Decision Tree Regression

Réalisé par :

EL AADRAOUI Yassine

KOUALIL Mohammed

MELLOUK Fatima Zahrae

Sous la supervision de :

Pr.OURDOU Amal

Master Big data & aide à la décision S2

27 février 2023

Contents

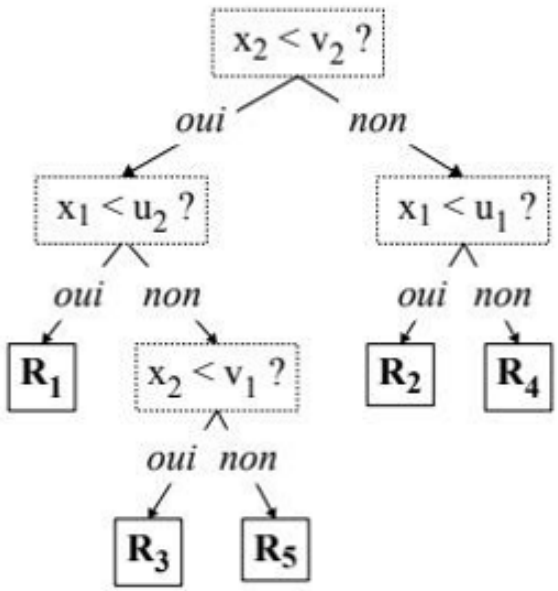
Introduction

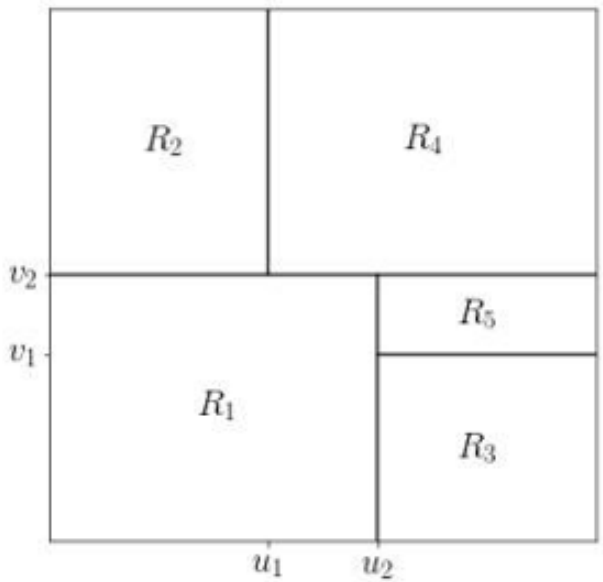
les arbres de décision sont des modèles hiérarchiques, qui se comportent comme une série successive de tests conditionnels, dans laquelle chaque test dépend de ses antécédents. Ils sont couramment utilisés en dehors du monde du machine Learning, par exemple pour décrire les étapes d'un diagnostic d'un choix de traitement pour un médecin, ou les chemins possibles dans un « livre dont vous êtes le héros ».

Définition

L'arbre de décision est composé d'une série de nœuds, chaque nœud représentant une décision ou une prédiction. Le modèle est construit en divisant les données en sous-ensembles plus petits en fonction de la valeur d'un certain attribut. Le processus est répété jusqu'à ce que toutes les données soient divisées en groupes homogènes et que la variable cible soit prédite.

Principe


$x_2 < v_2$?
oui
non
$x_1 < u_2$?
$x_1 < u_1$?
oui
non
oui
non
R_1
$x_2 < v_1$?
R_2
R_4
oui
non
R_3
R_5


R_2
R_4
v_2
R_5
v_1
R_1
R_3
u_1
u_2

Comment faire pousser un arbre ?

Pour entraîner un arbre de décision ,il existe plusieurs techniques performantes ,mais puisque on travaille sur la régression alors ,on va consacrer notre travail sur quelques techniques de ce type. La technique la plus connue c'est CART (Classification And Regression Tree).
Il s'agit d'un algorithme de partitionnement de l'espace par une approche gloutonne, récursive et divisive.

L'algorithme CART nécessite 3 composants :

- Définir un critère pour sélectionner la meilleure partition .
- PUne règle pour décider quand un nœud est terminal, c'est-à-dire qu'il devient un feuille.
- Tailler l'arbre pour éviter le sur-apprentissage.

