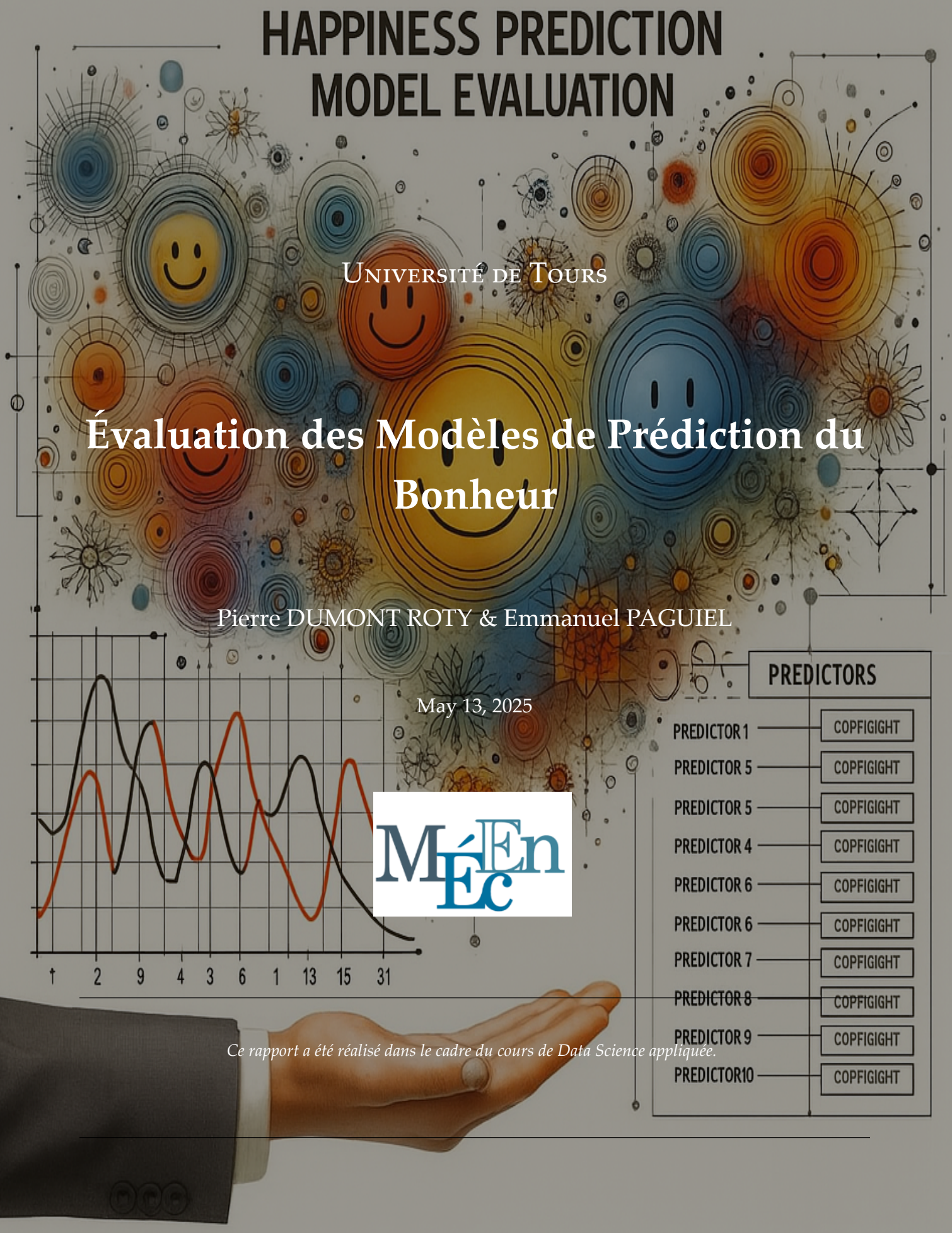


Université de Tours

Évaluation des Modèles de Prédiction du Bonheur

Pierre DUMONT ROTY & Emmanuel PAGUIEL

May 13, 2025

Ce rapport a été réalisé dans le cadre du cours de Data Science appliquée.

Ce rapport présente une approche de modélisation prédictive du bonheur à partir de données d'enquête. Nous commençons par décrire les variables socio-économiques utilisées comme prédicteurs. Ensuite, nous exposons la méthodologie de traitement des données, incluant le partitionnement et les recettes de prétraitement spécifiques à chaque modèle. L'objectif principal est de comparer l'efficacité de plusieurs algorithmes de classification (tels que l'arbre de décision, la forêt aléatoire, le SVM, etc.) pour prédire le niveau de bonheur déclaré. L'évaluation de la performance de ces modèles s'appuie sur des métriques clés comme l'AUC et le F1-score, permettant de déterminer les approches les plus aptes à discriminer les niveaux de bonheur et à fournir des prédictions équilibrées. Les résultats de cette comparaison mettent en lumière les modèles les plus prometteurs pour la compréhension et la prédiction du bien-être subjectif dans le contexte de cette enquête.

Contents

1 Problématique

Le bonheur est une aspiration universelle et un indicateur clé de notre qualité de vie. Savoir ce qui influence notre bonheur et pouvoir le comprendre permet d'agir plus consciemment sur nos choix de vie, d'identifier les facteurs qui nous épanouissent et de cultiver un bien-être durable.

Dans cette optique, ce rapport examine comment des outils d'analyse de données, appelés algorithmes de classification, peuvent nous aider à prédire si une personne se déclare "très heureuse" ou "pas très heureuse" à partir de réponses à une enquête.

En comparant la performance de ces outils d'analyse de données, nous visons à déterminer l'approche la plus pertinente pour la prédiction du bonheur dans le contexte des données étudiées.

2 Partie 1 : Les Données du Bonheur : Aperçu et Préparation

2.1 Préparation des Données :

L'ensemble de données initial comprenait 17 137 individus et 33 variables issues d'enquêtes. Après une phase de transformation et de sélection, notre analyse s'est concentrée sur un jeu de données réduit à 15 626 individus et 15 variables jugées plus pertinentes pour la prédiction du bonheur subjectif.

2.2 Variables Explicatives : Un Aperçu des Facteurs Potentiels

Pour analyser les multiples dimensions susceptibles d'influencer ce sentiment de bonheur, nous avons sélectionné un ensemble de variables d'intérêt, regroupées pour faciliter leur interprétation :

- **Contexte Socio-démographique et Temporel :**
  - ***year*** : L'année de collecte des données.
  - ***region*** : La région géographique de résidence (ex. : "Midwest", "Southwest").
  - ***female*** : Le genre de l'individu.
  - ***black*** : L'appartenance à la communauté noire.
- **Statut Socio-économique et Professionnel :**
  - ***workstat*** : Le statut professionnel (ex. : "Travailleur indépendant", "Étudiant").
  - ***prestige*** : Le score de prestige associé à la profession.
  - ***income***: Le niveau de revenu (ex. : "Faible revenu", "Revenu moyen", "Haut revenu").
  - ***unem10*** : L'expérience du chômage au cours des 10 dernières années (Oui/Non).
- **Relations Sociales et Familiales :**
  - ***attend*** : La fréquence de participation à des activités communautaires ou religieuses.
  - ***mothfath16*** : La présence des deux parents à l'âge de 16 ans (catégories à préciser).
  - ***DivWid*** : Le statut matrimonial (divorcé ou veuf).

  - ***kids*** : Le nombre d'enfants.

- **Comportements et Attitudes :**
  - ***tvhours*** : Le nombre d'heures consacrées au visionnage de la télévision.
  - ***owngun*** : La possession d'une arme à feu (Oui/Non).

Ces variables seront au cœur de notre analyse pour comprendre et prédire les niveaux de bonheur au sein de l'échantillon étudié.

2.3 Importance des Variables : Analyse par le Modèle LASSO

Afin d'identifier les principaux facteurs prédictifs du bonheur (vhappy), nous avons appliqué le modèle LASSO. Cette technique de régularisation effectue une sélection automatique des variables les plus influentes tout en contrôlant le surajustement.

