République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة بجاية
Tasdawit n Bgayet
Université de Béjaïa

Université Abderrahmane Mira de Bejaia
Faculté des Sciences Exactes
Département d'Informatique

MÉMOIRE DE MASTER RECHERCHE

Option : Systèmes d'Information Avancés

THÈME

Impact des Caractéristiques sur la Performance du Modèle Prédictif de la Rétinopathie Diabétique

Réalisé par :

BELLAL Ferroudja
BESSAA Tinhinane

Devant le jury composé de :

Président : Mr AMROUN Kamal
Examinateur : Mr MOKTFI Mohand
Encadrante : Mme BOUKERRAM Samira
Co-encadrante : Mme BOULAHROUZ Djamila
Membre invité : Dr TAFOUKT Rafik

Année universitaire : 2022/2023
Soutenu le 25 Juin 2023

Remerciements

Tout d'abord, nous remercions ***Allah*** *le Tout-Puissant de nous avoir donné le courage et la patience nécessaires pour mener ce travail à son terme.*

Nous tenons à remercier tout particulièrement notre encadrante, ***Mme. Boukerram Samira****, et notre co-encadrante,* ***Mme Boulahrouz Djamila****, pour l'aide immense qu'elles nous ont apportées, pour leur patience, leurs encouragements, la qualité de leur suivi ainsi que pour tous leurs conseils. Leur regard critique nous a été très précieux pour structurer notre travail et améliorer la qualité des différentes sections.*

Nous tenons également à remercier le docteur ***Tafoukt Rafik*** *pour son aide précieuse, ses conseils et les informations qu'il nous a prodigués avec un degré de patience et de professionnalisme sans égal.*

Nos vifs remerciements vont également aux membres du jury : ***M. Amroun Kamel*** *et* ***M. Moktfi Mohand*** *pour l'intérêt qu'ils ont porté à notre travail en acceptant de l'examiner et de l'enrichir par leurs propositions.*

Nous souhaitons aussi remercier l'équipe pédagogique et administrative du département informatique de l'université Abderrahmane Mira pour leurs efforts dans le but de nous offrir une excellente formation.

Enfin, nous remercions toutes les personnes ayant contribué de près ou de loin à la réalisation de ce travail, notamment nos parents, nos familles et nos amis.

Dédicaces

Je dédie ce modeste travail :

À mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études ;

À ma sœur de cœur Belynda pour sa présence, son soutien moral et son encouragement permanent ;

À ma binôme Tinhinane pour son dévouement et sa patience afin de mener à bien ce projet ;

À ma famille et mes amis pour leurs soutien ;

À Mme Boukerram et Mme Boulahrouz, pour la patience et le soutien dont elles ont fait preuve pendant toute la durée de ce travail et à qui je voudrais exprimer mes affections et mes gratitudes ;

Merci.

Ferroudja

Je dédie ce modeste travail :

À mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études ;

À mon frère Amirouche pour sa présence, son soutien moral et son encouragement permanent ;

À mon confident, mon pilier, une personne très spéciale qui a été d'un soutien inestimable. Sa présence et son encouragement constants ont été une force et une source d'inspiration pour moi ;

À ma binome Ferroudja pour son dévouement et sa patience afin de mener à bien ce projet ;

À ma famille et mes amis pour leurs soutien ;

À Mme Boukerram et Mme Boulahrouz, pour la patience et le soutien dont elles ont fait preuve pendant toute la durée de ce travail et à qui je voudrais exprimer mes affections et mes gratitudes ;

Merci.

Tinhinane

Résumé

La rétinopathie diabétique (RD) est une complication fréquente chez les patients atteints de diabète de type 2, entraînant une perte de vision significative, voire la cécité. La prédiction précoce de la RD joue un rôle crucial dans sa prévention et la réduction de sa progression.

L'objectif de cette étude est de mettre en évidence l'influence des caractéristiques du dataset sur la qualité du modèle prédictif, notamment l'historique de HbA1C, et proposer un nouveau modèle de prédiction de la RD avec de meilleures performances.

Ce modèle de prédiction est obtenu en ne prenant en considération que les caractéristiques les plus influentes et en optimisant la meilleure méthode parmi les 11 techniques d'apprentissage automatique utilisées. Afin d'améliorer les performances du modèle, nous avons appliqué une méthode de Bagging pour obtenir le modèle final le plus optimal.

Les résultats d'expérimentation obtenus sont prometteurs et ont montré l'importance effective de l'historique d'HbA1C dans la prédiction de la RD.

Mots Clés : Prédiction, Rétinopathie diabétique, Caractéristiques, HbA1c, Dataset, Apprentissage automatique

---

Abstract

Diabetic retinopathy (DR) is a common complication in patients with type 2 diabetes, leading to significant vision loss, even blindness. Early prediction of DR plays a crucial role in its prevention and reducing its progression.

The objective of this study is to highlight the influence of dataset characteristics on the quality of the predictive model, particularly the HbA1C history, and propose a new RD prediction model with improved performances.

This prediction model is obtained by considering only the most influential features and optimizing the best method among the 11 machine learning techniques used. To enhance the model's performance, we applied a Bagging method to obtain the optimal final model.

The obtained results are promising and have demonstrated the effective importance of HbA1C history in DR prediction.

Keywords : Prediction, Diabetic retinopathy, Characteristics, HbA1c, Dataset, Machine learning

---

Table des matières

Liste des tableaux

Table des figures

Liste des acronymes et des abréviations

6 :2 :2	*60% entrainement, 20% validation, 20% test indépendant*
8 :2	*80% entrainement, 20% test*
AAs	*Acides Aminées*
ab	*AdaBoost*
ACC	*Accuracy*
AUC	*Area Under the Curve*
AUROC	*"Area Under the Curve" of the "Receiver Operating Characteristic" curve*
BCE	*Entropie Croisée Binaire*
C	*Nombre de caractéristiques*
CART	*Classification And Regression Tree*
cb	*CatBoost*
CNKI	*China National Knowledge Infrastructure*
D1	*Dataset1*
D2	*Dataset2*
D3	*Dataset3*
DT	*Decision Tree*
FP	*Faux Positifs*
FN	*Faux Négatifs*
GBM	*Gradient Boosting Classifier*
HbA1c	*Hémoglobine A1C*
HGB	*Histogram Gradient Boosting Classifier*
IA	*Intelligence Artificielle*
KNN	*K Nearest Neighbors*
LASSO	*Least Absolute Shrinkage and Selection Operator*
LGBM	*Light Gradient Boosting Classifier*
LR	*Régression Logistique*
ML	*Machine Learning*

MLR	*Régression Logistique Multivariable*
P	*Nombre de patients*
Préc	*Précision*
RD	*Rétinopathie Diabétique*
RF	*Random Forest*
RFE	*Recursive Feature Elimination*
ROC	*Receiver Operating Characteristic*
SHAP	*SHapley Additive exPlanations*
SVM	*Support Vector Machine*
VP	*Vrais Positifs*
VN	*Vrais Négatifs*

Introduction générale

La rétinopathie diabétique (RD) est une complication sévère qui se manifeste chez les individus atteints de diabète, en particulier ceux souffrant de diabète de type 2. Elle est caractérisée par une détérioration des vaisseaux sanguins de la rétine, ce qui en fait l'une des principales causes de cécité dans le monde.

Le nombre de personnes atteintes du diabète et de la RD ne cesse d'augmenter, selon la Fédération Internationale du Diabète (FID) environ 425 millions d'adultes vivaient avec le diabète en 2017 et d'ici 2045, ce nombre devrait atteindre 629 millions. De plus, environ 212 millions de personnes atteintes de diabète ne sont pas diagnostiquées pour la RD, car cette affection est souvent asymptomatique jusqu'à un stade avancé [1].

La rétinopathie diabétique représente une lourde charge pour les spécialistes de la santé, pouvant parfois entraîner des diagnostics erronés. Cette situation met en évidence l'importance cruciale d'une prédiction précoce pour prévenir toute déficience visuelle irréversible [2]. La détection précoce de la RD revêt donc une importance capitale, permettant de prévenir les complications visuelles, de réduire les coûts élevés des traitements et d'alléger la charge de travail des professionnels de la santé [3, 4, 5].

Dans ce contexte, l'intelligence artificielle (IA) joue aujourd'hui un rôle majeur car elle permet d'effectuer des tâches intelligentes qui, dans le passé, ne pouvaient être effectuées que par des humains [6] . Elle permet également d'améliorer la collecte et le traitement de données. Il existe aujourd'hui plusieurs techniques et méthodes utilisées par l'IA notamment le Machine Learning (ML) et les réseaux de neurones.

Les chercheurs ont commencé alors à exploiter la puissance des modèles prédictifs pour faciliter le diagnostic et l'intervention précoces, et plusieurs méthodes d'apprentissage automatique ont été développées pour la prédiction de la rétinopathie diabétique (RD) permettant d'identifier les populations à haut risque et de développer des profils de risque et des soins et traitements personnalisés. Ces méthodes permettent d'établir un processus de prédiction automatisé, offrant ainsi une solution à faible coût pour fournir un traitement en temps opportun aux patients. Elles offrent également une opportunité d'analyser de grandes quantités de données sur les patients permettant ainsi d'identifier des associations subtiles et d'établir des relations causales entre les caractéristiques individuelles et le risque de rétinopathie qui auparavant pouvaient échapper aux approches traditionnelles.

Ces données et caractéristiques cliniques représentent un impact non négligeable et un facteur clé pour prédire le développement et la progression de la rétinopathie diabétique car elles permettent d'éclairer et de comprendre la relation complexe entre les caractéristiques du patient et le facteur de risque de développer une RD.

Parmi les caractéristiques déjà prises en considération, nous retrouvons souvent les informations basiques des patients telles que le sexe et l'âge, ce dernier est considéré comme facteur de risque important car la gravité de la RD augmente avec l'âge [4]. La durée du diabète et sa gestion, la glycémie et la pression artérielle contribuent également de manière significative à la prédiction de la rétinopathie. Il existe d'autres caractéristiques pertinentes dans la prédiction de la rétinopathie diabétique, telles que l'indice de masse corporelle (IMC), l'activité physique, les antécédents familiaux et médicaux, l'utilisation de certains médicaments, etc.

Les facteurs de risque de la rétinopathie diabétique (RD) ont fait l'objet d'études approfondies dans le passé, mais on ne sait toujours pas quels facteurs de risque sont davantage associés à la RD que d'autres. Si nous parvenons à détecter plus précisément les facteurs de risque liés à la rétinopathie diabétique, nous pourrons alors mettre en œuvre des stratégies de prévention précoce de la rétinopathie diabétique dans les populations les plus à risque.

Des recherches médicales [7, 8, 9] récentes menées par des experts ont montré qu'une réduction significative de l'hémoglobine glyquée (HbA1c) augmente considérablement le risque de développer une RD. Cependant, la littérature existante sur la prédiction de la RD ne prend en compte qu'une seule mesure de l'HbA1c.

L'objectif de notre travail est d'examiner l'importance de ces données cliniques dans la prédiction de la RD. Nous explorons ainsi comment l'exploitation de ces données cliniques, ainsi que d'autres facteurs pertinents, comme la présence d'au moins deux valeurs de l'hémoglobine glyquée contribuent au développement des modèles prédictifs. Une partie de ce travail a fait l'objet d'une communication dans une conférence internationale [10].

Dans cette étude, nous allons évaluer l'impact de l'utilisation d'au moins deux mesures de l'HbA1c sur les performances du modèle de prédiction de la RD.

Pour cela nous allons :

— Développer un modèle de prédiction de la RD en expérimentant plusieurs techniques d'apprentissage automatique, et en extraire le plus performant en utilisant les méthodes d'ensembles.
— Chacun des modèles sera entraîné sur trois types de datasets avec et sans la deuxième mesure de l'HbA1c.
— Mettre en valeur les caractéristiques influentes pour confirmer ou infirmer notre hypothèse.

Notre travail est divisé en trois chapitres comme suit :

Le premier chapitre sera consacré à l'intelligence artificielle et les différentes méthodes d'apprentissage automatique.

Dans le deuxième chapitre, nous verrons un aperçu sur l'anatomie de l'œil, la rétinopathie diabétique, ses types, ses causes et ses symptômes. Nous présenterons alors une revue des études antérieures et des travaux précédemment réalisés dans le contexte de la prédiction de la rétinopathie diabètique en utilisant les techniques d'apprentissage automatique vues dans le premier chapitre. Nous mettrons également en lumière l'impact des caractéristiques dans celle-ci.

Le troisième et dernier chapitre sera dédié au développement de notre modèle de prédiction nous permettant d'examiner le rôle des données cliniques en tant que source d'informations pour prédire la RD. Nous expliquerons comment les paramètres cliniques fondamentaux peuvent servir de prédicteurs fiables et contribuer à une évaluation complète du risque. Nous définirons ensuite les méthodologies et les techniques d'apprentissage automatique employées pour analyser les données cliniques dans nos modèles de prédiction de la rétinopathie. Les résultats de notre travail seront discutés dans le même chapitre.

Nous terminerons ce mémoire par une conclusion générale qui résume l'essentiel de notre travail.

Chapitre 1

Intelligence Artificielle et Machine Learning

Introduction

L'intelligence artificielle (IA) est un terme qui désigne l'étude de la façon dont les ordinateurs effectuent des tâches intelligentes qui, dans le passé, ne pouvaient être effectuées que par des humains [6]. Elle est basée sur des modèles analytiques qui génèrent des prédictions, des règles, des réponses, des recommandations ou des résultats similaires [11]. Ces dernières années, l'IA s'est développée rapidement et a changé les modes de vie des gens. Elle s'étale aujourd'hui sur différents domaines d'application dont la sécurité, le finance, la communication, l'éducation, la recherche scientifique, la santé. . .etc.

Notre travail s'intéresse à ce dernier ; le domaine de la santé, un domaine sur lequel l'IA fait aujourd'hui un travail remarquable notamment par rapport à la prédiction et la détection précoce de plusieurs maladies, la personnalisation des traitements et la surveillance des patients ainsi que la recherche médicale. Parmi ces maladies nous trouvons la rétinopathie diabétique, notre domaine de recherche durant ce travail.

Il existe aujourd'hui plusieurs techniques et méthodes utilisées par l'intelligence artificielle pour la prédiction ou la détection de la rétinopathie diabétique.

Dans ce travail nous allons exclusivement nous concentrer sur les différentes techniques de Machine Learning dans la prédiction de la rétinopathie diabétique.

1.1 Machine Learning

L'apprentissage automatique (ML), un sous-domaine de l'intelligence artificielle (IA), a évolué à partir du besoin d'enseigner aux ordinateurs comment apprendre automatiquement une solution à un problème (Essinger et Rosen, 2011). Il s'applique désormais à divers domaines de la vie quotidienne tel que les systèmes de reconnaissance vocale ou faciale, les assistants intelligents, les voitures autonomes, etc. [3]

Le ML est la technique visant à améliorer les performances du système en apprenant de l'expérience via des méthodes de calcul. Ses algorithmes traitent généralement des données

pour apprendre les modèles sur les individus, les entreprises, processus, transactions, événements, etc. [12]. Dans les systèmes informatiques, l'expérience existe sous forme de données, et la tâche principale de l'apprentissage automatique est de développer des algorithmes d'apprentissage qui construisent des modèles à partir de données. En alimentant les algorithmes d'apprentissage avec des données d'expérience, nous obtenons un modèle qui peut faire des prédictions [13].

Les données peuvent prendre différentes formes qui peuvent varier d'une application à l'autre dans le monde réel ; structurées, semi-structurées ou non structurées [14, 15].

Par conséquent, différents types d'algorithmes d'apprentissage automatique peuvent être utilisés en fonction de leurs capacités d'apprentissage. La figure qui suit résume le fonctionnement général d'un algorithme de machine learning. Voir Figure 1.1.

