la prédiction ou la classification de notre exemple.

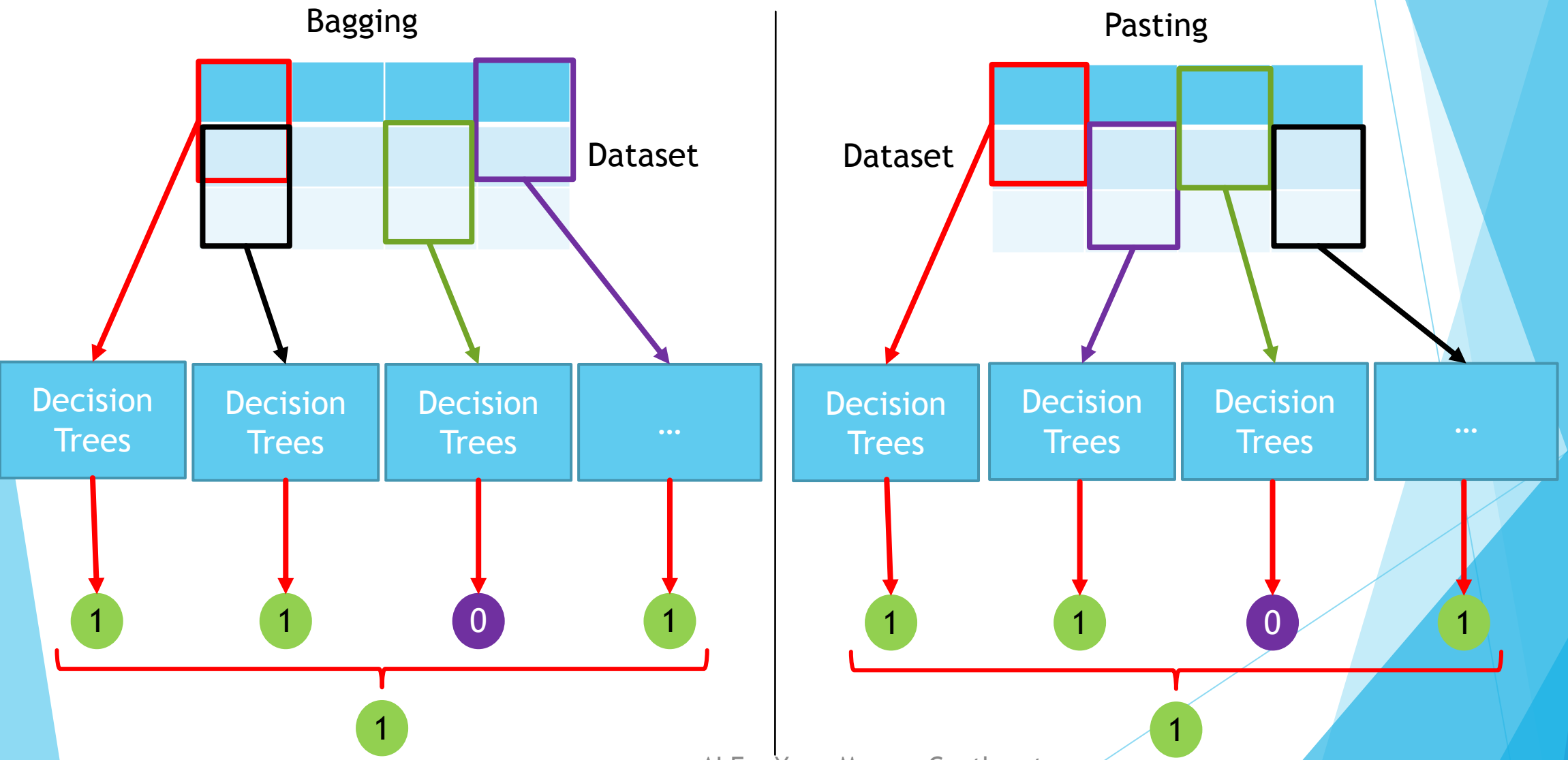
- Mais derrière ces qualités apparente on peut leur reprocher de ne pas être assez précis ni assez généralisable. En effet, l'arbre est performant avec le jeu d'entraînement mais il n'est pas assez flexible pour donner de bonnes performances sur des données de test. Or, but le d'un algorithme de machine learning est de pouvoir classifier des données dont on ne connaît pas encore la prédiction donc se sont des données autres que celle d'entraînement.

Random forest

- Pour palier à ce problème de généralisation, les random forest ont vue le jour. Cet algorithme permet d'utiliser plusieurs arbres de décision afin de créer une forêt et d'améliorer la généralisation de l'ensemble du modèle. On va emprunter la théorie de l'ensemble learning afin de faire fonctionner ses arbres ensemble.

