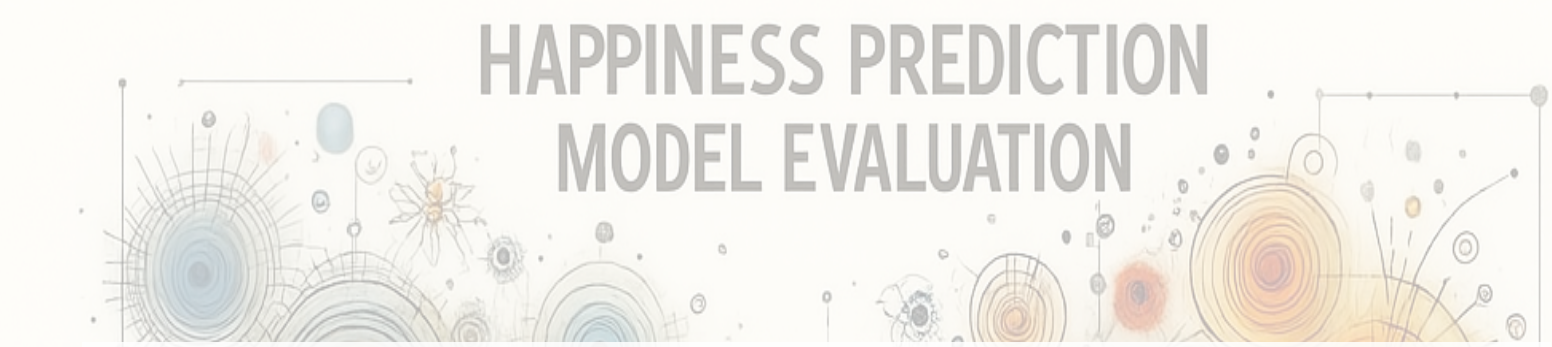


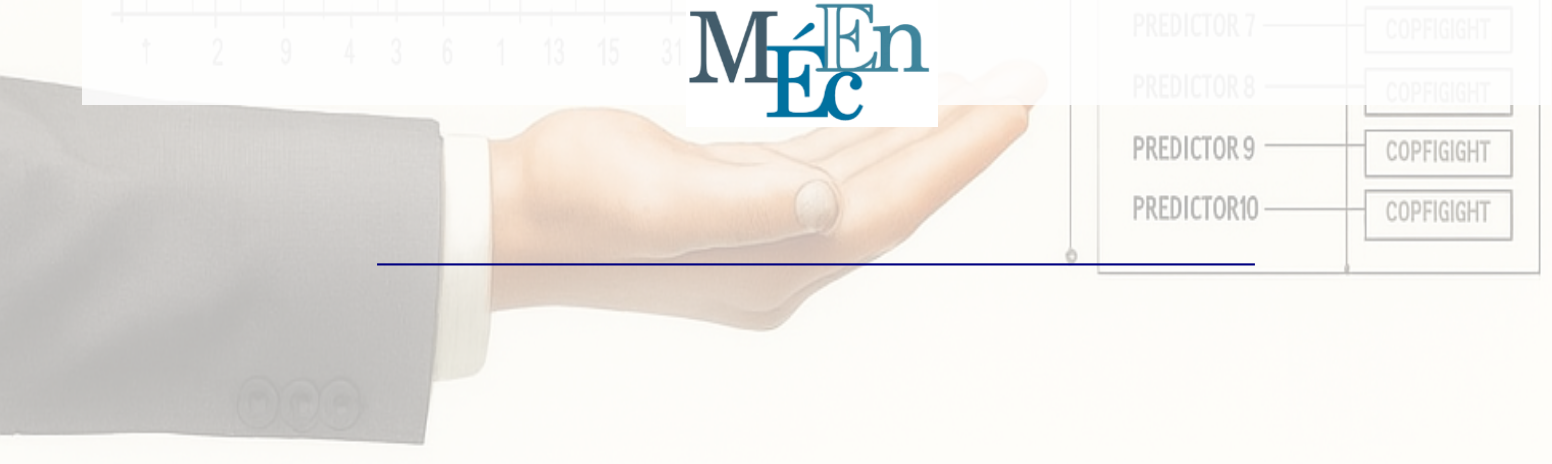
Université de Tours

UFR de Droit, Economie et Sciences Sociales

Évaluation Comparative des Modèles de Prédiction du Bonheur Subjectif

Pierre DUMONT ROTY & Emmanuel PAGUIEL

October 28, 2025



Rapport de projet réalisé dans le cadre du cours de

Data Mining

Master Economiste d'entreprise – Année Universitaire 2024-2025

Résumé

Ce rapport présente une approche de modélisation prédictive du bonheur à partir de données d'enquête. Après avoir identifié les variables socio-économiques les plus influentes via une analyse LASSO, nous mettons en œuvre une méthodologie de partitionnement des données avant tout traitement, prévenant ainsi les fuites d'information. Sept algorithmes de classification (LDA, QDA, SVM, KNN, arbre de décision, boosting et forêt aléatoire) sont comparés selon leurs capacités à prédire le niveau de bonheur déclaré. L'évaluation s'appuie sur les métriques AUC et F1-score, révélant des performances globalement modestes avec une supériorité du modèle LDA (AUC = 0,640). Les résultats soulignent la complexité de la prédiction du bonheur subjectif et identifient les approches les plus adaptées pour cette tâche spécifique, tout en mettant en lumière les défis persistants dans la détection des individus « très heureux ».

Table des Matières

1 Introduction **4**

2 Préparation et Identification des Déterminants **5**

2.1 Description du Jeu de Données et Variables d'Étude 5

2.2 Méthodologie de Préparation des Données 6

3 Analyse Comparative des Méthodes Prédictives **8**

3.1 Analyse Discriminante Linéaire (LDA) 10

3.2 Analyse Discriminante Quadratique (QDA). 11

3.3 La machine à vecteurs de support (SVM) 12

3.4 L'approche du KNN (K-Nearest Neighbors) 14

3.5 Analyse par Arbre de Décision 16

3.6 L'approche du Random Forest 19

3.7 Boosting par Gradient (XGBoost) 21

4 Partie 3: Comparaison des Modèles **23**

4.1 Synthèse des Résultats 23

4.2 **Analyse de l'Importance des Variables (Feature Importance)** 24

Annexes **26**

A Définitions Conceptuelles **26**

A.1 Définition Opérationnelle du Bonheur 26

B Métriques d'Évaluation des Performances **27**

B.1 Matrice de Confusion et Indicateurs Dérivés 27

B.2 Métriques de Performance 27

B.3 Tableau Récapitulatif des Métriques 29

B.4 Justification des Choix Méthodologiques 29

C Algorithmes de Classification : Fondements Théoriques **31**

C.1 Modèles Discriminants Linéaires 31

C.2 Machine à Vecteurs de Support (SVM) 32
C.3 K Plus Proches Voisins (KNN) 34
C.4 Méthodes Ensemblistes : Arbres de Décision 35

1 Introduction

Le bonheur, aspiration universelle définie dès Aristote comme finalité ultime de l'existence (*eudaimonia*), n'a émergé que récemment comme objet d'étude scientifique. Avec le développement de l'économie du bien-être et de la psychologie positive, les sciences sociales reconnaissent désormais le bien-être subjectif comme indicateur de développement à part entière, au même titre que la croissance économique.

Cette évolution soulève une question fondamentale : **dans quelle mesure un état subjectif aussi complexe peut-il être prédit à partir de variables socio-économiques objectives ?** Les conditions matérielles (revenu, emploi, éducation) suffisent-elles à expliquer les variations de bien-être, ou le bonheur conserve-t-il une part d'imprévisibilité échappant aux grilles d'analyse traditionnelles ?

Notre étude évalue **la capacité prédictive des approches de machine learning** à identifier les individus « très heureux » à partir de variables socio-démographiques, économiques et comportementales.

Notre question de recherche centrale est double : - Quel modèle de classification permet de prédire le plus efficacement le niveau de bonheur déclaré ? - Quelles variables socio-économiques se révèlent les plus influentes dans cette prédiction ?

Pour répondre à cette problématique, nous poursuivons trois objectifs :

1. **Identifier les déterminants statistiques du bonheur** : Par une analyse exploratoire et une régression LASSO, nous isolons les variables explicatives les plus pertinentes.
2. **Comparer systématiquement des algorithmes de classification** : Nous évaluons les performances de sept méthodes (LDA, QDA, SVM, KNN, arbre de décision, Random Forest et Boosting).
3. **Évaluer les limites de la prédictibilité du bonheur** : Au-delà des performances brutes, nous analysons les erreurs systématiques pour comprendre les frontières de l'approche quantitative.

2 Préparation et Identification des Déterminants

2.1 Description du Jeu de Données et Variables d'Étude

Notre analyse s'appuie sur un jeu de données issu d'une enquête nationale américaine (source : Wooldridge). Après nettoyage et sélection, l'échantillon final comprend 15 729 individus décrits par 15 variables explicatives couvrant quatre dimensions principales :

Table 1: Description détaillée du jeu de données

Variable	Type_R	Description	Prédicteur
vhappy	Factor (2 niveaux)	Variable cible : Niveau de bonheur déclaré (no/yes).	Cible (Classification)
year	Numérique	Année de l'enquête (1994–2014).	Contexte temporel
workstat	Factor (8 niveaux)	Statut professionnel (ex : working, retired).	(Socio-économique)
prestige	Numérique	Score de prestige de la profession.	Socio-économique
income	Factor (12 niveaux)	Niveau de revenu catégorisé (tranches).	Socio-économique
region	Factor (9 niveaux)	Région géographique de résidence.	Socio-démographique
attend	Numérique	Fréquence de participation à des activités communautaires ou religieuses.	Relations sociales
owngun	Factor (2 niveaux)	Possession d'une arme à feu (yes/no).	Comportement
tvhours	Numérique	Nombre d'heures consacrées au visionnage de la télévision.	Comportement
mothfath16	Numérique (binaire)	Présence des deux parents à l'âge de 16 ans.	Relations sociales
black	Numérique (binaire)	Appartenance à la communauté noire (1/2).	Socio-démographique
female	Numérique (binaire)	Genre (1=Femme, 2=Homme).	Socio-démographique
unem10	Factor (2 niveaux)	Expérience du chômage au cours des 10 dernières années (no/yes).	Socio-économique
DivWid	Factor (2 niveaux)	Statut matrimonial : divorcé ou veuf (yes/no).	Relations sociales
kids	Numérique	Membres du ménage de de 6 à 12 ans.	Relations sociales