Sélection de la meilleure partition

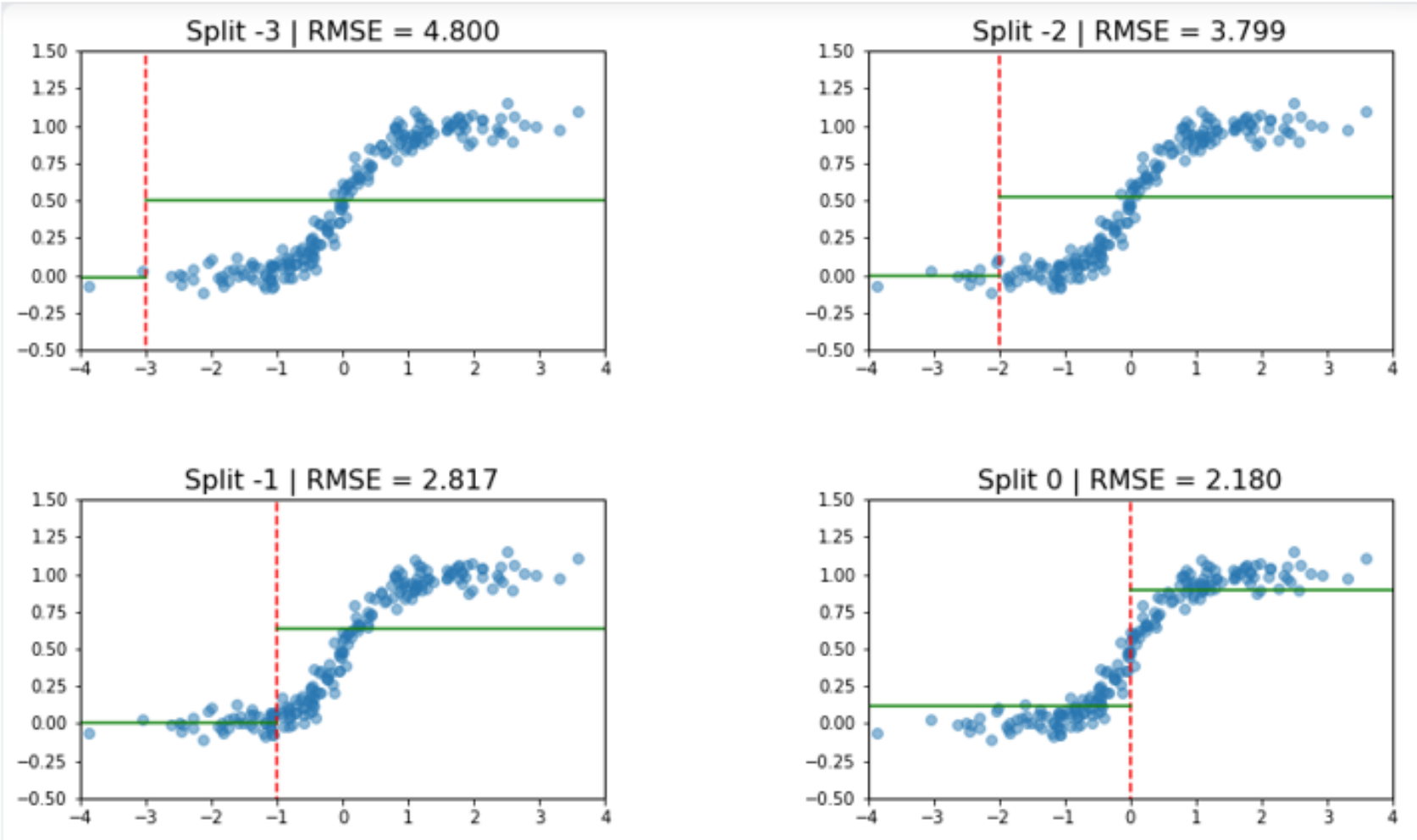
On appellera le nombre un split, c'est-à-dire la valeur où l'on sépare en deux l'espace.

Cette valeur est déterminée par CART en utilisant la MSE.c'est a dire on choisit la valeur qui minimise MSE.

La formule de la MSE est la suivante :

$$\sum_{j=1}^{j} \sum_{x_i \in R_m} (y_i - y_{Rj})^2 \qquad (1)$$

y_{Rj} :la réponse moyenne pour les observations d'entraînement dans la jème région.


Split -3 | RMSE = 4.800
Split -2 | RMSE = 3.799
Split -1 | RMSE = 2.817
Split 0 | RMSE = 2.180

Remarque

Dans le cas d'une variable continue x_j, si l'on suppose les valeurs prises par cette varibale dans D ordonnées :

$$x_j^1 \leq x_j^2 \leq ..., \leq x_j^n$$

alors les valeurs possibles de s sont $\frac{x_j^{i+1} - x_j^i}{2}$ pour toutes les valeurs de i telles que $x_j^{i+1} \neq x_j^i$

Règle d'arrêt

minsplit : pour éviter de créer des fractionnements qui petites feuilles, le nombre minimum d'observations qui doit exister dans un nœud pour qu'une scission soit tentée (minsplit = 20).

