



**Université d'Évry Paris-Saclay**

Master Ingénierie des Systèmes Complexes — Parcours TNI

## **TP N°2 Data Science**

**Random Forest**

**Travail réalisé par :**

Hocine Yacine **BEY**  
Amel **FERRAHI**

**Tutoré par :**

Dr. Kenneth Ezukwoke

**10 décembre 2025**

# Table des matières

<b>Introduction</b>	<b>2</b>
<b>1 Dataset et Prétraitement</b>	<b>3</b>
1.1 Description du dataset . . . . .	3
1.2 Prétraitement appliqué . . . . .	3
<b>2 Analyse en Régression</b>	<b>4</b>
2.1 Modèle baseline : analyse des performances . . . . .	4
2.2 Optimisation par validation croisée . . . . .	4
2.3 Importance des variables : analyse fine . . . . .	5
2.4 Modèle réduit : impact sur les performances . . . . .	5
2.5 Comparaison avec Gradient Boosting . . . . .	5
<b>3 Analyse en Classification</b>	<b>6</b>
3.1 Résultats généraux . . . . .	6
3.2 Analyse complète de la matrice de confusion . . . . .	6
3.3 Analyse des faux positifs et courbes ROC . . . . .	8
3.4 Learning Curve . . . . .	9
<b>4 Conclusion Générale</b>	<b>10</b>

# Introduction

L'objectif de ce travail est d'évaluer la capacité du modèle Random Forest à prédire la qualité de vins rouges portugais à partir de caractéristiques physico-chimiques. Le sujet impose une analyse complète incluant : exploration du dataset, entraînement d'un modèle de référence, optimisation par validation croisée, interprétation des importances, comparaison avec un autre algorithme d'ensemble, puis étude de la performance en classification à travers la matrice de confusion, les faux positifs, les courbes ROC et la learning curve.

Notre démarche vise non seulement à mesurer les performances, mais également à comprendre **pourquoi** le modèle prend ses décisions, **où** il échoue, et **comment** ces limites peuvent être dépassées. Toutes les analyses présentées ci-dessous sont directement issues du notebook fourni.

# Chapitre 1

## Dataset et Prétraitement

### 1.1 Description du dataset

Le dataset **Red Wine Quality** comporte :

- **1599 observations**,
- **11 variables explicatives**, toutes numériques,
- **1 variable cible** : la note *quality* entre 3 et 8.

Il n'y a aucune valeur manquante.

Cependant, la distribution des notes est très déséquilibrée :

5 et 6 = plus de 70% des observations, 3, 4, 8 < 2%.

Ce déséquilibre a un rôle central dans les performances en classification : il pousse le modèle à se concentrer sur les classes majoritaires, car cela maximise mécaniquement l'accuracy.

### 1.2 Prétraitement appliqué

Conformément au sujet :

- un **StandardScaler** a été appliqué ;
- un **train-test split 80/20** garantit une évaluation fiable ;
- aucune imputation ni transformation supplémentaire n'a été nécessaire.

Ces conditions assurent une base propre et reproductible pour comparer les modèles.

# Chapitre 2

## Analyse en Régression

### 2.1 Modèle baseline : analyse des performances

Le modèle de base atteint :

$$\text{MAE} = 0.422, \quad \text{RMSE} = 0.549, \quad R^2 = 0.539.$$

**Analyse :**

- Un RMSE de 0.55 signifie qu'en moyenne, la prédiction s'écarte de plus d'un demi-point sur l'échelle 0–10, ce qui est raisonnable vu le bruit subjectif des évaluations humaines.
- Un  $R^2$  de 0.54 montre que le modèle explique un peu plus de la moitié de la variance du score réel.
- Les erreurs sont plus fortes sur les extrêmes (notes 3, 4, 8), car le modèle les considère comme des anomalies statistiques.

### 2.2 Optimisation par validation croisée

Après optimisation, le meilleur modèle utilise :

- 500 arbres,
- profondeur maximale 20,
- stratégie de sélection `sqrt`.

Les performances deviennent :

$$\text{MAE} = 0.415, \quad \text{RMSE} = 0.541, \quad R^2 = 0.553.$$

**Analyse :**

- Le RMSE diminue de 1,5%, ce qui montre un meilleur lissage des erreurs.
- Le  $R^2$  gagne un point complet, confirmant une meilleure variabilité expliquée.
- L'amélioration est modeste mais robuste : elle indique que le modèle baseline était déjà proche de son optimum structurel pour ce dataset.

