Memora

Memora

Professeur référent : Ahmed Dhouibi

Ecole Supérieure Privée de Technologie et Ingénierie TEK-UP
ING-4-J-SDIA-A

Rapport pour le 24/05/2025

Nour Elhouda Zaabii
Mohamed Slim Chouaib
Mohamed Ali Bouhadja

Table des matières

Table des figures

Introduction Générale

La maladie d'**Alzheimer** est une maladie neurodégénérative chronique caractérisée par une perte progressive des capacités cognitives, notamment la mémoire, le langage et le raisonnement. Elle représente un **Défi majeur** de santé publique en raison de sa prévalence croissante et de l'impact important sur la qualité de vie des patients et de leurs proches.

Le diagnostic précoce est essentiel pour une prise en charge adaptée, mais les méthodes traditionnelles reposent souvent sur des évaluations cliniques et des examens complémentaires coûteux, invasifs et parfois tardifs. Par ailleurs, la complexité et l'hétérogénéité des données cliniques rendent difficile une identification rapide et fiable des patients atteints.

Dans ce contexte, l'**Apprentissage automatique** appliqué à des données cliniques et biologiques offre une opportunité prometteuse d'automatiser et d'améliorer la précision du diagnostic. Ce projet vise à développer et comparer plusieurs **Modèles de classification supervisée** (KNN, régression logistique, random forest, SVM) pour prédire la présence de la maladie d'**Alzheimer** à partir de données tabulaires.

Une démarche originale est d'intégrer une étape de **Clustering non supervisé** pour segmenter les patients en groupes homogènes, afin d'explorer si cette stratification peut améliorer la performance des modèles classiques. Ce travail s'inscrit dans une approche de recherche rigoureuse, visant à analyser et interpréter les résultats pour guider les futures applications d'**Intelligence Artificielle** en santé.

Chapitre 1

Compréhension du projet

1.1 Objectif métier

La maladie d'Alzheimer, en tant que pathologie neurodégénérative majeure, représente un enjeu crucial pour la santé publique. La détection précoce et précise de cette maladie permet non seulement d'optimiser les traitements et la prise en charge des patients, mais aussi de réduire les coûts associés aux soins sur le long terme.

L'objectif métier de ce projet est de concevoir un système d'aide à la décision capable de prédire la présence de la maladie d'Alzheimer à partir de données cliniques et biologiques. Ce système automatisé vise à accompagner les professionnels de santé en fournissant un diagnostic fiable, rapide, et reproductible, facilitant ainsi la détection précoce et la prise en charge adaptée des patients.

1.2 Problématique

Le diagnostic actuel repose sur des examens cliniques, cognitifs, et parfois d'imagerie, qui peuvent être coûteux, invasifs et sujets à des biais d'interprétation. Par ailleurs, la complexité et l'hétérogénéité des données cliniques, biologiques et démographiques compliquent la mise en place d'un modèle prédictif unique capable de généraliser efficacement à tous les profils de patients.

De plus, la population étudiée peut présenter une grande variabilité sous-jacente, avec des sous-groupes de patients aux caractéristiques cliniques et biologiques différentes. Il est donc pertinent de questionner l'impact d'une segmentation préalable des données, via des techniques de clustering non supervisé, pour mieux capturer cette diversité et potentiellement améliorer les performances des modèles supervisés de classification.

La problématique centrale de ce projet est donc de développer une démarche robuste combinant apprentissage supervisé et non supervisé, afin de maximiser la capacité prédictive tout en restant interprétable et applicable en contexte clinique.

1.3 Objectifs du projet

L'objectif principal de ce projet est de concevoir un système automatique performant pour la classification des patients atteints de la maladie d'Alzheimer à partir de données cliniques et biologiques.

Pour atteindre cet objectif, plusieurs étapes clés ont été mises en œuvre :

- Préparation et nettoyage rigoureux des données, incluant la gestion des valeurs manquantes, la suppression des variables constantes et la réduction de la redondance par élimination des variables fortement corrélées.
- Développement et optimisation de modèles d'apprentissage supervisé classiques (KNN, régression logistique, random forest, SVM) pour la classification binaire Alzheimer / Non Alzheimer, avec validation croisée et réglage des hyperparamètres.
- Intégration d'une étape de clustering non supervisé afin de segmenter les patients en groupes homogènes, suivie de l'inclusion de cette information comme variable explicative dans les modèles supervisés pour évaluer son impact.
- Mise en place d'une évaluation rigoureuse des modèles via des métriques multiples (accuracy, AUC, F1-score), permettant une comparaison complète des approches avec et sans clustering.
- Développement d'une API backend permettant l'intégration des modèles dans un environnement de production, capable de recevoir les données patient et de retourner des prédictions en temps réel.
- Création d'une interface utilisateur web intuitive, facilitant la saisie des données, la consultation des résultats et offrant une expérience fluide et accessible sur différents appareils.

Ainsi, ce projet vise à fournir une solution complète et opérationnelle pour assister le diagnostic médical de la maladie d'Alzheimer via l'intelligence artificielle, tout en assurant une facilité d'usage et une adaptabilité aux besoins réels des utilisateurs.

Chapitre 2

Compréhension des données

2.1 Sources de données

Le jeu de données regroupe des mesures cliniques, biologiques, démographiques et comportementales recueillies auprès de patients présentant divers profils cognitifs. Chaque patient est identifié par un ensemble de variables reflétant son état de santé, ses antécédents médicaux et ses habitudes de vie.

Ces données ont été collectées dans un cadre clinique strict, avec un diagnostic binaire indiquant la présence ou l'absence de la maladie d'Alzheimer, permettant de modéliser un problème de classification supervisée.

2.2 Description des données

Le dataset comprend des variables variées, réparties comme suit :

- **Variables démographiques** : âge (`Age`), sexe (`Gender`), ethnie (`Ethnicity`), niveau d'éducation (`EducationLevel`).
- **Indicateurs de mode de vie** : indice de masse corporelle (`BMI`), tabagisme (`Smoking`), consommation d'alcool (`AlcoholConsumption`), activité physique (`PhysicalActivity`), qualité du régime alimentaire (`DietQuality`), qualité du sommeil (`SleepQuality`).
- **Antécédents médicaux** : antécédents familiaux d'Alzheimer (`FamilyHistoryAlzheimers`), maladies cardiovasculaires (`CardiovascularDisease`), diabète (`Diabetes`), hypertension (`Hypertension`), traumatisme crânien (`HeadInjury`), dépression (`Depression`).
- **Mesures biologiques et physiologiques** : pression artérielle systolique et diastolique (`SystolicBP`, `DiastolicBP`), profils lipidiques (`CholesterolTotal`, `CholesterolLDL`, `CholesterolHDL`, `CholesterolTriglycerides`).
- **Fonctions cognitives et comportementales** : score Mini-Mental State Examination (`MMSE`), évaluation fonctionnelle (`FunctionalAssessment`), plaintes mnésiques (`MemoryComplaints`) troubles comportementaux (`BehavioralProblems`), activités de la vie quotidienne (`ADL`), confusion, désorientation, changements de personnalité, difficulté à accomplir les tâches (`Confusion`, `Disorientation`, `PersonalityChanges`, `DifficultyCompletingTasks`).
- **Variable cible** : diagnostic de la maladie (`Diagnosis`), codée en deux classes discrètes.

