Rapport TP2 ATDN :

Analyse de Données et Méthodes d'Ensemble

Partie 1 : Analyse exploratoire des données

Exercice 1 : Statistiques descriptives

1. Calcule de la moyenne, médiane, écart-type, variance et les quartiles pour les variables poids, nourriture et température.

Poids :

```
La moyenne, médiane, écart-type, variance et les quartiles pour la variable Poids_poulet_g
Moyenne Poids_poulet_g : 2509.58
Médiane Poids_poulet_g : 2481.5
Variance Poids_poulet_g : 807188.8176884422
Ecart-type Poids_poulet_g : 898.4368746263937
Quartiles Poids_poulet_g : 0.25    1810.75
0.50    2481.50
0.75    3356.50
Name: Poids_poulet_g, dtype: float64
```

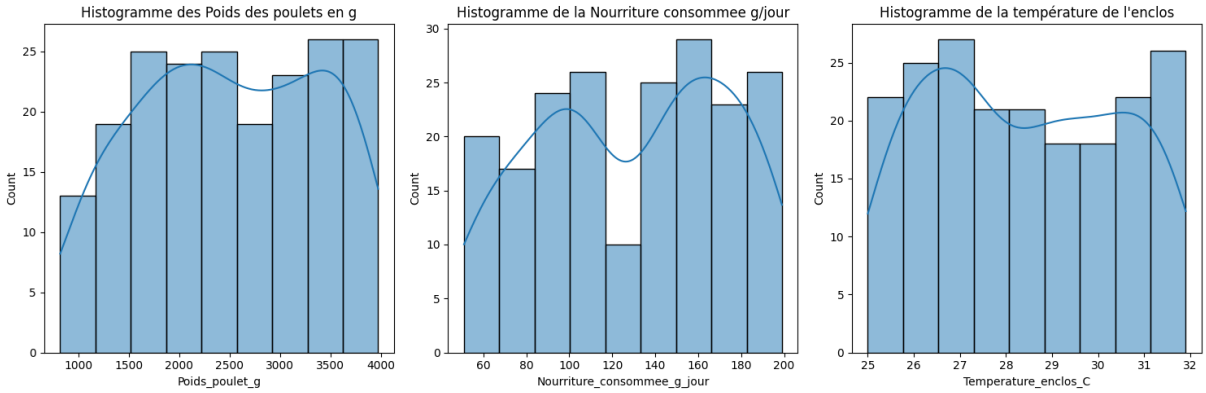
Nourriture :

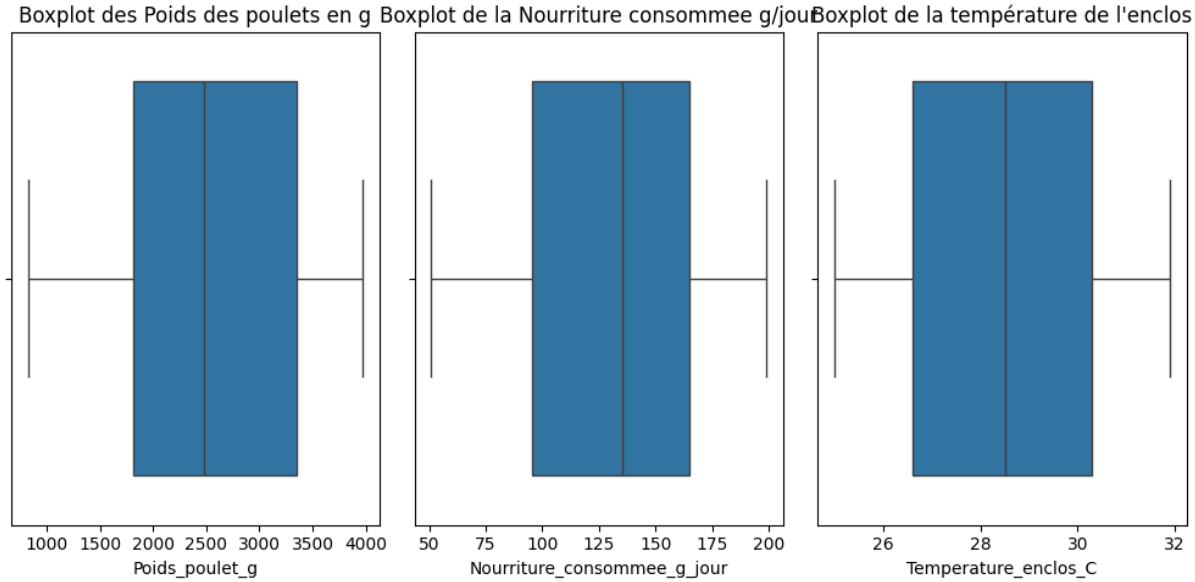
```
La moyenne, médiane, écart-type, variance et les quartiles pour la variable Nourriture_consommee_g_jour
Moyenne Nourriture_consommee_g_jour : 129.745
Médiane Nourriture_consommee_g_jour : 135.5
Variance Nourriture_consommee_g_jour : 1936.542688442211
Ecart-type Nourriture_consommee_g_jour : 44.00616648200808
Quartiles Nourriture_consommee_g_jour : 0.25     95.75
0.50    135.50
0.75    165.25
Name: Nourriture_consommee_g_jour, dtype: float64
...
```