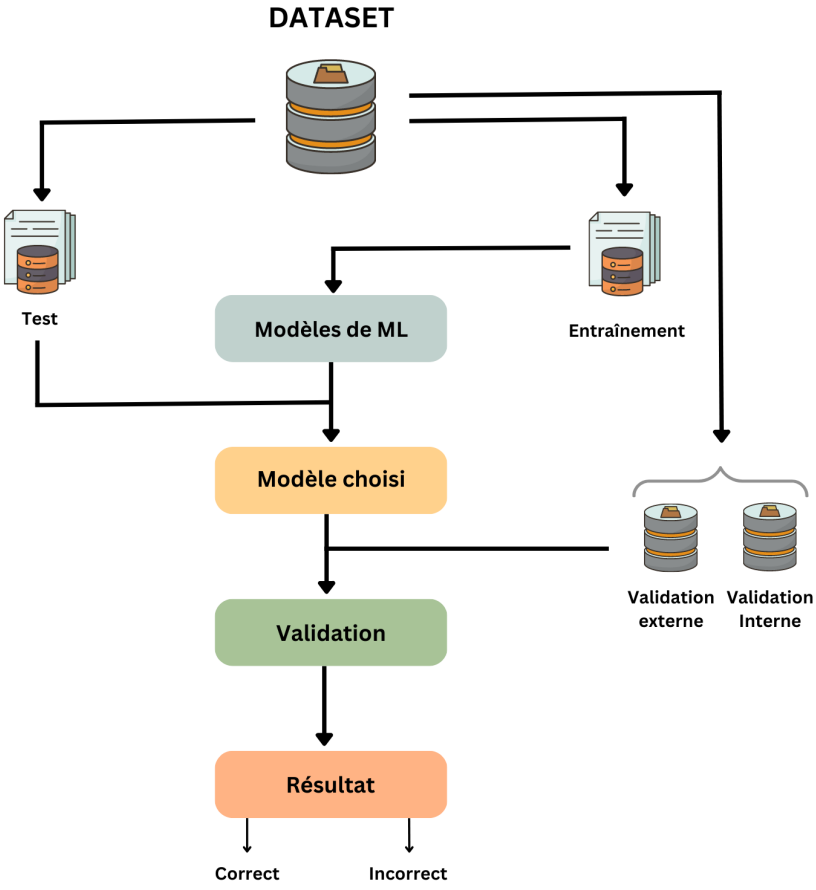

FIGURE 1.1 – Représentation du fonctionnement général de ML

1.2 Types d'algorithmes de ML

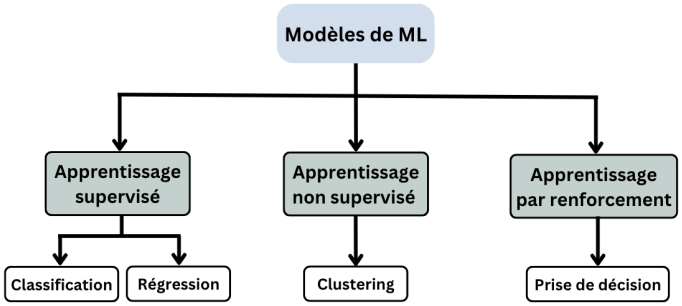

FIGURE 1.2 – Types d'algorithmes de ML [16]

1.2.1 Apprentissage supervisé

L'apprentissage supervisé consiste à apprendre une correspondance entre un ensemble de variables d'entrée X et une variable de sortie Y et en appliquant cette correspondance pour prédire les sorties [17]. Cet algorithme est alimenté avec des données étiquetées pour que le modèle soit au final capable de classer de nouvelles données [12]. Nous avons en guise d'exemple KNN, la régression logistique, SVM, Random Forest, l'arbre de décisions, AdaBoost, GBM, LGBM, XGBoost, HGBM et CatBoost.

KNN : (k Nearest Neighbors) est un algorithme d'apprentissage automatique utilisé pour les problèmes de classification et de régression, c'est un algorithme de généralisation des règles du plus proche voisin, considéré comme l'une des méthodes les plus simples dans l'exploration de données et l'apprentissage automatique. Le principe de l'algorithme kNN est que les échantillons qui sont les plus similaires sont très probablement de la même classe, et ce, dépendant du choix de k. Généralement, l'algorithme kNN trouve d'abord k voisins les plus proches d'une requête dans un ensemble de données d'apprentissage, puis prédit la requête avec la classe principale dans les k voisins les plus proches. Par conséquent, il a récemment été sélectionné comme l'un des 10 meilleurs algorithmes d'exploration de données [18, 19]

Régression Logistique : est un algorithme de machine learning utilisé pour les problèmes de classification binaire répandu en raison de sa simplicité, de son interprétabilité et de son efficacité. C'est l'un des modèles statistiques linéaires les plus couramment utilisés pour l'analyse discriminante [20]. Souvent utilisé pour traiter les problèmes de régression où la variable dépendante est une variable catégorielle [21]. L'algorithme fonctionne en ajustant une fonction logistique aux données d'entraînement basée sur les valeurs précédemment recueillies pour les variables prédictives (appelées covariables) et les résultats correspondants. La phase de formation est suivie d'une phase de test qui évalue la précision du modèle en comparant les caractéristiques d'entrée à la probabilité de sortie. Le modèle est dit valable lorsque la classification de la plupart des données correspond au résultat [22].

Support Vector Machine (SVM) : est un algorithme d'apprentissage supervisé utilisé à la fois pour la classification et la régression. Il vise à créer une frontière de décision entre deux classes qui permet la prédiction d'étiquettes à partir d'un ou plusieurs vecteurs de caractéristiques [23]. Cette frontière de décision est un hyperplan dans un espace de dimension

élevée ou infinie, qui maximise l'intervalle entre deux types d'échantillons de données [24]. C'est un algorithme puissant et largement utilisé en apprentissage automatique en raison de sa capacité à traiter des ensembles de données complexes et à obtenir de bonnes performances de généralisation.

Random Forest (RF) : C'est un algorithme d'apprentissage basé sur un ensemble qui est composé de n collections d'arbres de décision décorrélés. Random Forest utilise plusieurs arbres pour faire la moyenne (régression) ou calculer les votes majoritaires (classification) dans les nœuds feuilles terminaux lors d'une prédiction. Construits sur l'idée d'arbres de décision, les modèles de forêts aléatoires ont entraîné des améliorations significatives de la précision des prédictions par rapport à un arbre unique en augmentant le nombre d'arbres [20].

Decision Tree (DT) : Les arbres de décision sont l'une des méthodes puissantes couramment utilisées dans divers domaines, tels que l'apprentissage automatique, le traitement d'images et l'identification de modèles. C'est un modèle successif qui unit une série de tests de base de manière efficace et cohérente où une caractéristique numérique est comparée à une valeur seuil dans chaque test [25]. Les arbres de décision consistent en une structure arborescente où le nœud supérieur est considéré comme la racine de l'arbre qui est divisé de manière récursive en une série de nœuds de décision de la racine jusqu'au nœud terminal ou nœud de décision [20]. Un arbre de décision est un arbre où chaque nœud montre une caractéristique dans une catégorie à classer (attribut), chaque lien (branche) montre une décision (règle) et chaque feuille montre un résultat (valeur catégorique ou continue). Comme les arbres de décision imitent la pensée au niveau humain, il est simple de saisir les données et de faire de bonnes interprétations [25, 26].

1.2.2 Apprentissage non supervisé

Un processus piloté par les données [14] qui cherche à apprendre en l'absence d'une sortie préalablement identifiée [3], c'est-à-dire que l'information utilisée pour entraîner le modèle n'est ni classifiée ni étiquetée et ce, sans intervention humaine.
Le système ne connaît pas la sortie correcte avec certitude. Mais il tire des inférences des ensembles de données par rapport à ce que la sortie devrait être [27].

1.2.3 Apprentissage semi supervisé

L'apprentissage semi-supervisé est une hybridation des méthodes supervisées et non supervisées, qui consiste à utiliser des données étiquetées et non étiquetées pour effectuer certaines tâches d'apprentissage [28, 14, 29]. Il se situe entre l'apprentissage supervisé et non supervisé et permet d'exploiter des données non étiquetées étant disponibles dans de nombreux cas d'utilisation en combinaison avec des ensembles plus petits de données étiquetées qui pourraient être rares dans plusieurs contextes du monde réel[12, 30, 29].

1.2.4 Apprentissage par renforcement

Une méthode qui consiste à agir de manière itérative de sorte qu'il puisse évaluer et améliorer son comportement et son efficacité et ce, grâce aux réponses associées de l'environnement dans lequel il évolue [12, 31]. Chaque action a un impact sur l'environnement, et l'environnement fournit une rétroaction qui guide cet algorithme [32].

Cette rétroaction est basée sur la récompense ou la pénalité permettant d'augmenter la récompense ou de minimiser le risque, par conséquent l'agent apprend quelle action est la meilleure [27, 30].

1.2.5 Méthodes d'ensemble learning

Boosting

Le boosting est une méthode d'ensemble largement utilisée [33] qui cherche à modifier les données d'entraînement et à ajuster les modèles qui seront ajoutés à l'ensemble séquentiellement de sorte que chaque classificateur se concentre de plus en plus sur les instances mal classées par les classificateurs précédents et apprenne des erreurs de ses prédécesseurs [34, 35]. Le classificateur final est une somme pondérée des prédictions de l'ensemble [35]. Cette méthode combine des modèles dits apprenants faibles pour finalement obtenir un apprenant fort avec une haute performance [36, 34, 33]. Adaboost est considéré comme le premier à avoir démontré l'efficacité de la méthode d'ensemble learning [34]. La figure suivante montre un exemple de Boosting.

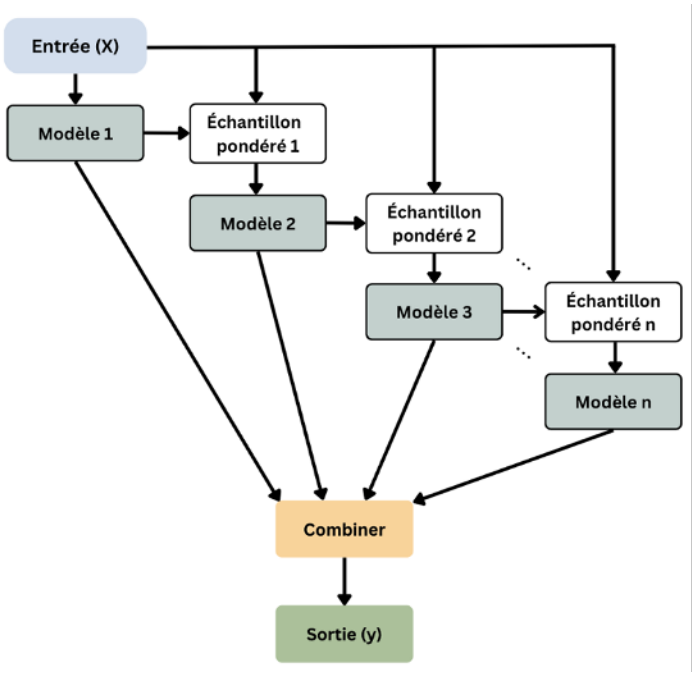

FIGURE 1.3 – Exemple de Boosting [34]

Les méthodes de Boosting les plus utilisées sont les suivantes :

— **Adaptive Boosting Classifier (AdaBoost) :** C'est un algorithme d'apprentissage automatique utilisé pour les problèmes de classification binaire. Il combine plusieurs classificateurs "faibles" en un classificateur "fort". Il s'adapte aux exemples difficiles du jeu de données et se caractérise par une plus grande attention aux échantillons mal catégorisés pendant l'entraînement [37].

— **Gradient Boosting Machine (GBM) :** C'est un algorithme d'optimisation numérique qui vise à trouver un modèle additif qui minimise la fonction de perte. Il amé-

liore progressivement l'erreur en ajoutant itérativement de nouveaux arbres de décision [38, 39].

— **Light Gradient Boosting Machine (LightGBM) :** C'est un framework d'apprentissage de gradient basé sur l'arbre de décision et l'idée de boosting. Il utilise des algorithmes basés sur des histogrammes pour accélérer le processus d'entraînement et réduire la consommation de mémoire [40].

— **Extreme Gradient Boosting (XGBoost) :** C'est un algorithme d'apprentissage automatique populaire qui appartient à la famille du gradient boosting. Il utilise une régularisation plus forte pour contrôler le surajustement et permet une sélection de variables plus efficace [41].

— **Histogram Gradient Boosting :** C'est une variante de l'algorithme traditionnel de gradient boosting spécialement conçue pour fonctionner avec des caractéristiques continues et discrètes. Il est plus rapide que le gradient boosting classique grâce à l'utilisation d'histogrammes pour la construction des arbres de décision [42].

— **CatBoost Classifier :** C'est un autre algorithme d'apprentissage automatique efficace qui gère automatiquement les caractéristiques catégorielles en les convertissant en caractéristiques numériques. Il utilise des arbres de décision binaires comme prédicteurs de base [43, 44].

Bagging (Bootstrap Aggregating)

Le Bagging est une méthode d'apprentissage d'ensemble qui implique l'entraînement du même algorithme plusieurs fois en utilisant différents sous-ensembles du même ensemble de données [36, 34, 35]. La prédiction finale est calculée en utilisant des statistiques simples, telles que le vote ou la moyenne des prédictions de tous les sous-modèles [36, 34, 35]. Le bagging permet d'augmenter la performance et d'améliorer la précision en réduisant la variance et les erreurs du modèle de base [36, 33, 35]. Voici un exemple de Bagging.

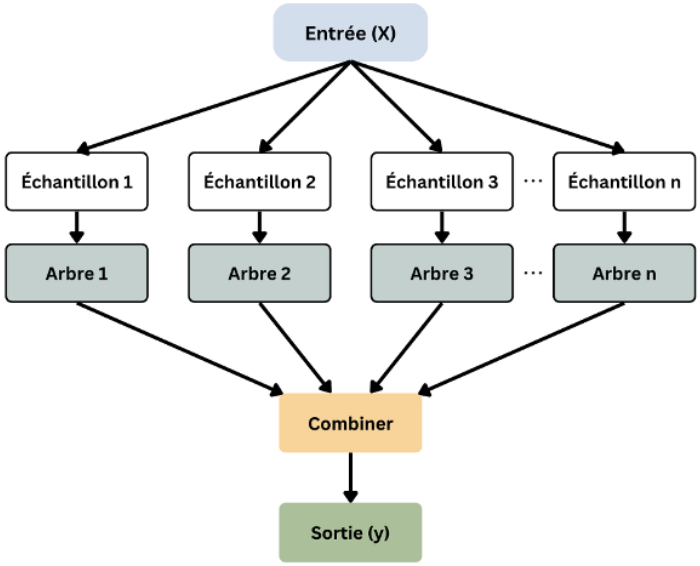

FIGURE 1.4 – Exemple de Bagging [34]

Stacking

Le stacking est une méthode d'ensemble qui combine différents modèles d'apprentissage automatique hétérogènes et diversifiés dans un seul modèle méta-apprenant [34] fonctionnant à des niveaux ou des couches [36, 35] . Le modèle à n niveaux utilise les prédictions des modèles à n-1 niveaux [36]. La prédiction finale est basée sur les prédictions de l'apprenant de base et tend à surpasser tous les apprenants de base individuels [35]. La figure suivante illustre un exemple de Stacking.

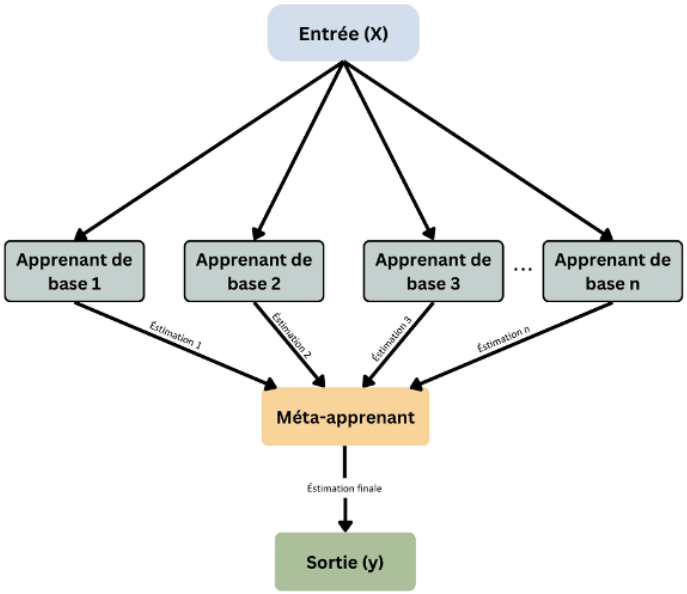

FIGURE 1.5 – Exemple de Stacking [34, 33]

Conclusion

Dans ce chapitre, nous avons introduit l'intelligence artificielle, plus particulièrement les techniques de Machine learning qui, par la suite, seront utilisées dans notre travail pour la prédiction de la rétinopathie diabétique.

Le chapitre suivant portera d'abord sur la rétinopathie diabétique afin de nous familiariser avec cette complication si fréquente du diabète, puis sur sa prédiction et les travaux et études antérieurs ainsi que les contributions les plus importantes dans ce domaine.

Chapitre 2

Rétinopathie - Prédiction et Impact des caractéristiques

Introduction

Le diabète est une maladie chronique grave due à un trouble de l'absorption, de l'utilisation et du stockage des sucres apportés par l'alimentation. Cela se traduit par un taux élevé de glucose dans le sang qu'on appelle hyperglycémie [45, 46, 47]. Il existe deux types de diabète : le diabète de type 1 et le diabète de type 2.

Le diabète peut engendrer certaines complications aux patients dont la **rétinopathie diabétique** qui affecte majoritairement les personnes atteintes de diabète de type 2.

Dans ce chapitre nous allons introduire cette complication du diabète, ses symptômes et ses types en commençant par l'anatomie de l'œil ; l'organe concerné par la RD.

2.1 Œil humain et son anatomie

L'œil est l'organe de la vision de l'être humain ; il lui permet de capter la lumière, distinguer les formes et les couleurs. Il est constitué d'un globe oculaire de 2,5 cm de diamètre et de 8 grammes de masse contenant une calotte sphérique transparente sur la partie antérieure appelée : la cornée, et sur le reste du globe on trouve ce qu'on appelle la sclère (Le blanc de l'œil).

Le globe oculaire est formé de trois tuniques, entourant une substance gélatineuse qui sert à maintenir la forme de l'œil appelée le corps vitré.
Les trois tuniques sont :

- La tunique externe : la sclérotique + la conjonctive + la cornée.
- La tunique moyenne : la choroïde + l'iris + la pupille + le corps ciliaire + le cristallin.
- La tunique interne : la rétine+ la macula + la tache aveugle, ou papille + la fovéa + le nerf optique [48].