Le sur-apprentissage

Le risque de sur-apprentissage (créer un arbre avec une très grande profondeur) est très élevé pour les modèles non-paramétriques, dont fait partie l'arbre de décision. i l'arbre décide d'effectuer une subdivision sur un seul point, on est dans le cas de sur-apprentissage, car ce point est très éloigné de l'ensemble des autres points, c'est un point extrême. On obtiendra des résultats très différents de la réalité si ce type de points est pris considération par le modèle. Pour limiter le sur-apprentissage, le choix d'hyper-paramètres est important.

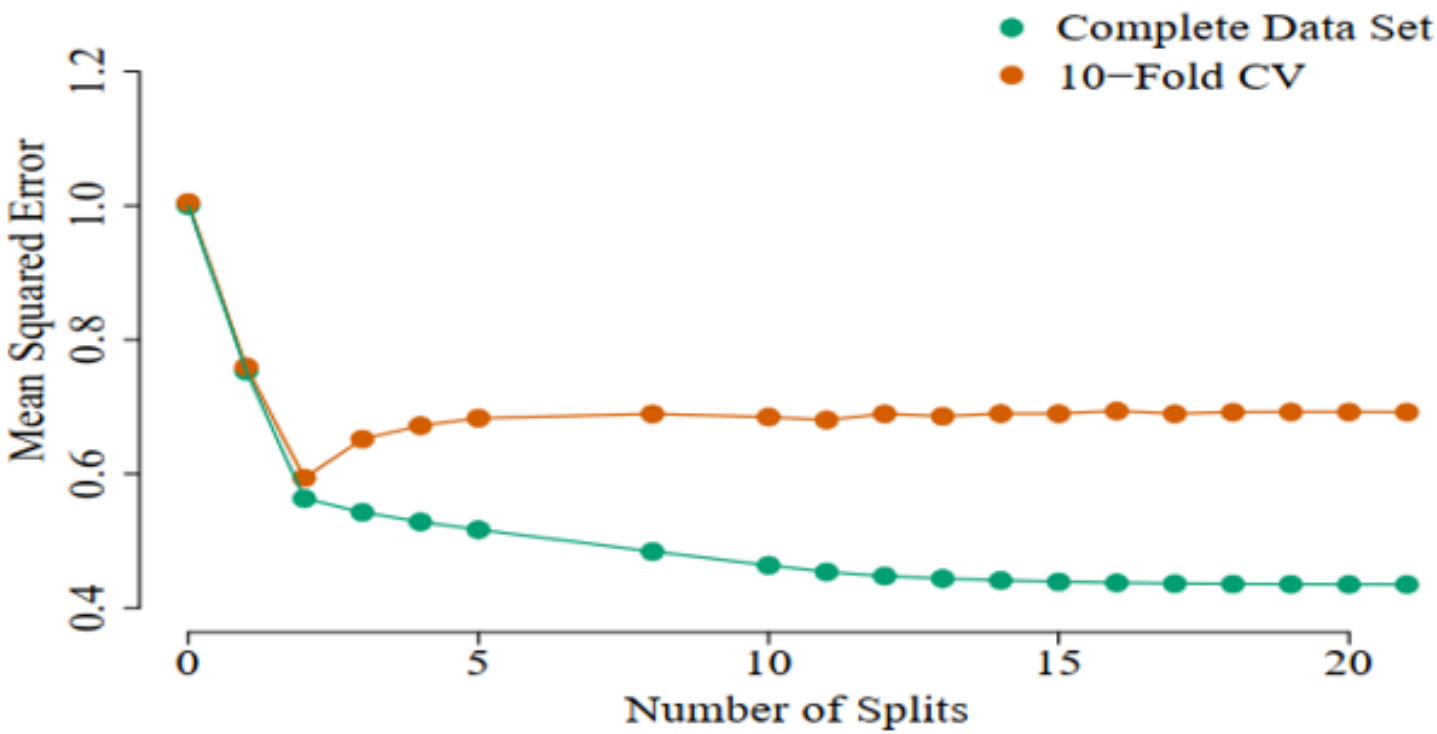
Ces hyper-paramètres limiteront le sur-apprentissage dans les arbres de décision une fois optimisés.