Table 1: Top 10 des Variables les Plus Influentes

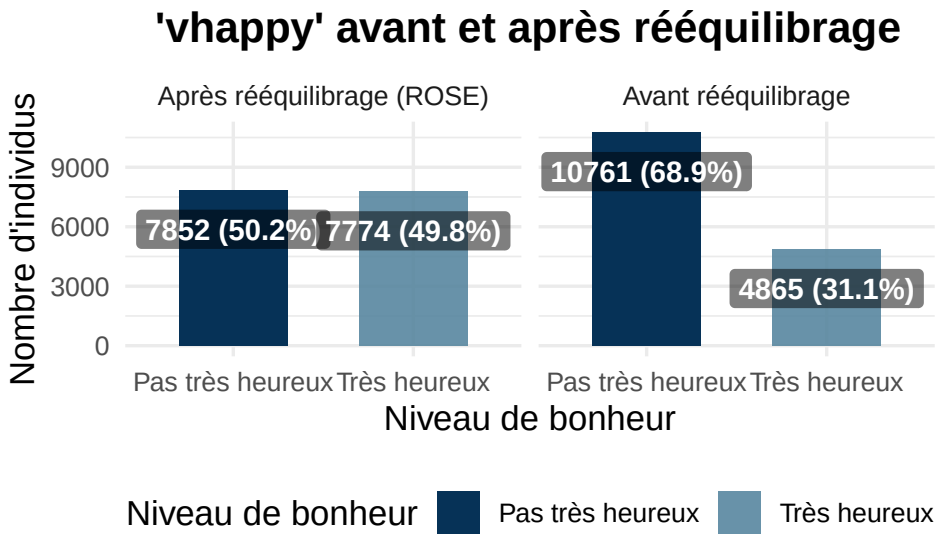
	Variable	Importance	Impact_on_Happiness
1	(Intercept)	-0.6743474	Augmente la probabilité d'être 'pas très heureux'
9	income$25000 or more	0.4895875	Augmente la probabilité d'être 'très heureux'
15	incomelt $1000	0.4742240	Augmente la probabilité d'être 'très heureux'
5	workstatunempl, laid off	-0.3937700	Augmente la probabilité d'être 'pas très heureux'
12	income$6000 to 6999	-0.3450531	Augmente la probabilité d'être 'pas très heureux'
27	unem10yes	-0.3087594	Augmente la probabilité d'être 'pas très heureux'
11	income$4000 to 4999	-0.2560614	Augmente la probabilité d'être 'pas très heureux'
4	workstatretired	0.2552339	Augmente la probabilité d'être 'très heureux'
3	workstatother	-0.2501351	Augmente la probabilité d'être 'pas très heureux'
23	owngunyes	0.2044907	Augmente la probabilité d'être 'très heureux'
26	black	-0.1887239	Augmente la probabilité d'être 'pas très heureux'

Notre analyse statistique LASSO a révèlé que le bonheur est influencé par un ensemble de facteurs : avoir un revenu élevé (25 000$ ou plus) et étonnamment très bas (moins de 1000$) , le statut professionnel (être retraité ou s'occuper de la maison, tandis qu'être au chômage est lié à moins de bonheur.). D'autres facteurs tels que la possession d'une arme à feu, la région de résidence, la présence des deux parents durant l'adolescence et l'absence de chômage récent sont associés à un plus grand bonheur, tandis qu'être noir est associé à une légère diminution de la probabilité d'être "très heureux. Cette analyse nous aide à mieux comprendre les éléments qui pourraient être liés au sentiment de bonheur des individus.

2.4 Ajustements:

La distribution initiale de notre variable cible 'vhappy' présentait un déséquilibre significatif (10761 'no' contre 4865 'yes'), ce qui aurait pu potentiellement introduire un biais en faveur de la classe majoritaire lors

de l'entraînement de nos modèles. Afin d'atténuer ce risque et d'améliorer la capacité prédictive pour la classe minoritaire ('yes'), nous avons opté pour l'application de la méthode 'both' de ROSE (Random Over-Sampling Examples). Cette technique combine une stratégie de sur-échantillonnage de la classe minoritaire avec un sous-échantillonnage de la classe majoritaire, visant à créer un ensemble de données d'entraînement plus équilibré (approximativement 7800 instances par classe). Notre intention à travers cet ajustement était d'accroître la sensibilité des modèles à la détection de la classe 'yes' sans compromettre excessivement leur spécificité.



3 Partie 2 : Modélisation Prédictive du Niveau de Bonheur

Afin d'identifier l'approche la plus efficace pour prédire le niveau de bonheur, cette section détaille la construction et l'évaluation de plusieurs modèles statistiques et algorithmiques.

- 1.**Partitionnement des Données**

L'ensemble de données a été divisé en une portion d'entraînement (80%) et une portion de test (20%). Cette séparation, réalisée avec stratification sur la variable cible "vhappy", assure que la distribution des niveaux de bonheur est préservée dans les deux ensembles, permettant une évaluation plus robuste et représentative des performances des modèles.

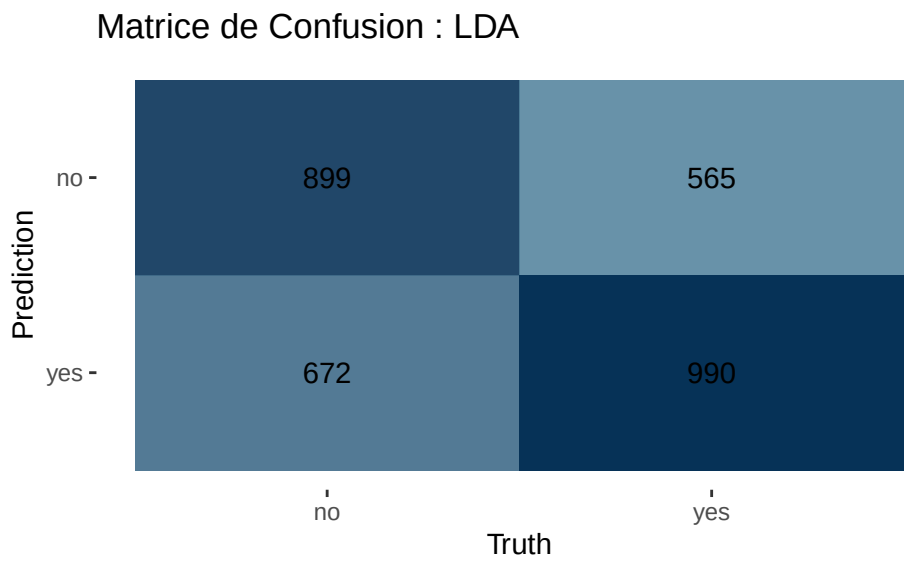
- 2.**Les Recettes de Prétraitement**

Afin d'assurer la qualité des données et leur adéquation aux différents algorithmes de modélisation, nous avons mis en œuvre un ensemble d'étapes de prétraitement gérées par les recettes du framework tidymodels. Ces étapes ont inclus l'imputation des valeurs manquantes (mode pour nominales, médiane pour

numériques), la gestion des niveaux inconnus et la suppression de la variance nulle. La création de variables dummy pour les prédicteurs nominaux et la normalisation des variables numériques ont également été largement appliquées, avec des exceptions notables comme l'exclusion des colonnes "unknown" de la normalisation pour la forêt aléatoire. Des ajustements spécifiques ont été réalisés pour certains modèles, tels que la suppression de colonnes dummy spécifiques de la variable income pour LDA/QDA et l'élimination des prédicteurs numériques fortement corrélés pour le boosting.

3.1 Analyse par Discriminante Linéaire (LDA)

- 1. **La Matrice de confusion:**



- La matrice de confusion du LDA montre que le modèle a correctement classé 899 personnes comme 'non heureuses' et 990 comme 'très heureuses'. Cependant, il a également produit 672 faux positifs (personnes réellement 'non heureuses' que le modèle a prédites 'très heureuses') et 565 faux négatifs (personnes réellement 'très heureuses' que le modèle a prédites 'non heureuses'). Ces erreurs indiquent une performance modérée du modèle dans la distinction entre les deux niveaux de bonheur.