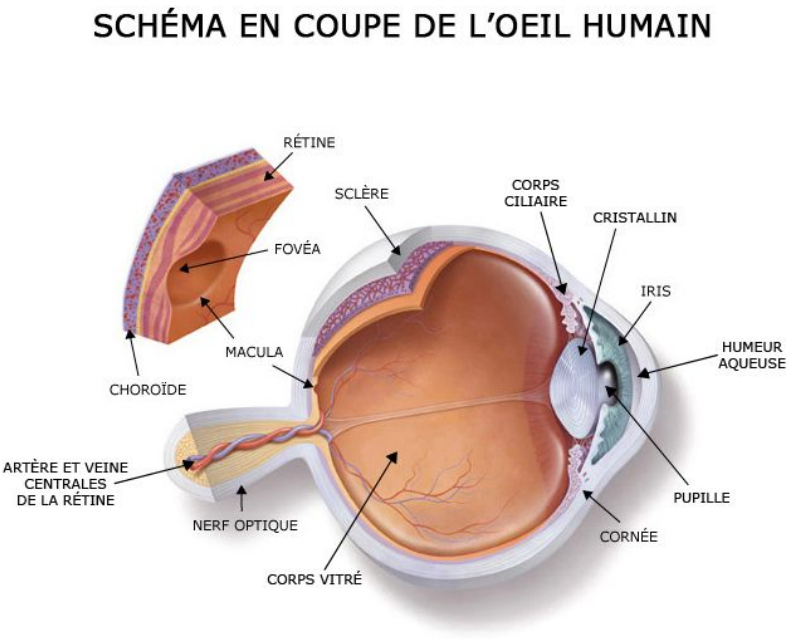

FIGURE 2.1 – Représentation de l'oeil humain [49]

Rétine

La rétine est une fine membrane tapissant le globe oculaire. Elle est chargée de capter les rayons lumineux pour les transmettre au système nerveux central. D'une épaisseur comprise entre 0,1 et 0,4 mm, elle est constituée de 10 couches : quatre de cellules photoréceptrices à l'extérieur et six de cellules nerveuses à l'intérieur [6]. Elle est formée de cellules sensorielles, les cônes (vision diurne) et les bâtonnets (vision nocturne), et de cellules nerveuses, les neurones[48, 50].

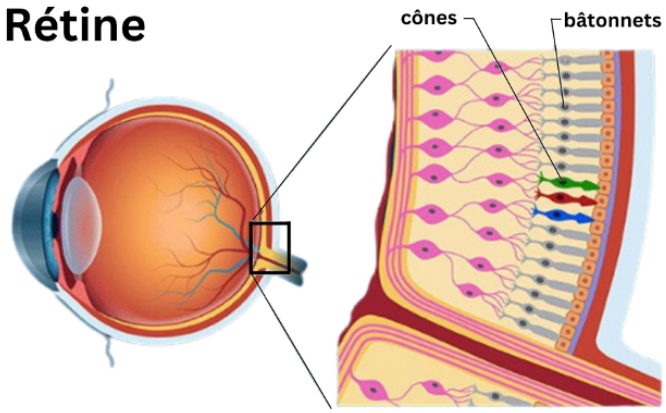

FIGURE 2.2 – Représentation zoomée de la rétine [49]

2.2 Rétinopathie et ses causes principales

La rétinopathie diabétique (RD) est une maladie qui apparait généralement chez les personnes âgées déjà atteintes de diabète ; diabète de type 2 pour être plus exact. Cette maladie cause une forte dégradation des vaisseaux sanguins de la rétine. Elle est donc considérée comme une maladie dégénérative et l'une des causes principales de la cécité si elle ne se fait pas traiter au plus vite [3].
Cette pathologie est connue pour être asymptomatique jusqu'au stade avancé, c'est pour cela qu'un examen oculaire régulier est fortement recommandé.
Néanmoins la rétinopathie peut être causée par de nombreux facteurs hormis le fait d'être déjà diagnostiqué diabétique, à savoir :

— Le niveau de glycémie.
— L'hypertension artérielle.
— Le tabagisme.
— L'ancienneté du diabète.
— L'instabilité du diabète.
— L'obésité.

Ces mêmes facteurs peuvent aussi augmenter la gravité de la maladie et accélérer sa propagation[51].

2.3 Type de rétinopathie diabétique

Il existe deux (02) types de rétinopathie :

1. **La rétinopathie non proliférative (RDNP) (Précoce) :** Également appelée rétinopathie diabétique de fond, considérée comme la forme la plus fréquente et la moins grave. Renvoie à la présence de lésions vasculaires intra rétiniennes avant le développement d'un tissu fibrovasculaire extra rétinien. Les lésions retrouvées dans la RDNP comprennent les micro-anévrysmes, les nodules cotonneux, les zones de non-perfusion capillaire, les anomalies microvasculaires intra rétiniennes (AMIR), les hémorragies intra rétiniennes, l'œdème rétinien, les exsudats lipidiques, les anomalies artériolaires et les dilatations et irrégularités veineuses [4].
   En résumé elle représente les symptômes suivants :
   — Microanévrysmes.
   — Taches floconneuses.
   — Microhémorragies.
   — Œdème maculaire.
2. **La rétinopathie proliférative (RDP) (Avancée) :** Connue pour être la forme de RD la plus sévère, elle est caractérisée par la présence d'une néovascularisation fragile ou d'une prolifération fibreuse sur le trajet du nerf optique ou ailleurs, avec des hémorragies pré-rétiniennes ou vitréennes, et un décollement rétinien. Ces saignements peuvent être une source majeure de cécité. En cas d'atteinte irienne, cela engendrerait un glaucome, pouvant ainsi endommager le nerf optique [52]. Autrement dit on y constate les symptômes suivants :
   — Fortes hémorragies.
   — Formation anormale des vaisseaux sanguins.
   — Diminution importante de l'irrigation sanguine.

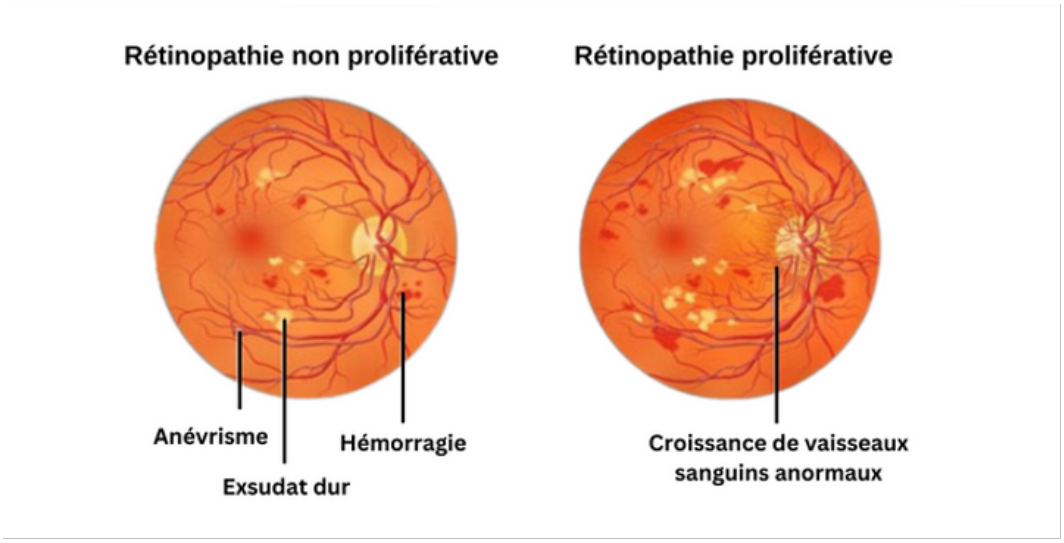

FIGURE 2.3 – Repésentation des types de rétinopathie [53]

2.4 Symptômes de la rétinopathie

Les symptômes les plus fréquents sont :
— Diminution du degré de sensibilité visuelle.
— Apparition de taches noires ou de point lumineux dans le champ de vision.
— Apparition de corps flottants.
— Vision floue.
— Perte de la vision des détails.

2.5 Application de l'apprentissage automatique pour la prédiction de la rétinopathie diabétique

La RD demeure la principale cause de cécité chez la population d'âge moyen atteinte du diabète de type 2. Un diagnostic tardif peut entraîner des complications pour les patients et des coûts de traitement supplémentaire ainsi qu'une charge de travail importante pour les spécialistes. C'est pour ces raisons que la prédiction de la rétinopathie est devenue un axe de recherche très actif et des progrès significatifs ont été réalisés dans ce domaine au cours des dernières années. Notamment dans le domaine du machine learning ; plusieurs méthodes d'apprentissage automatique sont développées.

Ces méthodes dépendent des caractéristiques cliniques, des antécédents médicaux et des données démographiques fournis, ces derniers jouent un rôle principal dans la prédiction de la RD et dans l'identification des facteurs de risque et des marqueurs prédictifs.

Dans ce qui suit nous présentons les caractéristiques les plus utilisées par les différentes techniques de ML.

2.5.1 Caractéristiques

La précision et l'efficacité des modèles prédictifs lors de la prédiction de la RD en utilisant

les méthodes de ML dépend de manière directe du choix des caractéristiques utilisées.

Le tableau suivant résume les caractéristiques les plus communes et les plus utilisées par les chercheurs dans la prédiction de la RD.

TABLE 2.1 – Résumé des caractéristiques les plus communes dans la prédiction de la RD

Caractéristique	
Sexe	Le sexe du patient
Âge	L'âge du patient
IMC	Indice de masse corporelle
Durée du diabète	La durée du diabète chez le patient
HbA1C	Hémoglobine glyquée (%)
HDL	Lipoprotéine de haute densité (g/l)
LDL	Lipoprotéines de basse densité (g/l)
Pression artérielle	La force exercée par le sang sur les parois des artères (mmHg)
Hypertension	Une pression artérielle élevée de manière chronique (mmHg)
Insuline	Si le patient est sous traitement à l'insuline ou non
TG	Triglycérides
Chol	Cholestérol
Cr	Créatine

Souvent nous retrouvons des données basiques telles que l'âge, le sexe, l'origine ethnique et les antécédents familiaux, qui sont d'une grande importance dans la prédiction de la rétinopathie diabétique. Accompagnées d'autres données de laboratoire, telles que les taux d'hémoglobine glyquée (HbA1c), les lipides sanguins (cholestérol, triglycérides), les marqueurs inflammatoires (protéine C réactive) et les marqueurs rénaux (créatinine, albumine).

Les mesures physiologiques, telles que la tension artérielle, les taux de glucose sanguin et l'indice de masse corporelle (IMC), jouent également un rôle majeur dans la prédiction de la RD. Ces derniers sont souvent associés à une progression plus rapide de la RD.

Les antécédents médicaux, tels que la durée du diabète, la présence de complications diabétiques antérieures et les traitements médicaux passés, peuvent également jouer un rôle principal dans la prédiction de la RD. Une durée prolongée du diabète et une gestion inadéquate de la glycémie ou des niveaux élevés d'HbA1c sont associées à un risque de développer une RD alors qu'une diminution de chaque 1% de HbA1C est égal à une réduction de 40% de la rétinopathie [54].

L'intégration de ces informations permet de mieux évaluer le risque et par conséquent d'adapter les interventions préventives.

Les méthodes d'apprentissage automatique développées se différencient selon les caractéristiques utilisées ou les approches développées pour trouver le meilleur modèle de prédiction. Dans ce qui suit nous allons présenter les méthodes les plus récentes en mettant en évidence les caractéristiques les plus importantes ainsi que la précision obtenue.

2.5.2 Techniques de prédictions de la RD - Etat de l'art

Les méthodes de machine learning exploitent les informations précédemment citées afin

d'améliorer la précision des prédictions et d'optimiser les stratégies de prévention et de gestion de la RD. La figure 2.4 donne un aperçu chronologique des techniques de ML utilisées au fil du temps.

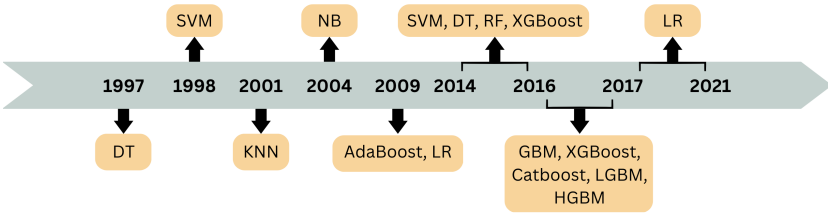

FIGURE 2.4 – Historique des modèles de ML utilisés pour la prédiction de la RD

Hsin-Yi Tsao et al. [55] ont appliqué différents algorithmes d'apprentissage automatique à savoir : SVM, les arbres de décision, la régression logistique ainsi que les réseaux de neurones artificiels dans le but d'obtenir la meilleure prédiction possible et d'identifier de nouvelles caractéristiques influentes dans la prédiction de la RD. Et ce, sur un dataset regroupant les informations de plusieurs patients sur une année ; de janvier 2012 à décembre 2012. Ce dataset a été pris sur une saison au hasard dans la base de données « DM Shared Care » dans un hôpital privé se trouvant au nord de Taïwan. Il contient un total de 536 patients dont 430 atteints de la RD et 106 sains. Ce dataset contient 10 caractéristiques dont le sexe, l'historique familial, le traitement à l'insuline, l'IMC, la Tension Artérielle Systolique et Diastolique...etc. Ils sont donc arrivés à dire que le déclenchement de la RD n'était pas seulement lié au taux élevé de HbA1C mais également à d'autres caractéristiques notamment l'utilisation d'insuline (le risque d'avoir une RD est multiplié par 3,561 pour les patients qui utilisent de l'insuline par rapport à ceux qui ne l'utilisent pas) et la durée du diabète (augmentation de 9,3% si la durée du diabète augmente d'un an). Le meilleur modèle est SVM avec une accuracy de 83.9% en utilisant une répartition de 80% pour l'entrainement et 20% pour le test. Deux nouvelles caractéristiques ont été identifiées comme importantes et influentes dans la prédiction de la RD à savoir l'utilisation de l'insuline et la durée du diabète. Cependant, cette étude se porte sur une seule ressource, il est alors recommandé que cette dernière subisse une validation externe afin d'éloigner la théorie que la généralisation des résultats peut être affectée.

Mo et al. [56] ont également fait une étude dans le but d'identifier les facteurs de risque associés à la RD et de construire un modèle prédictif pour présenter l'influence de ces facteurs de risque et ce, en utilisant la régression logistique multivariable et une méthode de sélection de caractéristiques. Cette étude a été faite sur un jeu de données comportant un total de 4170 patients tous atteints de diabète de type 2, provenant de six communautés différentes à Shangai de septembre 2015 à décembre 2018 selon des critères d'inclusion et d'exclusion. Ces données ont été rapportées à partir d'un questionnaire sur les informations basiques, un examen des indicateurs biochimiques, un examen physique et un examen de la rétine afin de détérminer si le patient est atteint de la RD ou non. Ces patients partagent 19 caractéristiques à savoir : âge, sexe, évolution de la maladie, glycémie à jeun (FBG), glycémie postprandiale (PBG), hémoglobine glyquée A1c (HbA1c), cholestérol total (TC), triglycérides (TG), lipoprotéines de basse densité (LDL), lipoprotéines de haute densité (HDL) , créatinine sérique (SCR), azote uréique sanguin (BUN), acide urique (UA), créatinine urinaire (UCR) et microalbumine urinaire (UMA). Ce jeu de données a été divisé en un ensemble d'entraînement comprenant

3130 patients et un ensemble de validation comprenant 1040 patients. La méthode de sélection de caractéristiques utilisée dans ce modèle de régression logistique est LASSO (The least absolute shrinkage and selection operator). Sept caractéristiques parmi les 19 précédemment citées ont été sélectionnées à savoir : âge, évolution de la maladie, PBG, HbA1c, UCR, UMA et SBP jugées les plus influentes lors de la prédiction de la RD car en effet, un âge plus jeune, une évolution plus longue de la maladie, un PBG et une HbA1c plus élevés (l'augmentation de HbA1c de 1% augmente le risque de RD par 1.53 fois), un RCU et une UMA plus élevés et une PAS plus élevée sont des risques grandement associés à la RD. La performance de ce modèle a été évaluée avec la métrique AUC qui était de 0.715, une performance considérée moyenne laissant croire qu'il existe davantage de caractéristiques influentes qui devraient être prises en considération dans des études ultérieures.