2.3 Nettoyage et préparation

Le dataset a fait l'objet d'un nettoyage rigoureux :

- Suppression des colonnes non informatives (ex. identifiants patients).

— Gestion des valeurs manquantes : imputation par la médiane pour les variables numériques, et par la catégorie la plus fréquente pour les variables catégorielles.
— Élimination des variables constantes et des variables numériques fortement corrélées (corrélation supérieure à 0.8), réduisant la redondance et améliorant la stabilité des modèles.

2.4 Analyse exploratoire

L'exploration initiale a permis d'observer une distribution équilibrée des classes diagnostiques, garantissant une base saine pour l'entraînement des modèles (cf. figure 2.1).

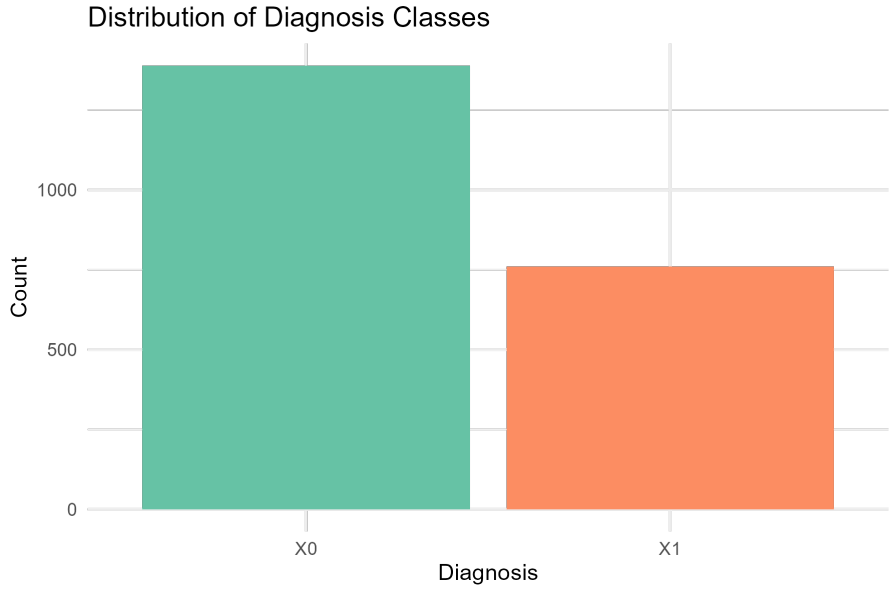

FIGURE 2.1 – Distribution des classes diagnostiques dans le jeu de données.

Une analyse des corrélations entre variables numériques est présentée à la figure 2.2. Cette visualisation permet d'apprécier les relations linéaires entre variables et d'identifier d'éventuelles redondances.

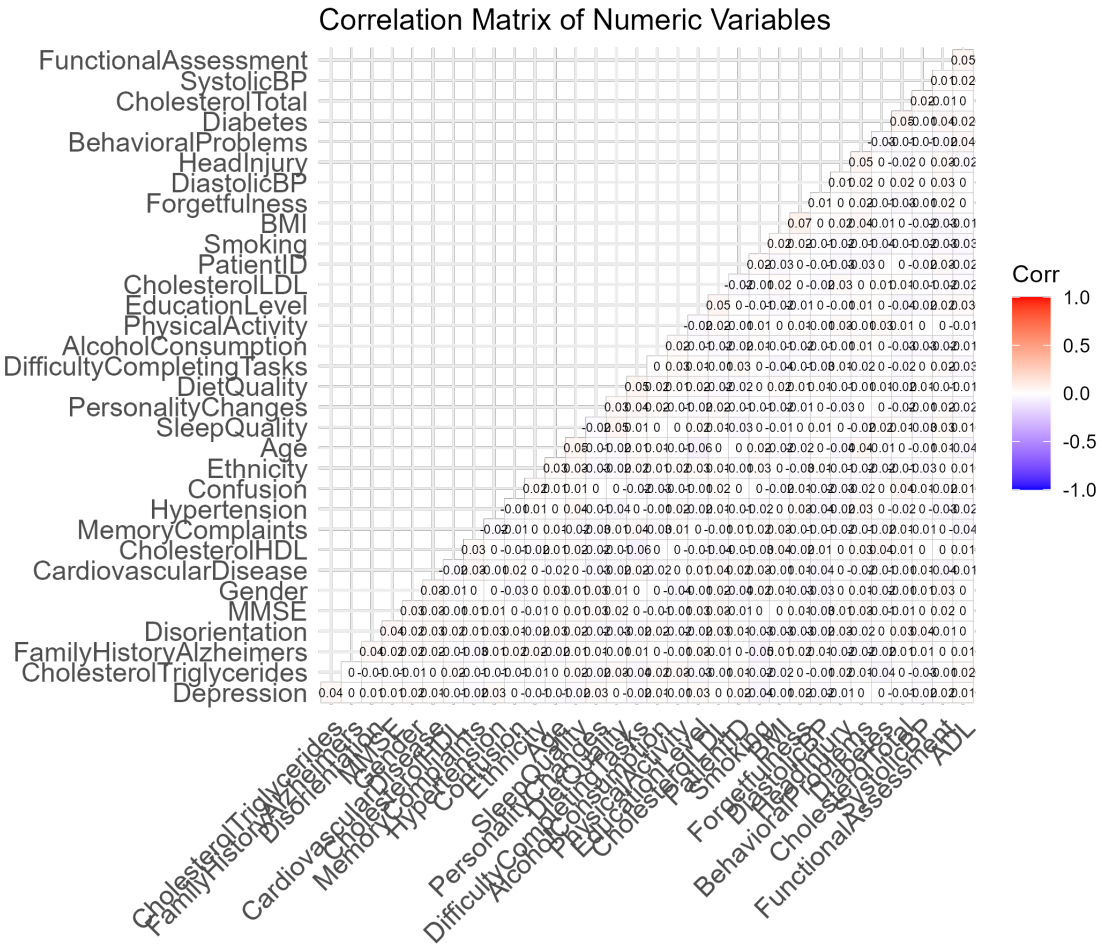

FIGURE 2.2 – Carte de corrélation des variables numériques du jeu de données.

La présence de groupes distincts dans les données, révélés par un clustering non supervisé via k-moyennes sur les variables numériques, met en lumière des sous-populations potentielles avec des profils cliniques spécifiques (figure 4.2).