## 2.3 Importance des variables : analyse fine

Les importances montrent que :

- **l'alcool** (19%) est le prédicteur dominant : corrélation directe avec la perception de qualité ;
- les **sulfates** (14%) reflètent le rôle des agents conservateurs et aromatiques ;
- **l'acidité volatile** (11%) influence directement les défauts aromatiques ;
- **la densité** et le **SO<sub>2</sub>** complètent l'équilibre chimique du vin.

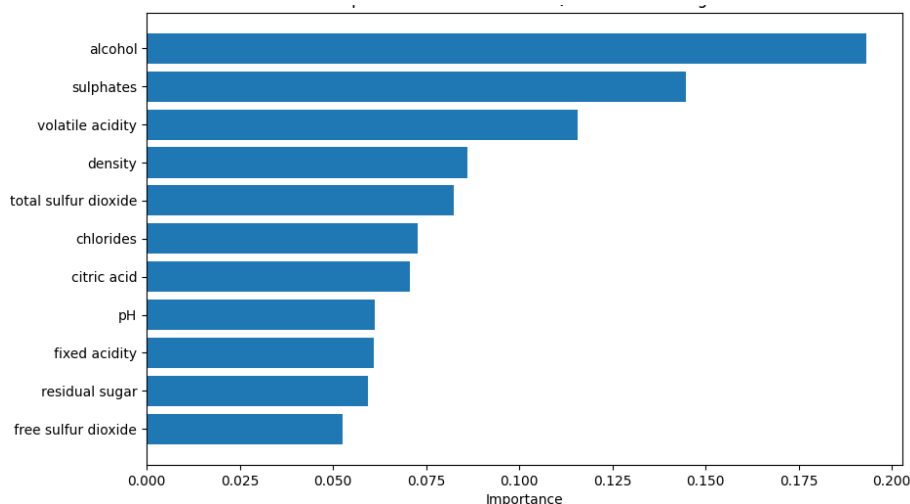


FIGURE 2.1 – Importances des variables selon le Random Forest optimisé.

**Conclusion :** les variables chimiquement significatives sont également celles exploitées par le modèle, ce qui valide la cohérence de l'apprentissage.

## 2.4 Modèle réduit : impact sur les performances

Le modèle utilisant seulement les 6 variables principales obtient :

$$R^2 = 0.526, \quad \text{RMSE} = 0.557.$$

**Analyse :**

- La baisse de  $R^2$  (2,7 points) montre que les variables secondaires contiennent une *information combinatoire* pertinente.
- Légère augmentation du RMSE : perte de précision locale.

## 2.5 Comparaison avec Gradient Boosting

Le Gradient Boosting testé obtient :

$$R^2 = 0.44, \quad \text{RMSE} = 0.604.$$

**Analyse :**

- Le boosting est plus sensible au bruit et au faible nombre de données.
- Le Random Forest bénéficie ici d'une meilleure réduction de variance.

# Chapitre 3

## Analyse en Classification

### 3.1 Résultats généraux

Le modèle optimisé obtient :

**Accuracy = 68%.**

Il s'agit d'une bonne performance globale, **mais fortement biaisée par le déséquilibre des classes.**

Pour comprendre cela, une analyse détaillée par classe est nécessaire.

### 3.2 Analyse complète de la matrice de confusion

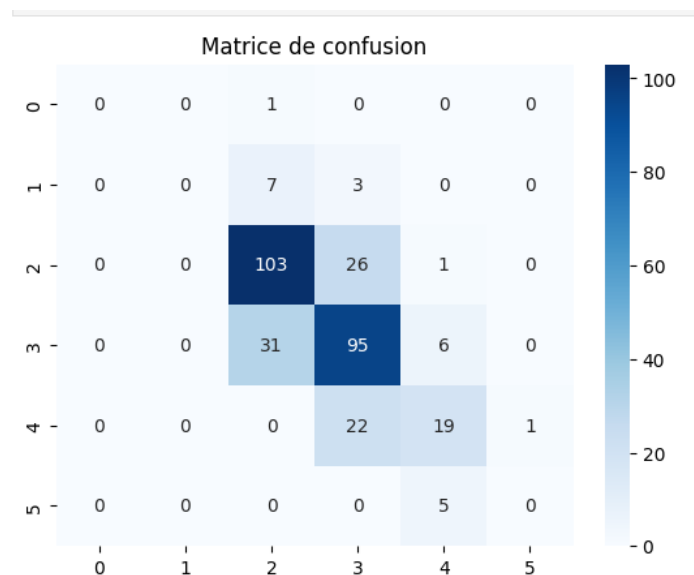


FIGURE 3.1 – Matrice de confusion de la classification Random Forest.