2.2 Méthodologie de Préparation des Données

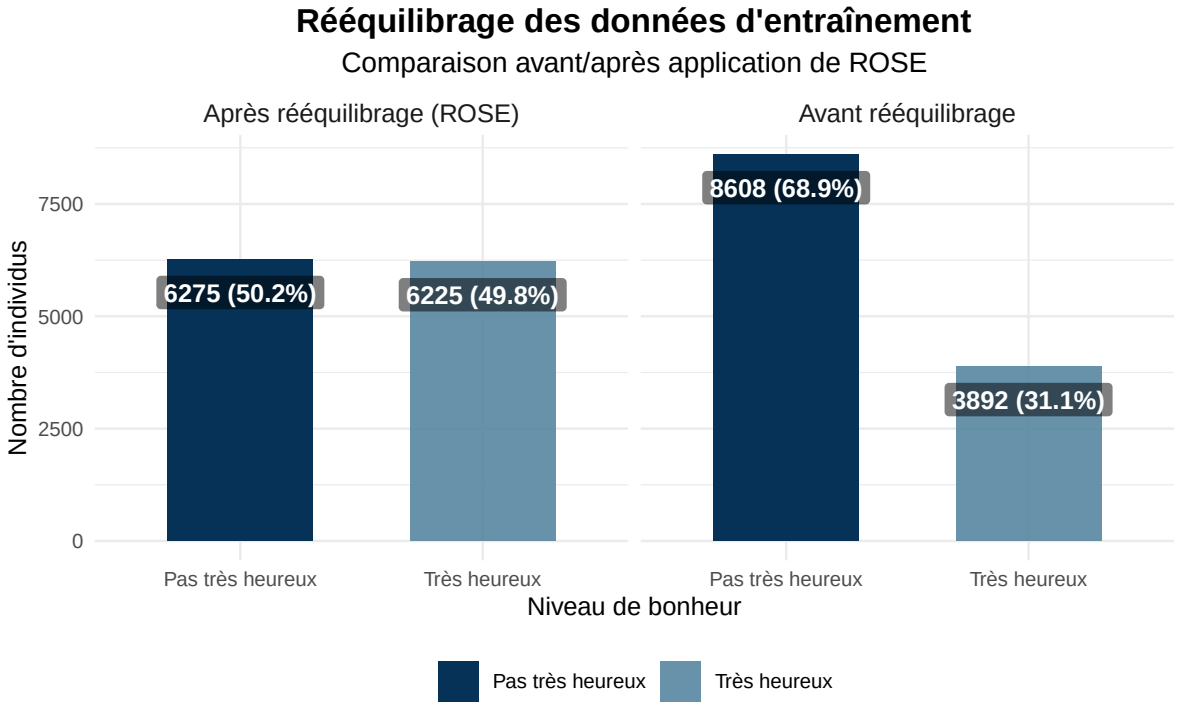
Phase 1 : Stratégie de Validation : Partitionnement des Données

Pour garantir la robustesse des modèles, le jeu de données a été stratifié en jeu d'apprentissage (80%) et jeu de test (20%) avant tout traitement. Cette stratification sur la variable cible (vhappy) préserve la distribution des classes dans les deux sous-ensembles et constitue une première mesure pour prévenir le risque de fuite d'information (data leakage).

Phase 2: Prétraitement des Données

Imputation des Valeurs Manquantes: La méthode MICE (Multiple Imputation by Chained Equations, avec PMM) a été appliquée exclusivement aux prédicteurs et indépendamment sur chaque partition (train/test). Cette approche évite toute contamination du jeu d'apprentissage par les informations du jeu de test ou de la variable cible

Rééquilibrage des Classes: La distribution initiale du bonheur présentait un déséquilibre significatif (environ 69% "pas très heureux" contre 31% "très heureux"). Pour améliorer la détection de la classe minoritaire d'intérêt, nous avons appliqué la technique ROSE (Random Over-Sampling Examples) en mode both, uniquement sur le jeu d'apprentissage. Cette méthode a permis de générer un jeu d'entraînement parfaitement équilibré, tandis que le jeu de test est resté dans sa distribution réelle pour une évaluation non biaisée.



Phase 3 : Analyse par Régression LASSO (Least Absolute Shrinkage and Selection Operator)

Avant de procéder à l'entraînement de modèles de classification complexes (tels que le Random Forest ou le Gradient Boosting), nous utilisons la **Régression Logistique pénalisée par LASSO** (régularisation L_1) pour deux raisons principales :

1. **Sélection de Variables** : Le LASSO force les coefficients des prédicteurs non pertinents à être exactement nuls, permettant d'identifier un sous-ensemble parcimonieux de variables.
2. **Interprétabilité** : Le modèle obtenu fournit une estimation linéaire de l'impact (positif ou négatif) de chaque facteur socio-économique restant sur la probabilité d'être très heureux.Les résultats suivants présentent les coefficients des variables sélectionnées par le LASSO pour le λ optimal, triés par ordre d'importance.

Table 2: Top 10 des variables les plus influentes selon le modèle LASSO

Rang	Variable	Coefficient
1	income$6000 to 6999	-0.5959411
2	incomelt $1000	0.5959372
3	workstatunempl, laid off	-0.5703698
4	income$25000 or more	0.4345188
5	workstatother	-0.4288222
6	income$7000 to 7999	-0.3322246
7	workstatschool	0.3317542
8	workstatretired	0.3296439
9	regione. sou. central	0.3028640
10	black	-0.2425392

Performance du modèle : AUC = 0.634 (λ optimal = 0.00178). **Analyse des coefficients :** Les résultats confirment que le bonheur est largement influencé par les facteurs socio-économiques. Le revenu et le statut professionnel demeurent les variables les plus déterminantes. Être au chômage réduit nettement la probabilité d'un niveau élevé de bonheur, tandis qu'un statut stable (retraité, étudiant, actif) ou un haut niveau de prestige social augmente significativement cette probabilité. Des variables comportementales (temps passé devant la télévision) et démographiques (région, genre, ethnicité) jouent un rôle complémentaire, bien que secondaire. Ces résultats servent de base à la comparaison des modèles de classification suivants.

3 Analyse Comparative des Méthodes Prédictives

L'analyse exploratoire par Régression LASSO a permis d'identifier les principaux facteurs d'influence socio-économiques. L'objectif de cette phase est de procéder à une **comparaison rigoureuse** des performances de sept algorithmes de classification (LDA, QDA, SVM, KNN, Arbre de décision, Forêt aléatoire, et Boosting).

Leur performance sera optimisée par validation croisée et mesurée sur l'ensemble de test non vu, afin de sélectionner la méthode qui offre le meilleur équilibre entre le **Score F1** et l'**AUC ROC** pour la prédiction du bonheur.

Les Recettes de Prétraitement

Pour garantir une comparaison équitable et exploiter au mieux les spécificités de chaque algorithme, nous employons **deux recettes de prétraitement distinctes** adaptées aux exigences des modèles basés sur la distance/régularisation et aux modèles basés sur les arbres.

Recette 1 : Pour les Modèles Linéaires et Basés sur la Distance (LDA, QDA, SVM, KNN)

Ces algorithmes sont sensibles à l'échelle des variables et nécessitent un encodage explicite des variables catégorielles. Le pipeline de prétraitement comprend les étapes suivantes :

1. **Imputation des valeurs manquantes** : mode pour les variables nominales, médiane pour les variables numériques
2. **Gestion des niveaux inconnus** dans les variables factorielles
3. **Encodage des variables nominales** en variables indicatrices (*dummy variables*)
4. **Normalisation** des variables numériques par standardisation (z-score)
5. **Élimination** des prédicteurs à variance nulle
6. **Suppression** des tranches de revenu intermédiaires

Cette approche garantit que les distances euclidiennes calculées par ces modèles ne soient pas dominées par les variables à grande échelle.

Recette 2 : Modèles Arborescents

Algorithmes concernés : Arbre de décision, Forêt aléatoire, Boosting (XGBoost)

Ces algorithmes effectuent des partitions récursives basées sur les valeurs brutes des prédicteurs et sont donc insensibles à l'échelle. Le prétraitement est simplifié :

1. **Imputation des valeurs manquantes** (mode/médiane)

2. **Gestion des niveaux inconnus** dans les variables factorielles
3. **Élimination** des prédicteurs à variance nulle
4. **Suppression** des tranches de revenu intermédiaires

L'encodage par variables indicatrices et la normalisation sont **intentionnellement omis** car ils fragmenteraient inutilement l'information catégorielle et n'apporteraient aucun bénéfice aux algorithmes arborescents.

