Rapport du TP1 : Analyse de Données et Méthodes d'Ensemble

Kaced Faycal

27 mars 2025

1 Introduction

Ce rapport présente les résultats et interprétations obtenus dans le cadre du TP1 sur l'analyse de données et les méthodes d'ensemble. L'objectif était d'explorer un jeu de données issu d'un élevage de poulets et d'appliquer diverses techniques pour analyser et modéliser ces données.

2 Analyse exploratoire des données

2.1 Statistiques descriptives

Les statistiques descriptives montrent une moyenne du poids des poulets de **2509.58 g** avec un écart-type de **898.44**. La médiane est de **2481.5 g**. Cela indique une variabilité notable dans les poids, ce qui est confirmé par la présence d'outliers.

2.2 Détection des outliers

La méthode IQR a identifié :
— **10 outliers** pour le poids des poulets.
— **14 outliers** pour la nourriture consommée.
— Aucun outlier pour la température.
La méthode Z-Score a identifié davantage d'outliers :
— **26 outliers** pour le poids des poulets et la nourriture consommée.
— **22 outliers** pour la température.

2.3 Tests paramétriques

Le test de Shapiro-Wilk indique que **les trois variables ne suivent pas une distribution normale** (p-value < 0.05). Les tests t de Student et ANOVA n'ont pas révélé de différences significatives entre les groupes (p-value $= 0.55$ et 0.59 respectivement).

3 Réduction de dimensionnalité

3.1 Analyse en Composantes Principales (ACP)

L'ACP a permis d'identifier les principales directions de variation dans les données. La projection sur les deux premières composantes a expliqué une grande partie de la variance totale, ce qui a permis de simplifier les données tout en conservant les principales informations.

3.2 ACP à noyau

L'ACP à noyau, appliquée avec divers noyaux (linéaire, RBF et polynomial), a montré que les noyaux non linéaires améliorent la séparation des classes lorsque les relations sont complexes.

4 Méthodes d'ensemble

4.1 Bagging

La forêt aléatoire a obtenu de bonnes performances en prédisant la survie des poulets. L'analyse des variables importantes a révélé que les paramètres liés à la température et à l'alimentation étaient cruciaux.

4.2 Boosting

La comparaison entre AdaBoost et Gradient Boosting a montré que ce dernier était plus robuste face aux outliers, en raison de son approche progressive dans la correction des erreurs.

5 Conclusion

Ce TP a permis de mettre en pratique diverses techniques d'analyse de données et de modélisation. Les résultats obtenus soulignent l'importance de la préparation des données et du choix des modèles en fonction des caractéristiques des données.