- Le random forest combine donc la simplicité des arbres de décision grâce à l'ensemble learning afin de gagner de la flexibilité et de la généralisation. Ce qui donne des modèles de bien meilleur qualité.

Bagging & Pasting



1) Création du data set

Data set du Random Forest

Nb de pièces	Surface	Garage	Prix
4	80	1	220 000
3	70	0	190 000
2	40	1	140 000
4	60	0	170 000
3	70	1	200 000

1) Création du data set

Data set du Random Forest

Nb de pièces	Surface	Garage	Prix
4	80	1	220 000
3	70	0	190 000
2	40	1	140 000
4	60	0	170 000
3	70	1	200 000

Data set de l'arbre de décision

Nb de pièces	Surface	Garage	Prix

1) Création du data set

Data set du Random Forest

Nb de pièces	Surface	Garage	Prix
4	80	1	220 000
3	70	0	190 000
2	40	1	140 000
4	60	0	170 000
3	70	1	200 000

Data set de l'arbre de décision

Nb de pièces	Surface	Garage	Prix
3	70	0	190 000

1) Création du data set

Data set du Random Forest

Nb de pièces	Surface	Garage	Prix
4	80	1	220 000
3	70	0	190 000
2	40	1	140 000
4	60	0	170 000
3	70	1	200 000

Data set de l'arbre de décision

Nb de pièces	Surface	Garage	Prix
3	70	0	190 000
4	60	0	170 000

1) Création du data set

Data set du Random Forest

Nb de pièces	Surface	Garage	Prix
4	80	1	220 000
3	70	0	190 000
2	40	1	140 000
4	60	0	170 000
3	70	1	200 000

Data set de l'arbre de décision

Nb de pièces	Surface	Garage	Prix
3	70	0	190 000
4	60	0	170 000
3	70	1	200 000

1) Création du data set

Data set du Random Forest

Nb de pièces	Surface	Garage	Prix
4	80	1	220 000
3	70	0	190 000
2	40	1	140 000
4	60	0	170 000
3	70	1	200 000

Data set de l'arbre de décision

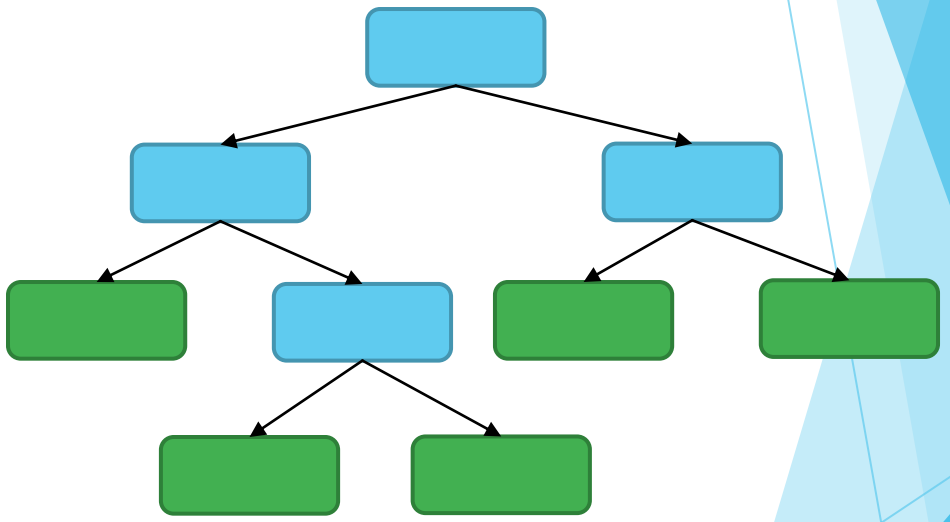
Nb de pièces	Surface	Garage	Prix
3	70	0	190 000
4	60	0	170 000
3	70	1	200 000
3	70	0	190 000

Pour maximiser la variété des arbres !

