2025

1

Rapport du Mini Projet ML-DL

ABDELHAK FADOUA

3LGL

Polytechnique Sousse

2024/202

1. Introduction

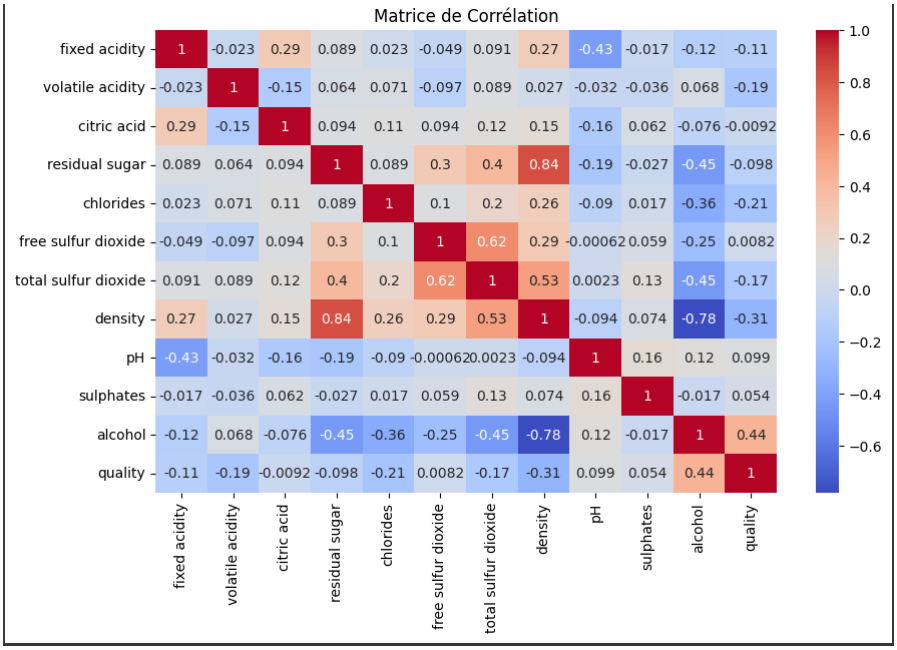
Ce projet vise à analyser un jeu de données sur la qualité des jus, en particulier leurs propriétés chimiques et leur qualité. L'objectif est de comprendre les relations entre les différentes caractéristiques chimiques des jus et leur qualité, puis de construire des modèles prédictifs pour estimer cette qualité. Ce projet s'inscrit dans mon intérêt pour les produits alimentaires sains et leur composition, car les jus sont souvent considérés comme des options saines et nutritives. Comprendre les facteurs qui influencent leur qualité peut aider à améliorer leur production et leur consommation.

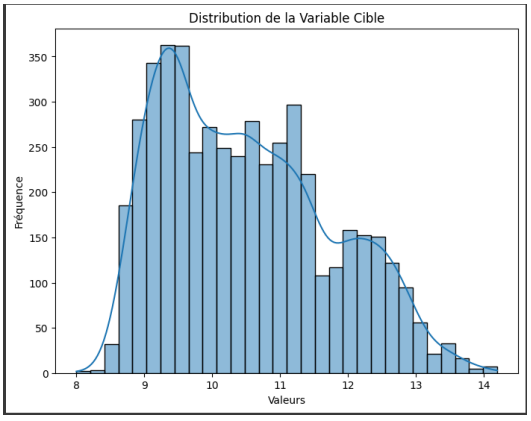
2. Description des Données

Le jeu de données utilisé, Juice Quality Dataset, contient plusieurs variables décrivant les propriétés chimiques des jus, ainsi que leur qualité. Les étapes suivantes ont été réalisées pour explorer et prétraiter les données :

Exploration des Données :

- Affichage des informations générales (df.info()) pour comprendre la structure des données.
- Statistiques descriptives (df.describe()) pour analyser la distribution des variables.
- Vérification des valeurs manquantes (df.isnull().sum()).
- Visualisation des Données :
  Histogrammes pour chaque variable afin de visualiser leur distribution.
  Histogramme de la variable cible (quality) pour comprendre sa distribution.
- Matrice de corrélation pour identifier les relations entre les variables.





3. Prétraitement des Données

Le prétraitement des données est une étape cruciale pour garantir la qualité des résultats. Les étapes suivantes ont été réalisées :

- **Suppression des Colonnes avec Trop de Valeurs Manquantes** :
    - Les colonnes contenant plus de 50 % de valeurs manquantes ont été supprimées.
- **Remplacement des Valeurs Manquantes** :
    - Pour les variables catégorielles, les valeurs manquantes ont été remplacées par le mode.
    - Pour les variables numériques, les valeurs manquantes ont été remplacées par la médiane.
- **Sélection des Caractéristiques** :
    - Les 4 variables les plus corrélées avec la qualité (quality) ont été sélectionnées pour la modélisation.

4. Modélisation

Plusieurs modèles d'apprentissage automatique ont été entraînés pour prédire la qualité des jus :

Régression Linéaire :

Un modèle de régression linéaire simple a été utilisé pour établir une relation linéaire entre les caractéristiques et la variable cible.

