RAPPORT DU PROJET WINE DATASET

Liadi, Alim Akorede

I- Problématique et difficultés rencontrées

L'objectif principal de ce projet était d'explorer différentes techniques de prétraitement, de réduction de dimension, de classification et de validation afin de mieux comprendre comment les données influencent la performance des modèles d'apprentissage supervisé. L'une des difficultés majeures rencontrées concernait l'intégration et l'interprétation des différentes transformations appliquées aux données, notamment la normalisation, la gestion des valeurs manquantes et la réduction de dimension. La comparaison entre plusieurs méthodes (PCA, LDA, KNN) a également demandé une compréhension plus approfondie de leur fonctionnement.

II- Choix techniques

Plusieurs choix techniques ont été effectués afin d'assurer la cohérence du projet. Pour la classification, l'algorithme KNN (k=5) a été choisi en raison de sa simplicité et de sa capacité à s'adapter aux données du dataset Wine, qui présente trois classes distinctes. Pour la réduction de dimension, les méthodes PCA et LDA ont été retenues : PCA pour capturer la variance globale des données et LDA pour maximiser la séparation entre les classes. La normalisation via StandardScaler a été appliquée systématiquement pour garantir une pondération équitable des caractéristiques, notamment avant PCA, LDA et KNN. Enfin, la validation croisée K-fold a été utilisée pour vérifier la robustesse des modèles, car elle est particulièrement adaptée aux petits ensembles de données.

III- Justification de l'implémentation et des choix

L'utilisation de KNN a été justifiée par sa capacité à fonctionner efficacement avec des données normalisées et par sa sensibilité directe aux distances, ce qui permet de bien évaluer l'effet des transformations. La gestion des valeurs manquantes a été réalisée à l'aide de deux méthodes (imputation par la moyenne et imputation par KNN), afin d'observer leur impact sur la classification et de déterminer la méthode la plus plausible pour ce type de données. Le choix de tester plusieurs dimensions avec PCA a permis d'évaluer la quantité minimale d'information nécessaire pour maintenir de bonnes performances. De même, LDA a été utilisé pour vérifier si une réduction supervisée améliorait la classification. Finalement, la validation croisée a été implémentée pour éliminer les risques de surapprentissage et obtenir une évaluation fiable.

VI- Discussion et explication des résultats

Les résultats montrent que la normalisation et la qualité des données jouent un rôle essentiel dans la stabilité des modèles. Sans réduction de dimension, le modèle KNN atteint une précision très élevée (près de 97–99%), ce qui confirme que les caractéristiques du dataset Wine sont suffisamment discriminantes. Après réduction par PCA, les performances varient

selon le nombre de composantes conservées : avec seulement deux composantes, la précision diminue, ce qui indique une perte d'information importante. Toutefois, lorsque 5 ou 7 composantes sont conservées, la performance se rapproche de celle du modèle original, ce qui montre que PCA est efficace à condition de ne pas réduire trop agressivement la dimension.

De son côté, LDA offre des résultats particulièrement intéressants. En réduisant les données à seulement deux dimensions (limite théorique liée aux trois classes du dataset), LDA atteint une précision supérieure ou égale à PCA dans la plupart des configurations. Cela s'explique par la nature supervisée de LDA, qui maximise directement la séparation entre les classes pendant la transformation. La validation croisée confirme également cette observation, puisque la variance des scores est faible avec LDA, indiquant une excellente robustesse. En revanche, PCA avec peu de dimensions montre une variance plus forte, révélant une sensibilité aux variations dans les données.

V- Améliorations futures

Plusieurs pistes d'amélioration peuvent être envisagées pour approfondir ce projet. L'utilisation d'autres classificateurs tels que SVM ou Random Forest pourrait permettre de comparer la performance de modèles plus complexes avec KNN. De plus, une analyse plus fine de la sélection de caractéristiques pourrait être intégrée afin d'identifier automatiquement les attributs les plus pertinents. Il serait également pertinent d'essayer des techniques avancées comme t-SNE ou UMAP pour la visualisation, ou encore des méthodes d'imputation plus sophistiquées pour gérer les valeurs manquantes. Enfin, l'optimisation systématique des hyperparamètres via GridSearchCV pourrait renforcer encore la robustesse des modèles testés.

VI- CONCLUSION

Ce projet a permis d'appliquer et de comparer plusieurs approches essentielles en apprentissage automatique. Les résultats montrent que les choix de prétraitement, d'imputation et de réduction de dimension influencent fortement la performance de la classification. La normalisation s'est révélée indispensable, tandis que LDA s'est démarquée comme la technique la plus efficace pour maintenir de bonnes performances même en dimension très réduite. Enfin, la validation croisée a confirmé la robustesse des modèles présentés. Ce travail offre une vision complète de l'impact des différentes étapes de préparation et de transformation des données sur un processus de classification supervisée.

Annexe

Voici le lien drive de la vidéo de présentation :
https://drive.google.com/file/d/138HvXo4Dz6fjvtZZnRWP4NltsjuzVylD/view?usp=sharing