Température :

```
La moyenne, médiane, écart-type, variance et les quartiles pour la variable Temperature_enclos_C
Moyenne Temperature_enclos_C : 28.389
Médiane Temperature_enclos_C : 28.5
Variance Temperature_enclos_C : 4.2672150753768845
Ecart-type Temperature_enclos_C : 2.0657238623245084
Quartiles Temperature_enclos_C : 0.25    26.6
0.50    28.5
0.75    30.3
Name: Temperature_enclos_C, dtype: float64
```

2. Tracé 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 ?

Interprétation des histogrammes et des boxplots :

Poids des poulets (g) :

- La distribution est asymétrique avec des valeurs élevées vers le droit donc présence

Nourriture consommée (g/jour) :

- La distribution est presque homogène avec une faible dispersion.

Température de l'enclos (°C) :

- Les valeurs sont regroupées autour de la médiane, donc on a peu de variabilité.
- Les températures sont homogènes.

Exercice 2 : Détection des outliers

3. Détection des outliers avec la méthode de l'écart interquartile (IQR) et la méthode du Z-Score

```
Outliers détectés avec IQR :
Poids_poulet_g: 10 outliers
Nourriture_consommee_g_jour: 14 outliers
Temperature_enclos_C: 0 outliers

Outliers détectés avec Z-Score : 159
```

Comparaison des résultats :

Pour détecter des outliers il faut rendre les deux méthodes plus strictes en baissant le seuil de l'écart-type (3 à 1) pour la méthode Z-Score et le facteur multiplicatif (1.5 à 0.5) de l'IQR pour la méthode interquartile.

IQR **vs** Z-Score :

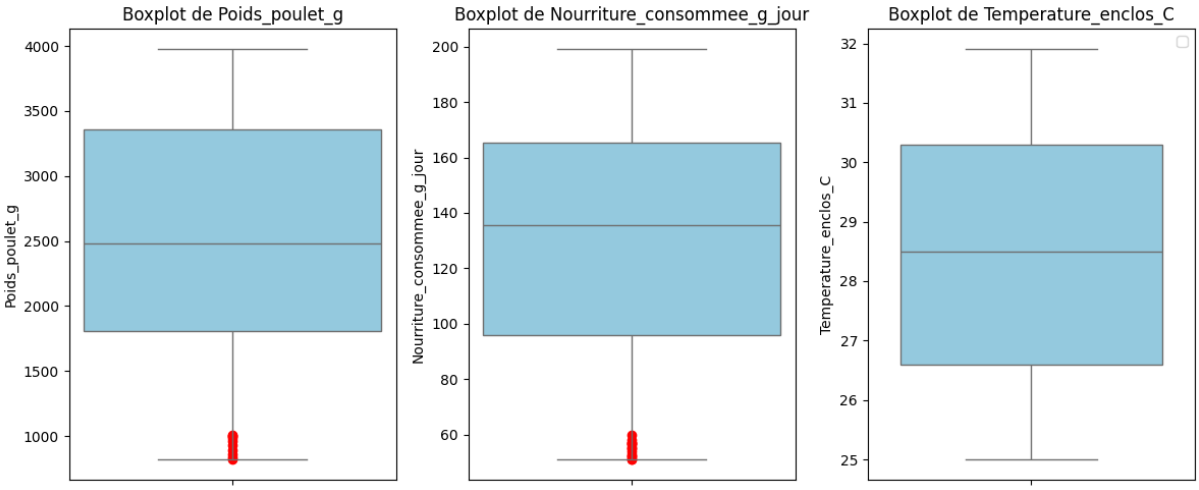
L'IQR est plus robuste car il n'est pas affecté par la distribution des données.

Le Z-Score est utile si les données suivent une loi normale, mais peut être trop sensible aux distributions asymétriques.

- Si une variable est très asymétrique, l'IQR peut ne pas bien fonctionner.

- Si elle est normalement distribuée, alors le Z-score est souvent plus efficace.