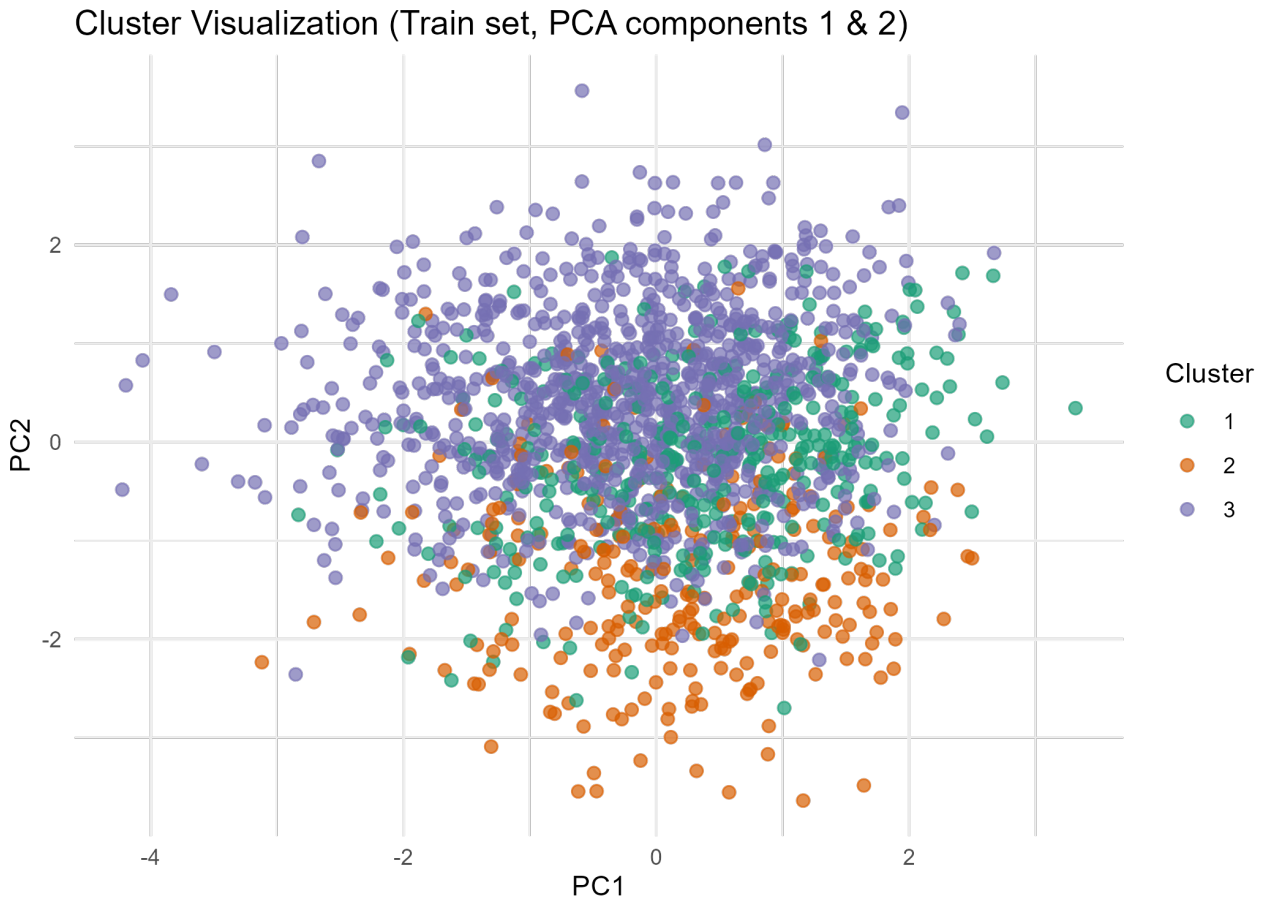

FIGURE 2.3 – Visualisation des clusters identifiés sur le jeu d'entraînement selon les deux premières composantes principales.

2.5 Conclusion

La richesse et la diversité des variables mesurées offrent une opportunité d'analyse fine et multidimensionnelle du profil des patients. Cette complexité nécessite des méthodes avancées de préparation, de sélection, et d'exploration, qui ont été intégrées dans le pipeline global du projet pour aboutir à des modèles performants et robustes.

Chapitre 3

Préparation des données

3.1 Nettoyage des données

Avant toute modélisation, une étape cruciale de nettoyage a été réalisée pour garantir la qualité et la cohérence des données. Cette phase comprend plusieurs actions :

- **Suppression des colonnes non informatives** : telles que les identifiants patients, qui n'apportent aucune information prédictive.
- **Gestion des valeurs manquantes** : Les variables contenant des données manquantes ont été traitées par imputation. Pour les variables numériques, l'imputation a été réalisée par la médiane afin de limiter l'influence des valeurs extrêmes. Pour les variables catégorielles, la catégorie la plus fréquente a été utilisée.
- **Suppression des variables constantes** : Ces variables ne fournissent aucune variabilité utile pour la modélisation.

3.2 Traitement des corrélations

Une analyse des corrélations entre variables numériques a été effectuée afin de détecter d'éventuelles redondances. Cependant, aucune variable n'a été supprimée car aucune corrélation trop forte n'a été détectée.

Pour mieux visualiser la structure des corrélations, la figure 3.1 présente la matrice de corrélation avec un regroupement hiérarchique des variables, mettant en évidence des groupes de variables fortement liées.