- 2. **Les mésures de performance du modèle:**

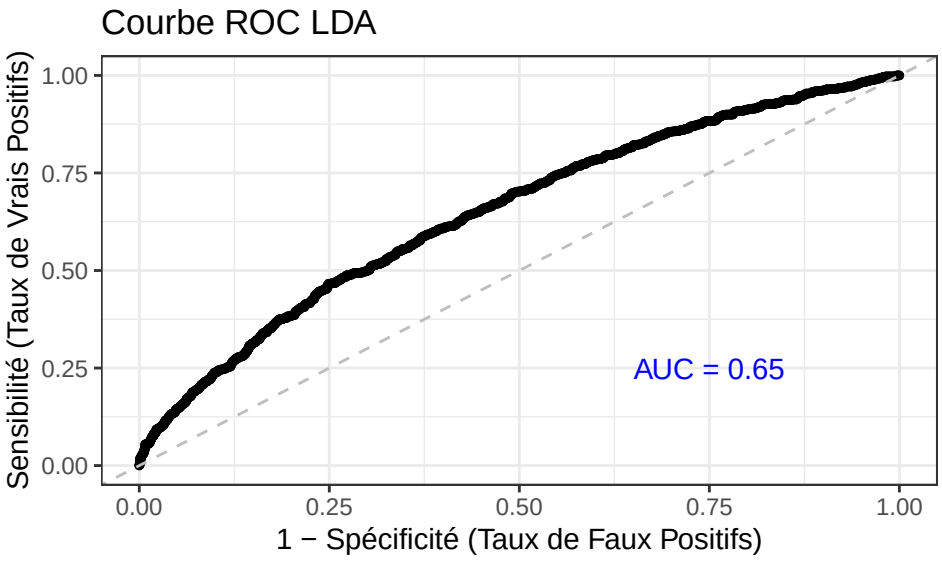
- Le modèle LDA a une précision de 60.43% avec une erreur de 39.57%. Il identifie correctement environ 60% des deux classes, mais commet aussi des erreurs dans les faux positifs et faux négatifs, résultant en un F1-score de 61.55%.

- 3. **La Courbe ROC**

L'analyse de la courbe ROC et de l'AUC révèle des informations sur la performance du modèle à séparer les classes de bonheur.

Table 2: Métriques de performance du modèle

Mesure de performance	Valeur
Précision (Accuracy)	60.43 %
Erreur globale	39.57 %
Vrais Négatifs (TN)	899
Faux Positifs (FP)	565
Vrais Positifs (TP)	990
Faux Négatifs (FN)	672
Sensibilité (Recall)	59.57 %
Spécificité	61.41 %
Précision	63.67 %
Score F1	61.55 %

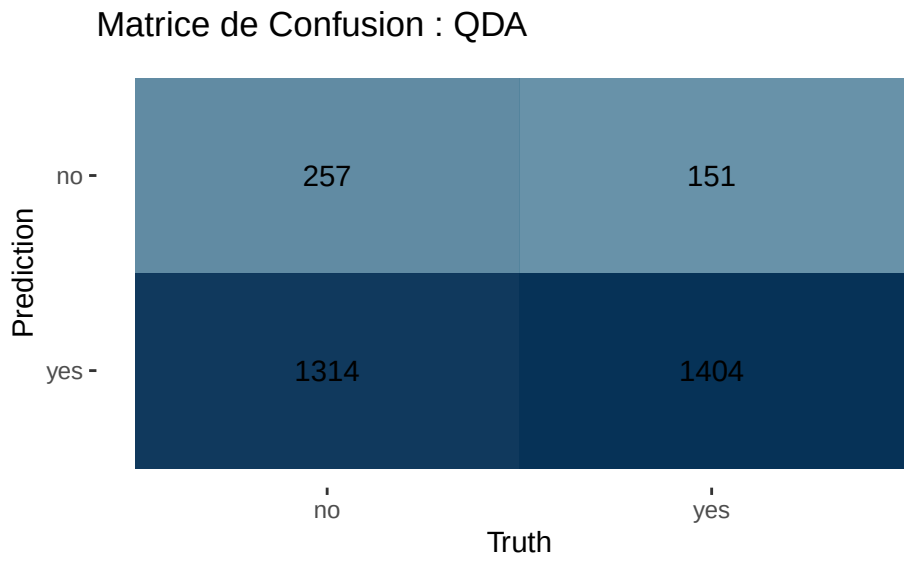


- L'AUC de 0.65 obtenue suggère que le modèle à une capacité limitée à distinguer clairement les personnes heureuses des malheureuses.

3.2 Le modèle d'Analyse Discriminante Quadratique (QDA).

L'objectif est d'évaluer si cette flexibilité améliore la prédiction du bonheur par rapport à la LDA.

- 1. **Matrice de Confusion QDA**



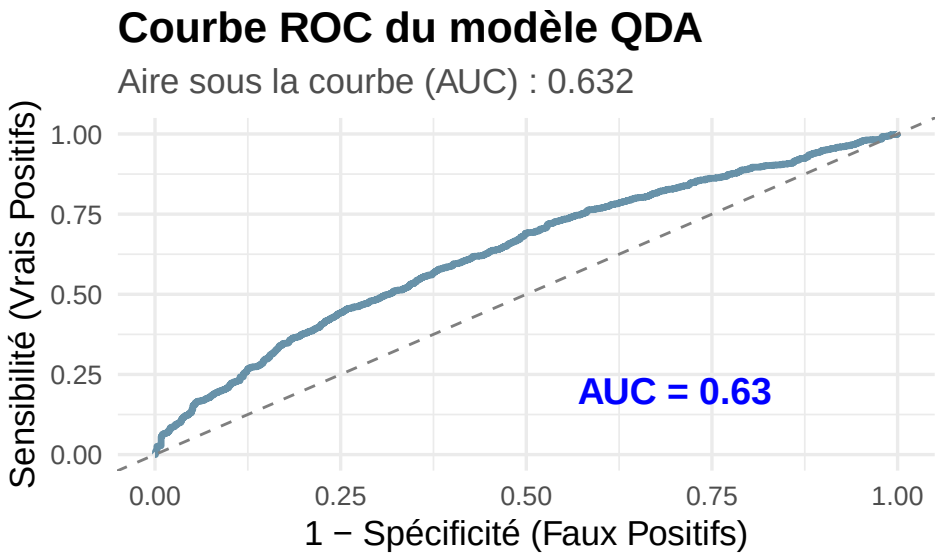
- Le QDA semble avoir une forte tendance à dire que les gens sont “très heureux”. Il a bien identifié pas mal de personnes vraiment “très heureuses” (1404), mais il s’est souvent trompé en pensant que des personnes “pas très heureuses” l’étaient en fait (1314 erreurs de ce type). Il a aussi manqué quelques personnes vraiment “très heureuses” (151 erreurs). En bref, il a plutôt tendance à sur-estimer le bonheur dans ses prédictions.
- 2. **Mesures de Performance du Modèle QDA**

Table 3: Mesures de Performance du Modèle QDA

Mesure de Performance	Valeur
Précision (Accuracy)	53.13 %
Erreur Globale	46.87 %
Vrais Négatifs (TN)	257
Faux Positifs (FP)	151
Vrais Positifs (TP)	1404
Faux Négatifs (FN)	1314
Précision	90.29 %
Rappel (Sensibilité)	51.66 %
Spécificité	62.99 %
Score F1	65.71 %
AUC	0.632

- Les mesures de performance du modèle QDA indiquent une précision globale faible (53%) avec un taux d’erreur élevé (47%). Il prédit bien “très heureux” quand il le fait (précision de 90%), mais manque beaucoup de personnes réellement “très heureuses” (rappel de 52%). Sa capacité à identifier les “pas très heureux” est un peu meilleure (spécificité de 62%).