4. Visualisation des outliers sur un boxplot annoté.



Analyse des outliers :

- Outliers réalistes (valeurs extrêmes mais logiques)

Par exemple, si un poulet beaucoup plus gros ou plus petit que la moyenne est détecté comme outlier, il peut être réaliste (poulet malade, croissance différente).

Si la nourriture consommée est beaucoup plus basse ou haute, cela peut être lié à une condition environnementale (température, alimentation).

Température de l'enclos : Une valeur extrême peut être due à une période de canicule ou de froid inhabituel, ce qui est possible.

Les outliers sont réalistes, donc on peut les conserver. Ils peuvent être analysés et expliqués pour en tirer des insights.

- Outliers non réaliste

Mais si les outliers ne sont pas réalistes ou ne nous servent pas dans notre analyse on doit les enlever pour pas qu'ils faussent les résultats.

Exercice 3 : Tests paramétriques

5. Teste de normalité des variables (poids, nourriture, température) avec le test de ShapiroWilk.

Le test de Shapiro-Wilk permet de vérifier si une variable suit une distribution normale. L'hypothèse nulle (H0) est que les données suivent une loi normale.

Interprétation :

Si la p-valeur > 0.05 -> On ne rejette pas H0, les données sont normalement distribuées.

Si la p-valeur < 0.05 -> On rejette H0, les données ne suivent pas une distribution normale.

```
Résultats des tests de normalité :
                     Variable  Statistique       p-value Normalité
0               Poids_poulet_g     0.956822  9.098264e-06       Non
1  Nourriture_consommee_g_jour     0.944871  6.230564e-07       Non
2         Temperature_enclos_C     0.943210  4.406064e-07       Non

Test t entre jeunes et agés:
t-statistique: 0.145, p-value: 0.885
Conclusion: Pas de différence significative

ANOVA entre groupes d'âge:
F-statistique: 0.131, p-value: 0.877

Corrélation Poids-Nourriture:
Corrélation: -0.082, p-value: 0.251
```

Observation :

Les résultats des tests de normalité montrent que les données sont clairement non normales (p-values très faibles).

6. Comparaison des moyennes de deux groupes avec le test t de Student, puis avec une ANOVA pour comparer les moyennes de plusieurs groupes.

```
Test t entre jeunes et agés:
t-statistique: 0.145, p-value: 0.885
Conclusion: Pas de différence significative

ANOVA entre groupes d'âge:
F-statistique: 0.131, p-value: 0.877

Corrélation Poids-Nourriture:
Corrélation: -0.082, p-value: 0.251
```

Test t entre jeunes et âgés:

- La p-value > 0.05 indique qu'il n'y a aucune différence statistiquement significative de poids entre les poulets jeunes (<60 jours) et agés (≥60 jours).
- La très faible valeur de t (proche de 0) montre que les moyennes des deux groupes sont quasi identiques.

Conclusion : L'âge ne semble pas influencer significativement le poids des poulets dans notre échantillon.

ANOVA entre groupes d’âge :

- La p-value > 0.05 confirme qu'il n'y a aucune différence significative de poids entre les 3 groupes d'âge (0-40j, 41-80j, 81-120j).
- La valeur F très faible (<1) indique que la variance entre groupes est négligeable comparée à la variance intra-groupe.

Conclusion : La stratification en 3 groupes ne révèle pas non plus d'effet de l'âge sur le poids.

Corrélation Poids-Nourriture :

La corrélation négligeable (proche de 0) et non significative (p > 0.05) montre que :

- Il n'y a pas de lien linéaire entre la quantité de nourriture consommée et le poids des poulets.
- La tendance légèrement négative (-0.08) est trop faible pour être interprétable.

Cela ne signifie pas nécessairement que la nourriture n'a aucun impact (cette relation peut être non linéaire ou médiée par d'autres variables).

Synthèse globale :