L'étude de Zun Shen et al. [57] avait pour objectif de construire et de proposer un modèle de prédiction de la RD basé sur l'ensemble learning c'est-à-dire combiner et fusionner plusieurs autres modèles de machine learning au lieu de procéder à un vote majoritaire, de prendre le meilleur ou de prendre la moyenne, mais également basé sur la sélection de caractéristiques dans le but de réduire la redondance des caractéristiques et d'améliorer la performance. Ce modèle est XGB-Stacking. Six modèles de machine learning ont été sélectionnés à savoir : KNN, AdaBoost, GBDT, XGBoost, LightGBM, and CatBoost. Ainsi qu'un méta-modèle, il s'agit de SVM. Cette étude est basée sur deux étapes principales : l'algorithme de sélection des caractéristiques « XGBIBS » et la méthode de fusion des modèles « Sel-Stacking ». Ce dernier ; « Sel-Stacking » combine KNN, GBDT, XGBoost, et CatBoost sachant que plusieurs combinaisons de plusieurs algorithmes ont été faites avant de sélectionner celle-ci. Lors de cette étude, Zun Shen et al. ont utilisé un dataset accessible au public, au moment de l'étude, qui provient du centre national chinois de données sur les sciences médicales cliniques. Ce dataset contient 2990 patients, 68 caractéristiques numériques et catégorielles ainsi qu'une sortie montrant si le patient est atteint de la RD ou non. Parmi ces caractéristiques nous retrouvons des informations basiques sur les patients (9) : nation, état civil, poids...etc. Des informations sur d'autres maladies (32) et des informations biochimiques (27) dont HBA1C, HDL, LDL, TG, TC...etc. Un prétraitement a été effectué sur le dataset original à savoir la suppression des patients avec des caractéristiques manquantes, le remplissage des valeurs manquantes et la suppression des valeurs aberrantes. Le dataset a été divisé en 6 :2 :2 (60% pour le l'entraînement, 20% pour la validation et 20% pour un test indépendant). Les résultats de la première étape à savoir l'algorithme de sélection des caractéristiques « XGBIBS » a pu en sélectionner dix jugées les plus importantes. Il s'agit de NEPHROPATHY, HEIGHT, HBA1C, CHD, LEADDP, OTHER_TUMOR, RESPIRATORY_SYSTEM_DISEASE, RENAL_FALIURE, HYPERLIPIDEMIA et GYNECOLGICAL_TUMOR. Les résultats de la deuxième étape ont montré que la combinaison et la fusion de plusieurs modèles donnent le meilleur résultat à savoir 9.7% de plus que les autres modèles donc l'utilisation d'un algorithme fusionnant différents modèles est finalement plus adapté pour la prédiction de la RD. Cette étude avait pour premier but d'appliquer un algorithme de sélection de caractéristiques appelé XGBIBS, ce dernier s'est vu comparé avec plusieurs autres algorithmes de sélection mais les résultats de ces derniers n'ont pas été mis en avant afin de voir quelles caractéristiques ont été jugées les plus influentes par chacun des algorithmes.

Wanyue Li et al. [58] avaient pour objectif d'étudier les facteurs de risque de la rétinopathie diabétique en utilisant les modèles d'apprentissage automatique et ce, à l'aide d'une très large base de données. Le modèle de prédiction de base est XGBoost, il s'est ensuite vu

comparer avec trois autres modèles à savoir la régression logistique, random forest et SVM. Les auteurs de cet article ont également utilisé la méthode d'explication additive de Shapley (SHAP) afin de visualiser les résultats mais également de définir les caractéristiques les plus importantes dans la prédiction de la rétinopathie diabétique. Wanyue Li et al. ont basé leur étude sur une très large base d'informations de patients hospitalisés atteints de diabète sucré de type 2 (T2DM) qui ont été extraits du système de dossier médical électronique de l'hôpital général chinois PLA du 1er janvier 2013 au 31 décembre 2017. Il contient 32452 patients dont 6.3% atteints de la rétinopathie diabétique, avec un total de 79 caractéristiques, qui, au fil du prétraitement se sont réduites à 60 après la suppression des caractéristiques dépassant 20% de valeurs manquantes...etc. Se voyant ensuite réduire à seulement 17 caractéristiques jugées les plus importantes grâce à l'élimination récursive de fonctionnalités (RFE). Parmi ces caractéristiques nous retrouvons : l'âge, le sexe, HbA1c, TG, TC, la nationalité, l'hypertension...etc. La méthode SHAP dans le modèle XGBoost a identifié les caractéristiques suivantes comme les plus influentes dans la prédiction de la RD : HbA1c, néphropathie, Créatinine sérique et le traitement à l'insuline. Ce qui veut dire qu'elles représentent des facteurs de risque importants pour la RD. XGBoost a réussi à avoir une performance remarquable à savoir 90% en accuracy, et par conséquent une fiabilité à distinguer les patients atteints de la RD et les patients sains, il peut ainsi évaluer les indicateurs de risque de la rétinopathie diabétique. Malgré l'avantage de la largeur de la base de données utilisée, les chercheurs n'ont utilisé qu'une seule validation (interne), l'utilisation d'une validation externe peut générer de différents résultats et par conséquent une possibilité d'amélioration. De plus, quelques caractéristiques importantes comme la durée du diabète n'ont malheureusement été communiqué qu'oralement par les patients, ce qui cause une incertitude et une possibilité d'erreur dans les résultats.

L'étude de Yazan Jian et al. [59] vise à prédire et à classer huit complications du diabète à savoir : le syndrome métabolique, la dyslipidémie, la neuropathie, la néphropathie, le pied diabétique, l'hypertension, l'obésité et la rétinopathie. Dans cette étude, plusieurs algorithmes de classification supervisée ont été appliqués notamment : la régression logistique, SVM, les arbres de décision (CART), forêt aléatoire, AdaBoost et XGBoost. Et ce, après un long prétraitement de données commençant par la suppression des données et des attributs inutiles ainsi que la suppression ou la gestion des valeurs manquantes par patient en utilisant la valeur la plus fréquente, la méthode de substitution moyenne, le modèle k-NN pour imputer les valeurs manquantes et en dernier, MissForest à l'aide des forêts aléatoires de manière itérative. Yazan Jian et al. sont ensuite passés à l'encodage des données catégorielles, à l'équilibrage et à la normalisation des données. Le dataset utilisé dans cette étude a été recueilli par le « Rashid Center for Diabetes and Research (RCDR) » situé à Ajman, UAE [60]. Cette base n'a jamais été utilisée avant cette étude, elle se compose de 884 patients avec 79 caractéristiques dont l'âge, le sexe, l'IMC, HbA1c, la tension artérielle, les triglycerides...etc. Étant donné que le datatset utilisé possède 79 caractéristiques, une méthode de sélection a été appliquée pour chaque modèle afin d'augmenter leurs performances et de savoir quelles caractéristiques sont les plus importantes. Il s'est avéré que le cholestérol total, la durée du diabète, le sexe, l'IMC et la pression artérielle sont les cinq caractéristiques les plus importantes, suivies de DT2, le poids, LDL, HDL et le rapport créatinine microalbumine pour former le top 10 des caractéristiques les plus importantes. Cependant, ces caractéristiques ne concernent pas uniquement la rétinopathie diabétique mais les huit complications précédemment citées ensemble. Le meilleur modèle étant eXtreme Gradient Boosting (XGBoost) a réussi à atteindre 87.2% d'accuracy prouvant que l'utilisation d'algorithmes d'ensemble basés sur des arbres est essentielle dans

la prédiction précoce de pathologies telles que la rétinopathie diabétique. Cette étude a permis d'avoir une amélioration de 10% comparée au score de précision des études précédentes. Néanmoins, Le nombre de valeurs manquantes et le problème du dataset déséquilibré peuvent biaiser les résultats.

Yuedong Zhao et al. [61] ont mené une étude des risques de rétinopathie diabétique chez les patients atteints de diabète de type 2 en utilisant cinq techniques de machine learning à savoir : RF, XGBoost, LR, SVM et KNN ainsi qu'une technique d'optimisation d'hyperparamètres : GridSearchCV et une validation interne à l'aide de la technique «fivefold cross-validation». Ils ont utilisé un dataset de 7943 patients provenant du département d'endocrinologie de l'Université médicale de Dalian Affilié à l'hôpital central de Dalian de janvier 2010 à septembre 2018. Cet ensemble a été divisé au hasard en un ensemble d'entraînement comprenant 5559 patients et un ensemble de test de 2384 patients. Ce dataset contient un total de 31 caractéristiques dont : HbA1c, Duration, temps de suivi, FBG, Âge, SUA, Hypertension, Insuline. . .etc. Les meilleures performances sont obtenues avec le modèle XGBoost avec une valeur de AUC de 0.913 et une accuracy de 79,9%. Yuedong Zhao et al. se sont intéressés au suivi annuel des patients et il s'est avéré que ce modèle prédit le premier diagnostic de RD en une moyenne de temps de 2,895 ans permettant ainsi aux cliniciens d'identifier avec précision la population à haut risque de RD. En plus des facteurs de risques classiques dont HbA1c, durée du diabète, FBG et âge. Cette étude a également identifié de nouvelles caractéristiques : acide urique sérique (SUA), cholestérol des lipoprotéines de basse densité (LDL-C), cholestérol total (TC), taux de filtration glomérulaire estimé (eGFR) et triglycérides (TG).

He-Yan Li et al.[62] avaient pour but d'évaluer la précision de la régression logistique ainsi que de trois autres modèles de machine learning à savoir KNN, Random Forest et SVM dans le pronostic et le diagnostic de la rétinopathie diabétique. Grid search a été utilisée avec tous les modèles afin d'ajuster leurs paramètres. He-Yan Li et al. ont utilisé un dataset provenant de NHANES (National Health and Nutrition Examination Survey) pour les cycles 2005-2006 et 2007-2008 [63]. Et ce, avec exactement les mêmes caractéristiques dont : l'âge, l'IMC, le sexe, la nationalité, etc. Le premier ; NHANES 2005-2006 a été utilisé pour l'entraînement des différents modèles ainsi que pour la validation interne. Ce dataset contient un total de 3056 patients, néanmoins, les données de seulement 757 patients ont été utilisées car jugées complètes en informations. Parmi ces 757 patients, 53 patients sont atteints de la rétinopathie diabétique. Quant au deuxième dataset, provenant de la même source ; NHANES 2007-2008 a été utilisé pour la validation externe. Dans cette étude, la régression logistique univariée a montré que le genre féminin, un HbA1C élevé, le taux de créatine sérique, le niveau de l'eGFR sont parmi les caractéristiques les plus importantes. De même, la régression multivariée a montré l'importance du genre et du HbA1C. Quant à la forêt aléatoire, elle a montré avec deux méthodes différentes à savoir IncMSE (increase in mean squared error) [64] et IncNodePurity (Incremental Node Purity) [64] l'importance des caractéristiques. La première méthode a sélectionné l'IMC, le genre, l'urée sanguine, le taux de créatine sérique et HbA1c comme caractéristiques influentes, tandis que la deuxième a sélectionné le taux de créatine sérique, HbA1c et l'IMC. KNN a été retenu comme meilleur modèle avec AUROC=98.4% et AUROC=98.2% en validation interne et externe respectivement. En revanche, l'étude des auteurs n'a pas suivi les patients dans le temps, par conséquent, une relation de cause ne peut pas être établie entre la rétinopathie diabétique et les facteurs de risque étudiés. Certaines données ont été obtenues par questionnaire d'auto-évaluation et datent de 2005-2008 ce qui pourrait ne pas refléter avec précision la situation actuelle de la prévalence de la rétinopathie diabétique et

des facteurs de risque. Cependant, nous remarquons également un grand déséquilibre dans le dataset utilisé car 53 patients sur 757 sont atteints de la rétinopathie contre 704 patients sains, ce qui peut fausser les performances des algorithmes et biaiser les prédictions souvent en faveur de la classe majoritaire [65].

Qingqing Xu et al. [2] ont fait une revue de quelques études antérieures tirées de PubMed et Google Scholar dont le thème est la prédiction de trois complications microvasculaires à savoir la RD, la néphropathie et la neuropathie. Six études ont été sélectionnées et trois études parmi les six ont porté sur la RD. Parmi les algorithmes de ML utilisés, on trouve l'arbre de classification et de régression (CART), la forêt aléatoire (RF), SVM, la régression logistique ainsi que les réseaux de neurones. Ces modèles ont été évalués avec AUC et Accuracy et certains ont eu recours à une validation externe. Certains échantillons de données ont subi des techniques de bootstrap, de suréchantillonnage et de validation croisée à cause de leur petite taille et des données déséquilibrées. La validation croisée a également été utilisée dans l'évaluation de la performance moyenne des modèles et ce, dans quatre études. Le modèle de prédiction de la première étude (Lagani et al. [66]) a retenu cinq caractéristiques à savoir : HbA1c, la durée du diabète post-pubère, l'IMC, l'état civil et le niveau de la rétinopathie. Le modèle de Aspelund et al. [67] a retenu trois caractéristiques : HbA1C, SBP et la présence de la NPDR. Quant au modèle de Skevofilakas et al. [68] sept caractéristiques ont été retenues à savoir l'âge, la durée du diabète (DT1), le cholestérol total, le niveau de triglycérides, l'hypertension et la durée du traitement. Les performances globales des modèles étaient modérées, à savoir des valeurs d'AUC inférieures à 0,8 et une précision d'environ 80%. En conclusion, les auteurs de cette revue ont souligné la non prise en considération de la variation des niveaux de HbA1c dans ces études alors que, médicalement, les chercheurs et les scientifiques ont constaté auparavant l'importance de cette variation dans le développement de la RD [7, 8, 9]. Ils ont par conséquent encouragé les études ultérieures à intégrer la variabilité de HbA1c et à la prendre en considération lors de l'évaluation ou de la prédiction d'une telle complication.

L'étude de Hong Pan et al. [69] avait pour objectif de construire un modèle de prédiction de la RD en prenant en considération les facteurs de risque c'est-à-dire les caractéristiques les plus importantes dans la prédiction de la RD. Et ce, en utilisant un jeu de données de 2,385 patients recueilli dans six centres de services de santé communautaires à Shanghai, d'octobre 2014 à avril 2015 par questionnaire, examens physiques et tests biochimiques. Ce jeu de données a été divisé en ensemble d'entraînement (70%) et ensemble de test (30%). Ensuite de comparer ce modèle utilisant les six caractéristiques les plus importantes avec un autre modèle d'une étude précédente (Mo et al. 2020 [56]) qui quant à lui en a utilisé sept. Et ce, en utilisant la régression logistique multivariée. Les modèles ont été évalués selon plusieurs métriques d'évaluation mais AUROC est la métrique prise en considération lors d'une classification ou d'une prédiction binaire. Les caractéristiques importantes ont été sélectionnées parmi 23 caractéristiques en utilisant quatre méthodes différentes à savoir : l'algorithme XGBoost, l'élimination récursive de caractéristiques par forêt aléatoire (RF-RFE), un réseau de neurones à rétropropagation (BPNN) et un modèle d'opérateur de rétrécissement absolu minimal (LASSO). Celles répétées plus de trois fois ont été prises en considération dans le Model I. À savoir : HbA1C, Course, PBG, Âge, SBP et ACR. Nous remarquons que le HbA1c prend la tête du classement mais sa variation n'est toujours pas prise en considération. Les auteurs ont également cité comme limite l'absence de certains aspects non physique qui peuvent être influents comme l'activité physique et l'alimentation des patients.

Chengjun Zhu et al.[70] ont mis en place un modèle optimisé en utilisant la régression

logistique. Pour mener à bien leur travail, les auteurs ont procédé à une récolte de données à travers des articles publié sur CNKI, WanFang, PubMed, Web of Science, Springerlink et VIP. Ils ont rassemblé un total de 2437 articles, qui ont ensuite été examinés et soumis à des critères d'inclusion et des critères d'exclusion selon le type de diabète, les facteurs de risque suivant : le sexe, la résidence, le tabagisme, l'alcool, l'alimentation, l'HbA1c, la chirurgie bariatrique, le contrôle glycémique, la durée de prise de médicaments hypolipidémiants, l'utilisation d'insuline, etc. Ainsi que d'autres critères. Après cette étape, uniquement huit articles datant tous de la période de création des bases de données utilisées à Mai 2019 ont été gardés, ces derniers incluent des patients de différentes nationalités ayant plus de 18 ans. Les facteurs d'influence comprenaient le sexe, la chirurgie bariatrique, la myopie, la résidence (zones urbaines ou rurales), le tabagisme et l'évolution du diabète, l'hypertension artérielle et la durée d'utilisation des médicaments hypolipémiants ainsi que quatre indicateurs physiologiques et biochimiques à savoir : la glycémie à jeun, l'HbA1c, le contrôle intensif de la glycémie et l'utilisation d'insuline. Cependant, la finalité de ce travail n'a pas été bien conclu et les auteurs ont précisé que la relation entre les facteurs et les performances du modèle n'a pas été étudiée dans ce travail et que des études ultérieures devraient traiter cette relation causale. Dans la majorité des études portant sur la prédiction de la RD en utilisant l'apprentissage automatique, la glycémie avait le rôle le plus important.