- La profondeur maximale (max_depth)
- Le nombre d'observations minimal dans un nœud pour effectuer un split.

- Nous cultivons d'abord le plus grand arbre possible Tmax puis le taillons retour afin d'obtenir un sous-arbre.
- Pour chaque valeur de , il existe un sous-arbre $T \subset Tmax$ qui minimise : $\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - y_{Rj})^2 + \alpha |T|$
- |T| indique le nombre de nœuds terminaux de l'arbre T, Rm est le rectangle correspondant à la même feuille, et ŷRm est la réponse prédite associée à Rm.
- α est choisi en utilisant la validation croisée v-fold.

$$CV(\alpha) = \frac{1}{5}\sum_{v=1}^{5} MSE_v^{(test)}$$

Over-fitting



Exemple

Prédiction des salaires des ligues majeures de baseball.

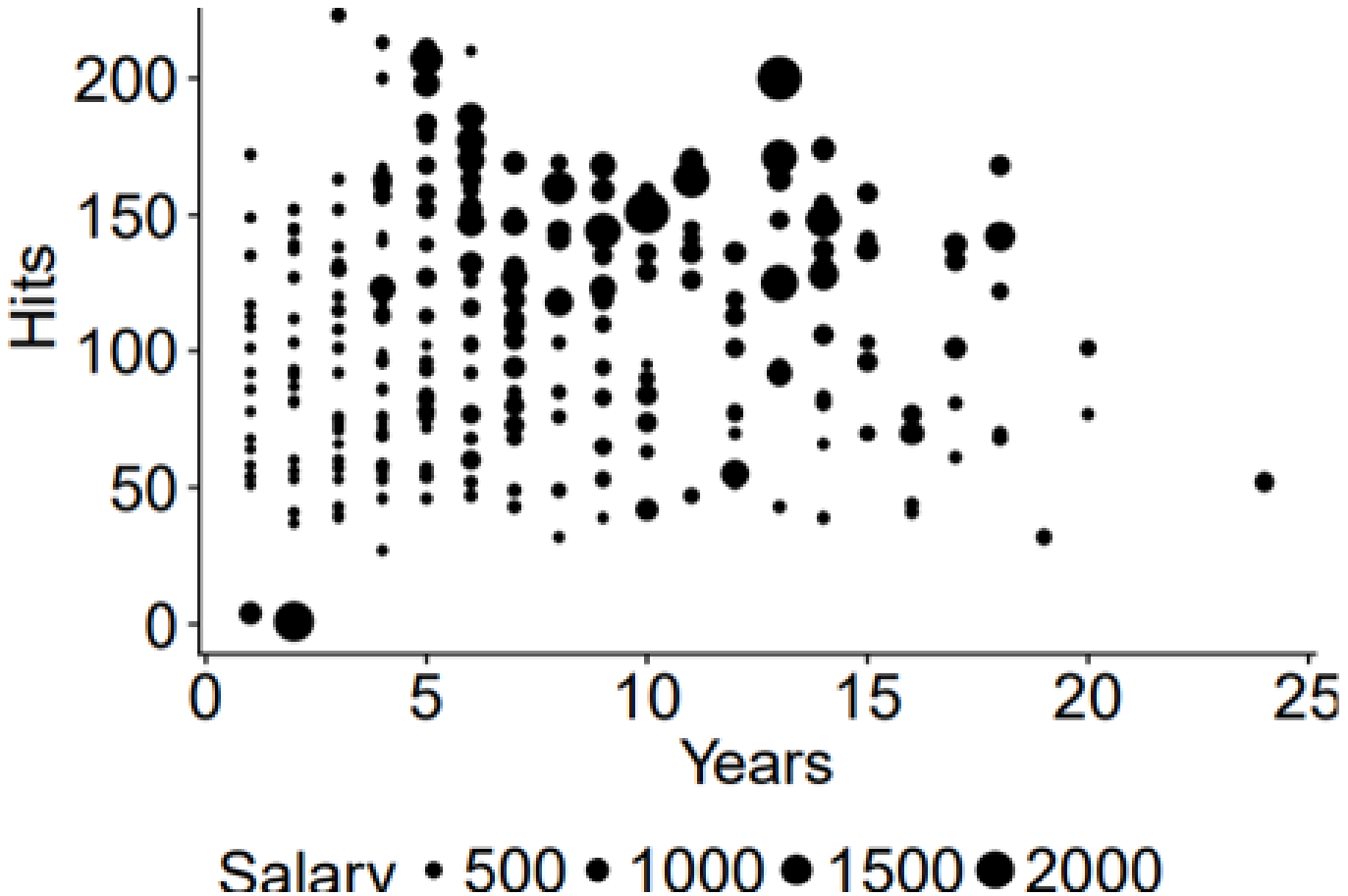
- variable cible
  - y : le salaire annuel en 1987 (variable cible y)
- Predictor variables :
  - X1 : le nombre d'années passées dans les ligues majeures
  - X2 : le nombre de coups (hits) réussis en 1986
- objectif prédire le salaire annuel au début de la saison de baseball de 1987 en utilisant les variables prédictives(years , hits).