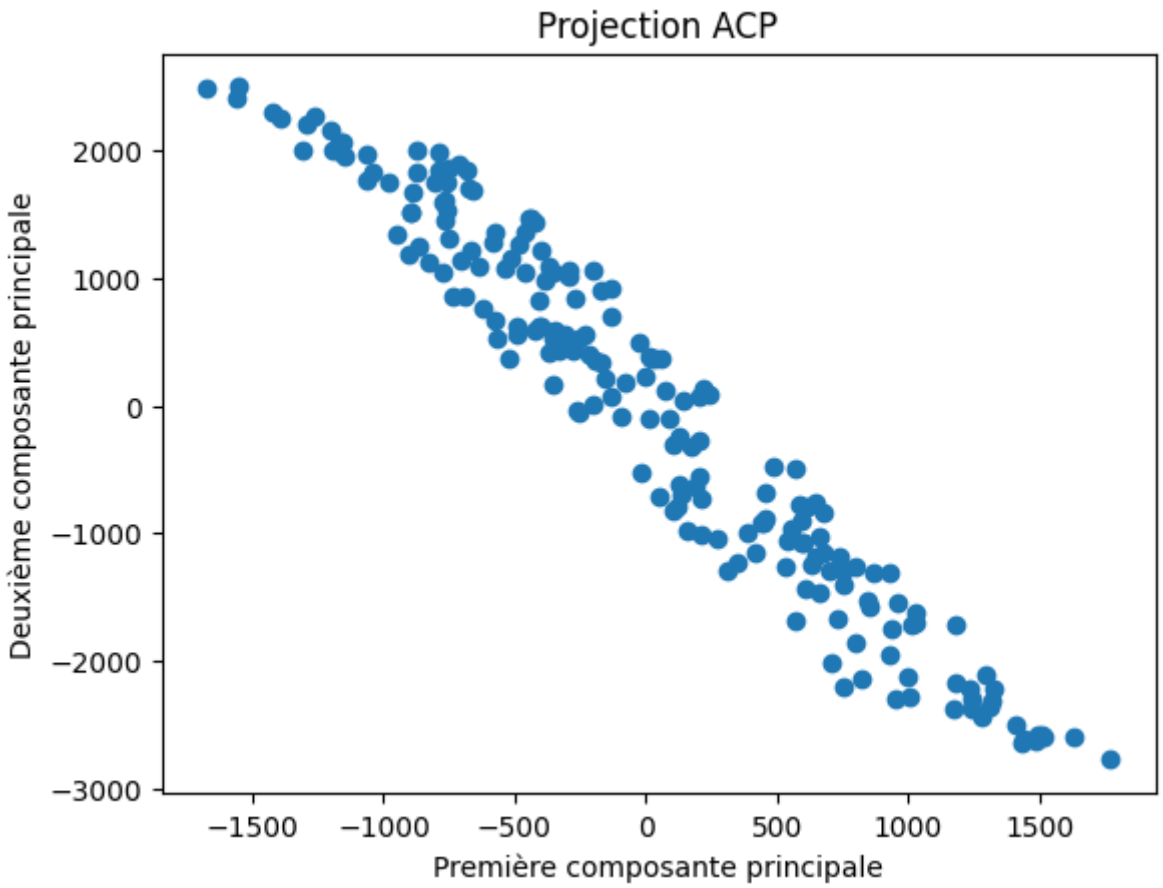
- Aucun des facteurs testés (âge, nourriture) ne montre d'impact significatif sur le poids des poulets dans notre dataset.
- Les résultats sont cohérents entre méthodes (test t, ANOVA, corrélation), renforçant leur fiabilité.