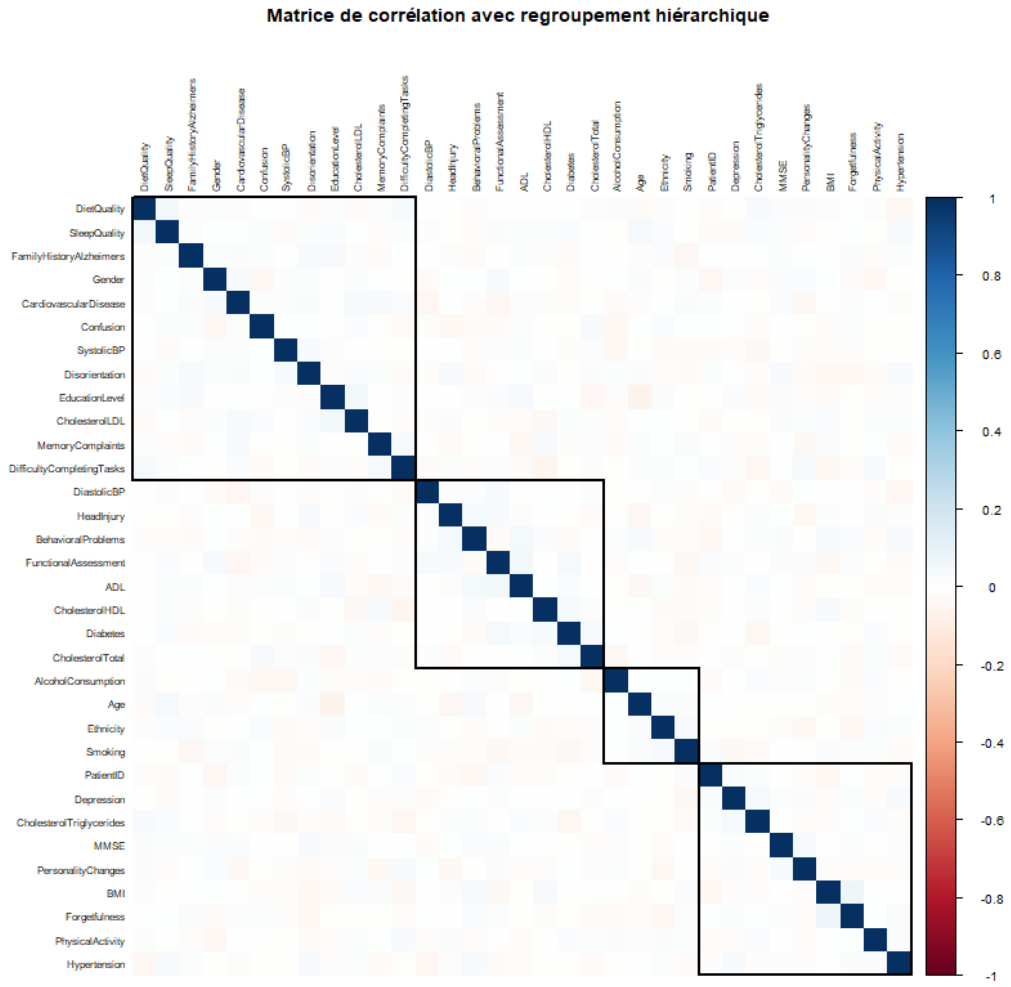

FIGURE 3.1 – Matrice de corrélation des variables numériques avec dendrogramme hiérarchique.

3.3 Mise à l'échelle des variables

Pour certains algorithmes sensibles à l'échelle des données (KNN, SVM, régression logistique), les variables numériques ont été centrées et réduites (standardisation). La figure 4.1 illustre la distribution des variables numériques après cette mise à l'échelle sur le jeu d'entraînement.

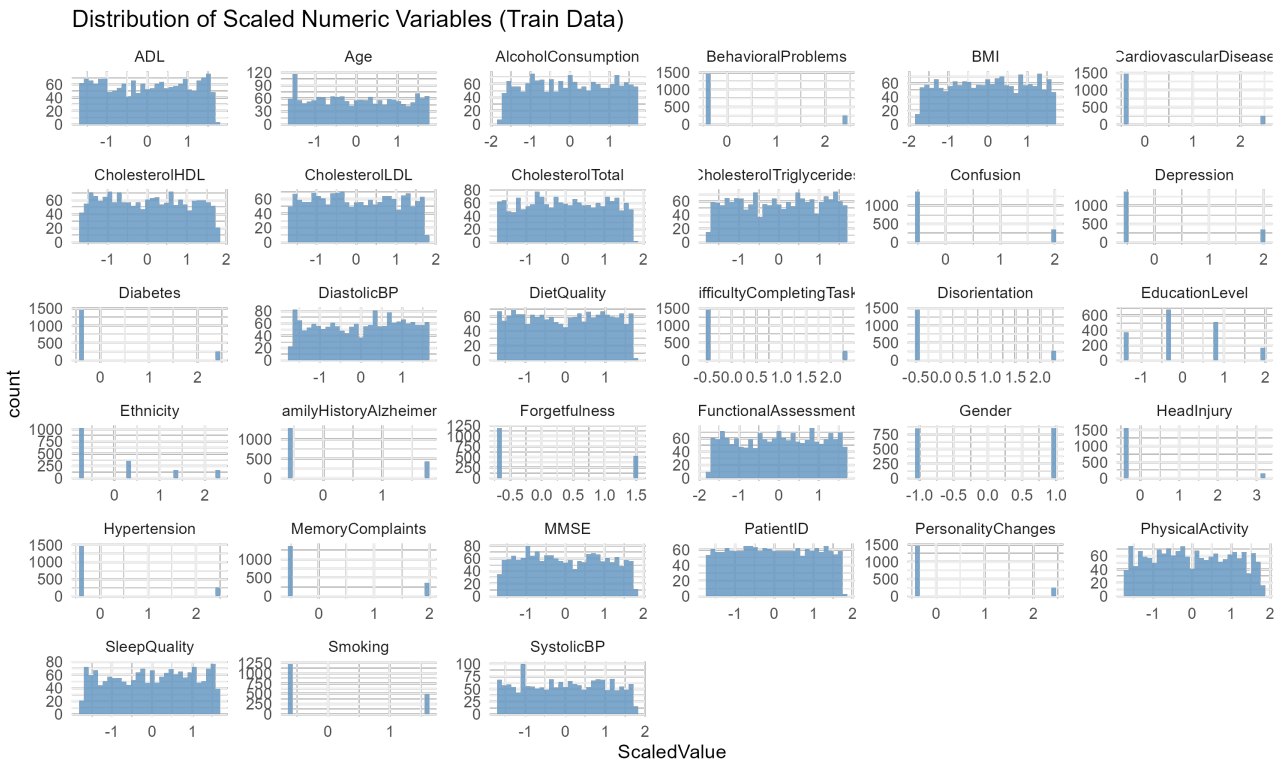

FIGURE 3.2 – Distribution des variables numériques après mise à l'échelle (jeu d'entraînement).

3.4 Partitionnement des données

Le jeu de données a été séparé en deux sous-ensembles :

— un jeu d'entraînement représentant 80% des données,

— un jeu de test correspondant aux 20% restants.

Cette séparation permet d'évaluer la performance des modèles sur des données non vues lors de l'entraînement, assurant une mesure fiable de leur généralisation.

Chapitre 4

Modélisation

4.1 Introduction

Dans le cadre du diagnostic assisté de la maladie d'Alzheimer, la modélisation prédictive joue un rôle central. Elle vise à construire des classificateurs capables de distinguer efficacement les patients malades des patients sains à partir de données cliniques, biologiques et comportementales complexes. Cette étape exige une approche rigoureuse, combinant des méthodes variées et une exploitation approfondie de la structure des données.