Les données

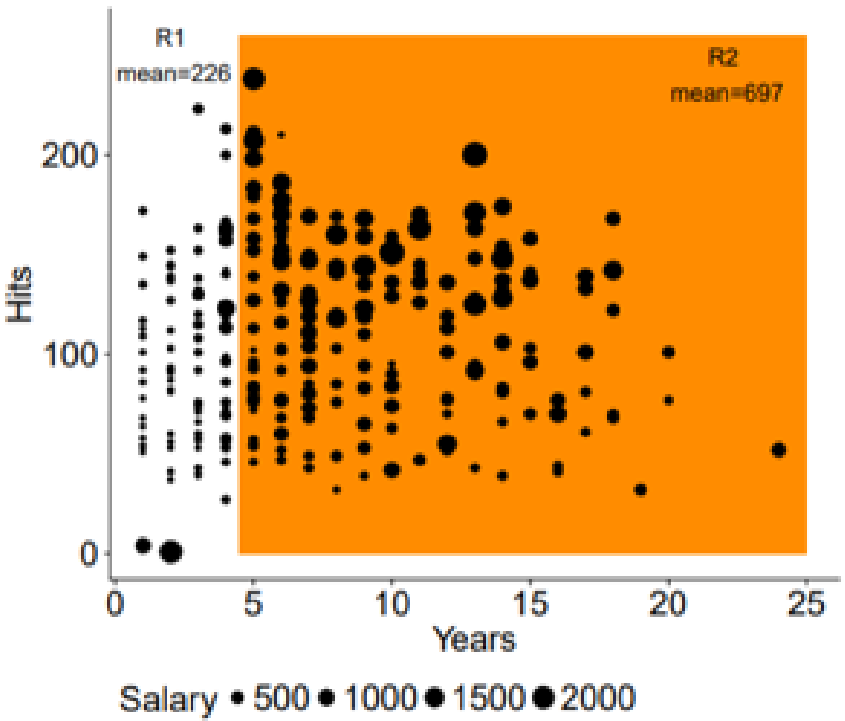
Un exemple de ce à quoi ressemblent les données.

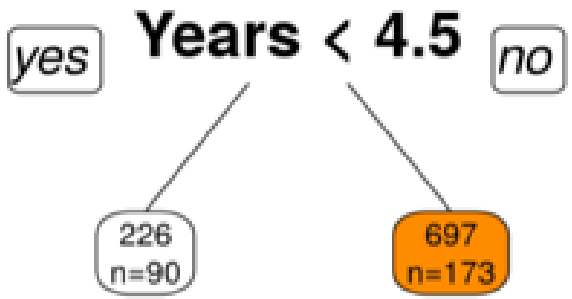
	Years	Hits	Salary
-Andre Dawson	11	141	500
-Andres Galarraga	2	87	92
-Barry Bonds	1	92	100
-Cal Ripken	6	177	1350
-Gary Carter	13	125	1926
-Joe Carter	4	200	250
-Ken Griffey	14	150	1000
-Mike Schmidt	2	1	2127
-Tony Gwynn	5	211	740