Spécifications Techniques des Recettes

Les deux recettes sont implémentées dans le cadre `tidymodels` de R, garantissant la traçabilité et la reproductibilité des transformations appliquées à chaque modèle.

```
# Recette A : Modèles sensibles à l'échelle
recipe_distance %
  step_impute_mode(all_nominal_predictors()) %>%
  step_impute_median(all_numeric_predictors()) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_normalize(all_numeric_predictors(),
                 -starts_with("unknown")) %>%
  step_zv(all_predictors()) %>%
  step_rm(matches("income_X\\.(10000|20000|15000)"))

# Recette B : Modèles arborescents
recipe_arbo %
  step_impute_mode(all_nominal_predictors()) %>%
  step_impute_median(all_numeric_predictors()) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_zv(all_predictors()) %>%
  step_rm(matches("income_X\\.(10000|20000|15000)"))
```

3.1 Analyse Discriminante Linéaire (LDA)

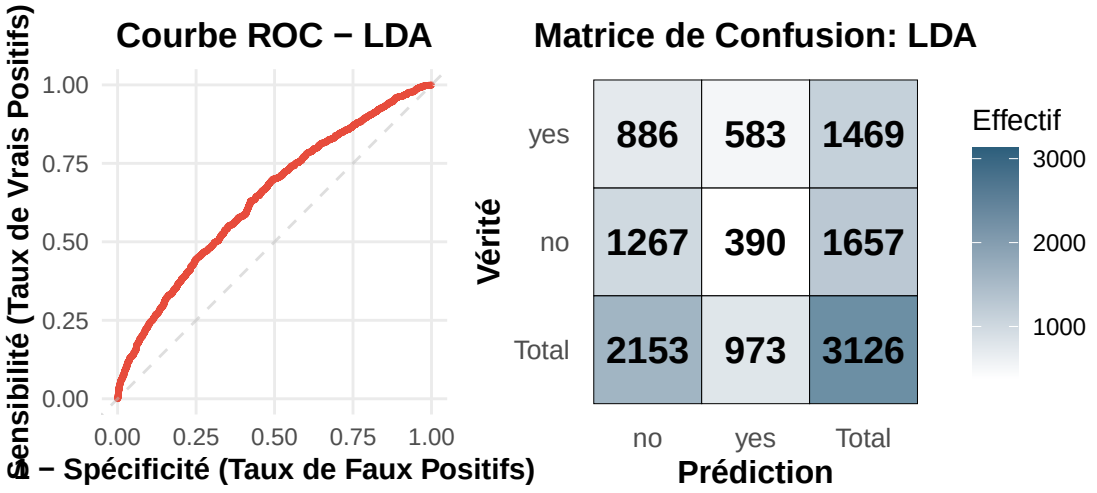


Table 3: Métriques de performance du modèle LDA

Métrique	Valeur	Description
Exactitude (Accuracy)	**0.592**	**Proportion totale de prédictions correctes**
Précision	0.397	Proportion de vrais positifs parmi les prédictions positives
Rappel (Sensibilité)	0.599	Proportion de vrais positifs correctement identifiés
Spécificité	0.588	Proportion de vrais négatifs correctement identifiés
Score F1	**0.477**	**Moyenne harmonique de la précision et du rappel**
AUC ROC	**0.640**	**Capacité discriminative du modèle**

Nous voyons que le modèle LDA présente une capacité de discrimination modérée (AUC de 0.640). Il offre un bon compromis entre Précision (39.7%) et Rappel (59.9%), ce qui lui permet d'identifier une proportion raisonnable d'individus très heureux, sans générer trop de faux positifs. Sa Spécificité (58.8%) reste correcte, mais l'Exactitude globale (59.2%) montre que le modèle reste perfectible.

3.2 Analyse Discriminante Quadratique (QDA).

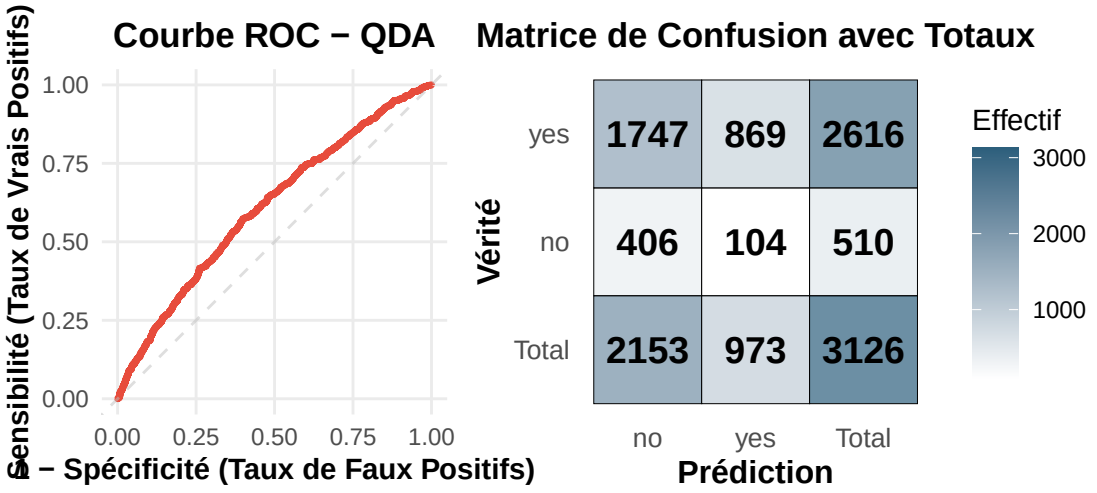
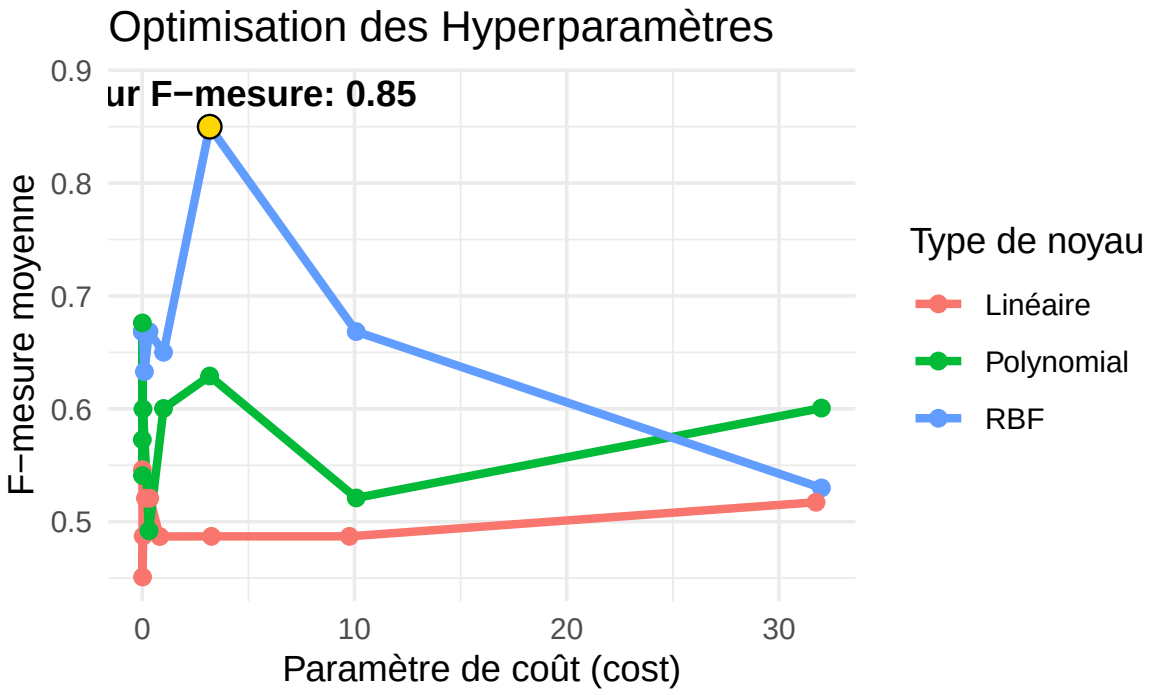


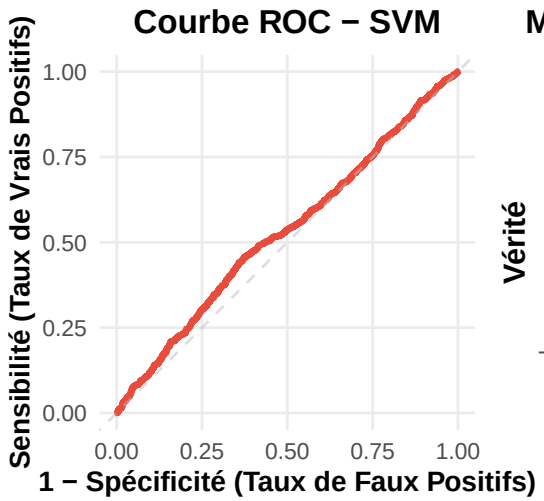
Table 4: Métriques de performance du modèle QDA

Métrique	Valeur	Description
Exactitude (Accuracy)	**0.408**	**Proportion totale de prédictions correctes**
Précision	0.332	Proportion de vrais positifs parmi les prédictions positives
Rappel (Sensibilité)	0.893	Proportion de vrais positifs correctement identifiés
Spécificité	0.189	Proportion de vrais négatifs correctement identifiés
Score F1	**0.484**	**Moyenne harmonique de la précision et du rappel**
AUC ROC	**0.608**	**Capacité discriminative du modèle**

Nous voyons que le modèle QDA présente une capacité discriminative faible (AUC de 0.608). Il est très sensible (Rappel de 89.3%) et parvient à identifier la majorité des individus très heureux, mais au prix d'une faible Précision (33.2%) et d'une Spécificité très basse (18.9%), ce qui indique un grand nombre de faux positifs. Ce modèle est donc trop permissif, ce qui limite sa fiabilité globale malgré un Score F1 modéré (0.484).