Zong et al.[71] avaient pour objectif de montrer que d'autres caractéristiques telles que les caractéristiques des métabolites dont les acides aminés (AAs) et les acylcarnitines (AcylCNs) avaient également un rôle principal malgré le bon suivi de la glycémie, et qu'elles pouvaient par conséquent représenter de nouveaux facteurs de risque de la RD. Dans ce but, Zong et al. ont utilisé une base d'informations de 1898 patients atteints du diabète de type 2 provenant de LMUFAH (Liaoning Medical University First Affiliated Hospital) de mai 2015 à août 2016. Au final, seulement 1031 patients agés de plus de 18 ans ont été sélectionnés pour cette étude selon des critères d'exclusion et d'inclusion. Ce jeu de données a été divisé en un ensemble d'entraînement (70%) et un ensemble de test (30%) et contient un total de 82 caractéristiques dont l'âge du patient, le sexe, la durée du diabète, l'IMC, pression artérielle systolique (PAS), tension artérielle diastolique (PAD), hémoglobine glyquée (HbA1c), triglycérides, HDL, LDL, cholestérol total, créatinine sérique, tabagisme actuel, consommation actuelle d'alcool, antidiabétiques, hypolipémiants, hypotenseurs, etc. Qui ont ensuite subi une sélection les réduisant à 15 en utilisant la méthode LASSO. Ce dataset a été utilisé afin de construire un modèle de prédiction utilisant la régression logistique et XGBoost avec deux méthodes différentes. GridSearch a été utilisée dans les deux afin d'avoir les meilleurs paramètres. Dans la première méthode (Model 1), 10 caractéristiques ont été utilisées parmi les 15 à savoir : l'âge, le sexe, la durée du diabète, l'IMC, PAS, PAD, triglycérides, HDL, LDL, et cholestérol total. Tandis que dans la deuxième (Model 2), 14 caractéristiques ont été utilisées dont 4 en commun avec le premier (âge, durée du diabète, PAS et cholestérol total). Ajoutées à celles-ci, 7 AAs (alanine, citrulline, glutamate, ornithine, phénylalanine, thréonine et tyrosine) et 3 AcylCNs (octacarbonylcarnitine (C18 : 1), 3-hydroxy-octadécylcarnitine (C18 : 1OH) et octadécadiénylcarnitine (C18 : 2)). Nous remarquons que le HbA1c ne figure dans aucun des modèles car il a été supprimé à cause des nombreuses valeurs manquantes. Les résultats ont montré que XGBoost avec le Model 2 en utilisant les caractéristiques métabolites a eu les meilleures performances. En effet, ce dernier a atteint une AUC de 0.82 et une accuracy de 88.39%. Tandis que LR Model 1, XGBoost Model 1 et LR Model 2 ont respectivement eu une AUC de 0.73, 0.64 et 0.78. La méthode SHAP a été utilisée avec le meilleur modèle à savoir XGBoost (Model 2)

afin de déterminer les caractéristiques les plus influentes dans ce modèle prédictif. Sept caractéristiques parmi les 14 ont été jugées les plus influentes à savoir : la durée du diabète, C18 : 1OH, phénylalanine, C18 : 1, thréonine, cholestérol total et tyrosine. Cette étude a réussi à démontrer que, en effet, les caractéristiques métabolites sont vaguement associées au risque de développement de la RD et jouent un rôle prépondérant dans sa prédiction. Cependant, nulle ne peut contredire la relation de cette complication du diabète avec la glycémie, le HbA1c a été supprimé en raison de son déficit excessif sachant qu'il s'agit de l'une des caractéristiques les plus importantes et cela a été démontré de multiples fois auparavant. Il est alors recommandé de le prendre en compte et de combiner les caractéristiques cliniques traditionnelles avec les caractéristiques métabolites dans des études ultérieures, cela peut amener à de meilleurs résultats.

2.6 Tableau récapitulatif des travaux précédents

Le tableau suivant représente un résumé des méthodes les plus récentes :

TABLE 2.2 – Tableau récapitulatif des travaux précédents

Article	Dataset(P/C)	Meilleur modèle	AUC	ACC	Limitations
Hsin-Yi Tsao et al. 2018 [55]	536/10	**SVM**	/		Peu de ressource et absence de validation
		Percentage Split(8 :2)	0.795	0.839	
		Three Way Data Split (6 :2 :2)	0.744	0.817	
Mo et al. 2020 [56]	4170/19	**MLR**	0.700	0.796	Performance moyenne
Zun Shen et al. 2021 [57]	2990/68	**Sel-Stacking**	/	0.839	Besoin de comparaison avec d'autres modèles et de tester sur d'autres datasets
Wanyue Li et al. 2021 [58]	32452/17	**XGBoost**	0.90	0.90	Absence de validation externe
Yazan Jian et al. 2021 [59]	844/87	**XGBoost**	/		Dataset déséquilibré
		Tous les attributs	/	0.771	
		K-fold (K=5)		0.804	
		K-fold (K=10)		0.872	
Yuedong Zhao et al. 2022 [61]	7943/31	**XGBoost**	0.913	0.799	Manque d'explications sur l'étape du suivi
Li, He-Yan et al. 2022 [62]	D1 : 757/7 D2 : 200/7	**KNN**	0.98	/	- Relation causale impossible à établir car il n'y a pas de suivi dans le temps - Données incomplètes et anciennes
Hong Pan et al. 2023 [69]	2385/23	**MLR**	0.703	0.796	Quelques caractéristiques qui peuvent être influentes n'ont pas été prises en considération
Zhu et al. 2023 [70]	/	**LR**	0.912	/	Manque d'informations sur le dataset utilisé
Zong et al. 2023 [71]	1898/82	**XGBoost Model 2**	0.82	0.884	Non prise en considération de la glycémie

Conclusion

Ce chapitre nous a servi d'introduction à la rétinopathie diabétique ; une complication du diabète qui ne cesse d'augmenter chez les personnes atteintes du diabète. Nous avons cité

ses symptômes qui, hélas, n'apparaissent que lorsqu'il est trop tard et que la maladie est déjà à un stade avancé, ce qui complique la tâche aux patients et aux spécialistes. C'est pour cette raison que l'on s'intéresse fortement aujourd'hui à la prédiction précoce de cette pathologie.

Nous avons également présenté une synthèse non exhaustive des travaux récents axés sur les techniques de prédiction de la rétinopathie diabétique (RD), en particulier ceux basés sur les techniques classiques de machine learning.

Il convient de noter que la comparaison directe entre différentes méthodes de prédiction de la RD peut être difficile en raison des différences dans les datasets utilisés. Certains datasets peuvent être petits mais comportent de nombreuses caractéristiques, tandis que d'autres peuvent être plus grands mais avec moins de caractéristiques, comme illustré dans le Tableau 2.2. Ces variations dans les datasets peuvent influencer les performances des modèles de prédiction.

De plus, il est important de noter l'absence d'un historique de HbA1c au niveau de tous les ensembles de données sur lesquels les études antérieures se sont concentrées.

C'est pourquoi dans le chapitre qui suit, nous proposerons un nouveau modèle de prédiction de la RD qui sera entrainé sur trois types de datasets différents et nous montrerons que le nombre de caractéristiques indépendantes influe sur l'obtention d'un bon modèle de prédiction. Et par-dessus tout, nous essayerons de mettre en évidence l'impact de la présence d'un historique de HbA1c sur la précision du modèle de prédiction de la RD.

Chapitre 3

Conception et réalisation d'un modèle de prédiction de la RD et étude de l'impact des caractéristiques notamment HbA1c sur le résultat

Introduction

Notre objectif est de développer un modèle de prédiction de la RD et de mettre en évidence que la présence de deux HbA1c parmi les caractéristiques joue un rôle important dans la précision et dans l'efficacité du modèle prédictif. En effet, plusieurs études en médecine ont montré son rôle prépondérant dans la prédiction de la rétinopathie diabétique [8, 9]. Parmi ces études, nous retrouvons celle menée par Alice Larroumet [7] selon qui une diminution rapide de 3% entre deux mesures consécutives de HbA1c représente un risque majeur de développement de la rétinopathie diabétique. Qingqing Xu et al. 2020 [2] ont vivement encouragé la prise en considération de cette variation, cependant, aucun modèle de prédiction ne l'a pris en considération auparavant et aucune étude ultérieure n'a appliqué leurs conclusions. C'est pour cette raison que nous mettons l'accent sur cette variable et soulignons l'importance d'intégrer au moins une deuxième mesure de HbA1c ou sa variation dans nos analyses.

Pour atteindre cet objectif, nous allons suivre le processus de développement d'un modèle de ML. Le schéma suivant montre la méthode adoptée afin de développer et de tester un modèle de prédiction performant. Voir Figure 3.1.

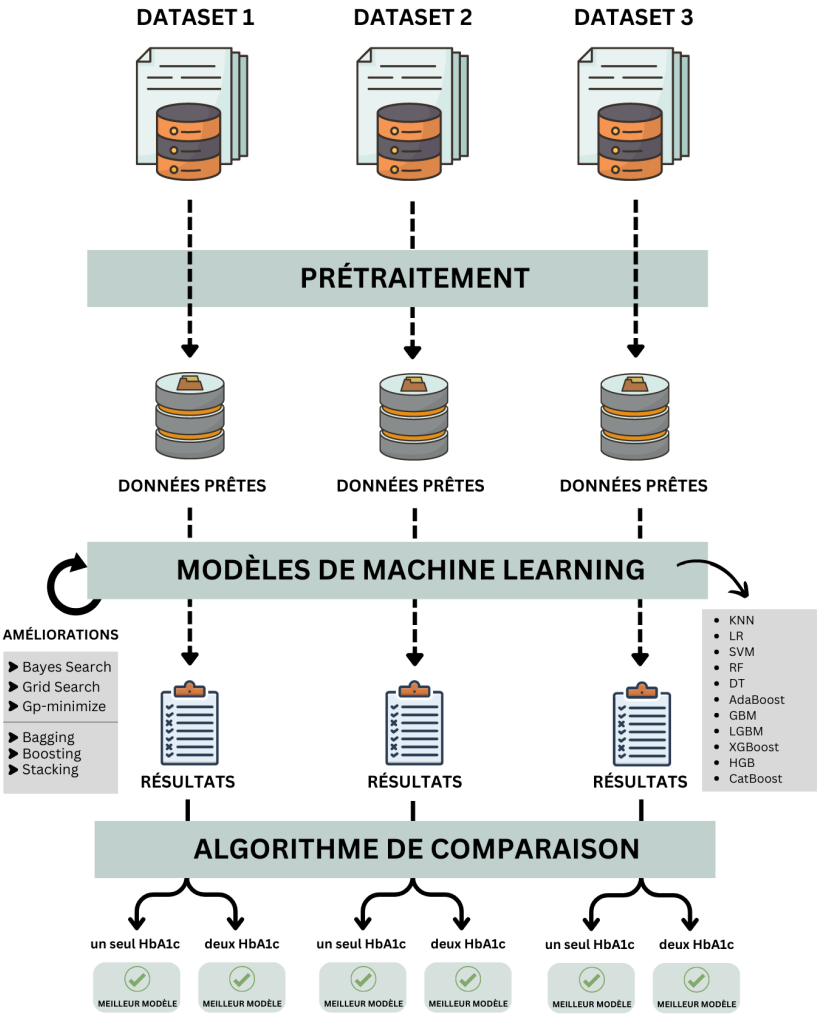

FIGURE 3.1 – Méthode adoptée

3.1 Définition des datasets utilisés et description des variables

Lors de l'application des méthodes de ML, les échantillons de données constituent les composants de base. Chaque échantillon est décrit par plusieurs caractéristiques et chaque caractéristique consiste en différentes valeurs.

Pour une étude précédente, nous avons utilisé deux datasets publics distincts ne contenant qu'une seule valeur d'Hb1AC et entraîné diverses techniques d'apprentissage automatique. En utilisant ces deux datasets, nous avons pu observer que le nombre et le type de caractéristiques peuvent influer sur les résultats de prédiction [10].

Pour cette étude, en plus de ces deux datasets, nous avons également utilisé un troisième dataset privé.

3.1.1 Premier dataset

Le premier dataset [72] nommé dataset1 contient un total de 844 patients dont 368

atteints de la RD et 476 sains, ainsi que 11 caractéristiques en entrée dont l'âge, l'urea, la créatine, l'hémoglobine glyquée, le cholestérol...etc. Et une seule sortie (Class). Table 3.1.

Ce dataset ne contient qu'une seule valeur de HbA1c.

TABLE 3.1 – Caractéristiques des patients (Dataset 1)

Variable	**Description**
Gender	Sexe du patient
AGE	Âge du patient
Urea	Déchet azoté éliminé dans l'urine
Cr	Créatine, acide organique azote
HbA1c	Hémoglobine glyquée
Chol	Cholestérol
TG	Triglycérides
BMI	Indice de masse corporelle
HDL	Lipoprotéine de haute densité
LDL	Lipoprotéines de basse densité
VLDL	Lipoprotéine de très basse densité

3.1.2 Deuxième dataset

Le deuxième dataset [73] nommé dataset2 est moins large en nombre de patients et plus riche en caractéristiques. Il contient 133 patients dont 42 atteints de la RD et 91 sains et il fournit 23 caractéristiques en entrée dont le sexe, l'âge, l'IMC, le type de diabète, sa durée, l'hémoglobine glyquée...etc. Et une sortie (Ret_pathy). Voir Table 3.2.

De même que le précédent, ce dataset ne contient qu'une seule valeur de HbA1c.

TABLE 3.2 – Quelques caractéristiques des patients (Dataset 2)

Description	**Variable**
Sexe	Sexe du patient
BMI	Indice de masse corporelle
DM_type	Type du diabète (1 ou 2)
DM_duration	Durée du diabète
FBS	Syndrome du rachis opéré
A1C	Hémoglobine glyquée
LDL	Lipoprotéines de basse densité
HDL	Lipoprotéine de haute densité
HDL	Lipoprotéine de haute densité
TG	Triglycérides.
Dose_Í2_if_BID	La dose de traitement.
Neu_pathy	Neuropathie
Neph_pathy	Néphropathie

3.1.3 Troisième dataset

Quant au troisième dataset que nous appellerons dataset3, il a été l'objet d'une récolte de donnée dans un cabinet médical privé [74]. Il contient 84 patients dont 34 atteints de la RD et 49 sains et il fournit 14 caractéristiques en entrée dont le genre, l'âge, l'IMC, la durée du diabète, les antécédents familiaux et personnels, un historique d'hémoglobine glyquée, la créatine...etc. Et une sortie (Rétinopathie). Voir Table 3.3.

Contrairement au deux premiers datasets, celui-ci contient deux valeurs de HbA1c ce qui nous permet de pouvoir valider notre approche.

TABLE 3.3 – Caractéristiques des patients (Dataset 3)

Description	**Variable**
Genre	Sexe du patient
Age	Âge du patient
IMC	Indice de masse corporelle
Durée Diabète	Durée du diabète
ATCD (F)	Antécédents familiaux
ATCD (P)	Antécédents personnels
HbA1C/HbA1c2	Hémoglobine glyquée
LDL	Lipoprotéines de basse densité
HDL	Lipoprotéine de haute densité
TG	Triglycérides
CT	Cholestérol total
Créat	Créatine, acide organique azote
ACR	Arrêt cardiorespiratoire

3.2 Métrique

Une fois qu'un modèle de prédiction est obtenu à l'aide d'une ou plusieurs techniques de ML, il est important d'estimer ses performances. L'analyse des performances de chacun des modèles proposés est mesurée en termes d'accuracy, d'aire sous la courbe (AUC), de précision, F1-score, etc. Comme illustré dans le Tableau 3.4. Dans notre cas, la métrique la plus adaptée et recommandée est AUC car elle permet d'évaluer et de visualiser la performance d'un modèle de prédiction binaire.

TABLE 3.4 – Métriques d'évaluation

Métrique d'evaluation	Description
AUROC (Area Under the Receiver Operating Characteristic)	Relation entre le taux de VP (sensibilité) et le taux de FP (1 - Spécificité)
Accuracy	$\frac{VP+VN}{VP+VN+FP+FN}$
Précision	$\frac{VP}{VP+FP}$
F1-Score	$\frac{2 \cdot Precision \cdot Rappel}{Precision+Rappel}$
Rappel	$\frac{VP}{VP+FN}$
Support	Nombre d'échantillons dans chaque classe réelle
Spécificité	$\frac{VN}{VN+FP}$
Matrice de confusion	Évaluer l'exactitude d'une classification en comparant ses prédictions avec les valeurs réelles des données de test
Courbe d'apprentissage	Évaluer la performance d'un modèle de machine learning et de représenter les scores d'entraînement et de test
Courbe de perte	Tracer la valeur de la fonction de perte. Formule mathématique de la fonction perte pour la prédiction binaire : $$BCE = -\left[y_{true} \cdot \log(y_{pred}) + (1 - y_{true}) \cdot \log(1 - y_{pred})\right]$$
Graphe d'importances	Permet de trier les caractéristiques de l'apprentissage par ordre d'importance

3.3 Prétraitement des données

Cette étape est considérée comme la première étape essentielle afin de réduire les fausses prédictions ou les résultats incorrects et améliorer la qualité globale des données. Elle joue un rôle essentiel dans l'analyse, l'étude et par conséquent les résultats de tout travail. Il arrive souvent que les données soient dupliquées, manquantes, aberrantes, biaisée ou non représentatives. Il est donc primordial de s'assurer que les données utilisées sont complètes, utiles et adaptées aux techniques que nous utilisons.

Afin d'assurer une analyse plus précise, il convient d'appliquer des étapes de prétraitement visant à modifier les données. Parmi ces étapes, nous incluons la suppression des valeurs manquantes et l'encodage des caractéristiques catégorielles.

De plus, nous avons effectué une sélection des caractéristiques essentielles afin de ne conserver que celles qui jouent un rôle prépondérant. Cette étape revêt une importance particulière dans notre travail, car elle nous permet d'identifier l'impact de ces caractéristiques dans la prédiction de la rétinopathie diabétique. Cette sélection permet d'améliorer les performances des algorithmes en réduisant la dimension des données.