Une représentation visuelle des donnée

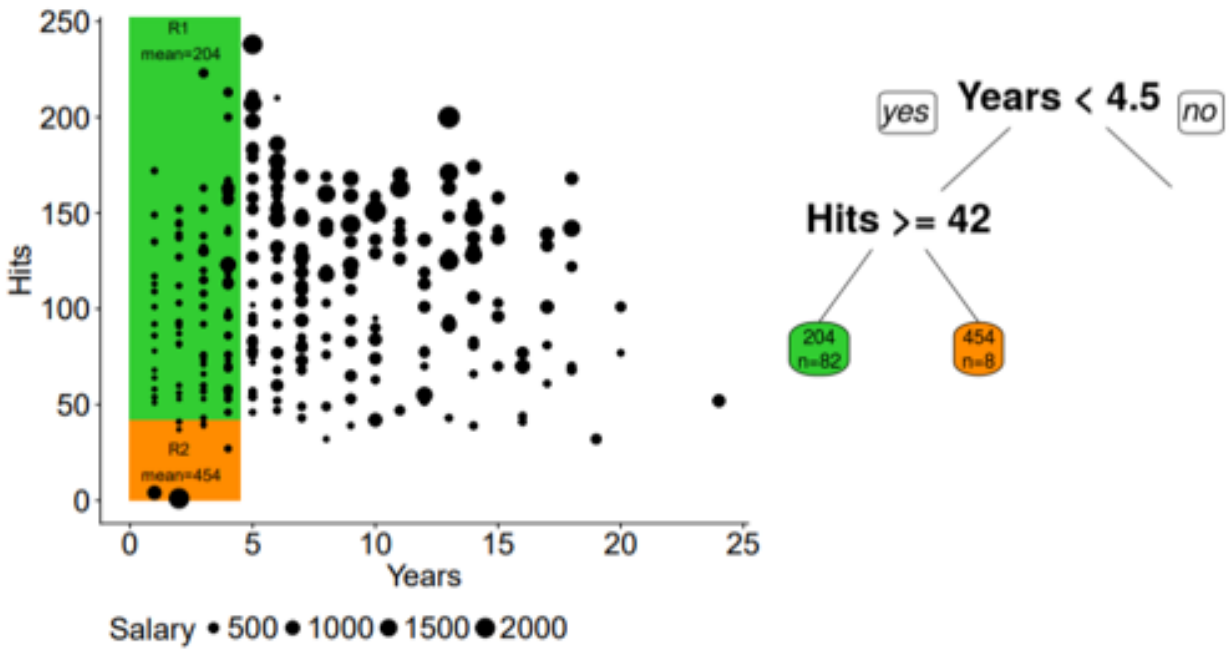


La Première division





Deuxième division



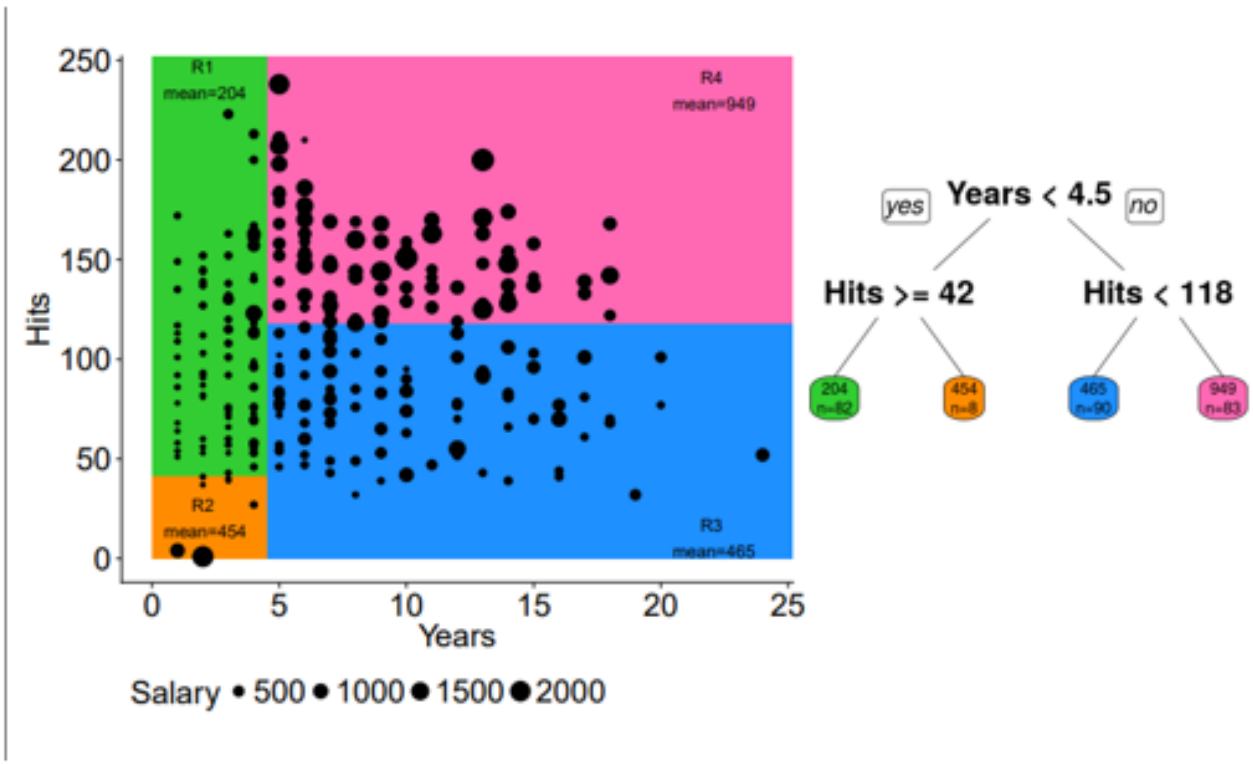
Une erreur dans les données

Les données indiquent qu'il a joué seulement 2 ans et a eu 1 coup sûr en 1987, ce qui est incorrect.