3.3 La machine à vecteurs de support (SVM)





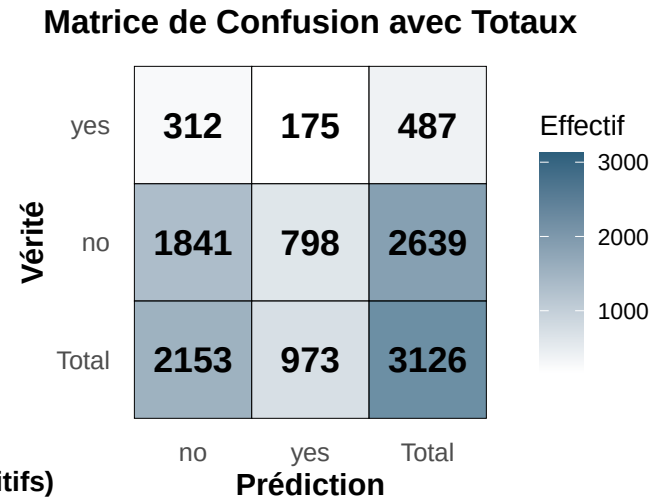


Table 5: Métriques de performance du modèle SVM

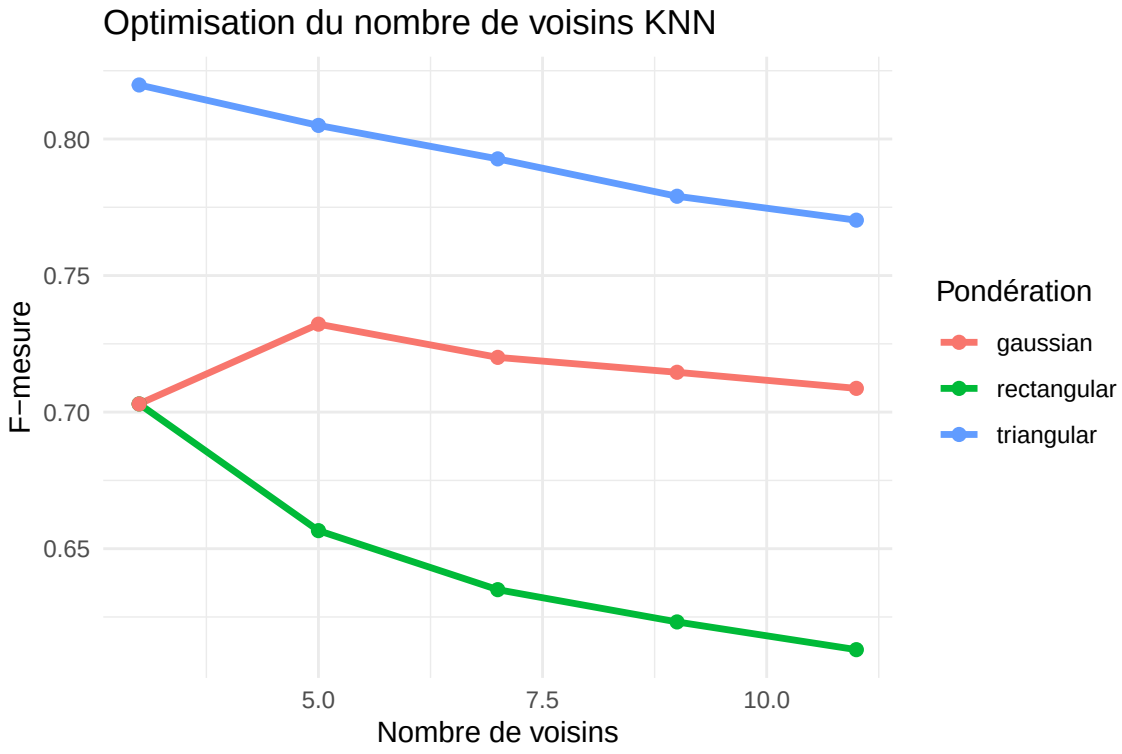
Métrique	Valeur	Description
Exactitude (Accuracy)	**0.645**	**Proportion totale de prédictions correctes**
Précision	0.359	Proportion de vrais positifs parmi les prédictions positives
Rappel (Sensibilité)	0.180	Proportion de vrais positifs correctement identifiés
Spécificité	0.855	Proportion de vrais négatifs correctement identifiés
Score F1	**0.240**	**Moyenne harmonique de la précision et du rappel**
AUC ROC	**0.529**	**Capacité discriminative du modèle**

Nous voyons que le modèle SVM RBF, bien que sélectionné pour ses performances internes élevées lors du 'tuning', a démontré une incapacité à généraliser. Le **Score F1 final sur le jeu de test est de seulement 0.240**, ce qui contraste fortement avec les ≈ 0.85 observés en validation croisée. Cet écart est le signe d'un **surapprentissage sévère** (*overfitting*).

En pratique, le modèle se montre particulièrement conservateur : il atteint une spécificité élevée (85.5%) et une exactitude globale correcte (64.5%), mais sa sensibilité est très faible (18.0%). Il privilégie la prédiction de la classe majoritaire ('non heureux') et **échoue à détecter efficacement les individus très heureux**.

Le modèle SVM est par conséquent inadapté à notre objectif de prédiction des cas positifs et doit être écarté.

3.4 L'approche du KNN (K-Nearest Neighbors)



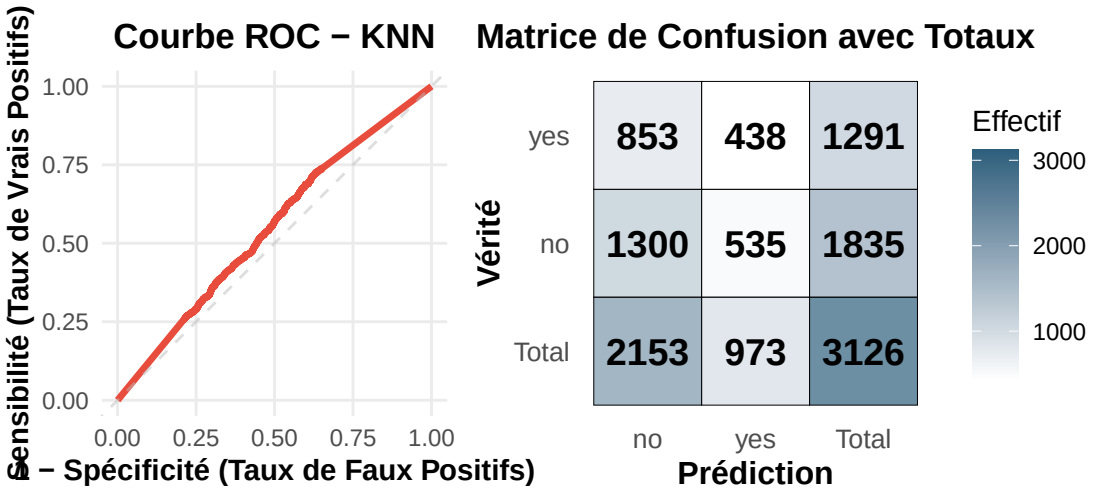
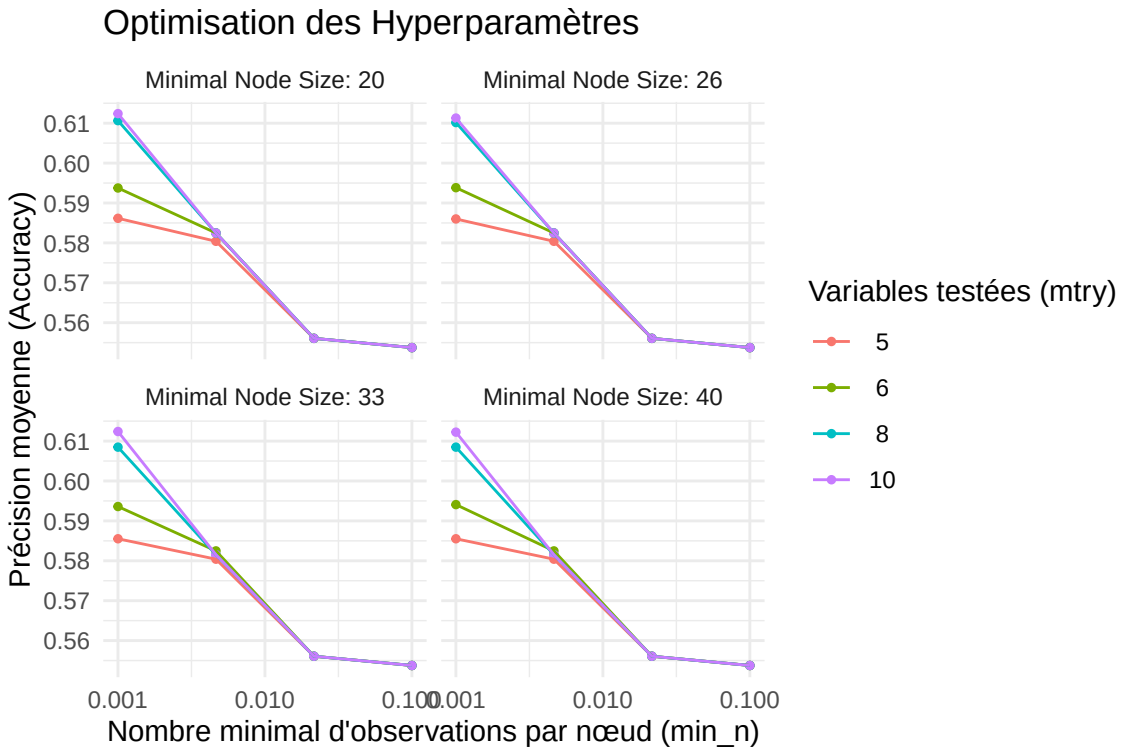


Table 6: Métriques de performance du modèle KNN (k = 3)