3.3.1 Prétraitement du Dataset1

Après l'importation de notre dataset1, Nous avons d'abord procédé à une suppression

des entrées à valeurs manquantes du dataset de manière à assurer la qualité des données utilisées dans notre modèle prédictif. Ensuite, nous avons effectué un encodage pour les caractéristiques catégorielles pour les préparer à l'apprentissage automatique. Voir Figure 3.2.

```
data = pd.read_csv('dataset1.csv') #importation du dataset
dff = data.copy() #copie du dataset
dff.dropna(axis=0, inplace=True) #suppression des valeurs manqauntes
dff = dff.drop(['HbA1c2','No_Pation'], axis=1) #suppression de "No_Pation" et "HbA1c2"
dff['Gender'].replace(['M','F'], [0,1], inplace=True) #encodage de "Gender"
```

FIGURE 3.2 – Importation du dataset1 et prétraitement

La figure 3.3 montre un aperçu du dataset après le prétraitement.

	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	Class
0	0	31	3.0	60	12.3	4.1	2.2	0.7	2.4	15.4	37.2	1
1	0	30	7.1	81	6.7	4.1	1.1	1.2	2.4	8.1	27.4	0
2	0	45	4.1	63	10.2	4.8	1.3	0.9	3.3	9.5	34.3	0
3	0	45	4.1	63	10.2	4.8	1.3	0.9	3.3	9.5	34.3	1
4	0	45	5.3	77	11.2	3.9	1.5	1.3	2.0	10.4	29.5	0
...	...	...	...	...	...	...	...	...	...	...	...	...
837	0	71	11.0	97	7.0	7.5	1.7	1.2	1.8	0.6	30.0	0
838	0	31	3.0	60	12.3	4.1	2.2	0.7	2.4	15.4	37.2	1
839	0	30	7.1	81	6.7	4.1	1.1	1.2	2.4	8.1	27.4	0
840	0	38	5.8	59	6.7	5.3	2.0	1.6	2.9	14.0	40.5	0
841	0	54	5.0	67	6.9	3.8	1.7	1.1	3.0	0.7	33.0	1

842 rows × 12 columns

FIGURE 3.3 – Aperçu du dataset1 après le prétraitement

Enfin, nous avons divisé nos données en des données d'entraînement représentant 80%, et des données de test représentant 20%. De plus, nous avons mis de côté 10 patients pour une validation interne afin d'évaluer les performances de notre modèle de manière indépendante. La Figure 3.5 illustre le prétraitement du dataset1.

Il est important de voir l'équilibre du dataset utilisé. Pour cela, nous avons affiché le nombre de patients atteints de la RD et sains comme suit. Voir Figure 3.4 :

```
a= dff['Class'].value_counts()
print("Nombre de patients atteints de la RD :", a[1])
print("Nombre de patients sains :", a[0])
```

```
Nombre de patients atteints de la RD : 367
Nombre de patients sains : 475
```

FIGURE 3.4 – Nombre de patients atteints de la RD et sains dans le dataset1

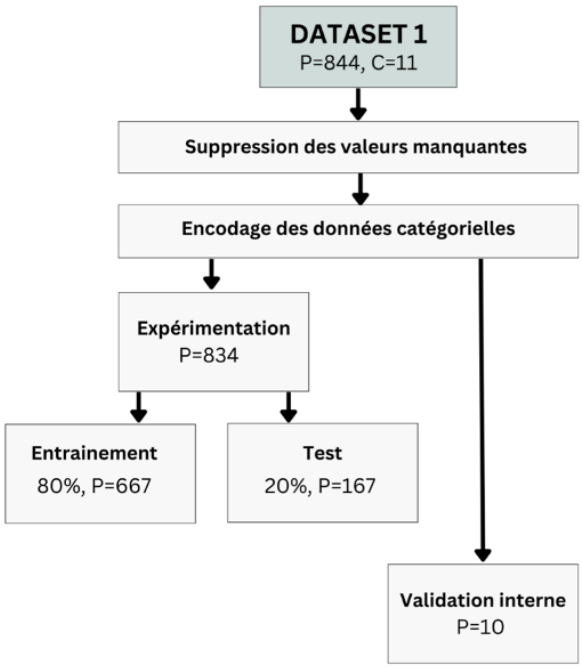

FIGURE 3.5 – Prétraitement du dataset1

Ajout d'une deuxième valeur de HbA1c

Afin de valider notre théorie et de démontrer l'importance d'un historique de HbA1c, nous avons ajouté une nouvelle colonne "HbA1c2" contenant une seconde valeur de l'hémoglobine glyquée en suivant l'étude de Alice Larroumet [7] dans laquelle il a été démontré qu'un déclin entre deux valeurs de HbA1c consécutives est lié à un risque majeur de développement de la RD. Nous nous sommes également inspirés du troisième dataset contenant des valeurs réelles afin de mieux se rapprocher de la réalité lors de l'ajout des valeurs. La figure 3.6 donne un aperçu du premier dataset avec une seconde valeur de HbA1c.

	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	HbA1c2	Class
0	0	31	3.0	60	12.3	4.1	2.2	0.7	2.4	15.4	37.2	7.9	1
1	0	30	7.1	81	6.7	4.1	1.1	1.2	2.4	8.1	27.4	6.5	0
2	0	45	4.1	63	10.2	4.8	1.3	0.9	3.3	9.5	34.3	9.2	0
3	0	45	4.1	63	10.2	4.8	1.3	0.9	3.3	9.5	34.3	13.3	1
4	0	45	5.3	77	11.2	3.9	1.5	1.3	2.0	10.4	29.5	11.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
837	0	71	11.0	97	7.0	7.5	1.7	1.2	1.8	0.6	30.0	8.0	0
838	0	31	3.0	60	12.3	4.1	2.2	0.7	2.4	15.4	37.2	8.6	1
839	0	30	7.1	81	6.7	4.1	1.1	1.2	2.4	8.1	27.4	7.2	0
840	0	38	5.8	59	6.7	5.3	2.0	1.6	2.9	14.0	40.5	8.5	0
841	0	54	5.0	67	6.9	3.8	1.7	1.1	3.0	0.7	33.0	11.8	1

842 rows × 13 columns

FIGURE 3.6 – Aperçu du dataset1 après l'ajout de HbA1c2

Nous avons suivi exactement les mêmes étapes afin de pouvoir effectuer une comparaison entre les performances des modèles prédictifs avec une seule valeur de HbA1c et avec deux valeurs de HbA1c. Les résultats sont également affichés dans la section suivante.

3.3.2 Prétraitement du Dataset2

Comme pour le premier dataset, les entrées contenant des valeurs manquantes ont été supprimées ainsi que des caractéristiques non pertinentes et incohérentes. De même, les caractéristiques catégorielles ont été encodées (voir Figure 3.7.

```
data = pd.read_excel('dataset2.xlsx') #importation du dataset
dff=data.copy() #copie du dataset
dff = dff.drop(['Name','DM treat','Male'], axis=1) #suppression "Name", "DM treat", "Male"
dff.dropna(axis=1, how='all', inplace=True) #suppression des valeurs manquantes
dff=dff.drop(0) #suppression de la première ligne
dff = dff.drop(range(134, 212)) #suppression des lignes 134-->212 car inutiles
dff['Sex'].replace(['Male','Female'],[0,1],inplace=True) #encodage de "Sex"
dff['DM type'].replace(['II','I'],[2,1],inplace=True) #encodage de "DM type"
dff['Neu-pathy'].replace(['Yes ','No'],[1,0],inplace=True) #encodage de "Neu-pathy"
dff['Neph-pathy'].replace(['Yes ','No'],[1,0],inplace=True) #encodage de "Neph-pathy"
dff['Ret-pathy'].replace(['Yes ','No'],[1,0],inplace=True) #encodage de "Neu-pathy"
dff['PVD'].replace(['Yes ','No'],[1,0],inplace=True) #encodage de "Ret-pathy"
dff['CDV'].replace(['Yes ','No'],[1,0],inplace=True) #encodage de "CDV"
dff['Ft ulcer'].replace(['Yes ','No'],[1,0],inplace=True) #encodage de "Ft ulcer"
dff['Dawn ef'].replace(['Yes ','No'],[1,0],inplace=True) #encodage de "Dawn ef"
dff['Dose ( Í2 if BID)'].replace(['No'],[0],inplace=True) #encodage de "Dose ( Í2 if BID)"
dff['Cum Atorvastatin Equ'].replace(['FALSE'],[0],inplace=True) #encodage de "Cum Atorvastatin Equ"
```

FIGURE 3.7 – Importation du dataset2 et prétraitement

La figure 3.8 montre le dataset 2 après le prétraitement.

	Sex	Age	BMI	DM type	DM duration	FBS	A1C	LDL	HDL	TG	...	Neph-pathy	Ret-pathy	PVD	CDV	Ft ulcer	Dawn ef	Sys BP	Dias BP	Cum Atorvastatin Equ	Real LDL
0	0	65.0	25.0	2	20.0	129.0	7.10	100.0	40.0	200.0	...	1	1	1	1	0	1	130.0	80.0	40.0	148.00
1	0	42.0	27.0	2	3.0	210.0	8.90	125.0	38.0	151.0	...	0	0	0	0	0	0	140.0	90.0	20.0	171.25
2	1	31.0	21.0	1	11.0	164.0	7.70	147.0	35.0	217.0	...	0	1	0	0	0	1	145.0	80.0	0.0	147.00
3	0	70.0	32.0	2	29.0	208.0	9.30	119.0	36.0	168.0	...	1	1	1	1	1	1	160.0	100.0	80.0	184.45
4	1	54.0	34.0	2	6.0	183.0	9.80	196.0	32.0	197.0	...	1	0	0	0	0	0	155.0	95.0	40.0	290.08
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
128	0	52.0	34.5	2	3.0	138.0	7.33	113.0	33.0	217.0	...	0	0	0	1	0	1	115.0	75.0	20.0	154.81
129	1	56.0	22.0	2	3.0	156.0	7.16	80.0	43.0	179.0	...	0	0	0	0	0	1	130.0	90.0	20.0	109.60
130	0	65.0	24.0	2	3.0	300.0	9.67	58.0	20.0	96.0	...	1	1	1	1	1	0	120.0	80.0	80.0	89.90
131	0	56.0	33.0	2	5.0	213.0	8.74	107.0	34.0	756.0	...	1	1	1	1	1	1	155.0	110.0	80.0	165.85
132	0	48.0	31.0	2	2.0	159.0	7.45	141.0	39.0	197.0	...	0	0	0	0	0	0	120.0	80.0	0.0	141.00

133 rows × 23 columns

FIGURE 3.8 – Aperçu du dataset 2 après le prétraitement

Pour ce dataset, nous avons réservé 5 patients pour une validation interne. Les données d'entraînement représentent 70% et celles du test représentent 30% (voir Figure 3.10).

Toujours dans le but de vérifier l'équilibre du dataset utilisé, nous avons affiché le nombre de patients atteints de la RD et sains comme la figure 3.9 suivante le montre :

```
a= dff['Ret-pathy'].value_counts()
print("Nombre de patients atteints de la RD :", a[1])
print("Nombre de patients sains :", a[0])
```

```
Nombre de patients atteints de la RD : 42
Nombre de patients sains : 91
```

FIGURE 3.9 – Nombre de patients atteints de la RD et sains du dataset2

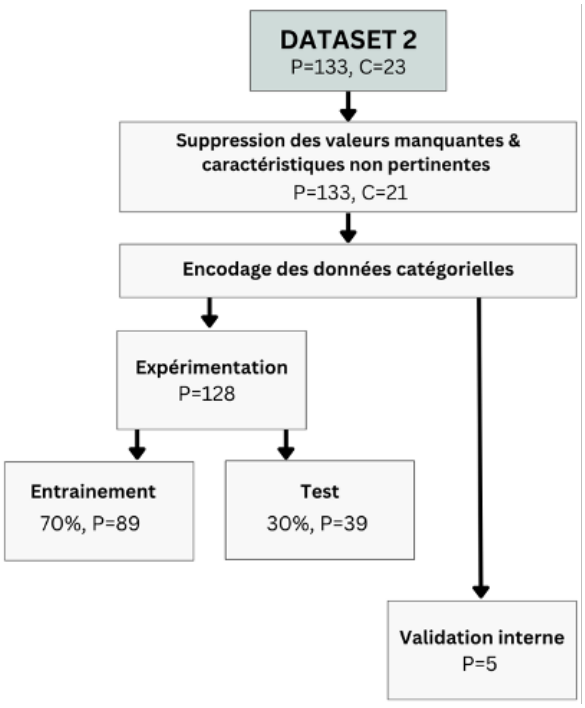

FIGURE 3.10 – Prétraitement du dataset2

Ajout d'une deuxième valeur de HbA1c

Dans le but de confirmer les résultats obtenus avec le premier dataset et par conséquent de confirmer notre hypothèse, nous avons également augmenté le deuxième dataset avec une seconde valeur de HbA1c en se basant toujours sur l'étude de Alice Larroumet [7] qui a prouvé que la majorité des patients atteints de la RD ont un déclin de plus de 3% dans la deuxième valeur de HbA1c comparée à la précédente. Cependant, nous avons fait une exception pour quelques patients. En ce qui est des patients sains, ceci reste aléatoire. Par conséquent, nous nous sommes inspiré des valeurs réelles du troisième dataset [74] qui reflètent mieux la réalité.

La figure 3.11 donne un aperçu du deuxième dataset avec une seconde valeur de HbA1c.

	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	HbA1c2	Class
0	0	31	3.0	60	12.3	4.1	2.2	0.7	2.4	15.4	37.2	7.9	1
1	0	30	7.1	81	6.7	4.1	1.1	1.2	2.4	8.1	27.4	6.5	0
2	0	45	4.1	63	10.2	4.8	1.3	0.9	3.3	9.5	34.3	9.2	0
3	0	45	4.1	63	10.2	4.8	1.3	0.9	3.3	9.5	34.3	13.3	1
4	0	45	5.3	77	11.2	3.9	1.5	1.3	2.0	10.4	29.5	11.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1447	0	71	11.0	97	7.0	7.5	1.7	1.2	1.8	0.6	30.0	8.0	0
1449	0	31	3.0	60	12.3	4.1	2.2	0.7	2.4	15.4	37.2	8.6	1
1451	0	30	7.1	81	6.7	4.1	1.1	1.2	2.4	8.1	27.4	7.2	0
1453	0	38	5.8	59	6.7	5.3	2.0	1.6	2.9	14.0	40.5	8.5	0
1455	0	54	5.0	67	6.9	3.8	1.7	1.1	3.0	0.7	33.0	11.8	1

842 rows × 13 columns

FIGURE 3.11 – Aperçu du dataset 2 après l'ajout de HbA1c2

Les mêmes étapes ont été suivies afin de comparer les performances des modèles avec une seule valeur contre deux valeurs de HbA1c. Ceci nous a permis de pouvoir valider notre approche. Les résultats se trouvent dans la section qui suit.

3.3.3 Prétraitement du Dataset3

Ce dataset a été spécialement conçu pour notre étude, il a été récolté auprès d'un cabinet médical [74] et il contient des valeurs correctes, cohérentes et complètes, par conséquent, son prétraitement consiste à seulement encoder les données catégorielles.

Contrairement aux deux premiers datasets, celui-là contient deux valeurs de HbA1c, nous avons alors dans un premier temps supprimé la deuxième valeur afin de garder le même plan de travail. Voir Figure 3.12.

```
data = pd.read_excel('dataset3.xlsx') #importation du dataset3
dff = data.copy() #copie du dataset
dff['Genre'].replace(['H','F'], [0,1], inplace=True) #encodage de "Genre"
dff.dropna(axis=0, inplace=True) #suppression des valeurs manquantes
dff=dff.drop('HbA1c2', axis=1) #suppression de "HbA1c"
dff['ATCD (F)'].replace(['DS2','X','DS2, HTA'], [0,1,2], inplace=True) #encodage de "ATCD (F)"
dff['ATCD (P)'].replace(['X','HTA'], [0,1], inplace=True) #encodage de "ATCD (P)"
dff['Rétinopathie'].replace(['Oui','Non'], [1,0], inplace=True) #encodage de "Rétinopathie"
```

FIGURE 3.12 – Importation du dataset3 et prétraitement

La Figure 3.13) montre un aperçu du dataset après le prétraitement.