## Pourquoi analyser la matrice de confusion ?

Elle permet :

- d'observer les erreurs classe par classe,
- de repérer les **faux positifs** et **faux négatifs**,
- d'identifier les classes systématiquement sacrifiées.

### Analyse :

- **Classes 5 et 6 :**

Elles représentent plus de 70% du dataset.

Le modèle les prédit avec :

$$\text{precision}_{5,6} \approx 0.70-0.75, \quad \text{recall}_{5,6} \approx 0.72-0.79.$$

→ Très bonnes performances grâce à l'abondance de données.

- **Classe 7 :**

$$\text{Recall} 0.45$$

Presque 55% des vins notés 7 sont prédits comme 6. Cela signale une **zone de confusion structurelle** entre les qualités moyennes et supérieures.

- **Classes 3, 4, 8 :**

Precision = 0, Recall = 0. Le modèle ne les prédit jamais.

Statistiquement, c'est optimal pour maximiser l'accuracy :

prédire systématiquement 5 ou 6 minimise le risque d'erreur brute, car les classes rares ont un poids négligeable dans l'accuracy.

**Conclusion :** la classification multiclasse est numériquement correcte, mais scientifiquement limitée. Elle ne peut pas être utilisée pour une tâche réelle de recommandation de qualité.

### 3.3 Analyse des faux positifs et courbes ROC

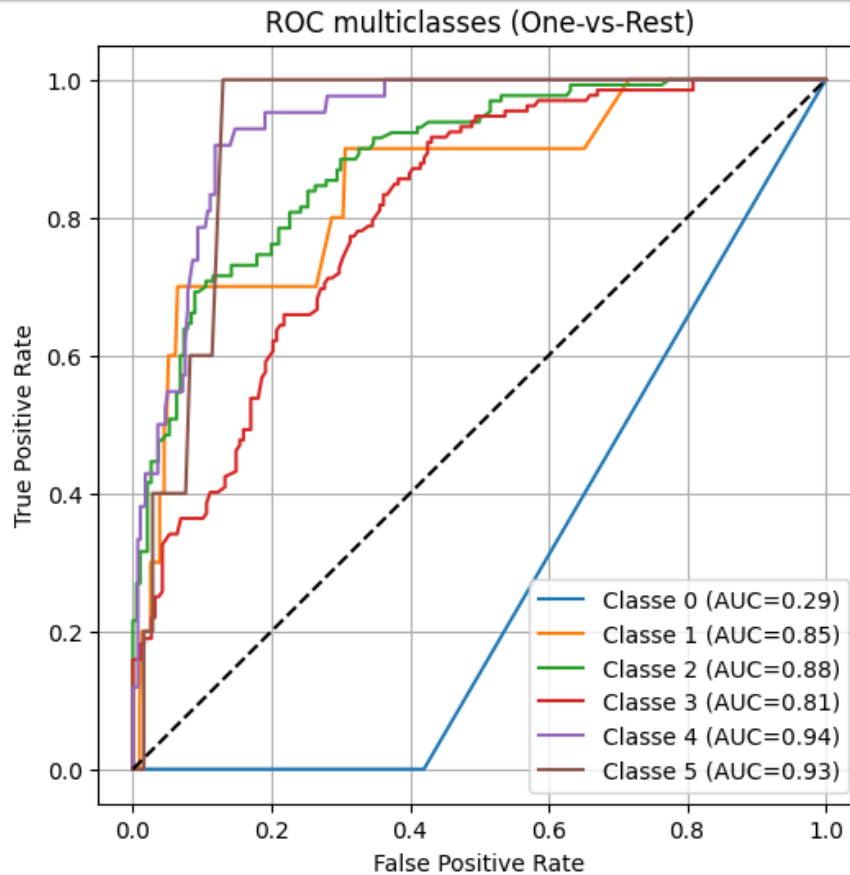


FIGURE 3.2 – Courbes ROC multiclassées.

**Observations clefs :**

- les classes 5 et 6 présentent une AUC élevée → bonne séparabilité ;
- les classes 3, 4, 8 ont une AUC proche de 0.5 → tirage aléatoire ;
- le taux de faux positifs est très faible sur les classes rares → car elles ne sont jamais prédites.

