Arbres de décision

Silhouettes of soft-drink bottles

ONNO Lilou
JAMIN Mathilde
M2 ECAP

April 20, 2025

Machine Learning et arbres de décision

1 Modifications majeures du script

Dans notre jeu de données, la variable d'intérêt est une variable quantitative discrète, contrairement à l'exemple étudié en cours, qui concernait une variable qualitative.

Pour adapter notre analyse, plusieurs modifications ont été apportées à notre code. En particulier, lors de la modélisation avec un arbre CART sur les données d'apprentissage, nous avons désormais utilisé la méthode ANOVA.

De plus, dans la section consacrée à la prédiction sur les ensembles d'apprentissage et de test, nous avons changé le type de classe en vecteur afin d'obtenir des résultats cohérents avec nos données.

Enfin, nous avons calculé les MSE, qui s'appliquent aux problèmes de régression quantitative, et non aux problèmes de classification qualitative, comme vu en cours. Cela a nécessité des ajustements dans le code pour évaluer correctement nos modèles de régression.

2 Choix de paramétrage des modèles et résultats observés

2.1 Arbre de décision CART

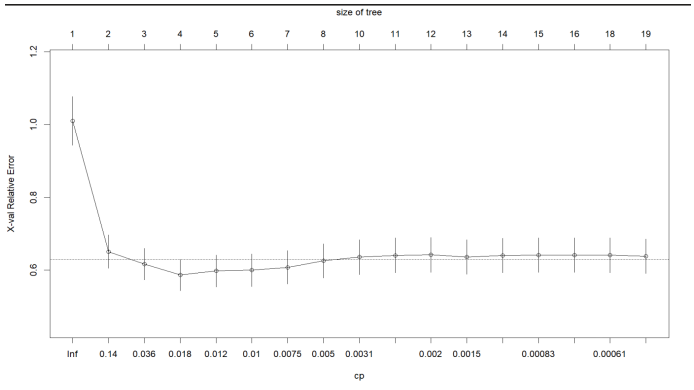

Figure 1: Arbre de décision de notre série

La Figure n°1, illustrant un arbre de régression modélisant notre variable d'interêt "Liking" à partir des autres variables explicative présentent dans notre base, met en évidence le paramètre de complexité optimal, qui correspond à l'erreur relative la plus faible. Dans notre analyse, nous constatons que la taille idéale de notre arbre, mesurée par la taille des nœuds, est de 3. Cette configuration permet de minimiser efficacement l'erreur de prédiction tout en évitant le surajustement.

2.2 Forêt aléatoire

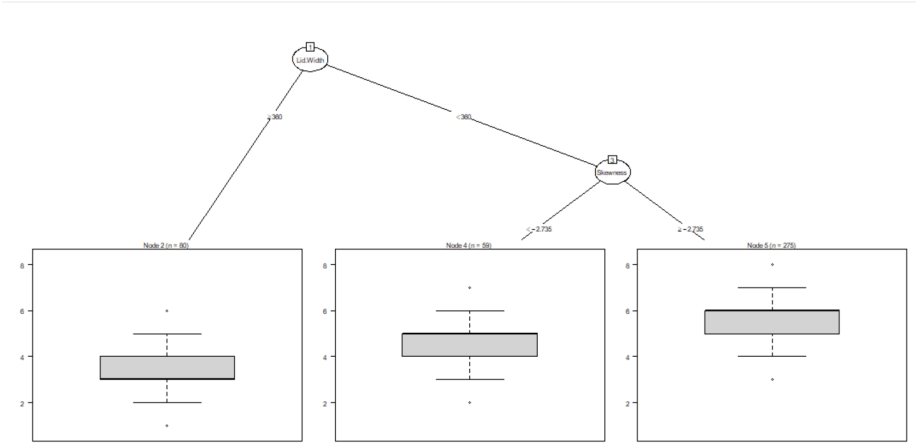

Figure 2: Forêt aléatoire

Ce graphique illustre un arbre de décision élagué, avec des boîtes à moustaches représentant les distributions des valeurs dans les nœuds feuilles. Le nœud racine, basé sur la variable Lid.Width (la largeur du bouchon), détermine la direction de l'arbre : si la largeur est inférieure à 360, les observations vont vers le nœud feuille gauche (80 observations, médiane d'environ 4 au vu des boxplot). Pour les valeurs supérieures ou égales à 360, une seconde décision est prise selon Skewness. Si cette dernière est inférieure ou égale à -2.735, les observations se dirigent vers le nœud feuille contenant 59 observations et une médiane similaire. En revanche, si Skewness est supérieur à -2.735, on obtient un nœud avec 275 observations et une médiane d'environ 6, indiquant une tendance plus élevée par rapport aux autres nœuds.