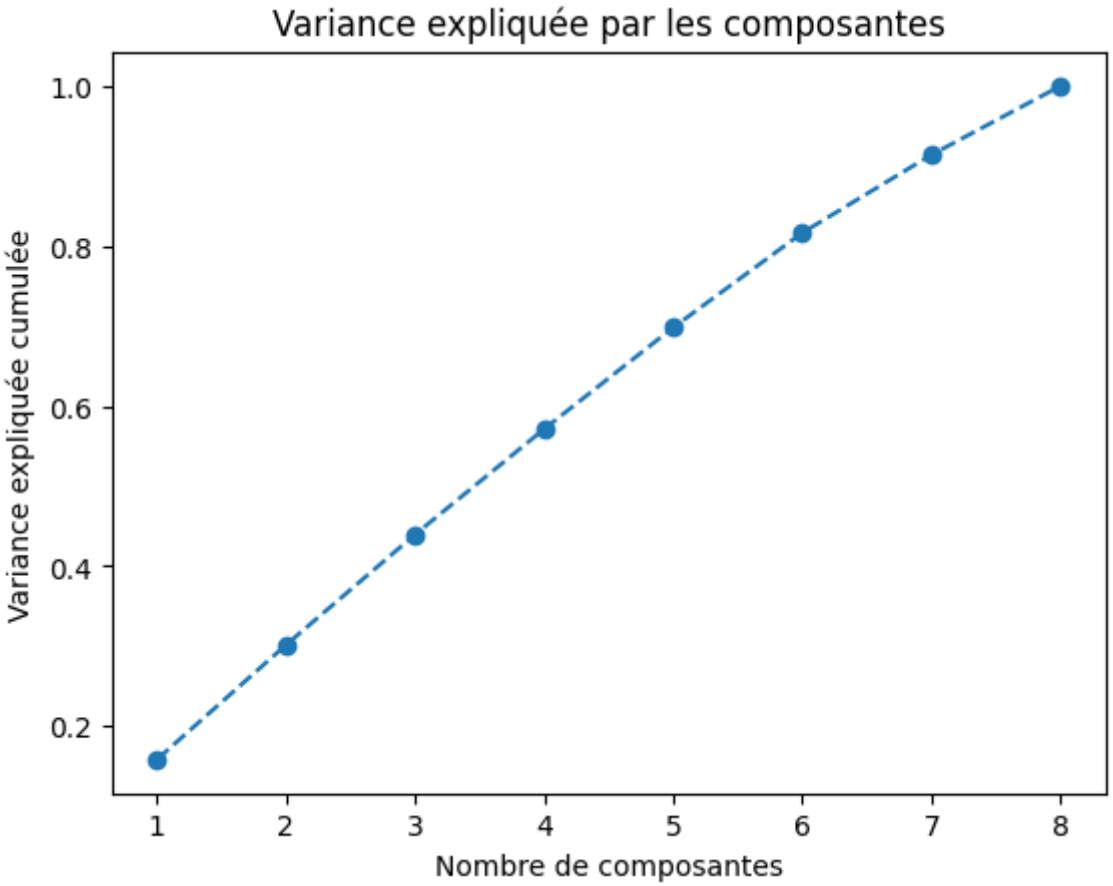
Partie 2 : Réduction de dimensionnalité

Exercice 4 : Analyse en Composantes Principales (ACP)

7. Implémentation d'une ACP sans scikit-learn (avec numpy). Calcule de la matrice de covariance, les valeurs propres et les vecteurs propres.

8. Projection des données sur les deux premières composantes principales et visualisation du résultat.


Projection ACP
Deuxième composante principale
Première composante principale
2000
1000
0
−1000
−2000
−3000
−1500
−1000
−500
0
500
1000
1500


Variance expliquée par les composantes
Variance expliquée cumulée
Nombre de composantes
1.0
0.8
0.6
0.4
0.2
1
2
3
4
5
6
7
8

Nombre de composantes :

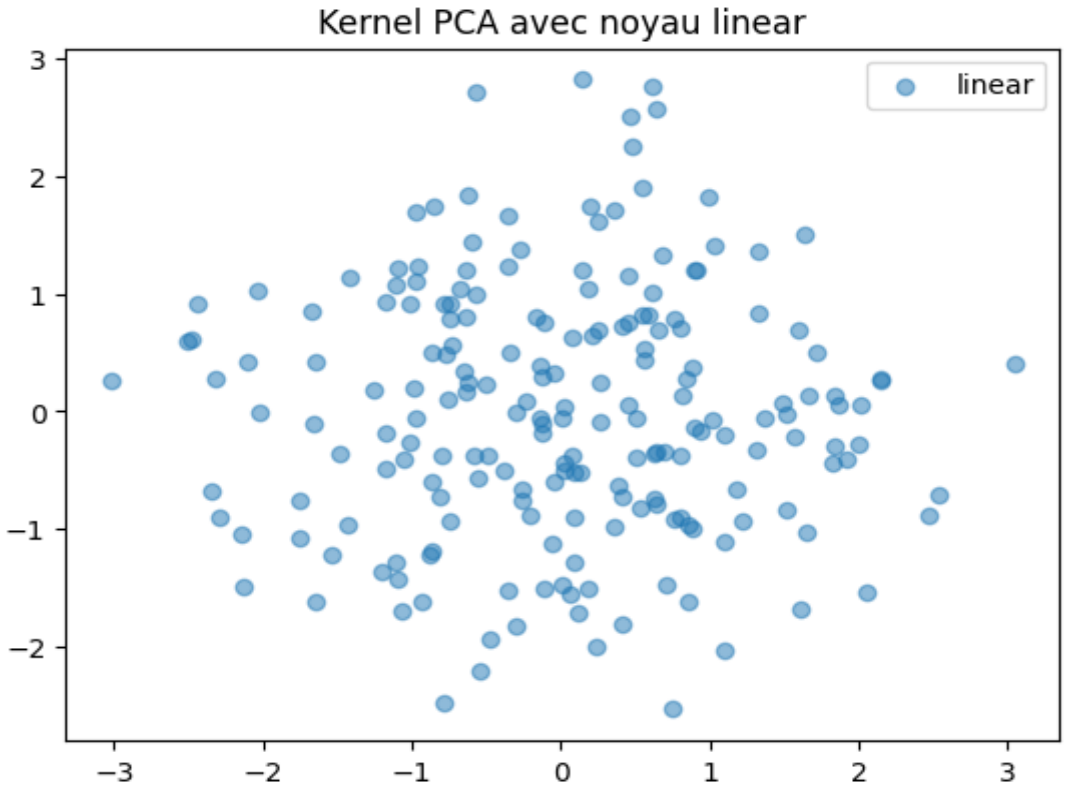
On conserve 5 composantes, car à partir de la cinquième, on observe une inflexion sur la courbe, et ce nombre permet de capturer 70 % de l'information, ce qui constitue un excellent résultat.