4.2 Choix des modèles

Afin de couvrir une large gamme de propriétés et d'aptitudes prédictives, nous avons sélectionné quatre algorithmes aux caractéristiques complémentaires :

- **K-Nearest Neighbors (KNN)** : Ce modèle, fondé sur la notion de proximité dans l'espace des variables, est intuitif et simple à mettre en œuvre. Il permet de classifier un individu selon les classes majoritaires de ses voisins les plus proches. Sa principale limite réside dans sa sensibilité à la dimensionnalité et à la mise à l'échelle des variables.
- **Régression logistique** : Modèle paramétrique linéaire qui estime la probabilité d'appartenance à une classe par une fonction logistique appliquée à une combinaison linéaire des variables explicatives. Ce modèle est facilement interprétable et constitue une bonne base pour la classification binaire.
- **Random Forest** : Méthode d'ensemble combinant plusieurs arbres de décision construits sur des échantillons bootstrapés et sous-ensembles aléatoires de variables. Elle est réputée pour sa robustesse face au bruit, sa capacité à modéliser des relations complexes non linéaires, et sa résistance au surapprentissage.
- **Support Vector Machines (SVM)** : Algorithme cherchant à maximiser la marge de séparation entre classes dans un espace potentiellement transformé par un noyau. Cette méthode est efficace pour gérer des frontières complexes mais requiert une bonne normalisation des données et un réglage soigné des hyperparamètres.

Cette diversité méthodologique permet d'explorer plusieurs perspectives analytiques, de la simplicité et interprétabilité à la puissance prédictive.

4.3 Standardisation des variables

La plupart des modèles choisis, notamment KNN, régression logistique et SVM, sont sensibles à l'échelle des variables. Par conséquent, nous avons appliqué une mise à l'échelle standard (centrage et réduction) des variables numériques pour :

— Éviter que des variables à grande amplitude dominent les mesures de distance ou d'influence,
— Faciliter la convergence des algorithmes d'optimisation,
— Améliorer la stabilité et la performance des modèles.

La figure 4.1 illustre la distribution homogène des variables numériques après standardisation.

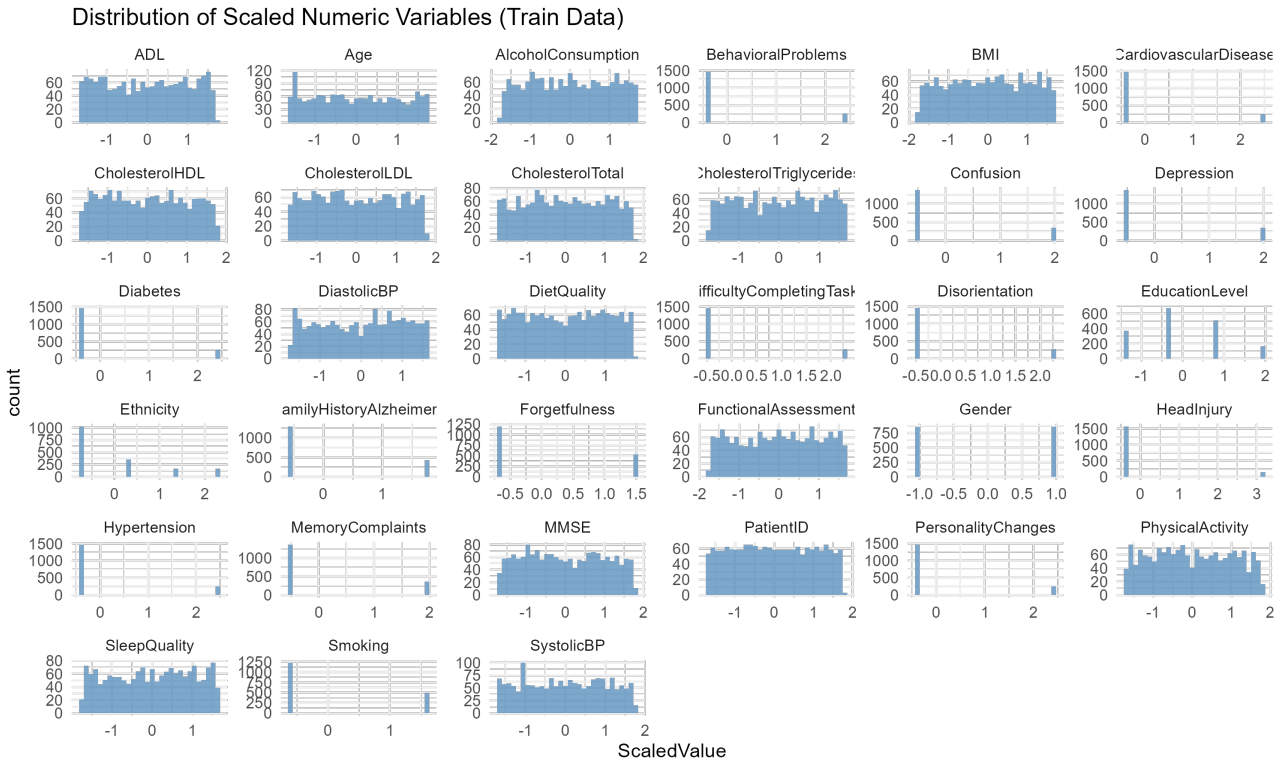

FIGURE 4.1 – Distribution des variables numériques après standardisation (jeu d'entraînement).

4.4 Clustering : méthode, justification et intégration

4.4.1 Pourquoi faire du clustering ?

Les données cliniques et biologiques présentent souvent des structures latentes, correspondant à des sous-populations aux caractéristiques homogènes. Identifier ces groupes peut :

— Révéler des profils patients distincts non visibles directement dans les variables originales,
— Améliorer la compréhension des données,
— Permettre d'intégrer cette information dans la modélisation supervisée pour améliorer la précision prédictive.

4.4.2 Méthode choisie : k-moyennes

Nous avons opté pour le clustering par *k-moyennes* pour plusieurs raisons :

— C'est une méthode simple, rapide et efficace pour segmenter des données numériques continues,
— Elle permet de partitionner les observations en k groupes en minimisant la variance intra-groupe,
— Elle est adaptée à notre jeu de données standardisé,
— Le nombre de clusters $k = 3$ a été choisi après expérimentation et observation des résultats.

4.4.3 Impact et intégration du clustering

La variable de cluster issue de cette partition est ajoutée comme variable catégorielle aux jeux d'apprentissage et de test. Cette nouvelle variable donne aux modèles une information supplémentaire sur la structure latente des données, pouvant améliorer la séparation entre classes.

Cette démarche hybride associe apprentissage non supervisé (clustering) et supervisé (classification), capitalisant sur les forces de chaque approche.

La figure 4.2 montre la projection des clusters sur les deux premières composantes principales, confirmant la cohérence des groupes détectés.