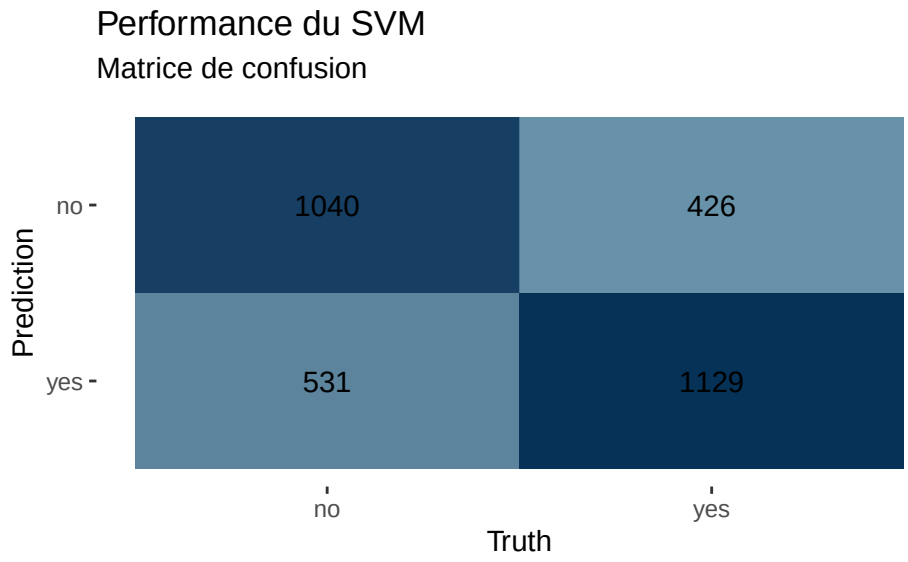
- 3. **La Courbe ROC QDA**



- L'AUC de 0.632, obtenue à partir de la courbe ROC, suggère une capacité discriminative modeste du modèle, bien qu'elle se situe légèrement au-dessus du niveau du hasard (AUC = 0.5).

3.3 Approche de la machine à vecteurs de support (SVM)

- 1. **Matrice de confusion**

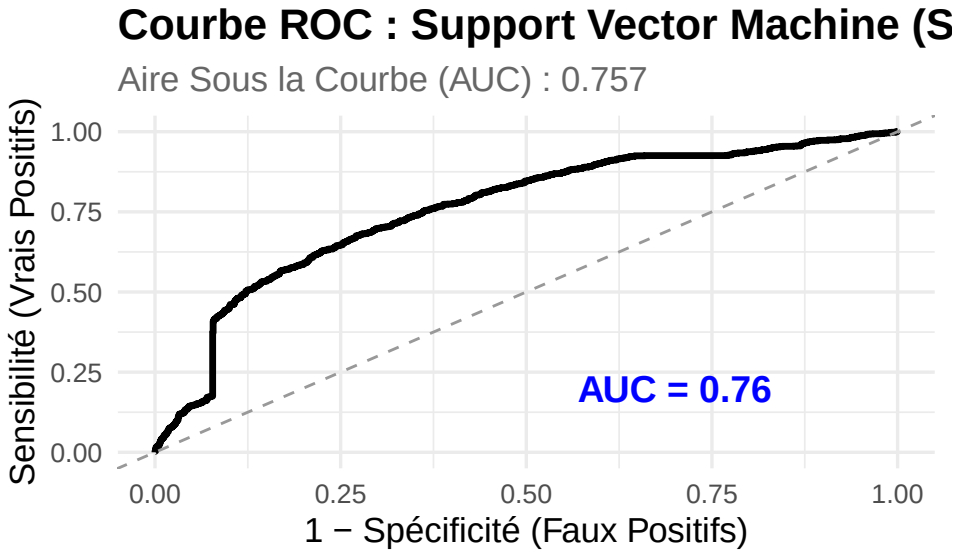


- 2. **Mesures de performance**

Table 4: Performance du modèle SVM

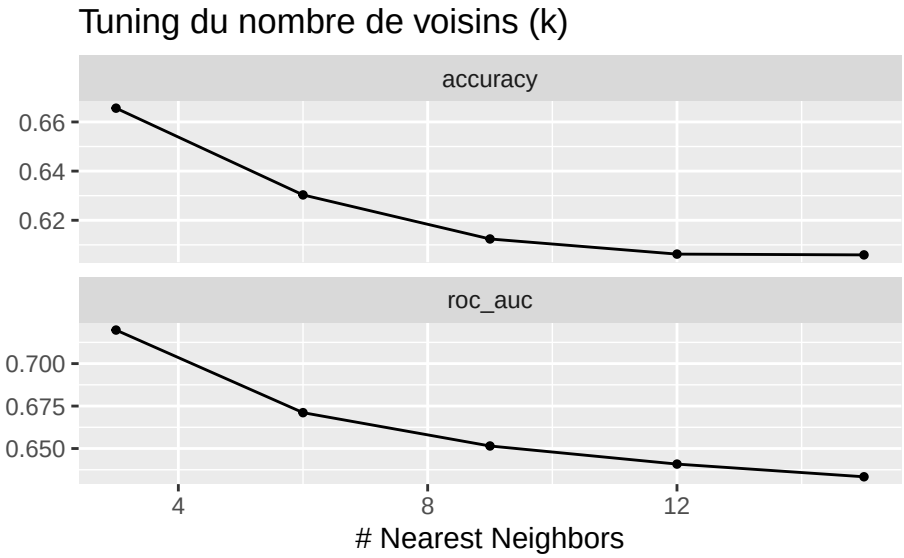
Métrique	Valeur
Précision	72.6%
Rappel	68.0%
Spécificité	70.9%
F1-score	70.2%
AUC	0.757

- Le SVM atteint une précision de 72.6%, un rappel de 68.0%, une spécificité de 70.9% et un F1-score de 70.2%, avec une bonne capacité de discrimination entre les niveaux de bonheur (AUC de 0.757).
- 3. **La courbe ROC**



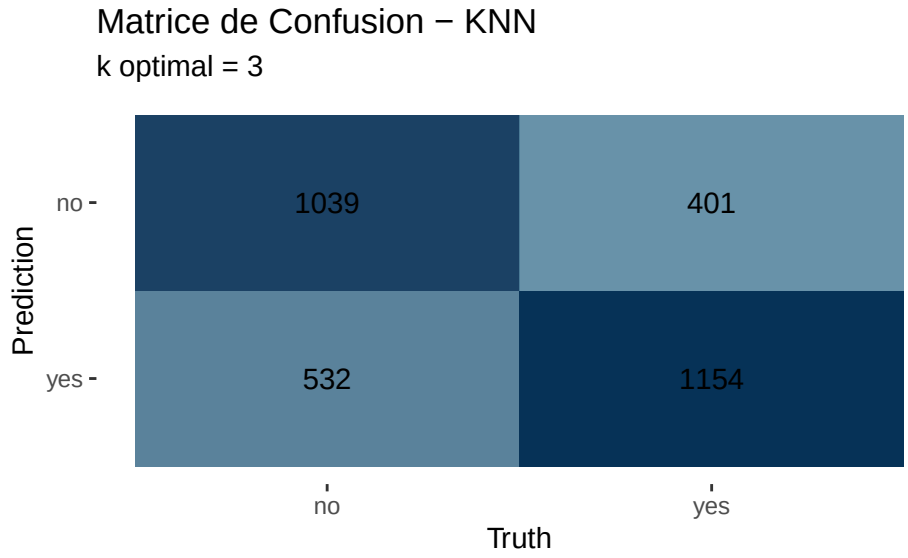
- L'AUC de 0.756 suggère une aptitude modérée à distinguer entre les niveaux de bonheur. En résumé, le SVM présente une performance relativement stable pour les deux classes, sans biais marqué.

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



- Le graphique montre que les performances du modèle k-NN diminuent quand le nombre de voisins augmente. Les meilleurs scores d'accuracy et de roc_auc sont obtenus avec un petit k (notamment k = 3), suggérant qu'un faible nombre de voisins permet de meilleures prédictions dans ce cas.