3 Interprétation générale

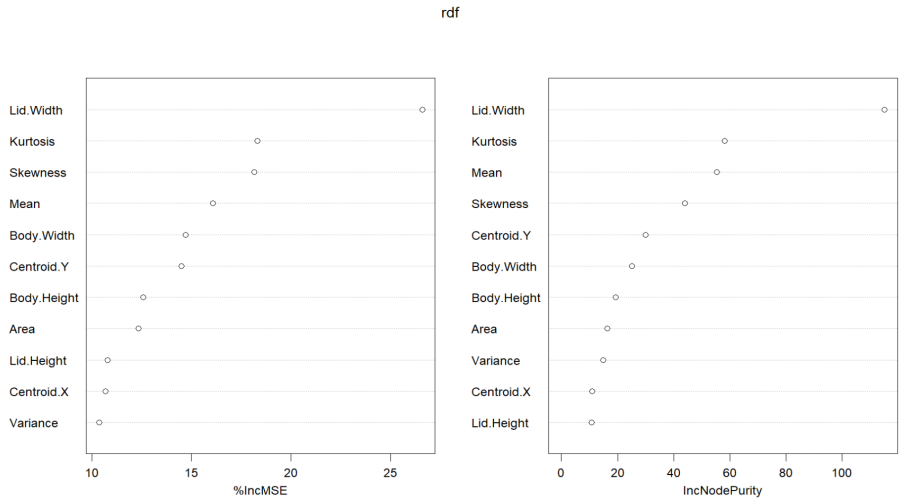

Figure 3: Importance des variables

Le graphique de gauche montre l'importance des variables en termes d'augmentation du MSE. Ce pourcentage reflète l'impact de chaque variable sur la précision du modèle. En permutant les valeurs d'une variable, nous observons de combien le MSE augmente. Une augmentation élevée indique que la variable est cruciale pour la performance du modèle.

Tandis que le graphique de droite présente l'importance des variables en termes d'amélioration de la pureté des nœuds (IncNodePurity). Cette mesure évalue combien une variable contribue à rendre les sous-ensembles de données plus homogènes lorsqu'elle est utilisée pour diviser un nœud dans les arbres de décision. Plus la valeur est élevée, plus la variable est importante.

Pour conclure les variables en haut du graphique, comme Lid.Width, Kurtosis, et Skewness, sont les plus importantes, car leur permutation provoque une augmentation significative du MSE. De plus, les variables comme Lid.Height et Centroid.X, en bas du graphique, ont un faible impact sur l'exactitude du modèle, car leur permutation augmente faiblement le MSE.

4 Conclusion

Les erreurs quadratiques moyennes (MSE) calculées sur les ensembles d'apprentissage et de test pour les deux modèles : d'arbres CART et de forêt aléatoire sont présentées dans le tableau ci-dessous. Nous pouvons observer une différence entre les MSE sur l'ensemble d'apprentissage et l'ensemble de test, ce qui indique un possible surapprentissage dans le premier modèle.

Modèle	**MSE (Apprentissage)**	**MSE (Test)**
Arbre	0.9932443	2.640711
Forêt aléatoire	0.8253913	0.95506225

Table 1: Erreurs quadratiques moyennes (MSE) pour les deux modèles

Interprétation des résultats de l'arbre de régression : Le MSE sur l'ensemble d'apprentissage est très faible (0.993), ce qui indique que le modèle ajuste bien les données d'entraînement. Cependant, cette performance élevée peut également signaler un sur-ajustement. En effet, le MSE sur l'ensemble de test est beaucoup plus élevé (2.641), suggérant que le modèle a plus de difficultés à prédire des données qu'il n'a jamais vues. L'écart entre le MSE d'entraînement et de test confirme potentiellement la présence d'un sur-ajustement.

Interprétation des résultats de la forêt aléatoire : Le MSE sur l'ensemble d'apprentissage est de 0.825, indiquant que le modèle s'ajuste bien aux données d'entraînement. Le MSE sur l'ensemble de test est légèrement supérieur (0.955), ce qui montre une bonne capacité de généralisation, sans signe évident de surapprentissage.

Les résultats montrent donc que la forêt aléatoire a de meilleures performances prédictives que l'arbre de décision sur l'ensemble de test, en raison de sa capacité à réduire la variance.

Nous avons également tester les méthodes du Gradien boosting qui se trouve à la fin du script R, cela nous montre des résultats plus performant en terme de MSE que les autres méthodes.