Pour obtenir 80-90% de la variance, on peut prendre jusqu'à 7 composantes.

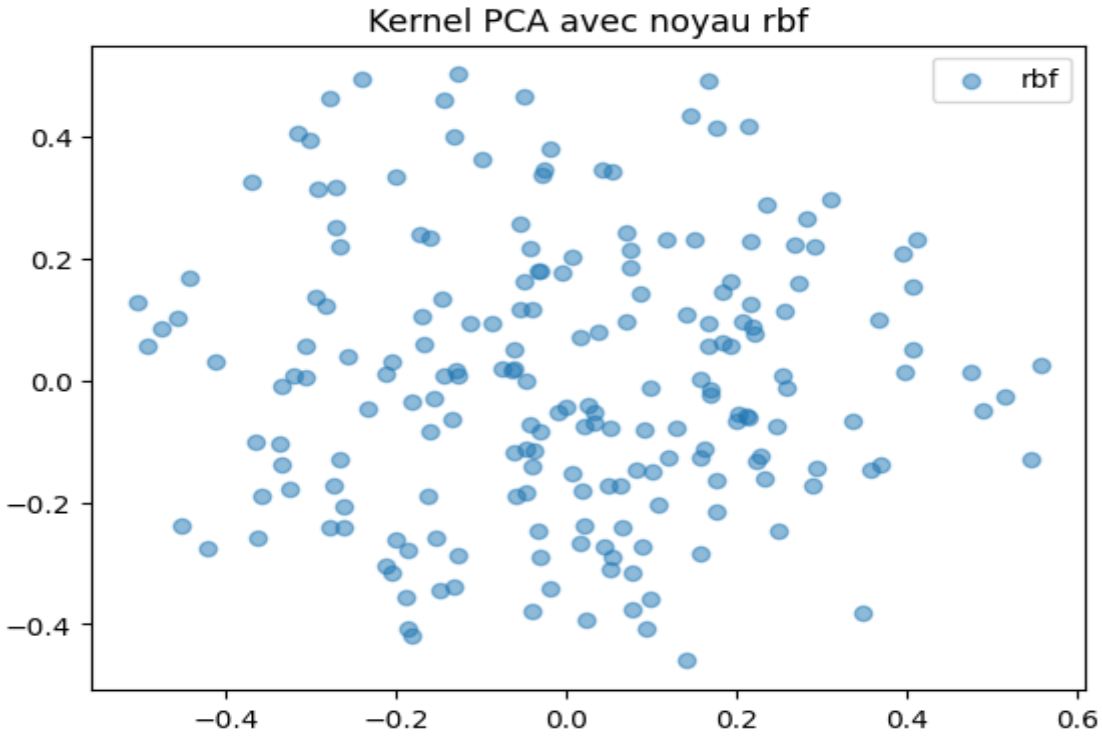
Exercice 5 : ACP à Noyau

9. Application de Kernel PCA (avec scikit-learn) sur les données et teste de différents noyaux (linéaire, RBF, polynomial).