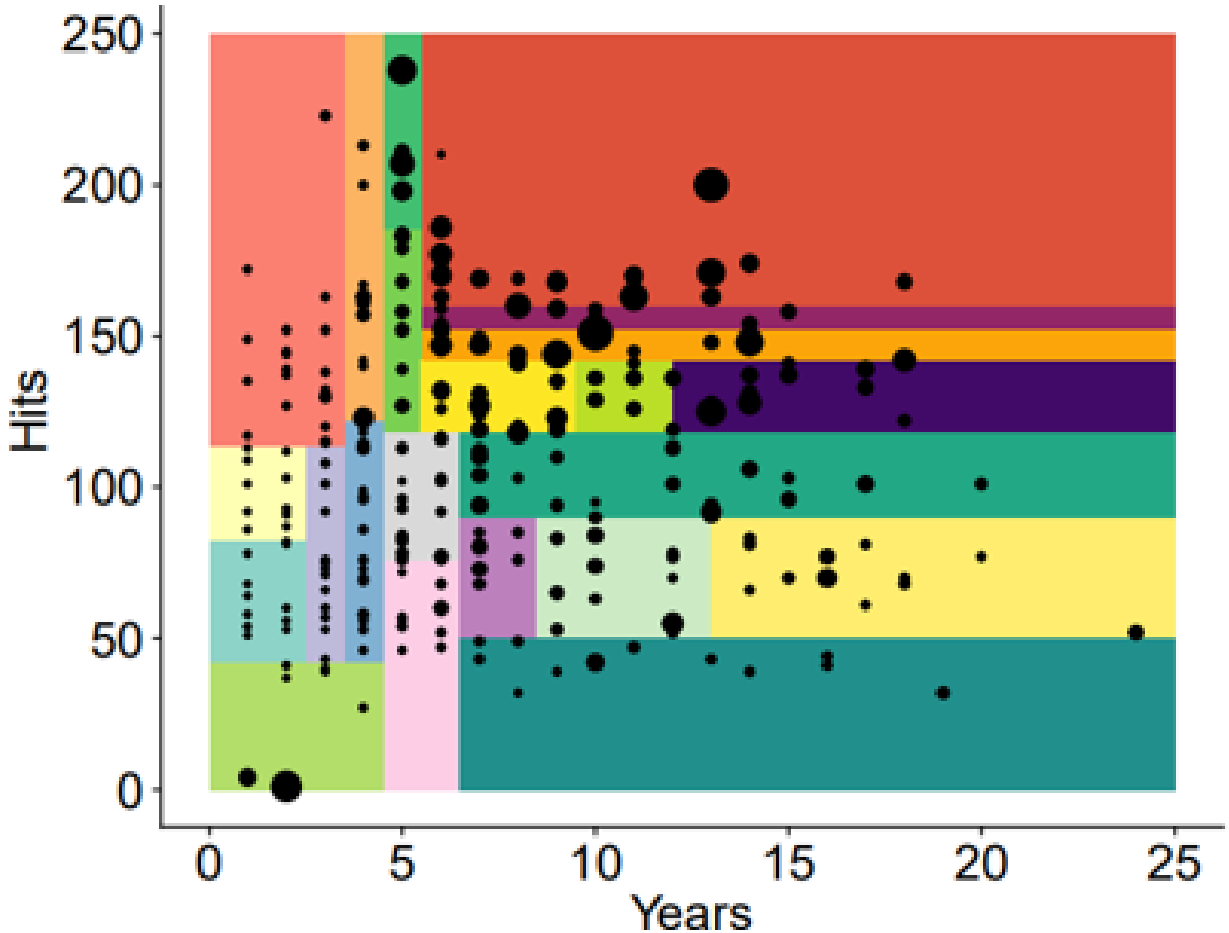
	Years	Hits	Salary
-Andre Dawson	11	141	500.00
-Andres Galarraga	2	87	91.50
-Barry Bonds	1	92	100.00
-Cal Ripken	6	177	1350.00
-Gary Carter	13	125	1925.57
-Joe Carter	4	200	250.00
-Ken Griffey	14	150	1000.00
-Mike Schmidt	2	1	2127.33
-Tony Gwynn	5	211	740.00

- Mike Schmidt started his career in 1972, and was inducted into the Baseball Hall of Fame in 1995.