2) Entraînement de l'arbre de décision

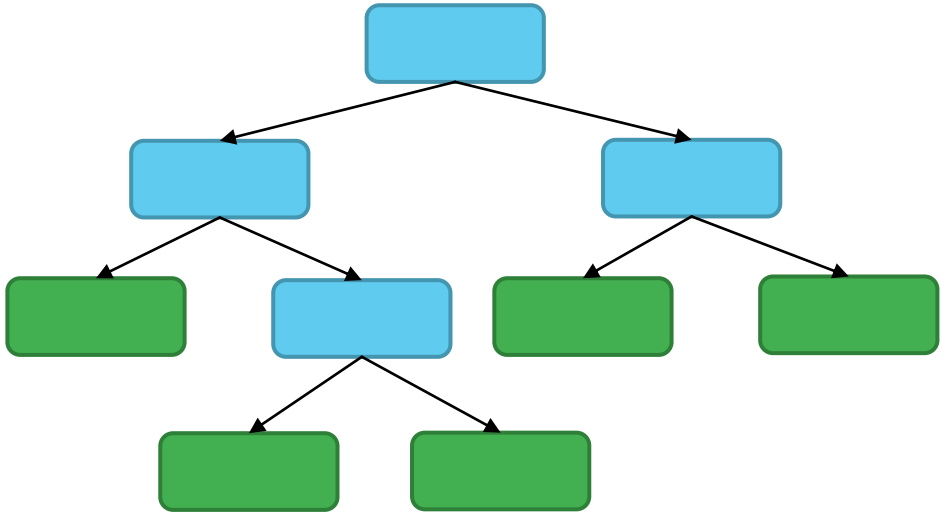
Data set de l'arbre de décision

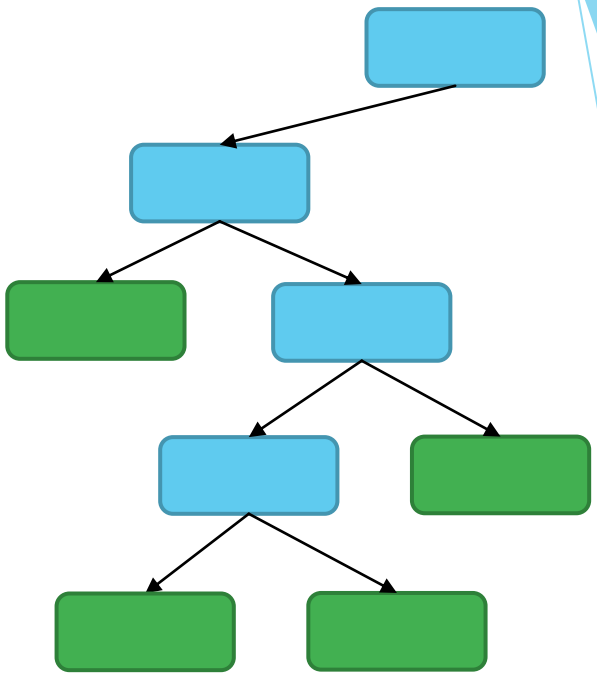
Nb de pièces	Surface	Garage	Prix
3	70	0	190 000
4	60	0	170 000
3	70	1	200 000
3	70	0	190 000

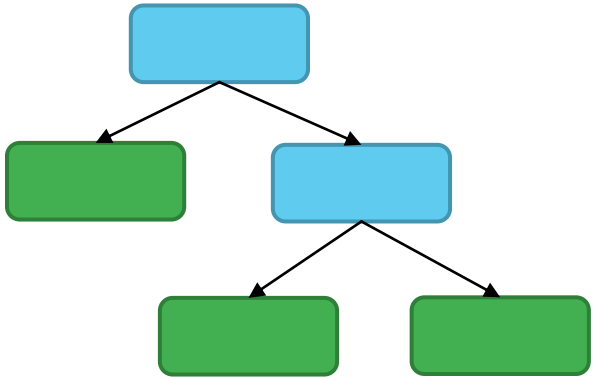


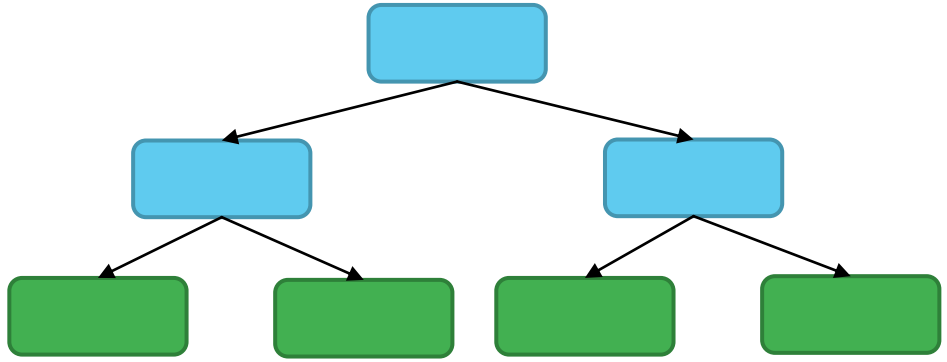
Pour maximiser la variété des arbres !