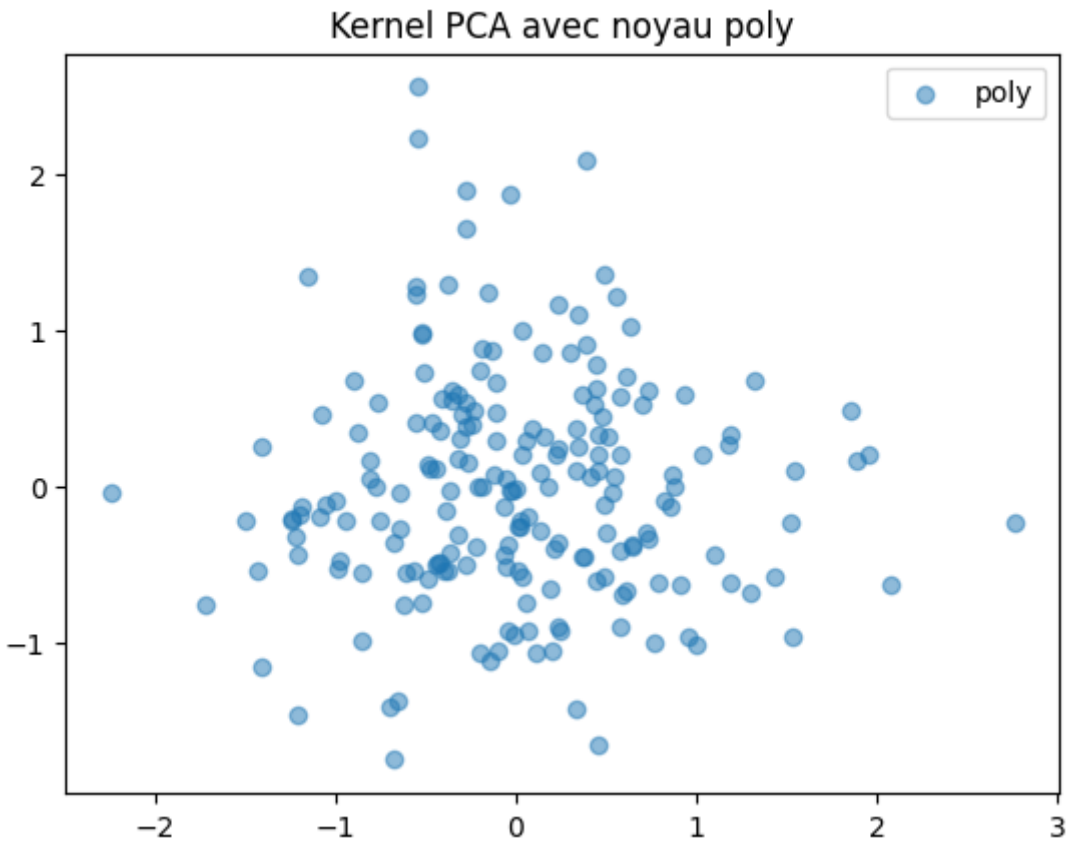
Kernel PCA avec Noyau linéaire :



Kernel PCA avec Noyau linéaire :



Kernel PCA avec Noyau Polynomial :



10. Comparaison des résultats avec l'ACP classique.

L'ACP à noyau peut mieux capturer les structures non linéaires des données (surtout avec un noyau RBF ou polynomial).

Si la relation entre les variables ne peut pas être bien capturée par une combinaison linéaire (comme dans l'ACP classique), Kernel PCA peut projeter ces données dans un espace de dimension supérieure où elles deviennent linéairement séparables.

Dans quels cas l'ACP à noyau donne-telle de meilleurs résultats ?

Elle est plus adaptée pour :

- Des données avec des relations non linéaires
- La classification de données complexes
- Des données de grande dimension avec des structures cachées
- La réduction de dimension pour des modèles non linéaires

Partie 3 : Méthodes d'ensemble

Exercice 6 : Bagging

11. Implémentation d'une forêt aléatoire (RandomForestClassifier) pour prédire la survie des poulets.

Analyse des performances (accuracy, F1-score).

```
Accuracy: 0.583
F1-score: 0.545
```

Accuracy :

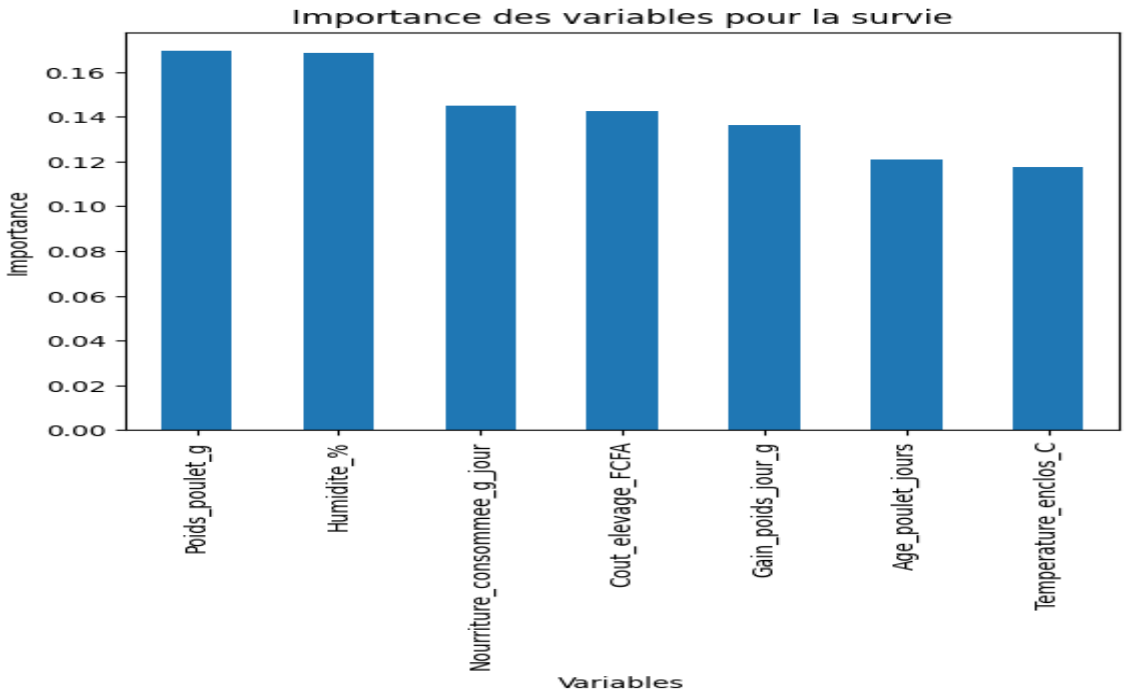
- Le modèle classe correctement 58.3% des poulets (survie vs non-survie).
- Légère amélioration, mais insuffisante pour des décisions critiques.