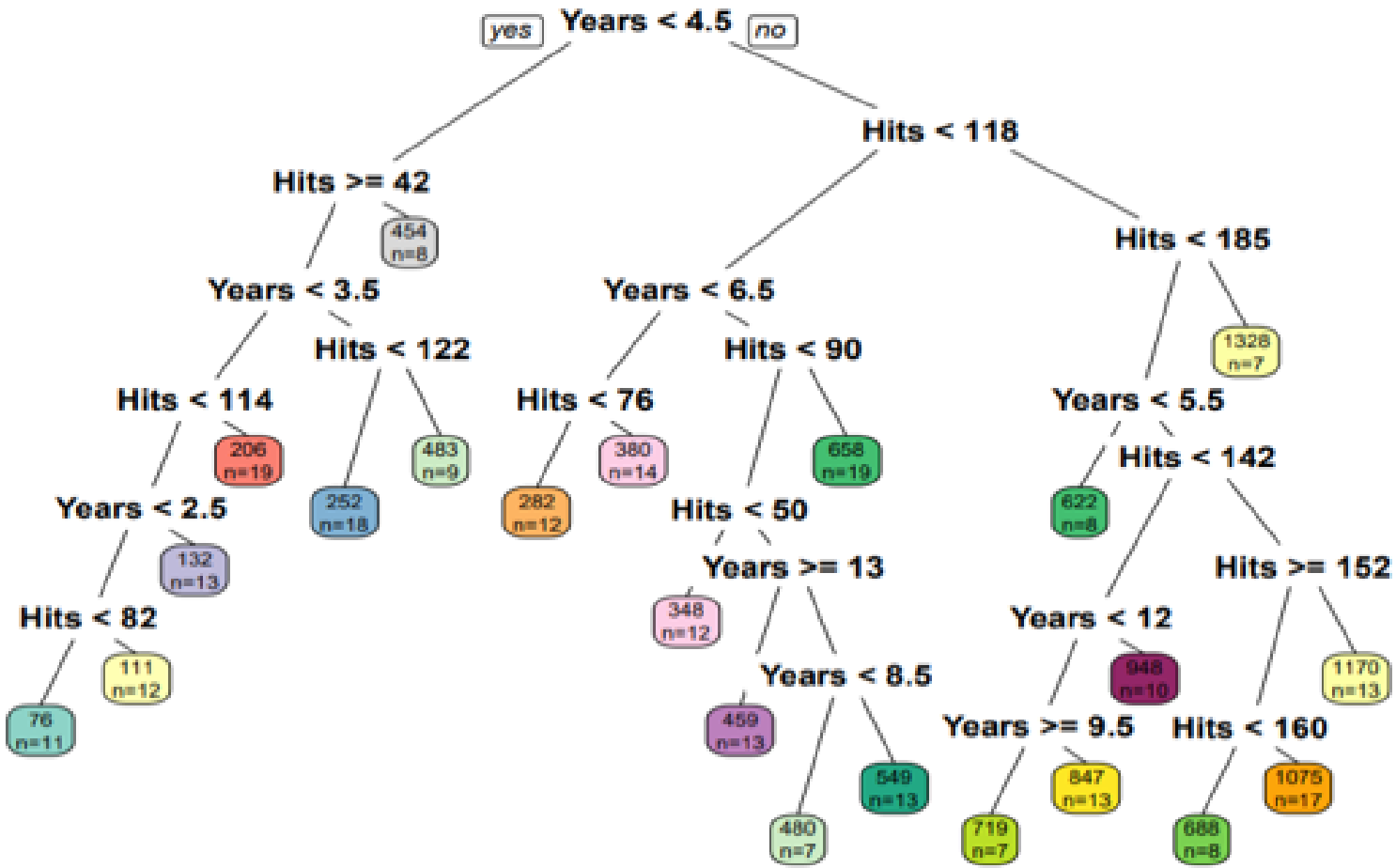
Deuxième division



Et si nous continuons...



Arrêter si le nombre d'observations est inférieur à 20



Conclusion

- Regression Arbres de Decision est une puissante technique d'apprentissage automatique qui peut être utilisée à la fois pour les problèmes de classification et de régression.
- Simple à comprendre et à interpréter, et être capable de gérer de grands ensembles de données avec des valeurs manquantes.

Merci pour votre attention !