	Genre	Age	IMC	Durée Diabète	ATCD (F)	ATCD (P)	HbA1c	CT	HDL	LDL	TG	Créat	ACR	Rétinopathie
0	1	33	30.86	4.0	0	0	7.5	2.03	0.38	1.35	1.48	7.16	5.00	0
1	0	48	28.63	6.0	0	0	6.5	2.02	0.44	1.20	1.89	9.32	4.10	0
2	0	69	23.94	2.0	0	1	10.4	1.25	0.32	0.83	0.48	7.09	138.00	1
3	1	66	22.66	19.0	1	0	6.9	2.22	0.48	1.46	1.38	6.09	5.15	0
4	0	63	27.13	11.0	2	1	7.0	1.87	0.53	0.97	1.85	6.93	28.00	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
75	1	70	33.59	12.0	1	1	6.3	1.13	0.39	0.60	0.69	7.79	12.00	0
76	1	69	25.81	10.0	0	1	7.6	1.19	0.38	0.55	1.32	9.99	111.00	1
77	1	60	27.64	17.0	1	1	11.1	1.98	0.45	0.53	0.53	13.00	7.16	1
78	1	58	30.41	9.0	0	0	7.0	1.85	0.32	1.27	1.28	5.66	11.60	0
79	1	57	34.11	20.0	0	1	7.3	1.43	0.30	0.76	1.86	5.99	12.96	0

80 rows × 14 columns

FIGURE 3.13 – Aperçu du dataset3 après le prétraitement

Après avoir gardé 4 patients pour la validation, celui-ci a été divisé en 85% de données d'entraînement et 15% de données de test à cause du nombre faible de patients. (Voir Figure 3.15). Ce dernier dataset contient 33 patients atteints de la RD et 47 patients sains comme affiché dans la figure 3.14 suivante :

```
a= dff['Rétinopathie'].value_counts()
print("Nombre de patients atteints de la RD:", a[1])
print("Nombre de patients sains :", a[0])
```

```
Nombre de patients atteints de la RD: 33
Nombre de patients sains : 47
```

FIGURE 3.14 – Nombre de patients atteints de la RD et sains

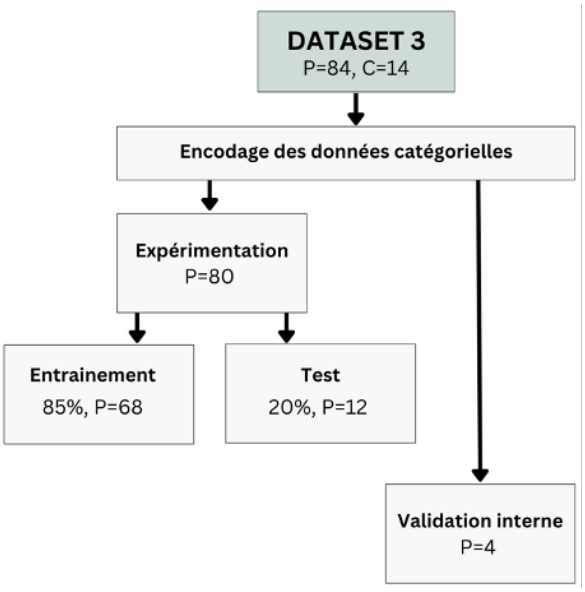

FIGURE 3.15 – Prétraitement du dataset3

Ajout de la deuxième valeur de HbA1c

Une fois que nous avons obtenus tous les résultats avec une seule valeur de HbA1c, nous avons refait la même expérimentation en gardant la deuxième valeur de HbA1c cette fois. La figure 3.16 montre un aperçu du dataset 3 avec les deux valeurs de HbA1c.

	Genre	Age	IMC	Durée Diabète	ATCD (F)	ATCD (P)	HbA1c	HbA1c2	CT	HDL	LDL	TG	Créat	ACR	Rétinopathie
0	1	33	30.86	4.0	0	0	7.5	6.6	2.03	0.38	1.35	1.48	7.16	5.00	0
1	0	48	28.63	6.0	0	0	6.5	6.6	2.02	0.44	1.20	1.89	9.32	4.10	0
2	0	69	23.94	2.0	0	1	10.4	6.1	1.25	0.32	0.83	0.48	7.09	138.00	1
3	1	66	22.66	19.0	1	0	6.9	7.1	2.22	0.48	1.46	1.38	6.09	5.15	0
4	0	63	27.13	11.0	2	1	7.0	7.4	1.87	0.53	0.97	1.85	6.93	28.00	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
75	1	70	33.59	12.0	1	1	6.3	6.4	1.13	0.39	0.60	0.69	7.79	12.00	0
76	1	69	25.81	10.0	0	1	7.6	7.8	1.19	0.38	0.55	1.32	9.99	111.00	1
77	1	60	27.64	17.0	1	1	11.1	10.2	1.98	0.45	0.53	0.53	13.00	7.16	1
78	1	58	30.41	9.0	0	0	7.0	6.9	1.85	0.32	1.27	1.28	5.66	11.60	0
79	1	57	34.11	20.0	0	1	7.3	6.9	1.43	0.30	0.76	1.86	5.99	12.96	0

80 rows × 15 columns

FIGURE 3.16 – Aperçu du dataset3 après l'ajout de HbA1c2

Cela nous a permis de constater la différence entre les performances en utilisant une seule valeur d'hémoglobine glyquée et en utiliser deux. Tous les résultats et les discussions sont résumés dans la section suivante.

3.4 Méthodologie

Dans le but d'étudier l'impact des caractéristiques, nous avons développé un modèle de

prédiction de la RD dont l'architecture est représentée par la Figure 3.17. Le modèle proposé se compose de 5 niveaux de traitement :

1. Trouver les meilleures combinaisons d'hyperparamètres pour chaque modèle afin de maximiser les performances de prédiction.
2. Évaluer les modèles en utilisant les métriques AUC, la précision et l'accuracy (ACC) et identifier le meilleur modèle selon les performances des modèles qui ont été regroupés dans un tableau récapitulatif.
3. Analyse de l'impact des caractéristiques dans la prédiction de la rétinopathie diabétique en affichant le graphe des importances pour les trois meilleurs modèles. Et supprimer ainsi les caractéristiques les moins influentes du meilleur modèle pour améliorer ses performances.
4. Construire un nouveau modèle en combinant les meilleurs modèles optimisés avec la technique d'ensemble learning "Bagging".
5. Validation du modèle obtenu avec un ensemble de patients dédié à cet effet.

La méthodologie adoptée est illustrée et résumée dans la Figure 3.17.

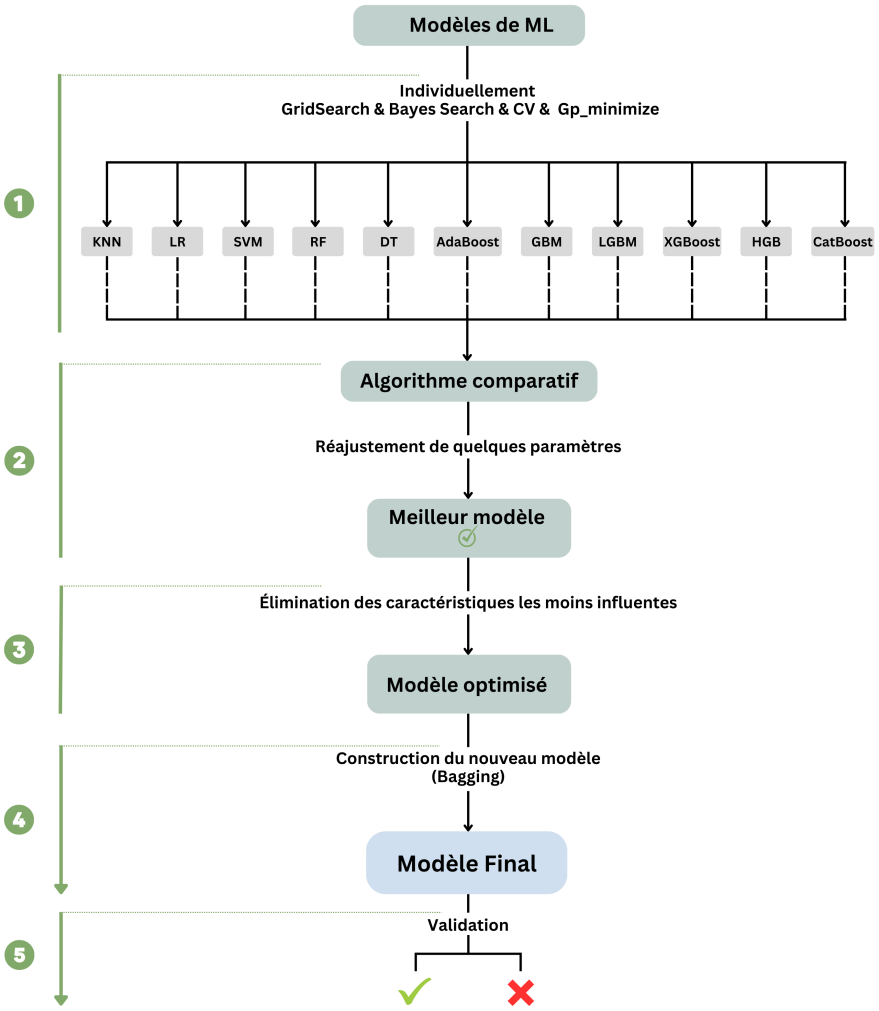

FIGURE 3.17 – Méthodologie adoptée

1. Dans le but de construire ce nouveau modèle de prédiction, nous avons en premier lieu appliqué, sur les trois datasets, 11 modèles d'apprentissage automatique dont des techniques d'ensemble learning : Adaboost, Catboost, DT, RF, GBM, HGB, KNN, LGBM, LR, SVM et XGBoost définis dans le chapitre précédent.
2. Chaque modèle a été tout d'abord entraîné et évalué individuellement et indépendamment des autres et ce, sur les trois datasets. Pour garantir les meilleurs résultats possibles des différents modèles, Nous avons eu recours à trois techniques d'optimisation d'hyperparamètres à savoir GridSearch, BayesSearch et Optimize (Gp_minimize). Ces techniques nous ont permis de trouver les meilleurs paramètres qui maximisent les performances de notre prédiction. Après l'obtention des meilleures performances et l'optimisation de tous les modèles individuellement, nous avons regroupé les méthodes et nous avons utilisé un algorithme de comparaison dans le but de les observer collectivement et d'effectuer une comparaison entre les performances. Quelques paramètres des modèles obtenus avec les techniques d'optimisation déjà citées ont été réajustés manuellement afin de garantir des performances encore meilleures. Toutes les métriques précédemment citées ont été calculées et toutes les courbes ont été tracées notamment la courbe ROC de tous les modèles simultanément. Les résultats des modèles ont été regroupés dans un tableau récapitulatif grâce à la bibliothèque "tabulate" et le meilleur modèle a été affiché avec le pourcentage de son AUC.
3. Après la sélection du meilleur modèle, nous entamons la troisième étape de notre travail à savoir l'étude et l'analyse de l'impact des caractéristiques dans la prédiction de la RD. Dans ce but, nous avons affiché le graphe d'importances sur les trois meilleurs modèles prédictifs, ce dernier montre l'influence des caractéristiques sur ces modèles. Nous avons ensuite sélectionné les caractéristiques qui se répètent le plus c'est à dire les moins influentes sur les trois meilleurs modèles, nous les avons supprimé dans le meilleur modèle et enfin, nous avons affiché le résultat.
4. La dernière étape consiste à construire un nouveau modèle en combinant notre meilleur modèle avec une des techniques d'ensemble learning à savoir le Bagging.
5. Enfin, nous avons effectué une validation avec les patients dédiés à ça avec le nouveau modèle optimisé. Tous les résultats sont affichés dans la section suivante (Résultats).

En parallèle à ce travail, nous avons testé, sur les trois datasets, une autre méthode d'ensemble learning à savoir le Stacking ainsi que la validation croisée en utilisant trois techniques : K-Fold, Shuffle Split et Stratified K-Fold. Mais, ces dernières n'ont pas amélioré les performances des modèles.

3.5 Résultats et discussions

Dans la suite de notre étude, nous présenterons les résultats de notre analyse comparative des différents modèles d'apprentissage automatique appliqués aux trois ensembles de données pour la prédiction de la rétinopathie diabétique. Pour cela nous allons procéder comme suit pour chaque dataset :

A- Première étape : Comme nous l'avons cité auparavant, nos modèles de ML ont subi plusieurs méthodes d'amélioration et d'optimisation de manière individuelle jusqu'à obtention de modèles performants. Ces derniers ont été ensuite intégrés à un algorithme de comparaison. Cette partie correspond alors aux résultats de cette étape.

B- Deuxième étape : Cette partie correspond aux résultats de l'élimination des caractéristiques non pertinentes dans le meilleur modèle et ce, en se basant sur les graphes d'importances des trois meilleurs modèles.

C- Troisième étape : Comprend les résultats obtenus avec le nouveau modèle proposé en combinant le meilleur modèle avec la méthode du Bagging.

D- Quatrième étape : S'agit des résultats de validation.

3.5.1 Dataset 1

Avant ajout de HbA1c2

A-Première étape

Dans ce qui suit les meilleures performances des modèles selon les méthodes d'amélioration ainsi que les courbes collectives.

TABLE 3.5 – Les performances des 11 méthodes sur le dataset 1 avec 1 HbA1c

Modèle	Méthode d'amélioration	AUC	Accuracy	Précision
Adaboost	Grid search	0.561235	0.622754	0.809524
CatBoost	Grid search + réajustement des paramètres	**0.615963**	0.580838	0.8
DT	Grid search + réajustement des paramètres	0.590732	0.622754	0.782609
GBM	Bayes Search	0.560946	0.568862	0.833333
HGB	Gp_minimize	0.569115	0.622754	0.842105
KNN	Grid search	0.464719	0.520958	0.4375
LGBM	Grid search + réajustement des paramètres	0.576634	0.580838	0.875
LR	Grid search	0.478022	0.526946	0.44
RF	Grid search	0.552921	0.610778	0.823529
SVM	Manuelle	0.563042	0.556886	1
XGBoost	Grid search	0.526027	0.610778	0.761905

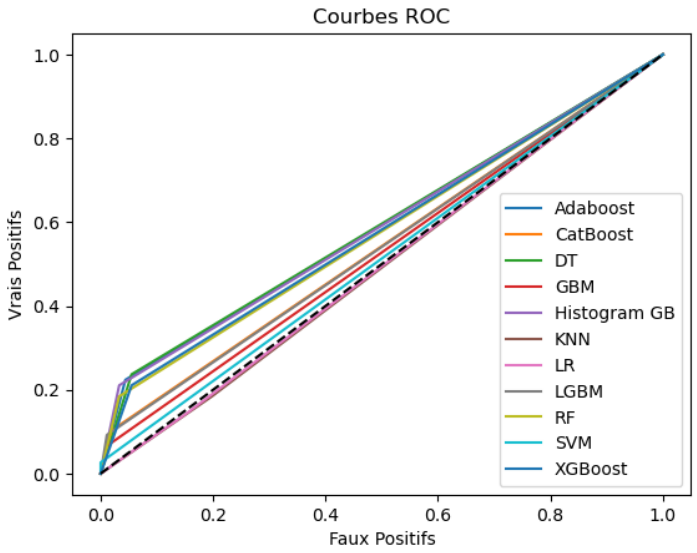

FIGURE 3.18 – Courbe ROC collective dataset1 avec 1 HbA1c

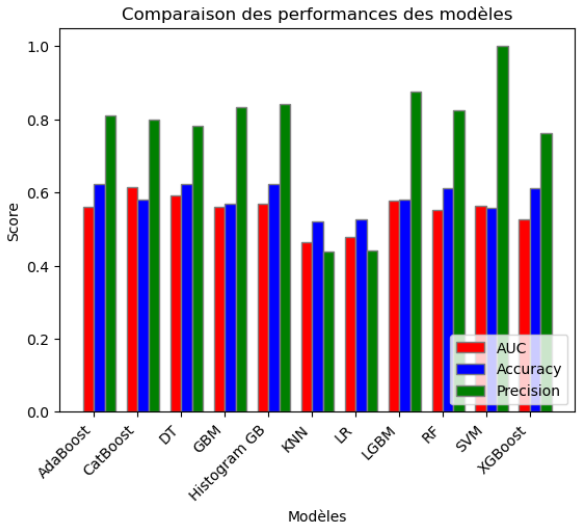

FIGURE 3.19 – Courbe des performances collectives du dataset1 avec 1HbA1c

Le meilleur modèle s'agit de CatBoost avec une AUC de 61.59%. La performance de ce modèle a été obtenue en utilisant la combinaison des paramètres présentés dans la Figure 3.20 et la Table 3.6.

```
cb=CatBoostClassifier(iterations=50,
                      learning_rate=0.01,
                      depth=3,
                      l2_leaf_reg=1,
                      random_strength=1,
                      loss_function='Logloss',
                      eval_metric='AUC')
```

FIGURE 3.20 – Paramètres du meilleur modèle du dataset1 avec 1 HbA1c

TABLE 3.6 – Explication des paramètres du modèle CatBoost

Paramètre	**Explication**
iterations	Sa valeur par défaut étant 500, ce paramètre représente le nombre maximal d'arbres pouvant être créés. Dans notre cas, c'est 50 [75]
learning_rate	Indique le taux d'ajustement ou de mise à jour des coefficients du modèle. Cela peut être interprété comme la vitesse à laquelle le modèle apprend [75]
depth	Représente la profondeur de l'arbre [75]
l2_leaf_reg	C'est le coefficient de la condition de régularisation L2 de la fonction de coût [75]
random_strength	Ce paramètre est utilisé pour éviter de surajouter le modèle. Il représente le degré du caractère aléatoire à utiliser pour la notation des divisions lorsque la structure arborescente est sélectionnée [75]
loss_function	Permet d'adapter la fonction de perte au type de problème de classification [75]
eval_metric	Sert à définir Métrique d'évaluation des données de validation. Il peut prendre trois valeur 'RMSE' pour une régression, 'MultiClass' pour une classification multiclasse et enfin ' AUC' pour une classification binaire comme la nôtre [75]

Dans ce qui suit toutes les courbes et les résultats affichées pour le meilleur modèle :