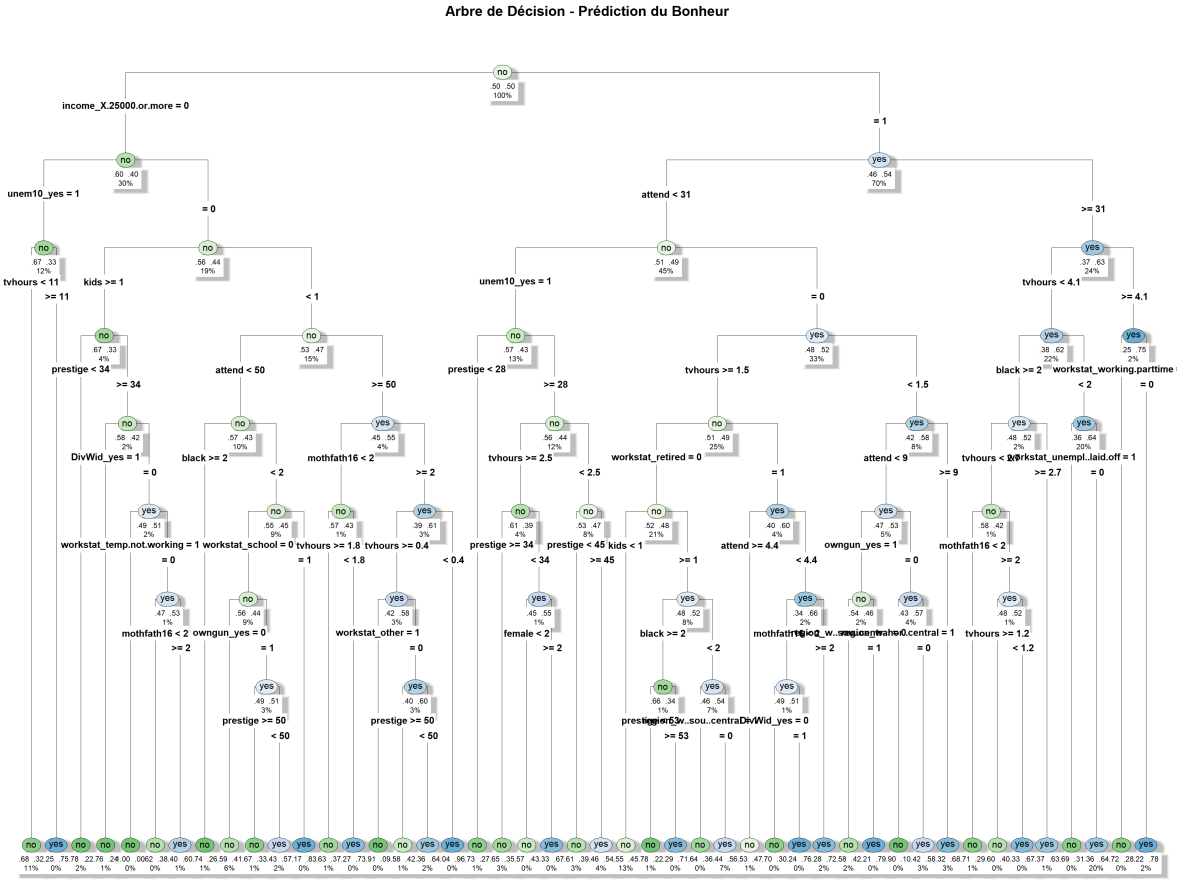
Métrique	Valeur	Description
Exactitude (Accuracy)	**0.556**	**Proportion totale de prédictions correctes**
Précision	0.339	Proportion de vrais positifs parmi les prédictions positives
Rappel (Sensibilité)	0.450	Proportion de vrais positifs correctement identifiés
Spécificité	0.604	Proportion de vrais négatifs correctement identifiés
Score F1	**0.387**	**Moyenne harmonique de la précision et du rappel**
AUC ROC	**0.548**	**Capacité discriminative du modèle**

Nous voyons que le modèle KNN (k = best_knn$neighbors) présente une performance finale modérée. Son AUC (≈0.548) et son F1-score (≈0.387) indiquent une capacité discriminative limitée sur les données non vues. Bien que l'optimisation ait atteint un F1-score interne maximal de ≈0.81 (pour k=3 avec la pondération triangular), l'écart significatif avec le F1 final démontre une difficulté à généraliser. Le modèle détecte une partie des individus très heureux (Recall≈45.0%) mais avec une précision faible (≈33.9%), générant un nombre important de faux positifs. Son équilibre global reste fragile, ce qui limite son utilité pour une prédiction fiable du bonheur.

3.5 Analyse par Arbre de Décision



Si l'on regarde l'optimisation des hyperparamètres, on peut voir que la performance diminue régulièrement à mesure que le paramètre de complexité du coût (Cost-Complexity Parameter) augmente, suggérant que les arbres trop pénalisés deviennent sous-ajustés. Les arbres plus profonds (profondeur 8 à 10) affichent des performances légèrement supérieures, notamment lorsque la taille minimale des nœuds reste faible (autour de 20 à 26 observations). Au-delà, l'augmentation de la taille minimale des nœuds entraîne une perte de précision et une simplification excessive du modèle.


Arbre de Décision - Prédiction du Bonheur
income_X.25000.or.more = 0
= 1
unem10_yes = 1
= 0
attend < 31
>= 31
tvhours < 11
>= 11
kids >= 1
< 1
unem10_yes = 1
= 0
tvhours < 4.1
>= 4.1
prestige < 34
>= 34
attend < 50
>= 50
prestige < 28
>= 28
tvhours >= 1.5
< 1.5
black >= 2
< 2
DivWid_yes = 1
= 0
black >= 2
< 2
mothfath16 < 2
>= 2
tvhours >= 2.5
< 2.5
workstat_retired = 0
= 1
attend < 9
>= 9
tvhours < 2.7
>= 2.7
workstat_unempl..laid.off = 1
= 0
workstat_temp.not.working = 1
= 0
workstat_school = 0
= 1
tvhours >= 1.8
< 1.8
tvhours >= 0.4
< 0.4
prestige >= 34
< 34
prestige < 45
>= 45
kids < 1
>= 1
attend >= 4.4
< 4.4
owngun_yes = 1
= 0
mothfath16 < 2
>= 2
mothfath16 < 2
>= 2
owngun_yes = 0
= 1
workstat_other = 1
= 0
female < 2
>= 2
black >= 2
< 2
>= 2
= 1
= 0
tvhours >= 1.2
< 1.2
prestige >= 50
< 50
prestige >= 50
< 50
>= 53
= 0
= 1

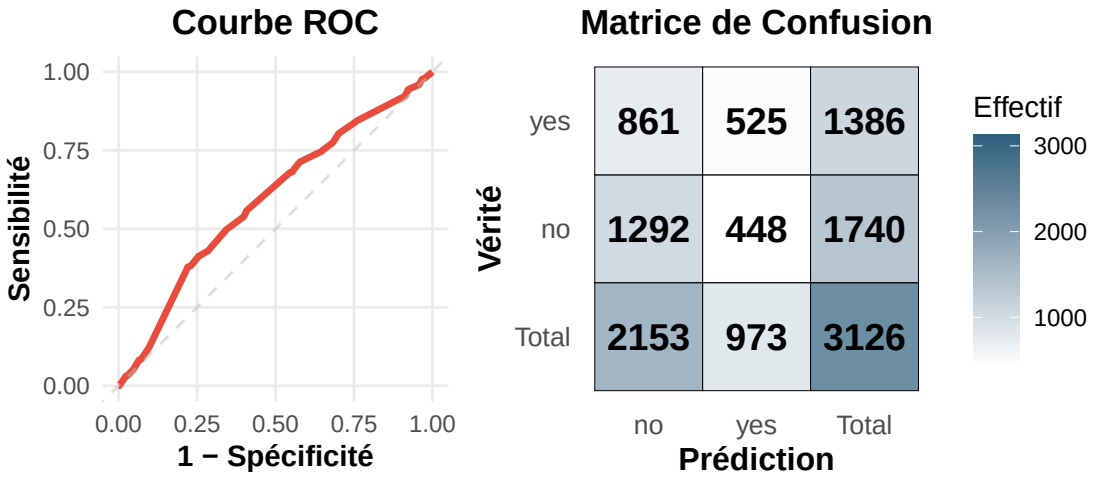
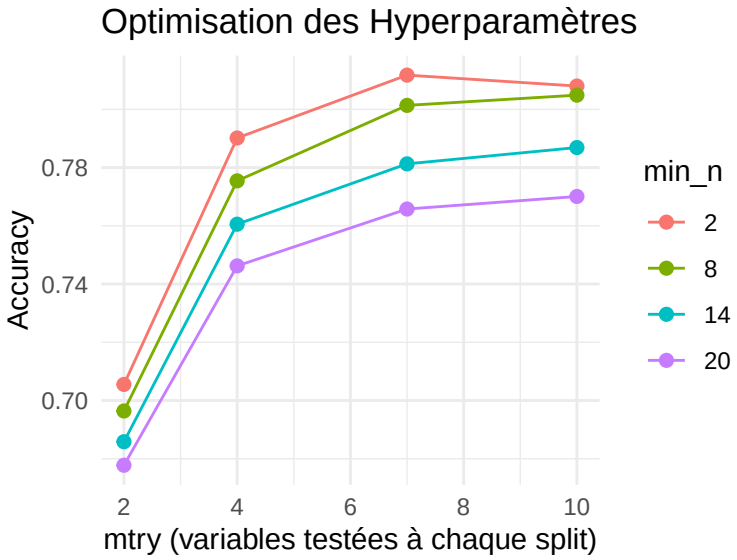


Table 7: Métriques de performance - Arbre de Décision

Métrique	Valeur	Description
Exactitude (Accuracy)	**0.581**	**Proportion totale de prédictions correctes**
Précision	0.379	Proportion de vrais positifs parmi les prédictions positives
Rappel (Sensibilité)	0.540	Proportion de vrais positifs correctement identifiés
Spécificité	0.600	Proportion de vrais négatifs correctement identifiés
Score F1	**0.445**	**Moyenne harmonique de la précision et du rappel**
AUC ROC	**0.591**	**Capacité discriminative du modèle**

Nous voyons que le modèle atteint une performance modeste et limitée. Son AUC (0.591) et son Score F1 (0.445) indiquent une faible capacité discriminative. Le modèle parvient à une exhaustivité modérée (Rappel 54.0%) mais avec une faible Précision (37.9%), générant un taux élevé de faux positifs. L'optimisation confirme que les arbres trop complexes ou trop pénalisés deviennent sous-ajustés, avec le meilleur compromis obtenu pour un arbre de profondeur intermédiaire (8 à 10) et de faible complexité de coût (0.001). Cependant, la fiabilité générale du modèle reste faible pour la prédiction du bonheur, justifiant le recours à des modèles d'ensemble.

3.6 L'approche du Random Forest



L'optimisation des hyperparamètres montre que la performance s'améliore nettement lorsque le nombre de variables testées à chaque division (mtry) augmente, avant de se stabiliser autour de mtry = 6 à 10. Les modèles avec un nombre minimal d'observations par feuille faible (min_n = 2 ou 8) offrent systématiquement les meilleures performances, traduisant un apprentissage plus flexible et mieux ajusté aux données. À l'inverse, des valeurs trop élevées de min_n (14 ou 20) réduisent la précision, en raison d'une moindre granularité dans la séparation des classes.