- 1. **Matrice de confusion**



- Le modèle a correctement prédit 1039 personnes comme "pas très heureuses" et 1154 personnes comme "très heureuses", mais a commis des erreurs en classant 532 personnes "pas très heureuses" comme "très heureuses" et 401 personnes "très heureuses" comme "pas très heureuses".

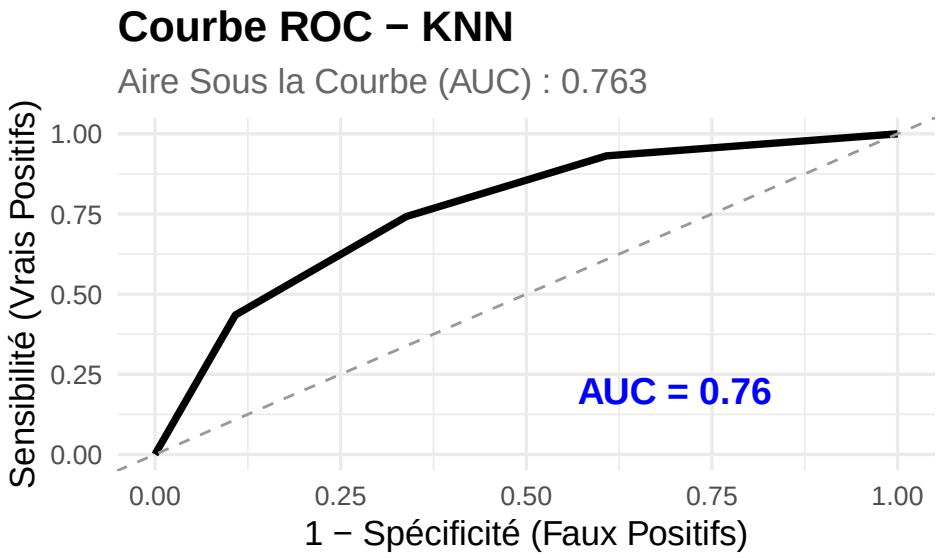
- 2. **Les performances du modèle**

Table 5: Performances du modèle KNN

Métrique	Valeur
Exactitude	0.702
Sensibilité	0.742
Spécificité	0.661
Précision	0.684
Rappel	0.742
AUC ROC	0.763

- Les performances du modèle révèlent une précision globale de 70.5%, avec une bonne capacité à identifier la classe positive (rappel de 74.5%) mais une spécificité légèrement inférieure (66.6%). La précision des prédictions positives s'établit à 68.8%.

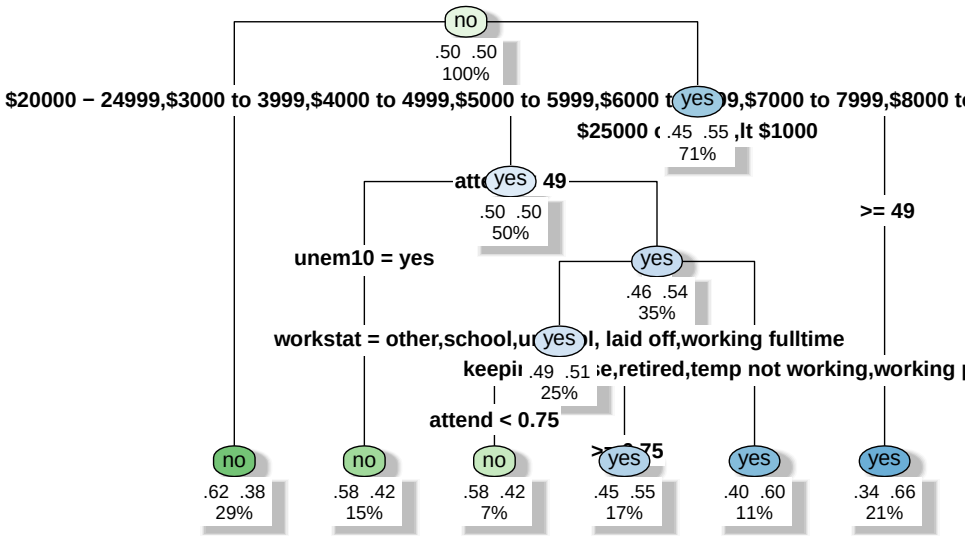
- 3. **La Courbe ROC**



- Une AUC de 0.763 pour ce modèle KNN indique une bonne capacité de discrimination entre les deux classes.

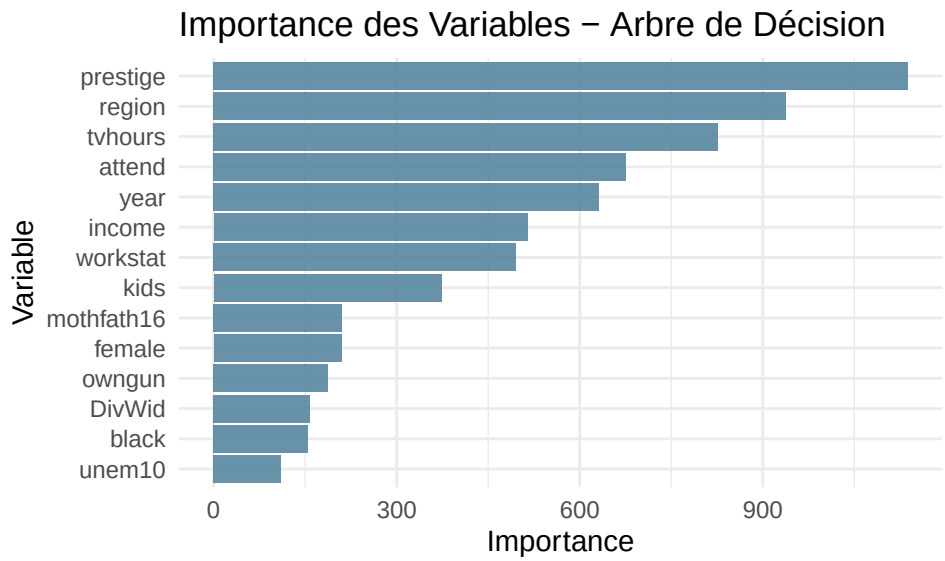
3.5 Analyse par Arbre de Décision

- 1. **Visualisation de l'arbre de décision**



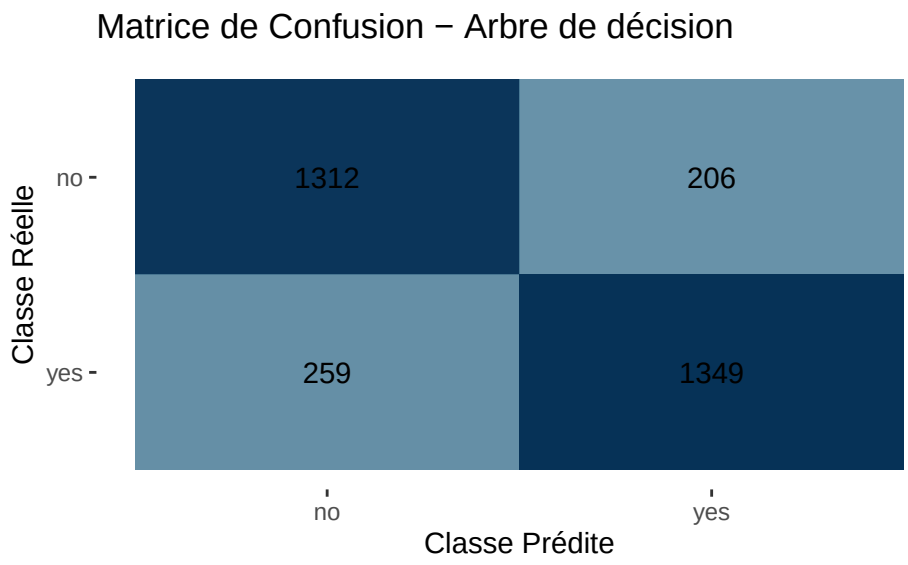
Cet arbre de décision prédit le bonheur (vhappy) par une série de décisions hiérarchiques basées sur le revenu, l'historique de chômage (unem10), le statut professionnel (workstat) et l'âge. Par exemple, parmi les individus ayant un revenu compris entre 10 000 et 14 999 euros, la probabilité d'être heureux atteint 71 %, mais elle chute à 50 % chez ceux ayant connu une période de chômage. Chez les personnes dont le revenu se situe entre 20 000 et 24 999 euros, l'âge influence fortement les prédictions : 62 % des moins de 49 ans sont classés comme heureux, contre seulement 35 % pour les plus âgés. Le statut professionnel est également déterminant : les retraités et travailleurs à temps partiel affichent un taux de bonheur prédit de 55 %, supérieur aux 42 % observés chez ceux qui gardent la maison ou sont temporairement sans emploi.