3) Répéter les étapes 1 et 2



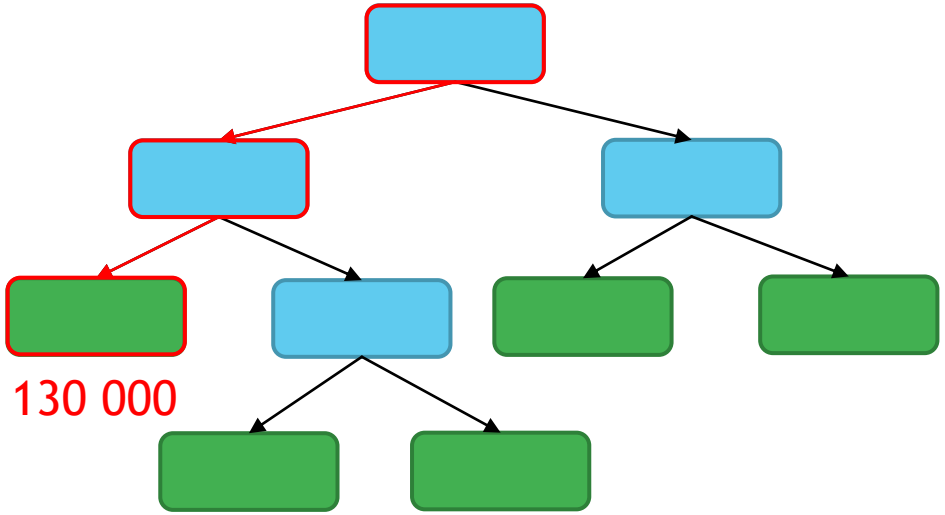




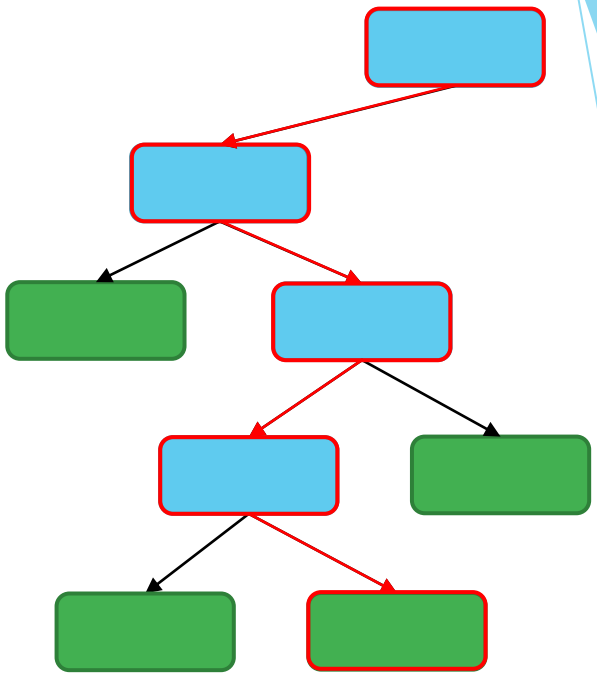


Prédiction

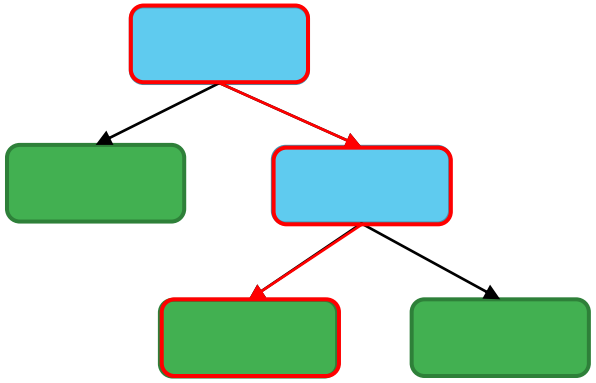
Prédiction moyenne : 134 500



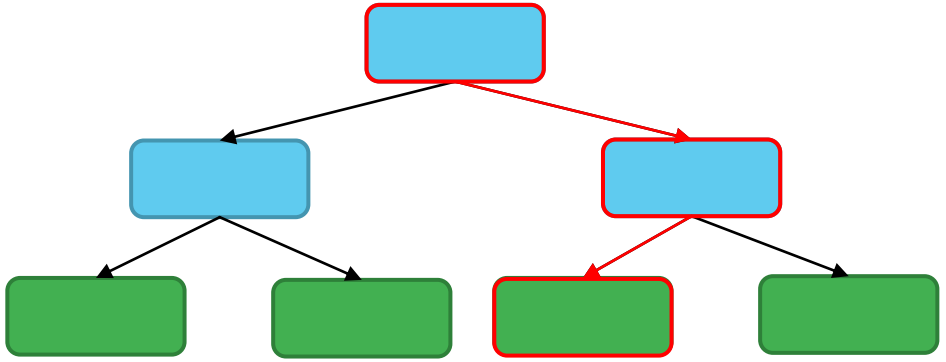
130 000



140 000



135 000



133 000