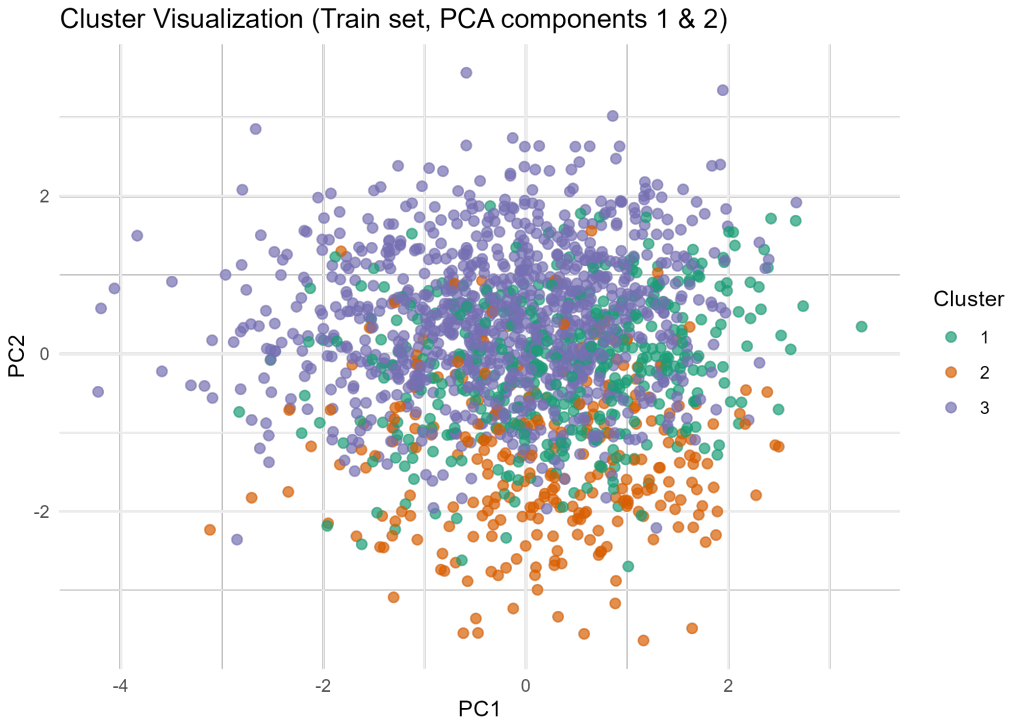

FIGURE 4.2 – Visualisation des clusters via analyse en composantes principales (ACP).

4.5 Entraînement et validation

Pour chaque modèle, deux versions ont été entraînées : sans et avec la variable cluster. Nous avons utilisé une validation croisée stratifiée à 5 plis, garantissant une optimisation fiable des hyperparamètres et une évaluation robuste.

L'aire sous la courbe ROC (AUC) a été la métrique principale utilisée pour la sélection des meilleurs modèles, car elle reflète la capacité à discriminer efficacement entre les classes.

4.6 Conclusion

Cette approche méthodologique complète, combinant prétraitements adaptés, clustering pertinent et modélisation variée, permet de maximiser les performances prédictives tout en conservant un cadre interprétable et robuste. Le clustering apporte une valeur ajoutée significative, notamment dans la captation de la diversité latente des profils patients.

Chapitre 5

Évaluation

5.1 Introduction aux métriques

Pour évaluer la performance des modèles, plusieurs métriques complémentaires ont été utilisées :

- **Accuracy (Exactitude)** : proportion d'observations correctement classées.
- **AUC (Area Under the ROC Curve)** : mesure la capacité globale à distinguer les classes sur tous les seuils possibles.
- **F1-Score** : moyenne harmonique entre précision et rappel, importante dans les contextes où l'équilibre entre faux positifs et faux négatifs est crucial.

5.2 Résultats globaux

Le tableau 5.1 récapitule les performances des modèles sur le jeu de test, avec et sans prise en compte du clustering.

Modèle	Clustering	Accuracy	AUC	F1-Score
KNN	Sans	0.762	0.839	0.560
KNN	Avec	0.765	0.840	0.567
Régression logistique	Sans	0.851	0.901	0.782
Régression logistique	Avec	0.851	0.901	0.782
Random Forest	Sans	0.935	0.983	0.903
Random Forest	Avec	0.939	0.983	0.910
SVM	Sans	0.858	0.918	0.787
SVM	Avec	0.860	0.918	0.792

TABLE 5.1 – Performances des modèles sur le jeu de test

5.3 Matrice de confusion

La figure 5.1 présente la matrice de confusion du modèle Random Forest avec clustering, qui affiche les meilleures performances.

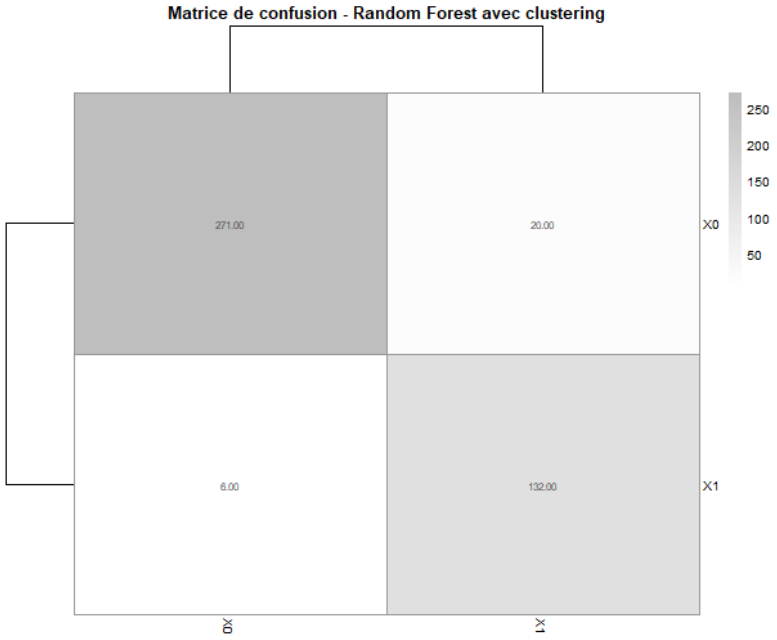

FIGURE 5.1 – Matrice de confusion du modèle Random Forest avec clustering sur le jeu de test

On y observe une forte proportion de vrais positifs et vrais négatifs, ainsi que des erreurs limitées (faux positifs et faux négatifs).

5.4 Courbes ROC

Les courbes ROC des modèles sans et avec clustering sont présentées respectivement en figures 5.2 et 5.4.