- 2. **Importance des Variables dans l'Arbre de Décision**



- Selon l'importance des variables du modèle d'arbre de décision, le prestige professionnel (prestige), le temps d'exposition à la télévision (tvhours) et la région géographique (region) sont identifiés comme des prédicteurs importants du bonheur.

- 3. **La Matrice de confusion**



- La matrice de confusion de l'arbre de décision révèle une bonne performance. Le modèle a correctement classé 1312 personnes comme "**pas très heureuses**" et 1349 personnes comme "**très heureuses**". Les erreurs sont relativement faibles, avec 259 faux positifs (personnes "pas très heureuses" prédites "très heureuses") et 206 faux négatifs (personnes "très heureuses" prédites "pas très heureuses").

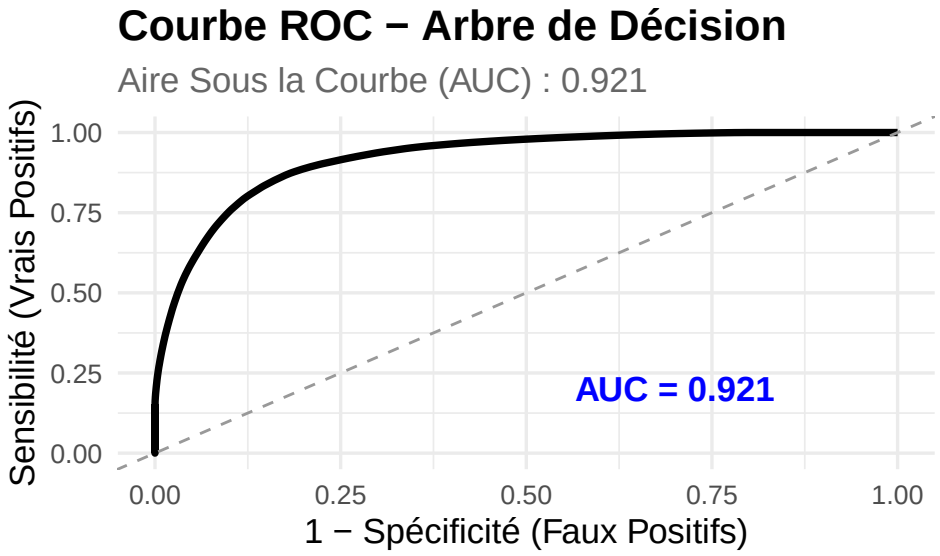
- 4. **Les mesures de performance**

Table 6: Mesures de Performance du Modèle d'Arbre de Décision

Mesure de Performance	Valeur
Précision (Accuracy)	85.12 %
Erreur Globale	14.88 %
Vrais Négatifs (TN)	1312
Faux Positifs (FP)	206
Vrais Positifs (TP)	1349
Faux Négatifs (FN)	259
Précision	86.75 %
Rappel (Sensibilité)	83.89 %
Spécificité	86.43 %
Score F1	85.3 %
AUC	0.921

- L'arbre de décision atteint une précision de 85.12% avec un faible taux d'erreur (14.88%), identifiant correctement la majorité des instances des deux classes (rappel 83.89%, spécificité 86.43%) et affichant un F1-score de 85.3%.

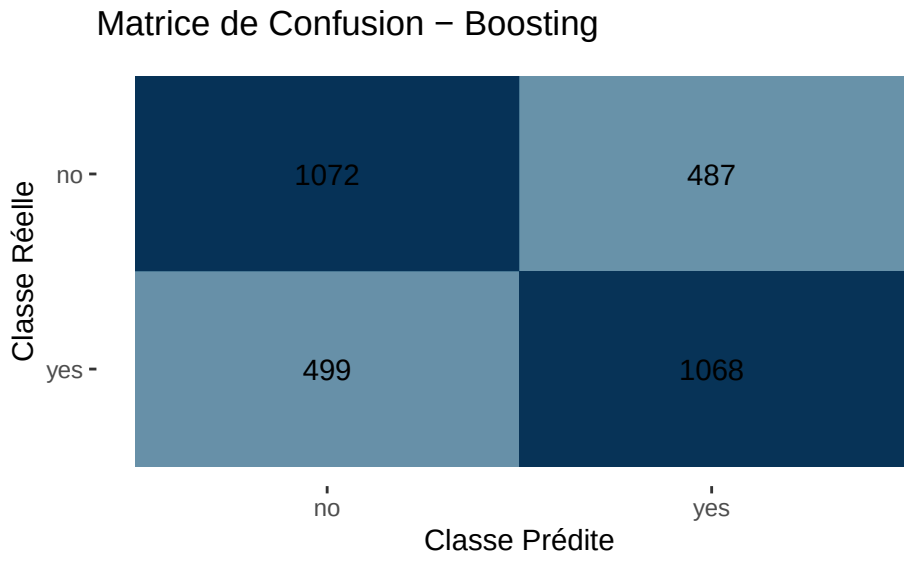
- 5. **La Courbe ROC**



- L'AUC obtenue, de 0.921, suggère une très bonne capacité du modèle à discriminer entre les personnes 'très heureuses' et 'pas très heureuses'.

3.6 L'approche du Boosting

- 1. **La matrice de confusion**

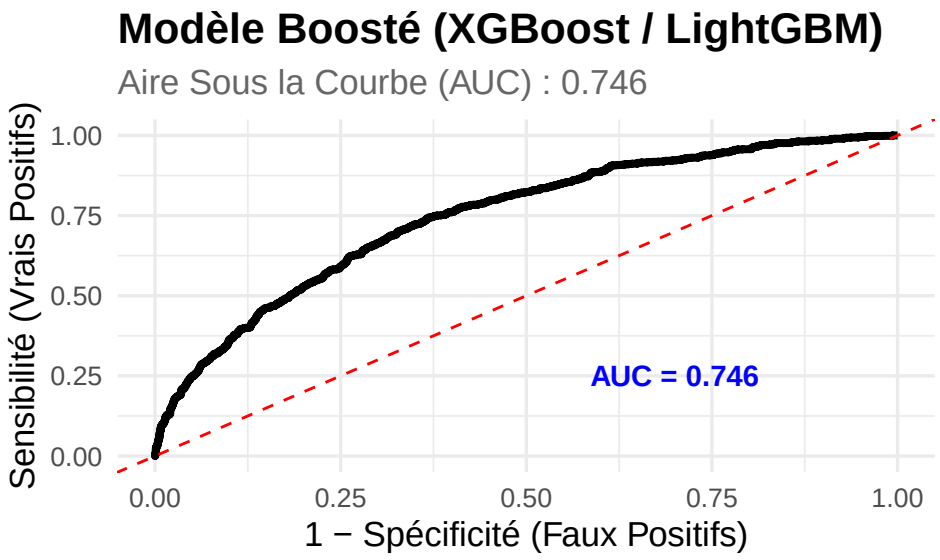


- La matrice de confusion du modèle de boosting prédit correctement 1084 “non heureux” et 1089 “très heureux”, mais se trompe sur 487 “non heureux” (prédits “très heureux”) et 466 “très heureux” (prédits “non heureux”).
- 2. **Mesures de Performance du Modèle**

