Découverte des données Iris à travers l'ACP et le KMeans

Bonjour, nous allons aujourd'hui nous plonger dans l'analyse du jeu de données Iris, une référence dans le domaine de l'apprentissage automatique. L'objectif est d'explorer la structure cachée des données sans nous baser sur les étiquettes préétablies. Pour cela, nous utiliserons deux méthodes : l'Analyse en Composantes Principales (ACP) pour simplifier et mieux visualiser les données, et le clustering KMeans pour tenter de regrouper les fleurs d'Iris selon leurs caractéristiques physiques, mimant ainsi une situation réelle de découverte de catégories au sein d'un ensemble de données non étiquetées.

Prétraitement des données

Avant de commencer notre analyse, il est essentiel de standardiser les données. Cette étape permet de s'assurer que chaque caractéristique est prise en compte de manière équitable, évitant ainsi que les caractéristiques les plus importantes dominent indûment les résultats. Cette standardisation est un gage de fiabilité pour nos résultats de clustering et améliore la précision de la représentation de la structure des données dans l'espace réduit.

Évaluation de notre démarche

Pour mesurer l'efficacité de notre approche, nous nous appuyons sur le score de silhouette et l'indice de Calinski-Harabasz. Le score de silhouette nous permet d'évaluer dans quelle mesure chaque fleur est bien placée dans son groupe par rapport aux autres groupes, témoignant ainsi de la pertinence de notre clustering. Quant à l'indice de Calinski-Harabasz, il mesure la validité du clustering en comparant la dispersion des données à l'intérieur des clusters avec celle entre les clusters. Des scores élevés sur ces deux mesures indiqueraient que notre clustering reflète fidèlement les structures naturelles présentes dans le jeu de données, validant ainsi notre méthode d'apprentissage non supervisé.

Conclusion

L'application de l'ACP et du KMeans au jeu de données Iris ne démontre pas seulement la capacité de l'apprentissage non supervisé à révéler des informations significatives à partir de données non étiquetées. Elle illustre également l'utilité pratique de ces techniques dans des contextes où les étiquettes claires ne sont pas disponibles. Notre analyse a permis de mettre en lumière des motifs et des regroupements au sein des données