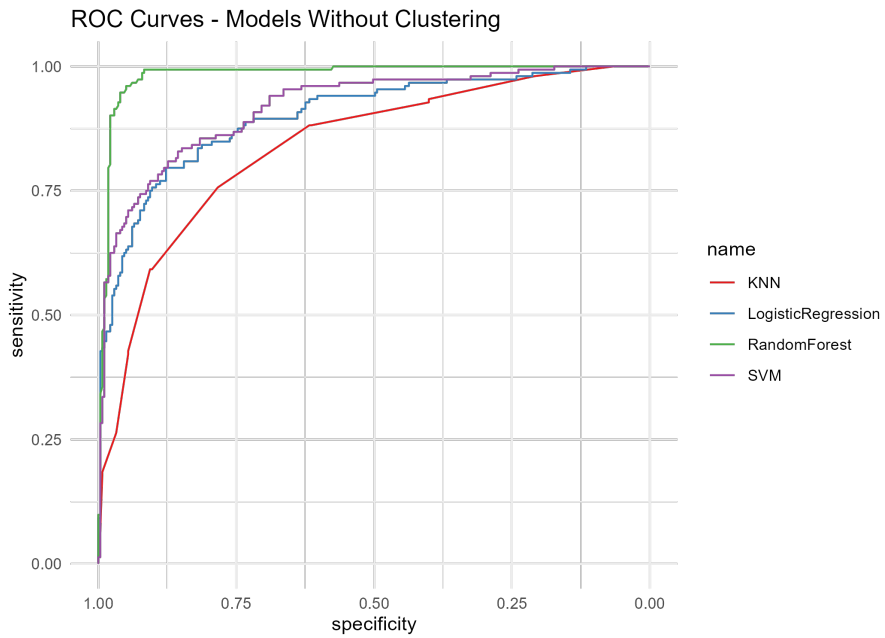

FIGURE 5.2 – Courbes ROC des modèles sans clustering

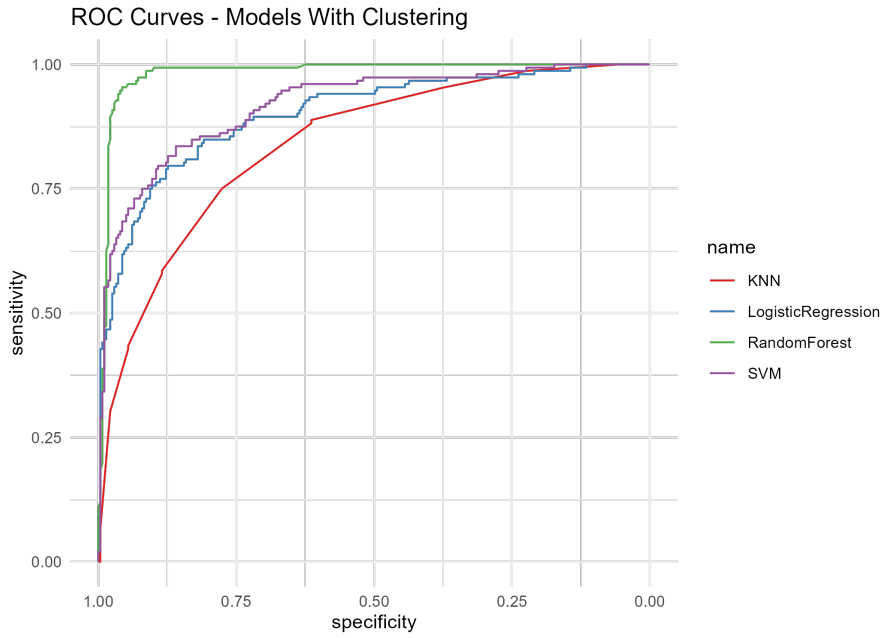

FIGURE 5.3 – Courbes ROC des modèles avec clustering

5.5 Interprétation des résultats

— Le modèle **Random Forest** est le plus performant, avec une précision de 93.5 % sans clustering, légèrement améliorée à 93.9 % avec clustering. Son AUC proche de 0.98 indique une très bonne capacité discriminante, et son F1-score reflète un bon équilibre entre précision et rappel.
— Les modèles **KNN** et **SVM** montrent des performances correctes mais plus modestes. L'ajout du clustering améliore légèrement leurs scores, notamment le F1, suggérant que l'information de groupe latente est utile.
— La **régression logistique** offre une performance stable sans variation notable avec le clustering, ce qui peut traduire que sa nature linéaire limite sa capacité à exploiter cette information additionnelle.
— Globalement, l'ajout de la variable cluster améliore les performances dans la majorité des cas, confirmant l'intérêt de combiner apprentissage supervisé et non supervisé.

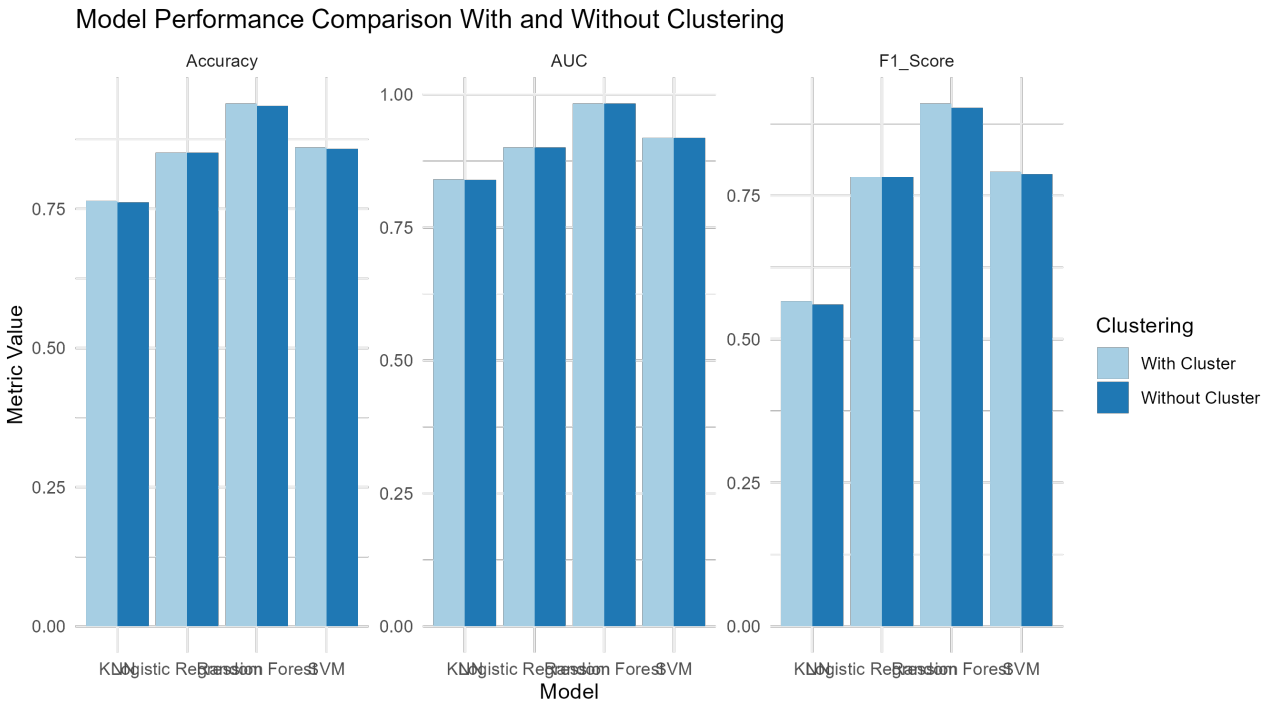

FIGURE 5.4 – Comparaison des performances des modèles

5.6 Conclusion

Cette évaluation confirme la robustesse et la pertinence des modèles développés, en particulier de Random Forest enrichi par le clustering. L'approche hybride retenue montre un réel bénéfice, en particulier dans un contexte biomédical complexe.