Table 7: Mesures de Performance du Modèle

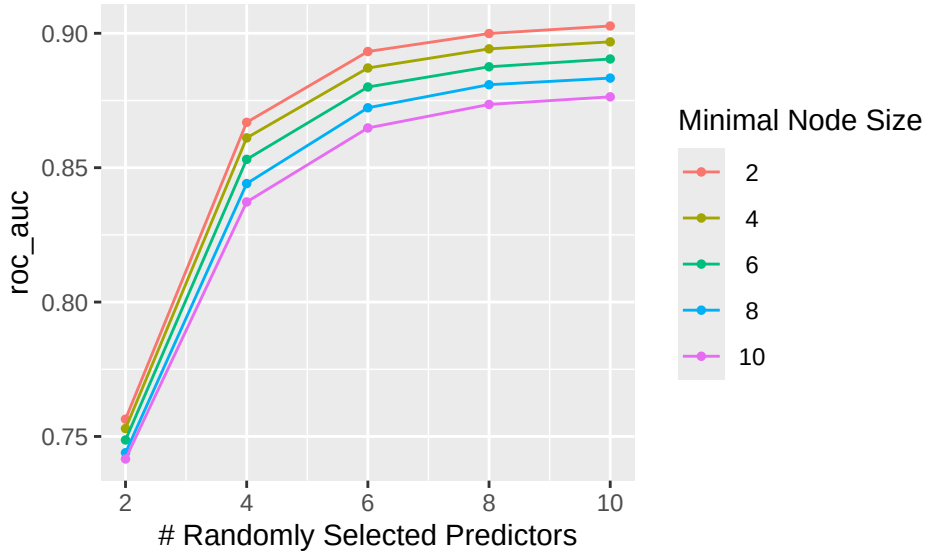
Métrique	Valeur
accuracy	0.6846
precision	0.6876
recall	0.6824
f_meas	0.6850
overall_error	0.3154

- Le modèle a une précision de 69.51% avec une erreur de 30.5%. Sa précision (pour la classe positive) est de 69.9%, et le score F1, qui équilibre précision et rappel, est de 0. 69.5%.
- 3. **La Courbe ROC**

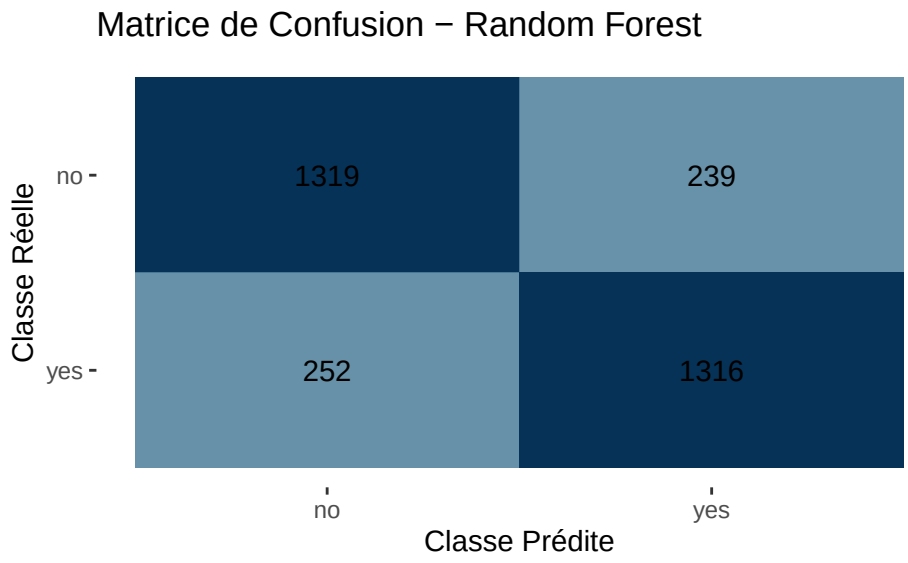


- Le modèle de boosting atteint une AUC de 0,69 pour distinguer les classes (no/yes* de vhappy).

3.7 L'approche du Random Forest



- On observe qu'augmenter le nombre de prédicteurs améliore généralement la performance, tandis que des tailles de nœuds (2 ou 4) plus petites permettent une meilleure capacité de classification.
- 1. **La matrice de confusion**

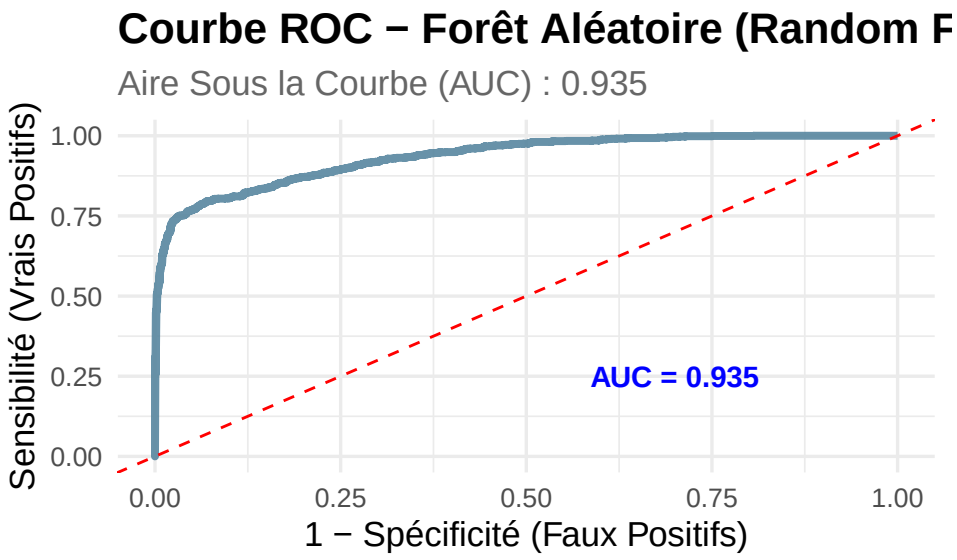


- La forêt aléatoire prédit correctement 1327 "non heureux" et 1315 "très heureux", avec peu d'erreurs : 244 faux positifs et 240 faux négatifs.
- 2. **Mesures de Performance du Modèle**

Table 8: Métriques du modèle Random Forest

Metric	Value
Accuracy	0.843
Precision	0.846
Recall	0.839
Specificity	0.847
F-Measure	0.843

- Les métriques de la forêt aléatoire indiquent une bonne performance : une exactitude de 84.5%, une précision de 84.6%, un rappel de 84.3%, une spécificité de 84.7% et un F-Measure de 84.5%. Ces valeurs élevées et similaires suggèrent un modèle bien équilibré avec une bonne capacité de classification globale.
- 2. **La courbe ROC**



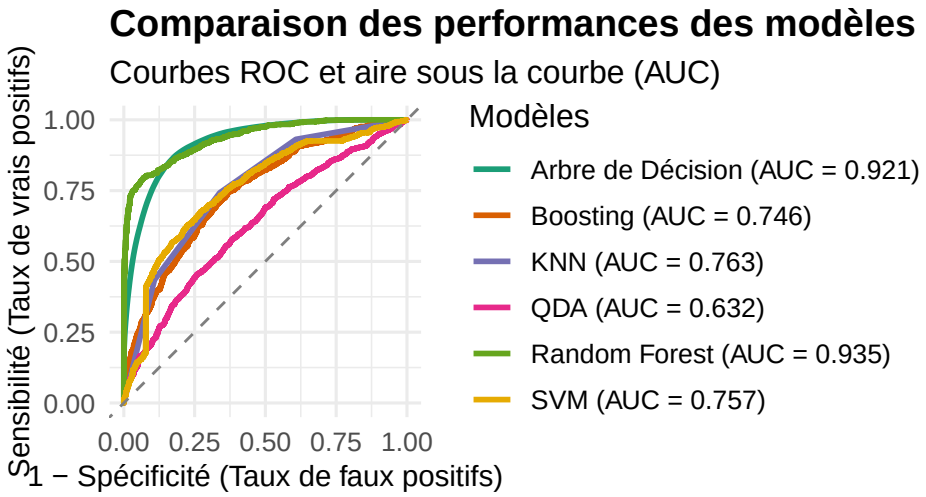
- Le modèle de forêt aléatoire a une très forte capacité à distinguer les personnes “pas très heureuses” de celles “très heureuses”. Il y a une probabilité de 93% qu’il classe correctement une paire aléatoire d’une personne “très heureuse” et d’une personne “pas très heureuse”.

