Ml OIVM / TP1 ATDN

Analyse de Données et Méthodes d'Ensemble

Les objectifs pédagogiques de ce TP sont les suivants :

- Maîtriser les bases de l'analyse exploratoire des données (statistiques descriptives, détection des outliers, tests statistiques)
- Comprendre les méthodes de réduction de dimensionnalité (ACP, ACP à noyau)
- Implémenter des méthodes d'ensemble (Bagging, Boosting)
- Appliquer ces techniques sur des données réelles issues d'un élevage de poulets
- Comparer les performances des différentes approches

Attention : Ce que je note, ce sont **vos** travaux à **vous**, votre compréhension du problème et des solutions que **vous** proposez (et non ce qui est généré par LLM : Gemini, GPT et consorts).

Partie 1 : Analyse exploratoire des données

Exercice 1 : Statistiques descriptives (1 points)

1. (0.5 pts) Calculez la moyenne, médiane, écart-type, variance et les quartiles pour les variables poids, nourriture et température.
2. (0.5 pts) Tracez des histogrammes et des boxplots pour visualiser la répartition des données. Que pouvez-vous déduire de ces graphiques ? Les données semblent-elles homogènes ou dispersées ?

Exercice 2 : Détection des outliers (3 points)

3. (1,5 pts) Détectez les outliers avec la méthode de l'écart interquartile (IQR) et la méthode du Z-Score. Comparez les résultats.
4. (1,5 pts) Visualisez ces outliers sur un boxplot annoté. Les outliers détectés sont-ils réalistes ou issus d'erreurs de mesure ? Faut-il les exclure ou les garder ? Justifiez votre choix.

Exercice 3 : Tests paramétriques (4 points)

5. (2 pts) Testez la normalité des variables (poids, nourriture, température) avec le test de Shapiro-Wilk. Expliquez ce que vous observez.
6. (2 pts) Comparez les moyennes de deux groupes avec le test t de Student, puis utilisez une ANOVA pour comparer les moyennes de plusieurs groupes. Interprétez les résultats.

Partie 2 : Réduction de dimensionnalité

Exercice 4 : Analyse en Composantes Principales (ACP) (3 points)

7. (1,5 pts) Implémentez une ACP sans scikit-learn (avec numpy). Calculez la matrice de covariance, les valeurs propres et les vecteurs propres.
8. (1,5 pts) Projetez les données sur les deux premières composantes principales et visualisez le résultat. Combien de composantes gardez-vous ? Justifiez.

Exercice 5 : ACP à Noyau (3 points)

9. (1,5 pts) Appliquez KernelPCA (avec scikit-learn) sur les données et testez différents noyaux (linéaire, RBF, polynomial).
10. (1,5 pts) Comparez les résultats avec l'ACP classique. Dans quels cas l'ACP à noyau donne-t-elle de meilleurs résultats ?

Partie 3 : Méthodes d'ensemble

Exercice 6 : Bagging (3 points)

11. (1,5 pts) Implémentez une forêt aléatoire (RandomForestClassifier) pour prédire la survie des poulets. Analysez les performances (accuracy, F1-score).
12. (1,5 pts) Identifiez les variables les plus importantes. Quels attributs influencent le plus la survie des poulets ? Pourquoi ?

Exercice 7 : Boosting (3 points)

13. (1,5 pts) Comparez AdaBoost et Gradient Boosting sur la prédiction du gain de poids. Analysez leurs performances.
14. (1,5 pts) Les deux algorithmes réagissent-ils différemment aux outliers ? Expliquez pourquoi.