F1-score :

- Combinaison de la précision et du rappel.
- Indique que le modèle a des difficultés à équilibrer les prédictions positives/négatives.

12. Les variables les plus importantes.

Attributs influencent le plus la survie des poulets



Les attributs qui influencent le plus la survie des poulets sont le poids des poulet et l'humidité.

- Le poids est souvent un indicateur de la santé et du développement des poulets. Un poids trop faible peut signaler une malnutrition ou une mauvaise croissance, ce qui peut rendre les poulets plus vulnérables aux maladies et réduire leur espérance de vie.

- L'humidité joue un rôle crucial dans l'environnement de vie des poulets. Un taux d'humidité trop élevé ou trop bas peut affecter leur bien-être, leur respiration et leur capacité à réguler leur température corporelle. Une humidité excessive peut également favoriser la croissance de moisissures ou de bactéries, augmentant le risque de maladies respiratoires ou d'infections.

Exercice 7 : Boosting

13. Comparaison AdaBoost et Gradient Boosting sur la prédiction du gain de poids.

Analyse de leurs performances.

```
MSE AdaBoost : 0.250
MSE Gradient Boosting : 0.281
R² AdaBoost : -0.004
R² Gradient Boosting : -0.130
```

Les deux modèles obtiennent des scores très proches.

AdaBoost est légèrement meilleur, mais la différence n'est pas significative.

On a un R2 négatif, ce qui signifie que les deux modèles sont moins performants qu'un simple modèle qui prédirait la moyenne des valeurs cibles.

Ceci peut s'expliquer par le fait que les variables n'expliquent pas bien la variable cible, donc les modèles ne pourront pas apprendre correctement.

14. Les deux algorithmes réagissent-ils différemment aux outliers ?

Oui les deux modèles réagissent différemment aux outliers :

AdaBoost : Sensible aux outliers

- Donne plus de poids aux erreurs à chaque itération.
- Les outliers mal prédits reçoivent des poids élevés, et ça peut fausser le modèle.
- Si un point a une erreur de prédiction importante, AdaBoost va s'obstiner à le corriger au détriment des autres points.

Gradient Boosting : Plus robuste

- Minimise les résidus (erreurs) étape par étape sans pondération explicite.
- Utilise des gradients pour corriger progressivement les erreurs, ce qui atténue l'impact des outliers.

- Un outlier influencera moins la direction globale de l'optimisation.

Conclusion

Dans ce TP, nous avons analysé les facteurs influençant la croissance et la survie des poulets à travers des méthodes statistiques et d'apprentissage automatique.

L'analyse descriptive nous a aidés à identifier tendances et anomalies, tandis que les tests statistiques (Shapiro-Wilk, test t, ANOVA) ont montré que ni l'âge ni la nourriture n'avaient d'impact significatif sur le poids.

L'ACP et l'ACP à noyau ont facilité la représentation des données, et les modèles de Bagging et Boosting ont révélé que le poids et l'humidité étaient des facteurs clés de survie. Enfin, nous avons constaté que le Gradient Boosting était plus robuste aux outliers qu'AdaBoost même si les données actuelles ne permettent pas de faire une bonne prédiction.

Ce TP nous a permis de mieux comprendre l'analyse statistique, la réduction de dimension et les modèles prédictifs, des compétences essentielles en data science.