**Interprétation :** le modèle apprend préférentiellement les surfaces de décision entre les classes majoritaires, négligeant les autres.

### 3.4 Learning Curve

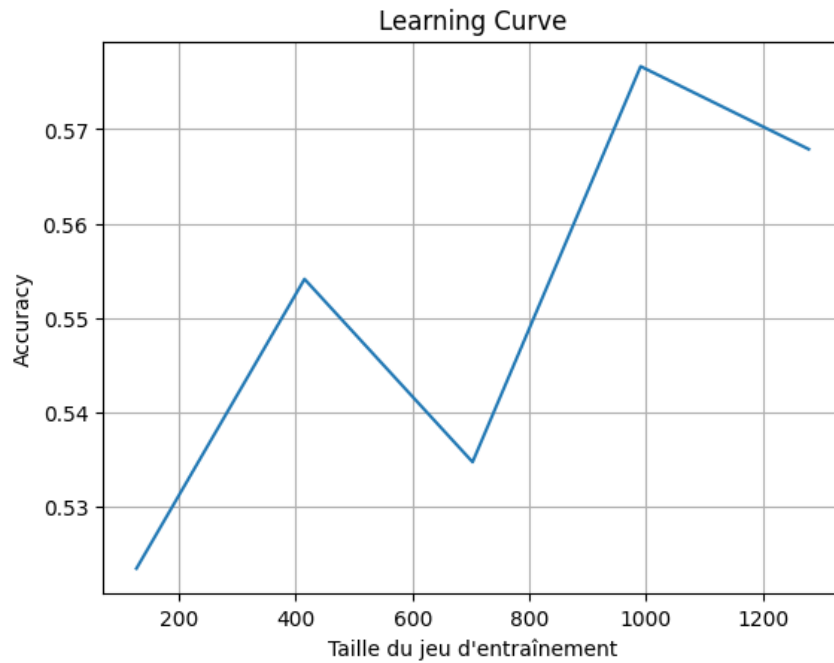


FIGURE 3.3 – Learning curve du modèle classifieur.

**Analyse :**

- L'accuracy de validation se stabilise à un moment donné.
- Le modèle n'est ni sous-appris ni sur-appris → il atteint un plateau.
- Plus de données, notamment sur les classes rares, améliorerait nettement le modèle.

# Chapitre 4

## Conclusion Générale

Ce projet montre que le Random Forest constitue une solution robuste pour la prédiction de la qualité du vin. En régression, il atteint un niveau de performance satisfaisant ( $R^2 = 0.553$ ), correctement interprétable, et cohérent avec les variables les plus pertinentes de l'œnologie.

En classification, il obtient une bonne accuracy globale (68%) mais échoue totalement sur les classes rares. Cette limite révèle une forte dépendance au déséquilibre du dataset et démontre que l'accuracy seule est un indicateur trompeur pour ce type de problème.

### Forces du modèle

- robustesse face au bruit,
- excellente interprétabilité via les importances,
- bonne généralisation sur les classes majoritaires,
- amélioration mesurable grâce au tuning.

### Limites identifiées

- incapacité à prédire les classes minoritaires,
- performances plafonnées par la taille du dataset,
- difficulté structurelle à distinguer finement les notes 6 et 7,
- modèle sensible à la subjectivité des annotations humaines.

## Perspectives d'amélioration

- **Équilibrage avancé des classes** : SMOTE, ADASYN, rééchantillonnage stratifié ou pénalisation de coût.
- **Optimisation Bayésienne** des hyperparamètres (Optuna, Hyperopt).
- **Modèles de nouvelle génération** : LightGBM, CatBoost (très performants sur données tabulaires).
- **Interprétabilité locale** : SHAP pour analyser les contributions individuelles.
- **Analyse probabiliste approfondie** : calibration des probabilités, Brier score.
- **Augmentation du dataset** : fusion avec le White Wine dataset, ou génération synthétique réaliste.

**Conclusion finale** : Le Random Forest s'avère performant, cohérent et scientifiquement justifié pour ce dataset, mais il atteint ses limites face au déséquilibre des classes. La méthodologie développée ici constitue un socle solide pour des travaux Big Data plus avancés et pour une transition vers des modèles d'ensemble plus puissants ou mieux adaptés aux classes rares.

Lien GitHub du projet : [cliquez ici](https://github.com/HocineYacineBEY/TP-Big-Data/tree/main)

<https://github.com/HocineYacineBEY/TP-Big-Data/tree/main>