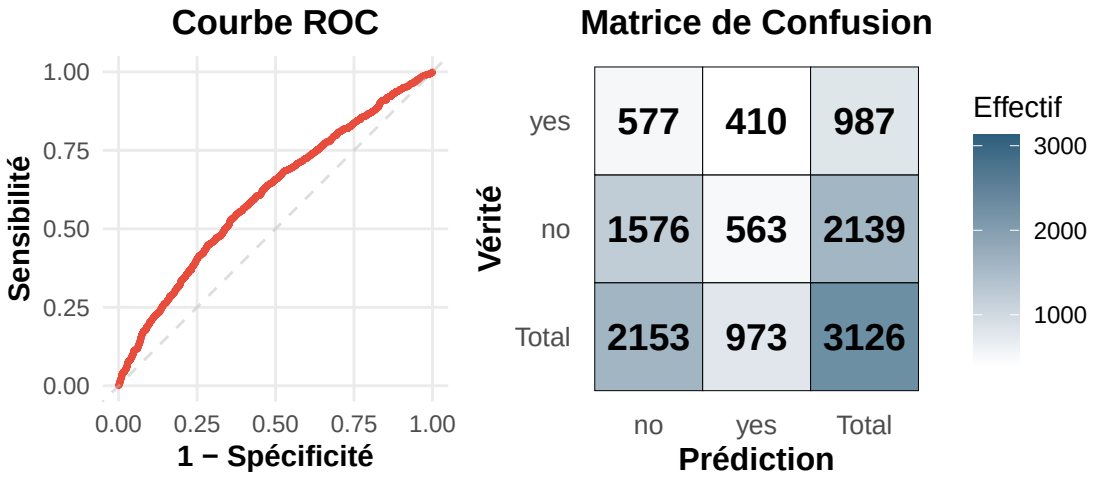
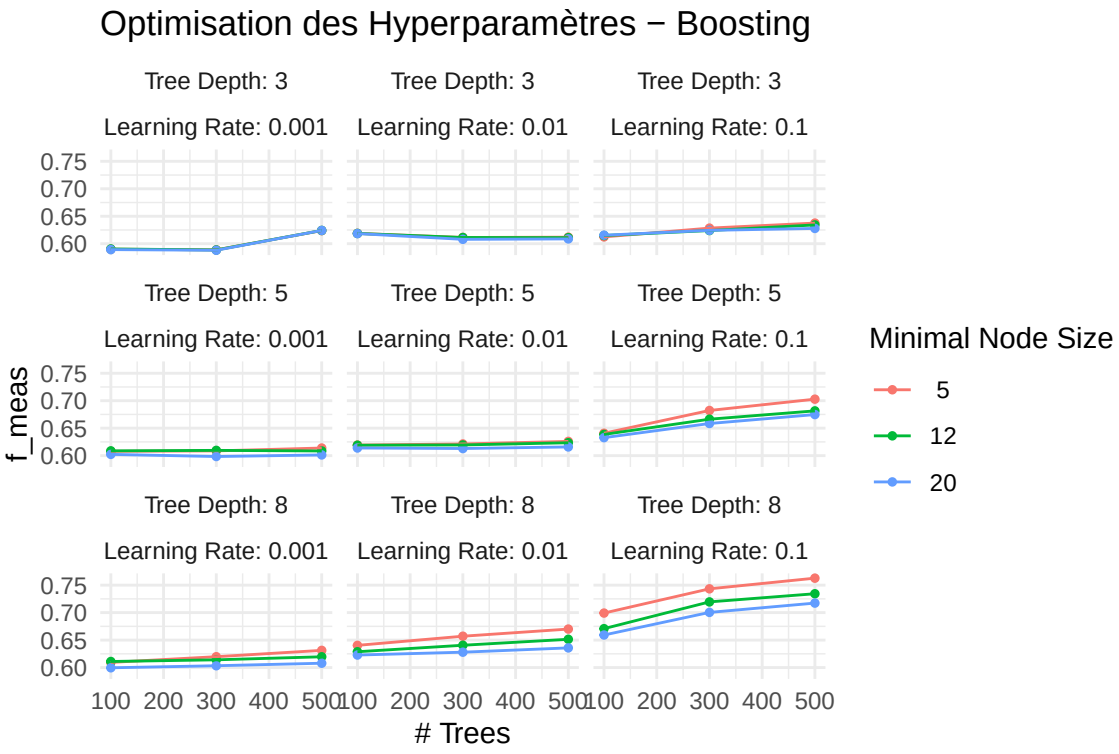


Table 8: Métriques de performance - Random Forest

Métrique	Valeur	Description
Exactitude (Accuracy)	**0.635**	**Proportion totale de prédictions correctes**
Précision	0.415	Proportion de vrais positifs parmi les prédictions positives
Rappel (Sensibilité)	0.421	Proportion de vrais positifs correctement identifiés
Spécificité	0.732	Proportion de vrais négatifs correctement identifiés
Score F1	**0.418**	**Moyenne harmonique de la précision et du rappel**
AUC ROC	**0.607**	**Capacité discriminative du modèle**

Nous voyons que le modèle présente une performance modérée mais reste limité. Son AUC (0.607) et son Score F1 (0.418) sont faibles, indiquant une capacité discriminative insuffisante pour prédire le bonheur de manière fiable. Le modèle montre un équilibre fragile entre le Rappel (42.1%) et la Précision (41.5%), manquant plus de la moitié des vrais cas positifs. L'optimisation a montré que la meilleure Exactitude est obtenue avec un nombre intermédiaire de variables (mtry 7) et les arbres les plus flexibles (min_n = 2).

3.7 Boosting par Gradient (XGBoost)



L'optimisation des hyperparamètres révèle que la performance du modèle s'améliore avec l'augmentation du nombre d'arbres, particulièrement lorsque la profondeur est intermédiaire (5 à 8) et que le taux d'apprentissage (learning rate) est modéré (≈ 0.1). Les arbres peu profonds et un taux d'apprentissage trop faible (0.001) conduisent à un sous-apprentissage, avec des gains de performance limités. À l'inverse, une taille minimale de nœuds réduite (≈ 5) permet au modèle de mieux capter les interactions complexes entre variables.

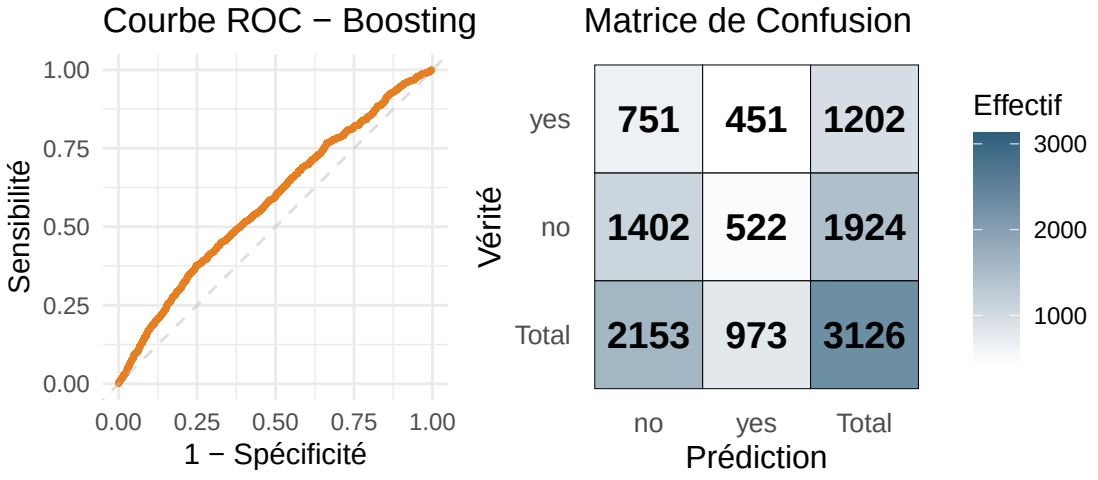


Table 9: Métriques de performance - Boosting

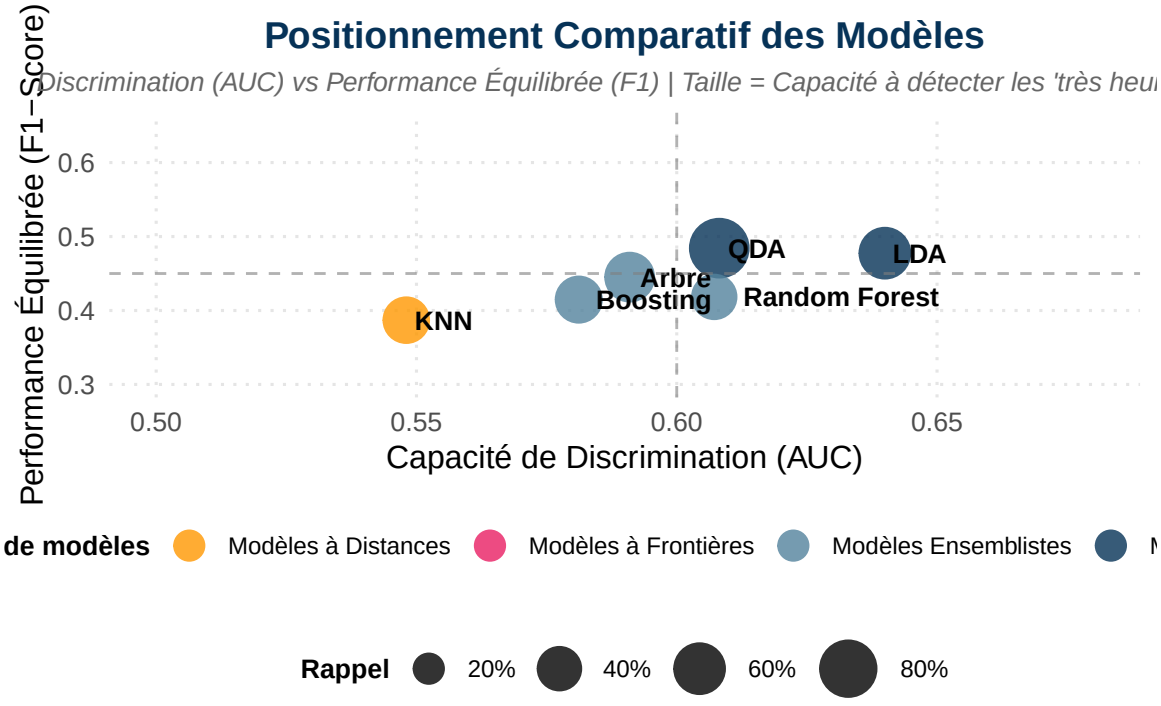
Métrique	Valeur	Description
Exactitude (Accuracy)	**0.593**	**Proportion totale de prédictions correctes**
Précision	0.375	Proportion de vrais positifs parmi les prédictions positives
Rappel (Sensibilité)	0.464	Proportion de vrais positifs correctement identifiés
Spécificité	0.651	Proportion de vrais négatifs correctement identifiés
Score F1	**0.415**	**Moyenne harmonique de la précision et du rappel**
AUC ROC	**0.581**	**Capacité discriminative du modèle**

Nous constatons une performance décevante, avec une AUC (0.581) et un Score F1 (0.415) qui sont parmi les plus faibles de tous les algorithmes testés. Il est à peine plus performant que le hasard (AUC > 0.50). Le modèle souffre d'un déséquilibre : le Rappel (46.4%) est obtenu au détriment d'une Précision faible (37.5%), générant un grand nombre de faux positifs. Le Boosting est inefficace pour ce problème de classification.