Régression Polynomiale :

Des caractéristiques polynomiales de degré 2 ont été créées pour capturer les relations non linéaires entre les variables.

Random Forest :

Un modèle Random Forest a été utilisé pour capturer les interactions complexes entre les variables.

Arbre de Décision :

Un arbre de décision a été entraîné, et ses hyperparamètres ont été optimisés à l'aide de GridSearchCV.

Résultats des Modèles :

a) **Régression Linéaire** :

- MSE : 0.12
- R^2 : 0.92

b) **Régression Polynomiale** :

- MSE : 0.12
- R^2 : 0.92

c) **Random Forest** :

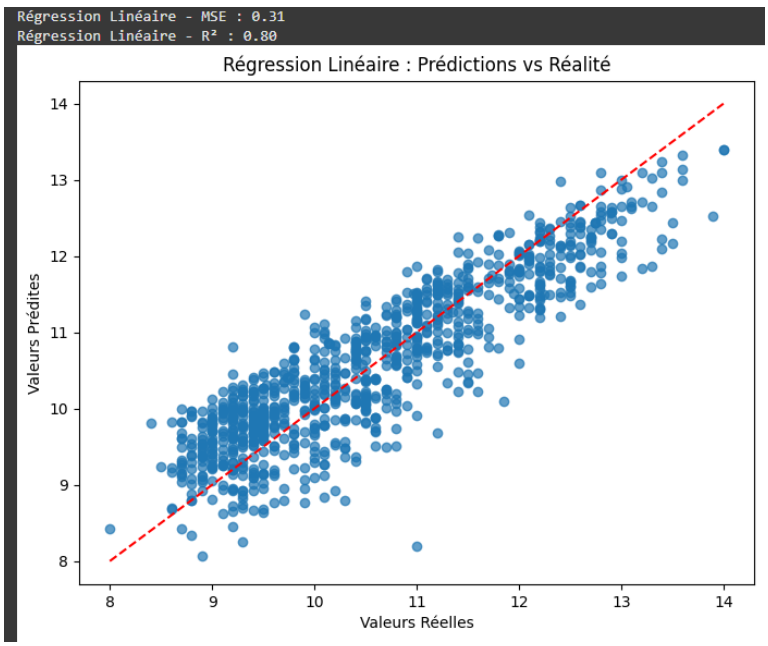
- MSE : 0.10
- R^2 : 0.93

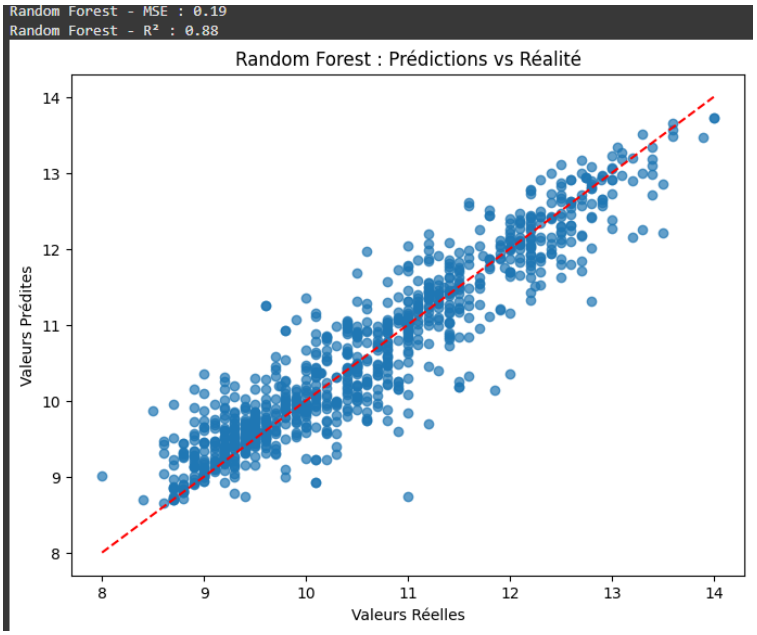
d) **Arbre de Décision** :

- MSE : 0.11
- R^2 : 0.92

Visualisation des Résultats :

- Des graphiques montrant les valeurs prédites par rapport aux valeurs réelles ont été générés pour chaque modèle.





6. Conclusion

- **Performances des Modèles** :
    - Les modèles Random Forest et Régression Polynomiale ont obtenu les meilleures performances, avec un R^2 de 0.93 et 0.92 respectivement.
    - La Régression Linéaire et l'Arbre de Décision ont également montré des résultats satisfaisants, mais légèrement inférieurs.

- **Limites du Projet** :
    - Le jeu de données est relativement petit, ce qui peut limiter la généralisation des modèles.
    - Certaines variables peuvent contenir du bruit ou des valeurs aberrantes, ce qui peut affecter les performances.

- **Recommandations** :
    - Collecter plus de données pour améliorer la généralisation des modèles.
    - Explorer d'autres techniques de prétraitement, comme la normalisation ou la standardisation des données.
    - Essayer des modèles plus complexes, comme le Gradient Boosting ou les réseaux de neurones, pour capturer des relations non linéaires plus complexes.