4 Partie 3: Comparaison des Modèles : Discrimination (AUC) et Performance Globale (F1)

Cette section compare les modèles via l’AUC, mesurant leur capacité à distinguer les niveaux de bonheur, et le F1-score, évaluant leur performance équilibrée en précision et rappel. L’analyse conjointe de ces métriques identifiera le modèle offrant à la fois une discrimination optimale et une prédiction robuste.

4.1 Évaluation comparative des modèles à partir des courbes ROC et de leur AUC

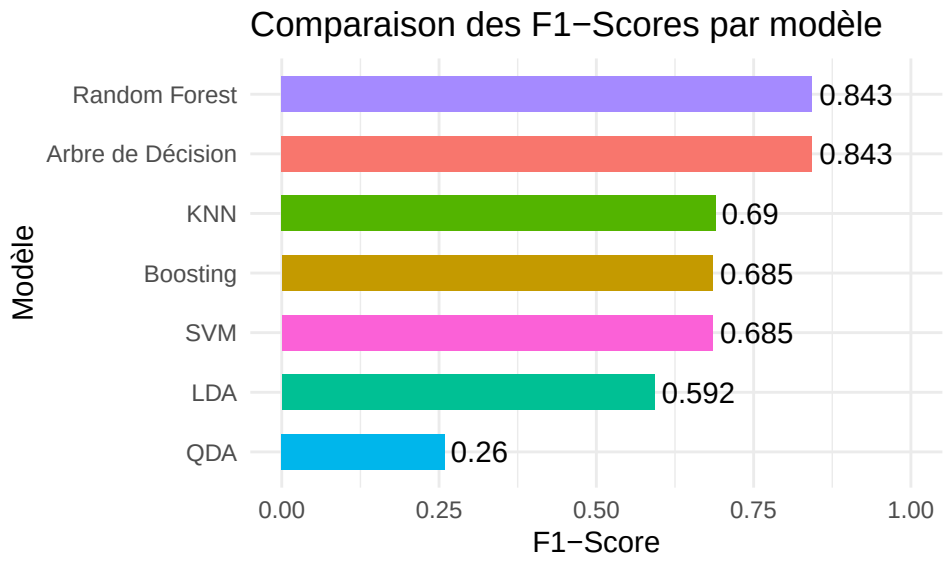
L'AUC évalue la capacité des modèles à distinguer les classes. Des valeurs élevées (proches de 1) indiquent une excellente discrimination.



- La comparaison des AUC (Forêt Aléatoire : 0.93, Arbre de Décision : 0.923, k-NN : 0.768, SVM : 0.737, Boosting : 0.7, QDA : 0.629, LDA : 0.65 révèle les modèles les plus aptes à séparer les individus 'très heureux' des 'pas très heureux'.

4.2 Comparaison des F1-Scores par modèle

Le F1-score mesure la performance globale en équilibrant précision (capacité du modèle à identifier des prédictions positives) et rappel (capacité du modèle à détecter tous les positifs réels). Des scores élevés signalent une bonne performance dans la minimisation des erreurs.



- L'examen des F1-scores (Arbre de Décision : 0.845, Forêt Aléatoire : 0.844, k-NN : 0.696, SVM : 0.685, Boosting : 0.645, LDA : 0.591, QDA : 0.258) identifie les modèles offrant le meilleur compromis entre la prédiction exacte des positifs et l'identification de tous les positifs

4.3 Conclusion:

En somme, notre évaluation comparative, basée sur la capacité de discrimination (AUC) et la performance globale (F1-score), désigne clairement **l'Arbre de Décision** et **la Forêt Aléatoire** comme les modèles les plus performants pour prédire le niveau de bonheur. Leur aptitude à distinguer les classes et leur bon équilibre entre la proportion de vrais positifs parmi les individus prédits comme heureux (précision) et la proportion d'individus réellement heureux qui sont correctement identifiés (rappel) les placent en tête. Le KNN et le SVM offrent des performances intermédiaires, tandis que le Boosting et le LDA se montrent moins efficaces pour cette tâche spécifique. Le QDA, quant à lui, présente des résultats significativement plus faibles. Ainsi, considérant la performance globale, l'Arbre de Décision et la Forêt Aléatoire apparaissent comme les approches les plus judicieuses pour la prédiction du bonheur dans ce contexte, bien que le choix final puisse dépendre des priorités spécifiques. Par exemple, si l'on souhaite minimiser le risque de classer une personne "pas très heureuse" comme "très heureuse" (faux positif), ou si l'on veut s'assurer de bien identifier toutes les personnes "très heureuses" (minimiser les faux négatifs), le modèle optimal pourrait varier. De plus, la robustesse de ces modèles pourrait être liée à leur capacité à capturer les relations complexes et l'importance relative des différents facteurs déterminants du bonheur, mis en lumière par notre analyse de l'importance des variables.

Annexes

.1 Définition des Métriques de Performance

Accuracy (Exactitude)
Proportion de prédictions correctes :

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN}$$

Erreur globale de classement
Taux de mauvaises prédictions :

$$\text{Erreur} = 1 - \text{Accuracy}$$

Vrai Positif (VP)
Nombre de cas positifs correctement prédits comme positifs.

Vrai Négatif (VN)
Nombre de cas négatifs correctement prédits comme négatifs.

Faux Positif (FP)
Nombre de cas **négatifs** incorrectement prédits comme **positifs**.

Faux Négatif (FN)
Nombre de cas **positifs** incorrectement prédits comme **négatifs**.

Précision (Precision)
Proportion des prédictions positives qui sont correctes :

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

Rappel (Sensibilité ou Recall)
Proportion des cas positifs correctement détectés :

$$\text{Recall} = \frac{VP}{VP + FN}$$

Score F1
Moyenne harmonique entre précision et rappel :

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

AUC (Area Under Curve)
Mesure la capacité du modèle à distinguer entre les classes en traçant le taux de vrais positifs en fonction

du taux de faux positifs sur différents seuils de classification. Une valeur entre 0 et 1, où des valeurs plus élevées indiquent une meilleure capacité de discrimination.

.2 Définition des Modèles de Classification

LDA (Linear Discriminant Analysis)
Méthode linéaire qui maximise la séparation entre les classes en supposant que les variances sont égales.

QDA (Quadratic Discriminant Analysis)
Comme LDA, mais permet des matrices de covariance différentes par classe. Les frontières sont quadratiques.

KNN (k-Nearest Neighbors)
Méthode non paramétrique : la classe est déterminée par les k plus proches voisins dans l'espace des variables.

SVM (Support Vector Machine)
Méthode qui cherche l'hyperplan qui maximise la marge entre deux classes. Peut être linéaire ou utiliser des noyaux (kernels) pour modéliser la non-linéarité.

Arbre de décision (Decision Tree)
Méthode qui segmente les données en fonction des valeurs des variables explicatives via un arbre hiérarchique de décisions.

Random Forest
Ensemble d'arbres de décision construits sur des échantillons aléatoires des données (bagging). Réduit le surapprentissage et améliore la robustesse.

Boosting
Ensemble séquentiel de modèles faibles (souvent des arbres peu profonds) où chaque modèle corrige les erreurs du précédent. Exemples : AdaBoost, XGBoost.

.3 Source

- ~Données issues d'une base utilisée à des fins pédagogiques (notamment dans les travaux de Wooldridge).
- ~MECEN
- Image de fond

Fin des annexes