Rapport de Projet : Analyse en Composantes Principales (ACP)

---

1. Introduction

Ce projet explore l'application de l'Analyse en Composantes Principales (ACP) sur l'ensemble de données *Iris*, un ensemble classique en apprentissage machine. L'objectif principal est de réduire la dimensionnalité des données tout en conservant leur structure essentielle, facilitant ainsi leur interprétation et leur visualisation.

L'analyse repose sur l'utilisation de bibliothèques Python standard telles que **NumPy**, **Pandas**, **Matplotlib**, et **Scikit-learn**. Une comparaison entre une implémentation personnalisée de l'ACP et celle fournie par Scikit-learn est également présentée. Ce rapport détaille les étapes suivies, les résultats obtenus et les conclusions tirées.

---

2. Description de la méthodologie

a) Chargement des données

Les données *Iris* sont chargées à l'aide de la fonction `load_iris` de Scikit-learn. Elles comprennent 150 échantillons appartenant à trois classes différentes : *Setosa*, *Versicolor*, et *Virginica*. Les caractéristiques incluent :

- La longueur des sépales,
- La largeur des sépales,
- La longueur des pétales,
- La largeur des pétales.

Les données sont converties en un DataFrame Pandas pour simplifier leur manipulation. Les caractéristiques (épaisseur des pétales, longueur des sépales, etc.) sont stockées dans la variable **X**, tandis que les classes des échantillons sont conservées dans **y**.

b) Application de l'ACP

- **Prétraitement des données** : Avant de procéder à l'ACP, les données sont centrées et réduites afin d'éviter qu'une caractéristique domine les autres en raison de son échelle.

- **Modèle personnalisé** :

    1. La matrice de covariance est calculée pour capturer les relations linéaires entre les caractéristiques.
    2. Les valeurs propres et vecteurs propres sont extraits pour déterminer les directions principales de variation.
    3. Les données sont projetées dans l'espace des composantes principales.

- **Modèle Scikit-learn** : L'ACP est également effectuée à l'aide de la classe `PCA` de Scikit-learn, qui automatise ces calculs et permet de vérifier la cohérence des résultats.

c) Visualisation

Les données transformées sont projetées sur les deux premières composantes principales. Des visualisations suivantes sont réalisées :

- **Nuage de points 2D** : Les échantillons sont colorés selon leurs classes pour analyser leur séparabilité.
- **Graphique de variance expliquée** : Une courbe montre la proportion de variance capturée par chaque composante principale.
- **Visualisation 3D** : Les données sont représentées dans un espace tri-dimensionnel pour une meilleure interprétation.

---

3. Résultats et analyse

a) Variance expliquée

L'analyse montre que les deux premières composantes principales expliquent environ 95 % de la variance totale des données. Cela indique que ces deux composantes sont suffisantes pour représenter les données tout en minimisant la perte d'information.

b) Visualisation des données

- **Nuage de points 2D** : Les classes *Setosa*, *Versicolor*, et *Virginica* sont bien séparées le long des deux premières composantes, en particulier *Setosa* qui est facilement distinguable.
- **Visualisation 3D** : Cette représentation permet d'observer des patterns supplémentaires, bien que la troisième composante ajoute peu d'information supplémentaire.

c) Validation

La comparaison entre l'implémentation personnalisée et celle de Scikit-learn montre des résultats identiques. Cela valide la méthode mise en œuvre et démontre la compréhension approfondie des concepts mathématiques derrière l'ACP.

---

4. Conclusion

Ce projet illustre l'utilité de l'ACP comme outil d'analyse exploratoire et de réduction de dimensionnalité. Les résultats obtenus mettent en évidence les relations entre les

caractéristiques et permettent de visualiser efficacement les données dans un espace à dimension réduite.

La validation par un modèle prédéfini renforce la crédibilité des résultats, tandis que l'implémentation personnalisée montre une maîtrise des aspects mathématiques et pratiques de l'ACP. Cette expérience pourrait être étendue à d'autres ensembles de données ou à des problèmes où la réduction de dimensionnalité est essentielle.