4 Partie 3: Comparaison des Modèles

4.1 Synthèse des Résultats

Cette étude avait pour ambition de déterminer dans quelle mesure le bonheur subjectif, pouvait être prédit à partir de variables socio-économiques objectives. À travers une méthodologie combinant analyse exploratoire, sélection de variables par LASSO et comparaison systématique de sept algorithmes de classification, plusieurs enseignements majeurs se dégagent.



Principaux constats:

Premièrement, notre analyse LASSO a identifié le **revenu**, le **statut professionnel** (notamment le chômage), et le **prestige social** comme les déterminants les plus influents du bonheur déclaré. Ces résultats confirment l'ancrage du bien-être subjectif dans des réalités socio-économiques tangibles, tout en révélant l'importance des effets non linéaires sur le revenu, où les extrêmes sont associés à une probabilité accrue d'être "très heureux".

Deuxièmement, la comparaison systématique des modèles prédictifs révèle que les **modèles linéaires/discriminants** surpassent les approches sophistiquées. Le **LDA (Linear Discriminant**

Analysis) émerge comme l'approche la plus performante avec la meilleure capacité de discrimination (AUC = **0.640**). Le **QDA** le suit de près en AUC, mais offre le **meilleur équilibre global** (F1-Score = **0.484**). Cette supériorité des modèles simples sur les approches d'ensemble (Random Forest, Boosting) et les SVM suggère que les relations sous-jacentes sont en grande partie capturables par des structures linéaires.

Table 10: Tableau récapitulatif des performances des sept modèles

Modèle	AUC	F1-Score	Rappel	Exactitude
LDA	**0.640**	**0.477**	**0.599**	**0.592**
QDA	0.608	0.484	0.893	0.408
SVM	**0.529**	**0.240**	**0.180**	**0.645**
KNN	0.548	0.387	0.450	0.556
Arbre	0.591	0.445	0.540	0.581
Boosting	0.581	0.415	0.464	0.593
Random Forest	0.607	0.418	0.421	0.635

Troisièmement, une **limite fondamentale** traverse l'ensemble des modèles : leur difficulté à identifier correctement les individus "très heureux". Les scores de Rappel, ainsi que l'AUC maximale de seulement **0.640**, témoignent de la complexité intrinsèque de la prédiction du bonheur subjectif. Cette observation soulève une question méthodologique cruciale : les variables socio-économiques, bien qu'influentes, ne sont pas suffisantes pour capturer l'intégralité des déterminants du bien-être (facteurs psychologiques, génétiques, relationnels). Le rejet catégorique du SVM (F1 $= 0.240$) illustre de plus un échec de généralisation pour les modèles trop sensibles.

Choix Final : Le **Modèle QDA** est l'approche la plus judicieuse, car il maximise l'équilibre entre la détection des individus "très heureux" et la minimisation des faux positifs. Cependant, la faible performance globale de tous les modèles (AUC maximale de 0.640) indique une limite inhérente aux prédicteurs sociodémographiques seuls.

4.2 Analyse de l'Importance des Variables (Feature Importance)

L'évaluation des modèles de classification a désigné les modèles discriminants (LDA/QDA) comme les plus efficaces en termes d'équilibre, mais l'analyse de l'importance des variables par les modèles arborescents (Random Forest, XGBoost) est essentielle pour identifier les facteurs les plus déterminants de la prédiction du bonheur.

Prédominance des Facteurs Sociaux et Subjectifs : Contrairement à une attente axée uniquement sur le revenu, la Participation communautaire (mesurant l'engagement social) est unanimement le

Table 11: Comparaison des importances des variables entre les trois modèles arborescents

Variable	Arbre	Random Forest	XGBoost
Participation communautaire	113.5	0.149	0.1
Revenu ≥ 25 000 $	109.01	0.086	0
Prestige professionnel	67.53	0.135	0.2
Heures de télévision	64.87	0.119	0.2
Chômage (10 ans)	37.8	0.069	0
Enfants en bas âge	26.98	0.054	0.1
Origine ethnique (noir)	26.44	0.026	0
Structure familiale	20.39	0.046	0
Divorcé/Veuf	18.09	0.037	0
Retraité	11.59	0.03	—
Année d'enquête	—	0.093	0.1
Possession d'arme	—	0.049	0
Genre (féminin)	—	0.044	0
Travail à temps plein	—	0.039	0
Région Sud-Atlantique	—	0.024	0
Région Pacifique	—	—	0

facteur le plus important dans l'Arbre de Décision et le Random Forest. Ceci renforce l'idée que les variables comportementales ou liées au capital social sont plus prédictives du bonheur subjectif que les variables démographiques brutes.

Rôle du Statut Socio-économique : Le Revenu et le Prestige professionnel se positionnent immédiatement derrière la participation communautaire. Ils forment le socle socio-économique identifié précédemment par la Régression LASSO. Il est notable que le Prestige professionnel est même jugé le plus important par le XGBoost (0.2), soulignant son rôle crucial dans la satisfaction personnelle.

Indicateurs Comportementaux : La variable **Heures de télévision** est très influente.

Considérations Demographiques : Les variables liées à la situation de vie (Chômage, Enfants en bas âge, Divorcé/Veuf) maintiennent une importance constante, bien que modérée, à travers tous les modèles. Le Chômage reste l'un des principaux facteurs de risque identifié.

Variables Marginales : De nombreux facteurs démographiques simples (Genre, Possession d'arme, Régions) ont une influence très faible ou nulle (coefficient de 0 pour le XGBoost), suggérant qu'une fois les facteurs socio-économiques et sociaux pris en compte, leur capacité à affiner la prédiction est minime.

Annexes

A Définitions Conceptuelles

A.1 Définition Opérationnelle du Bonheur

Dans cette étude, le terme **"bonheur"** désigne le *bien-être subjectif auto-déclaré*, mesuré par une question directe issue de l'enquête nationale américaine (General Social Survey) :

> *"Taken all together, how would you say things are these days – would you say that you are very happy, pretty happy, or not too happy?"*
> **Traduction :** "Pris dans son ensemble, diriez-vous que vous êtes très heureux, assez heureux, ou pas très heureux ?"

Opérationnalisation

Nous avons procédé à une **dichotomisation** de cette échelle ordinale à trois niveaux :

- **Classe positive (vhappy = yes)** : Individus se déclarant "très heureux" (*very happy*)
- **Classe négative (vhappy = no)** : Individus se déclarant "assez heureux" ou "pas très heureux" (*pretty happy + not too happy*)

B Métriques d'Évaluation des Performances

B.1 Matrice de Confusion et Indicateurs Dérivés

La **matrice de confusion** constitue le fondement de l'évaluation en classification binaire. Pour un problème à deux classes (Positif / Négatif), elle se présente ainsi :

		Prédiction	
		Positif	Négatif
Vérité	Positif	VP	FN
	Négatif	FP	VN

Table 12: Structure de la matrice de confusion

Légende :

- **VP (Vrais Positifs)** : Cas positifs correctement prédits comme positifs
- **VN (Vrais Négatifs)** : Cas négatifs correctement prédits comme négatifs
- **FP (Faux Positifs)** : Cas négatifs incorrectement prédits comme positifs (*erreur de type I*)
- **FN (Faux Négatifs)** : Cas positifs incorrectement prédits comme négatifs (*erreur de type II*)

B.2 Métriques de Performance

1. Exactitude (Accuracy)

Proportion de prédictions correctes parmi l'ensemble des prédictions :

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN} = \frac{\text{Prédictions correctes}}{\text{Total des prédictions}}$$

Interprétation : Métrique globale adaptée aux classes équilibrées. **Attention :** Trompeuse en cas de déséquilibre de classes (un modèle naïf prédisant toujours la classe majoritaire peut atteindre une accuracy élevée).

2. Précision (Precision)

Proportion de vraies détections parmi les prédictions positives :

$$\text{Precision} = \frac{VP}{VP + FP}$$

Question associée : "Parmi tous les individus prédits comme *très heureux*, quelle proportion l'est réellement ?"

Utilité : Crucial quand le coût des faux positifs est élevé (ex : diagnostic médical).

3. Rappel / Sensibilité (Recall / Sensitivity)

Proportion de cas positifs correctement détectés :

$$\text{Recall} = \frac{VP}{VP + FN} = \frac{VP}{\text{Total réellement positifs}}$$

Question associée : "Parmi tous les individus *réellement très heureux*, quelle proportion a été correctement identifiée ?"

Utilité : Critique quand le coût des faux négatifs est élevé (ex : détection de fraude).

4. Spécificité (Specificity)

Proportion de cas négatifs correctement détectés :

$$\text{Specificity} = \frac{VN}{VN + FP} = \frac{VN}{\text{Total réellement négatifs}}$$

Complémentarité : $\text{Specificity} = 1 - \text{Taux de Faux Positifs}$

5. Score F1 (F1-Score)

Moyenne harmonique entre précision et rappel, pénalisant les déséquilibres :

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot VP}{2 \cdot VP + FP + FN}$$

Interprétation : Le F1-Score atteint son maximum (1.0) uniquement si précision et rappel sont tous deux parfaits. Il pénalise fortement les modèles déséquilibrés (ex : un modèle avec Precision = 0.9 et Recall = 0.3 obtient F1 = 0.45, révélant sa faiblesse).

