Université du Québec en Outaouais

Département d’informatique et d’ingénierie

Projet: Analyse de vins (Wine dataset)

INF1823 – Introduction à la science des données

Enseignant: Étienne St-Onge

Membres du groupe :

Yamako Tiani Yoann Oliver YAMY88260005

Hien Florian Yann Ollo HIEF75340305

2025-11-25

1. Problématique et difficultés rencontrées

L'objectif du projet était d'explorer, d'analyser et de modéliser le jeu de données Wine à l'aide des outils Python présentés en classe. Les principales difficultés que nous avons rencontrées concernaient la compréhension des questions du projet et la traduction des consignes en opérations concrètes dans Scikit-learn. Par exemple, comprendre comment combiner la corrélation avec la transformée de Fourier nécessitait une interprétation logique, puisque cette dernière s'applique rarement à des données non temporelles.

D'autres difficultés ont concerné notre choix d'approches (quelle méthode utiliser pour la réduction de dimension ou l'imputation des données) ainsi que la gestion de la syntaxe de certaines librairies, notamment TSNE, LDA et KNNImputer.

2. Choix techniques

Les choix techniques ont été guidés par nos compétences actuelles et par la recherche d'un niveau d'implémentation avancé. Nous avons choisi les librairies NumPy, Pandas, Matplotlib, SciPy et Scikit-learn, qui offrent des outils robustes pour le traitement, la visualisation et la modélisation des données.

Pour la classification, nous avons utilisé le KNeighborsClassifier (simple, flexible et efficace pour les frontières complexes) et, pour la réduction de dimension, les méthodes PCA, LDA et t-SNE, afin de comparer les approches supervisées et non supervisées.

La validation croisée a été mise en œuvre pour assurer le juste-apprentissage et éviter le sur-apprentissage. Ce choix démontre l'intégration d'un concept avancé du cours et justifie la mention « niveau d'implémentation avancé ».

3. Justification de l'implémentation et des choix

Chaque section du projet a été conçue pour appliquer une technique du cours à un cas concret.

- Pour la corrélation, nous avons choisi deux variables présentant un coefficient supérieur à 0.5, ce qui montre une relation linéaire forte. La transformée de Fourier a été ajoutée pour visualiser la fréquence de variation des mesures, même si son interprétation reste mathématique plutôt que physique.
- Pour la classification, l'imputation des valeurs manquantes via KNNImputer nous a permis de conserver la structure locale des données sans déformer la distribution.
- Pour la réduction de dimension, la comparaison PCA / LDA / t-SNE illustre la perte d'information après compression et montre que LDA conserve mieux la séparation entre classes.
- Enfin, pour la robustesse, nous avons utilisé la validation croisée à 5 folds avec une régression logistique (modèle polynomial d'ordre 1) afin de garantir un juste-apprentissage

4. Explication des résultats (discussion)

a. Corrélation et classification

Les mesures d'alcool et de couleur dans le jeu de données Wine présentaient une corrélation de 0.69, indiquant une relation linéaire forte. Graphiquement, la pente positive de la régression confirme cette dépendance : lorsque l'intensité colorimétrique augmente, le taux d'alcool augmente aussi. La transformée de Fourier appliquée sur la série d'alcool illustre simplement les composantes fréquentielles, sans signification temporelle.

Lors de la classification, le modèle RandomForest a obtenu une précision supérieure à 90%, confirmant que les attributs du vin permettent de bien distinguer

les classes. Les valeurs imputées par KNN pour la nouvelle observation étaient cohérentes avec la moyenne des voisins, ce qui rend ces prédictions plausibles.

b. Réduction de dimension et robustesse

Après compression des données en deux dimensions avec PCA, la précision moyenne a légèrement diminué (d'environ 4%), ce qui montre une perte d'information liée à la compression. LDA, quant à lui, a conservé une meilleure performance car il maximise la séparation inter-classes. t-SNE a produit une excellente visualisation mais reste moins stable pour la prédiction.

La validation croisée a montré que la variance des scores restait faible (écart-type < 0.03), indiquant un modèle robuste. Après réduction, la performance moyenne baisse légèrement mais reste dans la marge du juste-apprentissage. Notre analyse confirme que la réduction de dimension doit être utilisée avec prudence, car elle simplifie le modèle au prix d'une compression parfois excessive de l'information.

5. Améliorations futures

Une amélioration possible serait d'ajouter un pipeline automatisé pour combiner le prétraitement, la réduction de dimension et la classification en un seul flux reproductible. Nous aimerions aussi intégrer des réseaux de neurones légers (MLPClassifier) pour comparer leur performance avec les méthodes linéaires. Une autre piste serait d'appliquer ces techniques à un jeu de données réel plus complexe (par exemple, la prédiction de vins à partir d'images ou de données capteurs) afin de tester la généralisation du modèle. Enfin, le projet pourrait être enrichi par une interface visuelle interactive (Dash / Streamlit) pour explorer dynamiquement la corrélation et les projections PCA/LDA.