Chapitre 6

Déploiement et interface utilisateur

6.1 Sauvegarde et export du modèle

Une fois le modèle Random Forest entraîné et validé sur le jeu de données Alzheimer, il a été sauvegardé au format RDS, permettant ainsi sa réutilisation dans un environnement de production sans nécessiter de nouvel entraînement. Ce fichier contient également les centres des clusters issus d'un k-means effectué sur les variables numériques, permettant de segmenter les patients en groupes homogènes pour améliorer la précision des prédictions.

6.2 Développement de l'interface Shiny

Pour rendre le modèle accessible aux utilisateurs non techniques, une interface web a été développée avec `Shiny`. Cette application permet à l'utilisateur de saisir ses données cliniques au format texte libre, en suivant un format prédéfini. Une fonction Python d'extraction, intégrée via `reticulate`, analyse ce texte et extrait les variables nécessaires à la prédiction. L'utilisateur reçoit ensuite en sortie la prédiction du modèle (diagnostic Alzheimer ou non) ainsi que la probabilité associée, via une interface intuitive et responsive.

Les fonctionnalités principales de l'interface sont :

- Saisie des variables cliniques dans un champ texte libre, avec aide à la saisie sur le format attendu.
- Extraction automatique et validation des données grâce à une fonction Python dédiée.
- Attribution dynamique d'un identifiant patient fixe pour assurer la cohérence des clusters.
- Application d'un clustering k-means sur les données saisies, compatible avec les clusters calculés lors de l'entraînement.
- Prédiction du diagnostic à l'aide du modèle Random Forest intégré.
- Affichage clair et stylisé du résultat et de la confiance associée.

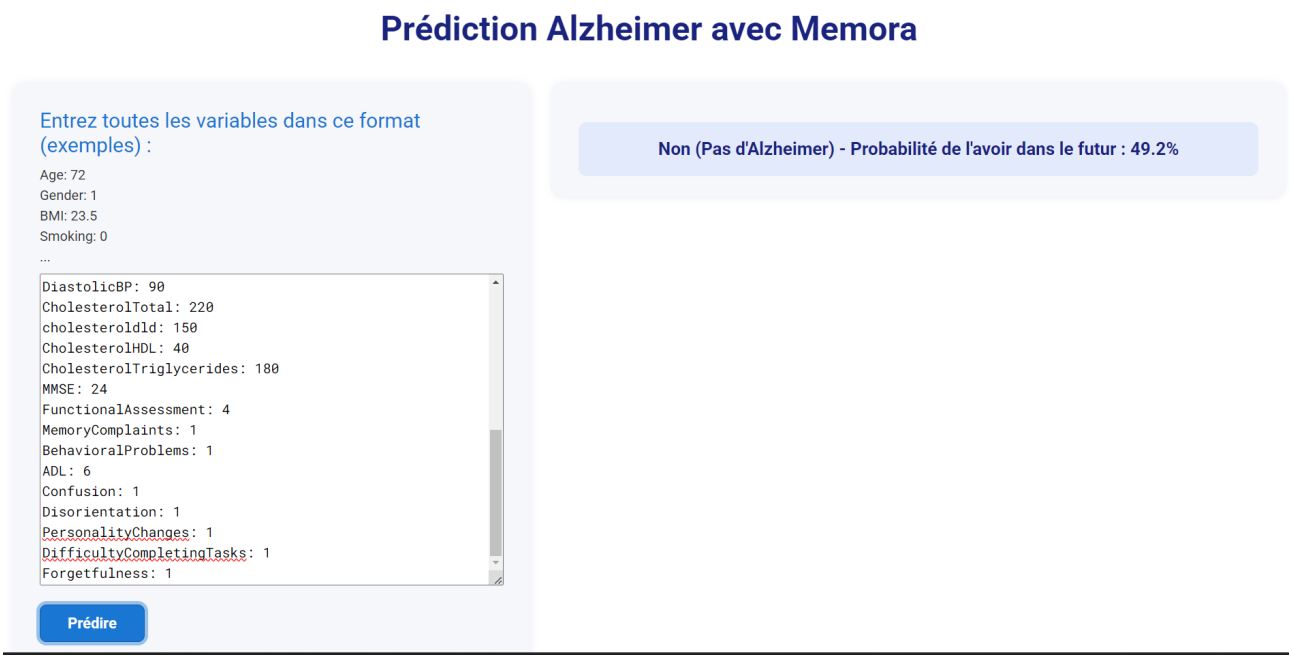

FIGURE 6.1 – Interface utilisateur Shiny pour la prédiction de la maladie d'Alzheimer

6.3 Organisation technique de l'application

L'application Shiny repose sur un backend R qui charge le modèle sauvegardé (`model_rf_clust.rds`) et les centres du clustering. La fonction Python d'extraction des variables utilise des expressions régulières pour parser le texte libre de l'utilisateur et extraire les valeurs numériques des différentes caractéristiques cliniques.

Le pipeline de traitement dans Shiny est le suivant :

1. Réception du texte utilisateur via un champ `textarea`.
2. Appel à la fonction Python d'extraction via `reticulate::source_python()`.
3. Conversion des données extraites en `data.frame` et mise en forme (types, facteurs).
4. Ajout d'un `PatientID` fixe pour le clustering.
5. Application de la mise à l'échelle et attribution du cluster.
6. Prédiction via le modèle Random Forest.
7. Affichage du diagnostic avec la probabilité associée.

6.4 Conclusion

Cette solution permet de démocratiser l'accès à un modèle prédictif complexe en fournissant une interface web accessible, simple d'utilisation et robuste. Elle prépare également le terrain pour une intégration future dans une API ou une application mobile, facilitant ainsi le déploiement en milieu clinique.

Chapitre 7

Conclusion

Ce projet a permis de concevoir et d'évaluer un pipeline complet pour la classification automatique de la maladie d'Alzheimer à partir de données cliniques et comportementales. Après un nettoyage rigoureux et une exploration approfondie des données, plusieurs modèles supervisés ont été entraînés.

L'ajout d'une étape de clustering non supervisé a enrichi la représentation des patients, permettant d'améliorer les performances, notamment celle du modèle Random Forest, qui atteint une précision proche de 94

Ces résultats démontrent le potentiel des méthodes combinant apprentissage supervisé et non supervisé pour soutenir le diagnostic médical. Les travaux futurs viseront à enrichir le dataset, explorer d'autres approches de modélisation, et déployer un système opérationnel accessible aux professionnels de santé.