6. Aire Sous la Courbe ROC (AUC-ROC)

L'**AUC (Area Under the ROC Curve)** mesure la capacité du modèle à discriminer entre les classes, indépendamment du seuil de classification choisi.

Construction de la courbe ROC :

1. Pour chaque seuil possible de probabilité prédite (de 0 à 1)
2. Calculer le *Taux de Vrais Positifs* (TPR = Recall) et le *Taux de Faux Positifs* (FPR = 1 - Specificity)
3. Tracer TPR en fonction de FPR

$$\text{TPR} = \frac{VP}{VP + FN} \quad ; \quad \text{FPR} = \frac{FP}{FP + VN}$$

Interprétation de l'AUC :

- **AUC = 1.0** : Discrimination parfaite (séparation complète des classes)
- **AUC = 0.9 - 1.0** : Excellence (modèle très performant)
- **AUC = 0.8 - 0.9** : Bonne performance
- **AUC = 0.7 - 0.8** : Performance acceptable
- **AUC = 0.6 - 0.7** : Performance faible mais supérieure au hasard
- **AUC = 0.5** : Modèle équivalent au hasard (classification aléatoire)
- **AUC < 0.5** : Pire que le hasard (inversion des prédictions améliorerait le modèle)

Avantage majeur : L'AUC est insensible au déséquilibre de classes et au choix du seuil de décision, ce qui en fait une métrique robuste pour comparer des modèles.

B.3 Tableau Récapitulatif des Métriques

B.4 Justification des Choix Méthodologiques

Dans notre étude, nous avons privilégié deux métriques complémentaires :

1. **AUC-ROC** comme métrique principale : pour sa robustesse au déséquilibre de classes et sa capacité à évaluer la discrimination globale

Métrique	Formule	Plage	Quand privilégier ?
Accuracy	$\frac{VP+VN}{\text{Total}}$	[0, 1]	Classes équilibrées
Precision	$\frac{VP}{VP+FP}$	[0, 1]	Coût élevé des FP
Recall	$\frac{VP}{VP+FN}$	[0, 1]	Coût élevé des FN
Specificity	$\frac{VN}{VN+FP}$	[0, 1]	Détecter les négatifs
F1-Score	$\frac{2 \cdot P \cdot R}{P+R}$	[0, 1]	Équilibre P et R
AUC-ROC	Aire sous courbe	[0, 1]	Comparaison globale

Table 13: Synthèse des métriques d'évaluation

2. **F1-Score** comme métrique secondaire : pour évaluer l'équilibre entre la précision des prédictions positives et la capacité à détecter tous les cas positifs

Cette double évaluation permet d'éviter les biais d'interprétation : un modèle peut avoir une AUC élevée mais un F1 faible si le seuil de décision est mal calibré, ou inversement.

””

C Algorithmes de Classification : Fondements Théoriques

C.1 Modèles Discriminants Linéaires

Analyse Discriminante Linéaire (LDA)

Principe : Le LDA cherche à projeter les données dans un espace de dimension réduite qui maximise la séparation entre les classes. Il repose sur trois hypothèses fondamentales :

1. Les données suivent une **distribution normale multivariée** dans chaque classe
2. Les classes partagent la **même matrice de covariance** Σ
3. Les a priori de classe sont estimés par les proportions empiriques

Formulation mathématique :

La règle de décision de Bayes classe une observation $\mathbf{x}$ dans la classe k qui maximise la fonction discriminante :

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k)$$

où :

- μ_k : vecteur des moyennes de la classe k
- Σ : matrice de covariance commune (estimée par pooling)
- π_k : probabilité a priori de la classe k

Frontière de décision : Hyperplan linéaire défini par $\delta_1(\mathbf{x}) = \delta_2(\mathbf{x})$

Avantages :

- Modèle parcimonieux (peu de paramètres)
- Très efficace sur petits échantillons
- Interprétable (coefficients linéaires)
- Robuste au surapprentissage

Limites :

- Hypothèse forte de normalité multivariée
- Frontière linéaire uniquement (inadaptée aux relations complexes)
- Sensible aux valeurs aberrantes

Analyse Discriminante Quadratique (QDA)

Relaxation de l'hypothèse : Le QDA autorise des matrices de covariance **différentes par classe** (Σ_k au lieu de Σ).

Fonction discriminante :

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k) + \log(\pi_k)$$

Frontière de décision : Surface quadratique (courbe, ellipsoïde)

Compromis biais-variance :

- **QDA** : plus flexible (biais faible) mais plus de paramètres (variance élevée)
- **LDA** : moins flexible (biais plus élevé) mais plus parcimonieux (variance faible)

Règle pratique : Privilégier QDA si n (taille échantillon) $\gg p^2$ (nombre de paramètres), sinon LDA.

C.2 Machine à Vecteurs de Support (SVM)

Principe fondamental : Le SVM cherche l'**hyperplan de marge maximale** qui sépare au mieux les deux classes. La marge est la distance minimale entre l'hyperplan et les points de données les plus proches (appelés *vecteurs de support*).

SVM Linéaire

Pour des données linéairement séparables, le problème d'optimisation s'écrit :

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{sous contrainte} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1,\ \forall i$$

où :

- $\mathbf{w}$: vecteur normal à l'hyperplan
- b : biais (intercept)
- $y_i \in \{-1, +1\}$: étiquette de classe

SVM Non Linéaire avec Noyau RBF

Pour capturer des frontières complexes, le SVM utilise l'**astuce du noyau** (*kernel trick*) qui projette implicitement les données dans un espace de dimension supérieure.

Noyau Gaussien (RBF) utilisé dans notre étude :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$$

où γ (sigma) contrôle la flexibilité :

- γ élevé : frontière très flexible (risque de surapprentissage)
- γ faible : frontière plus lisse (risque de sous-apprentissage)

Hyperparamètre C (coût) : Pénalité pour les erreurs de classification

- C élevé : marge dure (peu d'erreurs tolérées, risque de surapprentissage)
- C faible : marge souple (erreurs tolérées, meilleure généralisation)

Avantages :

- Très performant en haute dimension
- Robuste au surapprentissage (régularisation implicite)
- Flexibilité via les noyaux

Limites :

- Sensible au choix des hyperparamètres
- Temps de calcul élevé sur grands échantillons
- Interprétabilité limitée (boîte noire)

C.3 K Plus Proches Voisins (KNN)

Principe : Méthode non paramétrique basée sur la similarité locale. Une observation est classée selon le vote majoritaire de ses k plus proches voisins dans l'espace des features.

Algorithme :

1. Calculer la distance entre $\mathbf{x}_{\text{new}}$ et tous les points d'entraînement
2. Sélectionner les k points les plus proches
3. Assigner la classe majoritaire parmi ces k voisins

Distance euclidienne (utilisée) :

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{p=1}^{P} (x_{ip} - x_{jp})^2}$$

Choix de k :

- k petit : frontière complexe, sensible au bruit (variance élevée)
- k grand : frontière lisse, sous-apprentissage (biais élevé)
- **Règle empirique :** $k = \sqrt{n}$ comme point de départ

Avantages :

- Simplicité conceptuelle
- Pas d'hypothèse sur la distribution des données
- Adapté aux frontières non linéaires

Limites :

- **Fléau de la dimension** : performance dégradée en haute dimension
- Sensible à l'échelle des variables (nécessite normalisation)
- Coût computationnel élevé en prédiction (calcul de toutes les distances)
- Stockage de tout l'échantillon d'apprentissage

C.4 Méthodes Ensemblistes : Arbres de Décision

Arbre de Décision (CART)

Principe : Partitionnement récursif de l'espace des features selon des règles de décision binaires.

Critère de split (impureté de Gini) :

$$\text{Gini} = 1 - \sum_{k=1}^{K} p_k^2$$

où p_k est la proportion de la classe k dans le nœud.

Algorithme de construction :

1. Sélectionner la variable et le seuil minimisant l'impureté
2. Diviser le nœud en deux nœuds enfants
3. Répéter récursivement jusqu'à un critère d'arrêt

Élagage (*pruning*) : Suppression de branches pour éviter le surapprentissage, contrôlée par le paramètre de complexité α.

Forêt Aléatoire (Random Forest)

Principe : Agrégation de multiples arbres de décision entraînés sur des échantillons bootstrap avec sélection aléatoire de features.

Algorithme :

1. Pour $b = 1, \dots, B$ (nombre d'arbres) :
   - Tirer un échantillon bootstrap de taille n
   - Construire un arbre en sélectionnant aléatoirement m features à chaque split
   - Faire croître l'arbre sans élagage
2. Prédiction finale : vote majoritaire des B arbres

Hyperparamètres clés :

- `ntree` : nombre d'arbres (typiquement 500-2000)
- `mtry` : nombre de features candidates à chaque split ($\sqrt{p}$ par défaut)
- `min_n` : taille minimale des nœuds terminaux

Avantages :

- Très robuste au surapprentissage (moyennage)
- Performance élevée sans tuning intensif
- Gestion native des non-linéarités et interactions
- Mesure d'importance des variables

Gradient Boosting (XGBoost)

Principe : Construction séquentielle d'arbres faibles, chacun corrigeant les erreurs du précédent.

Algorithme simplifié :

1. Initialiser $\widehat{f}(\mathbf{x}) = \bar{y}$
2. Pour $m = 1, \dots, M$:
   - Calculer les résidus $r_i = y_i - \widehat{f}(\mathbf{x}_i)$
   - Ajuster un arbre h_m sur les résidus
   - Mettre à jour : $\widehat{f}(\mathbf{x}) \leftarrow \widehat{f}(\mathbf{x}) + \eta \cdot h_m(\mathbf{x})$

Hyperparamètres critiques :

- η (`learn_rate`) : taux d'apprentissage (0.01-0.3)
- `max_depth` : profondeur maximale des arbres (3-10)
- `n_estimators` : nombre d'itérations de boosting

Régularisation XGBoost :

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(h_m)$$

où Ω pénalise la complexité des arbres (nombre de feuilles, normes des poids).

Avantages :

- Performance state-of-the-art sur données tabulaires
- Régularisation intégrée
- Gestion native des valeurs manquantes
- Parallélisation efficace