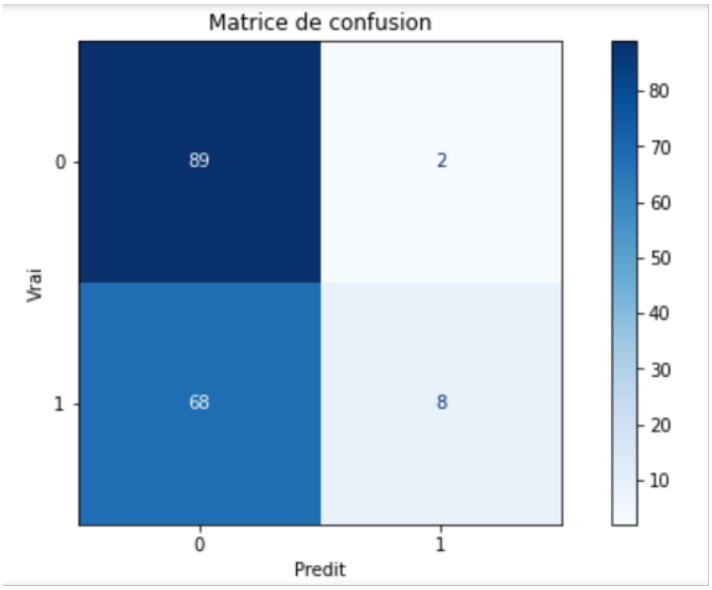

FIGURE 3.21 – Matrice de confusion du meilleur modèle du dataset1 avec 1 HbA1c

	precision	recall	f1-score	support
0	0.57	0.98	0.72	91
1	0.80	0.11	0.19	76
accuracy			0.58	167
macro avg	0.68	0.54	0.45	167
weighted avg	0.67	0.58	0.48	167

FIGURE 3.22 – Rapport de classification du meilleur modèle du dataset1 avec 1 HbA1c

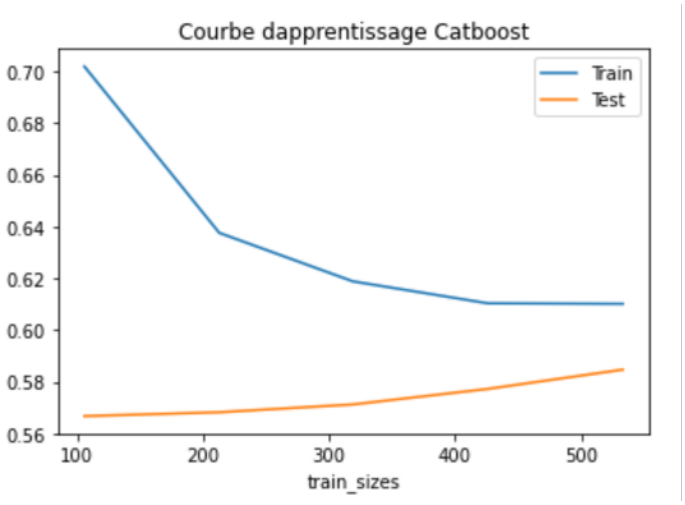

FIGURE 3.23 – Courbe d'apprentissage du meilleur modèle du dataset1 avec 1 HbA1c

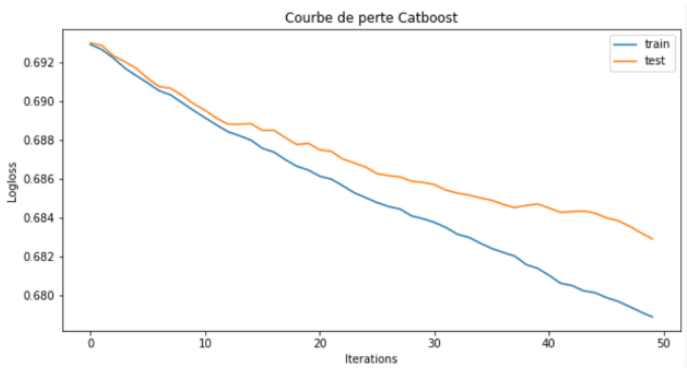

FIGURE 3.24 – Courbe perte du meilleur modèle du dataset1 avec 1 HbA1c

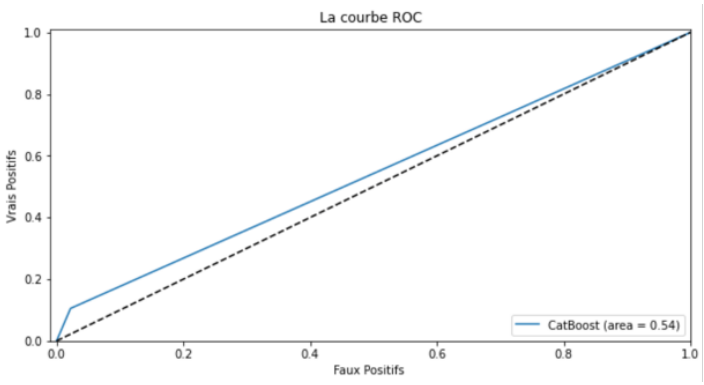

FIGURE 3.25 – Courbe ROC du meilleur modèle du dataset1 avec 1 HbA1c

B-Deuxième étape

Les trois meilleurs modèles sont : CatBoost, Decision Tree et Histogram Gradient Boosting. HGB ne possède pas de graphe d'importances, par conséquent, nous avons eu recours à la méthode SHAP afin de visualiser la contribution des caractéristiques sur les résultats de prédiction. Les graphes sont les suivants :

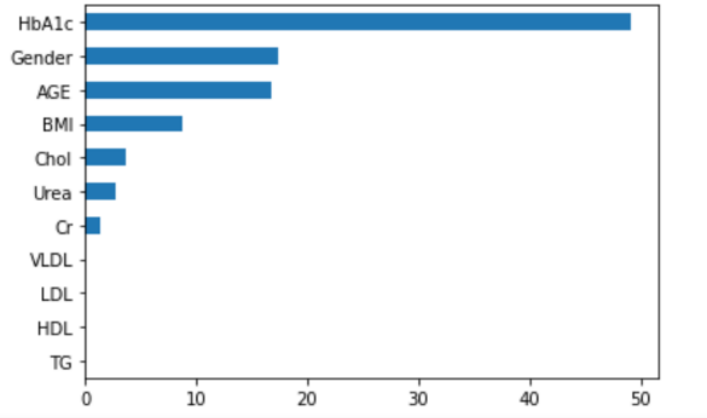

FIGURE 3.26 – Graphe d'importance du modèle CatBoost du dataset1 avec 1 HbA1c

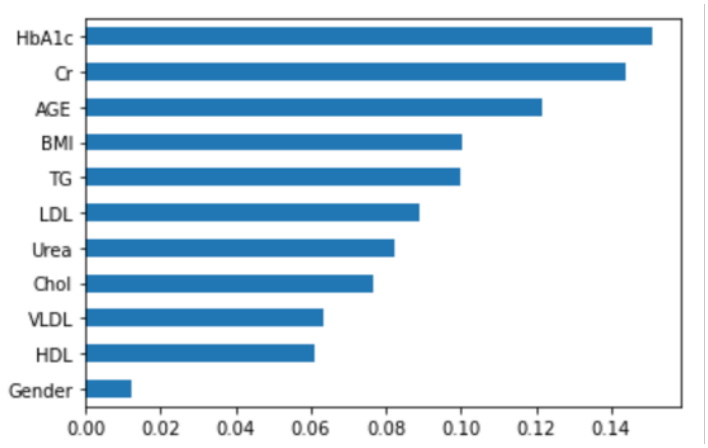

FIGURE 3.27 – Graphe d'importance du modèle DT du dataset1 avec 1 HbA1c

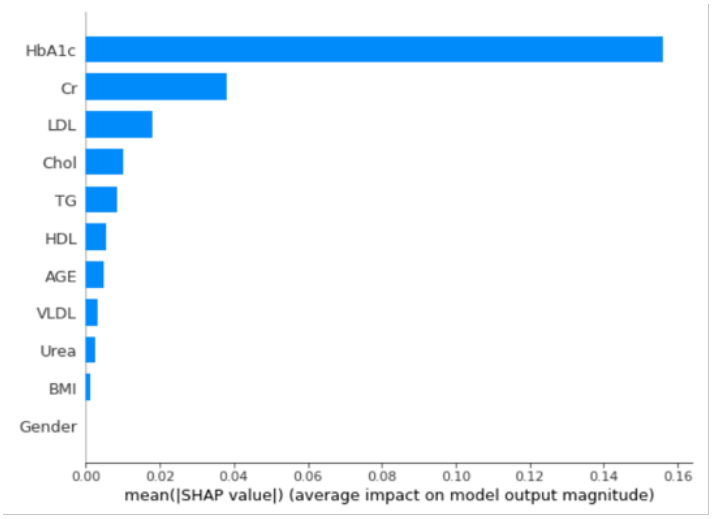

FIGURE 3.28 – Graphe SHAP du modèle HGB du dataset1 avec 1 HbA1c

Nous remarquons que les caractéristiques : "TG", "LDL", "VLDL" ET "HDL" sont celles qui se répètent le plus comme caractéristiques non influentes dans les trois meilleurs modèles. Par conséquent, elles ont été supprimées dans le meilleur modèle c'est à dire CatBoost. La table suivante montre les résultats obtenus.

TABLE 3.7 – Performances du modèle optimisé du dataset1 avec 1 HbA1c

Modèle optimisé	**AUC**	**Accuarcy**	**Précision**
CatBoost	0.6243	0.605	0.917

Nous remarquons que les performances se sont améliorées.

C-Troisième étape

Notre modèle est désormais optimisé et les facteurs de risques sont identifiés. Nous passons

alors à la construction d'un nouveau modèle en le combinant avec la technique du Bagging comme suit :

```
bagging_cb = BaggingClassifier(base_estimator=cb1, n_estimators=10)
bagging_cb.fit(x, y)
y_pred = bagging_cb.predict(x_test)
accuracy = accuracy_score(y_test, y_pred)
y_prob_test=bagging_cb.predict_proba(x_test)[:,1]
auc_test= roc_auc_score(y_test,y_prob_test)
Test_precision = precision_score(y_test,y_pred)
print('Accuracy = ' + Style.BRIGHT + str(accuracy) + Style.RESET_ALL +
      '\nAUC = ' + Style.BRIGHT + str(Test_precision) + Style.RESET_ALL +
      '\nPrécision = ' + Style.BRIGHT + str(cb_precision) + Style.RESET_ALL+'\n')
```

FIGURE 3.29 – Code Bagging du dataset1 avec 1 HbA1c

Voici les nouvelles performances de notre modèle :

TABLE 3.8 – Performances du modèle final du dataset1 avec 1 HbA1c

Modèle final	**AUC**	**Accuarcy**	**Précision**
CatBoost + Bagging	0.857	0.635	0.8

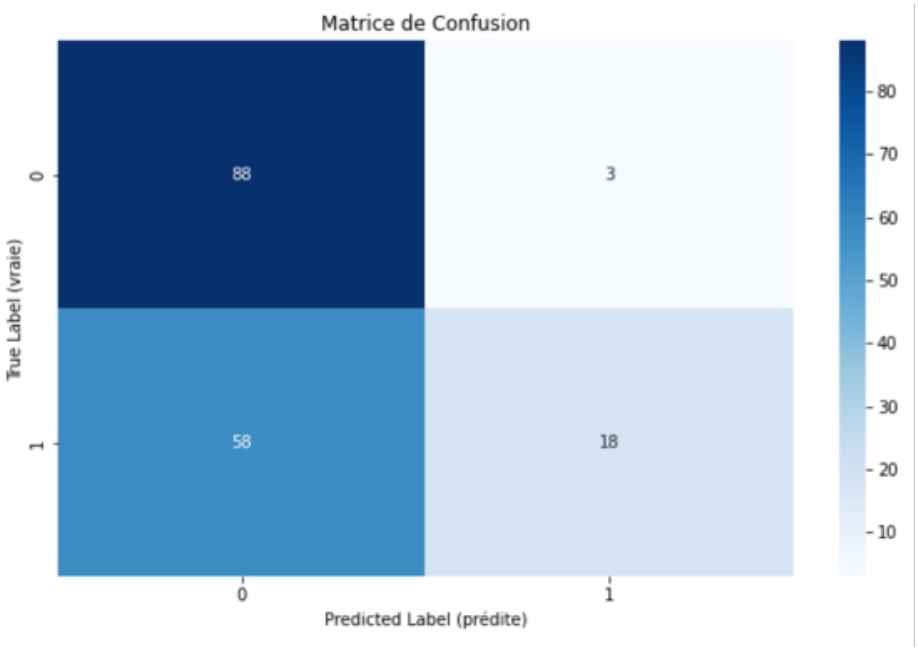

FIGURE 3.30 – Matrice de confusion du modèle final du dataset1 avec 1 HbA1c

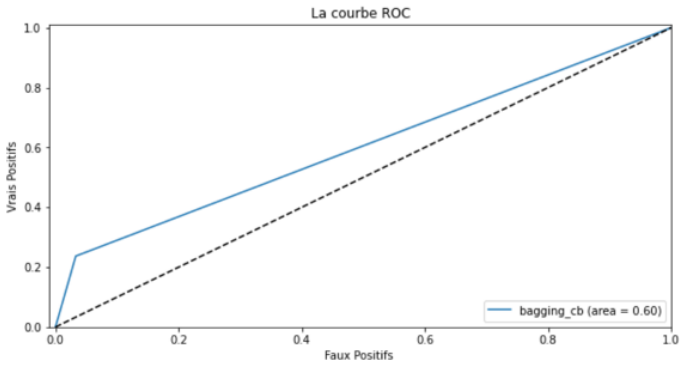

FIGURE 3.31 – Courbe ROC du modèle final du dataset1 avec 1 HbA1c

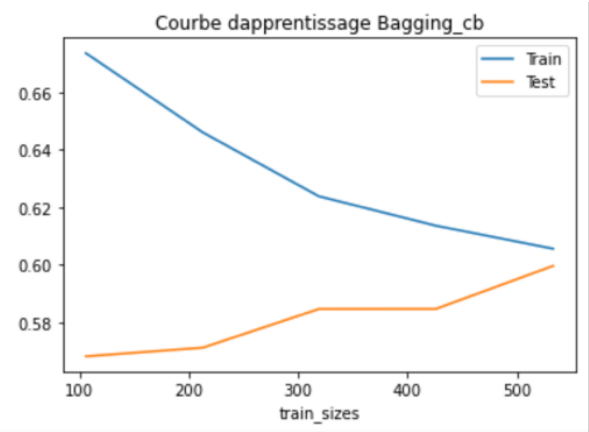

FIGURE 3.32 – Courbe d'apprentissage du modèle final du dataset1 avec 1 HbA1c

Notre modèle arrive à mieux prédire et distinguer entre les patients sains et les patients atteints ou susceptible d'avoir la RD.

D- Quatrième étape

Dans ce qui suit deux exemples de validation :

	Gender	AGE	Urea	Cr	HbA1c	Chol	BMI	Class
0	1	57	4.1	70	9.3	5.3	29.0	0
1	1	55	4.1	34	13.9	5.4	33.0	1
2	0	55	3.1	39	8.5	5.0	27.0	0
3	0	28	3.5	61	8.5	4.5	37.0	0
4	0	69	10.3	185	7.7	4.9	37.0	1
5	0	71	11.0	97	7.0	7.5	30.0	0
6	0	31	3.0	60	12.3	4.1	37.2	1
7	0	30	7.1	81	6.7	4.1	27.4	0
8	0	38	5.8	59	6.7	5.3	40.5	0
9	0	54	5.0	67	6.9	3.8	33.0	1

FIGURE 3.33 – Dataset de validation du dataset1 avec 1 HbA1c

Exemple 1

```
def patient1 (bagging_cb,Gender=1,AGE=57,Urea=4.1,Cr=70,HbA1c=9.3,Chol=5.3,BMI=29.0):
    x= np.array([Gender,AGE,Urea,Cr,HbA1c,Chol,BMI]).reshape(1,7)
    print(bagging_cb.predict(x))
patient1(bagging_cb)
```

```
[0]
```

FIGURE 3.34 – Exemple 1 de validation du dataset1 avec 1 HbA1c

Le résultat de la validation avec le premier exemple est correct.

Exemple 2

```
def patient2 (bagging_cb,Gender=1,AGE=55,Urea=4.1,Cr=34,HbA1c=13.9,Chol=5.4,BMI=33.0):
    x= np.array([Gender,AGE,Urea,Cr,HbA1c,Chol,BMI]).reshape(1,7)
    print(bagging_cb.predict(x))
patient2(bagging_cb)
```

```
[1]
```

FIGURE 3.35 – Exemple 2 de validation du dataset1 avec 1 HbA1c

De même, le résultat de la validation avec le deuxième exemple est correct.

Après ajout de HbA1c2

A- Première étape

De même, dans ce qui suit les résultats de tous les modèles après l'optimisation des paramètres ainsi que le réajustement manuel :

TABLE 3.9 – Performances des 11 méthodes sur le dataset 1 avec 2 HbA1c

Modèle	Méthode d'amélioration	AUC	Accuracy	Précision
Adaboost	Manuelle	0.961828	0.904192	0.954545
CatBoost	Grid search + réajustement des paramètres	**0.989879**	0.958084	0.972603
DT	Grid search	0.89611	0.898204	0.927536
GBM	Grid search	0.985107	0.928144	0.957143
HGB	Manuelle	0.985975	0.928144	0.957143
KNN	Manuelle	0.573091	0.57485	0.54386
LR	Grid search	0.565789	0.580838	0.588235
LGBM	Manuelle	0.984673	0.91018	0.955224
RF	Grid search	0.85895	0.754491	0.843137
SVM	Grid search	0.89987	0.850299	0.849315
XGBoost	Manuelle	0.988143	0.928144	0.984848

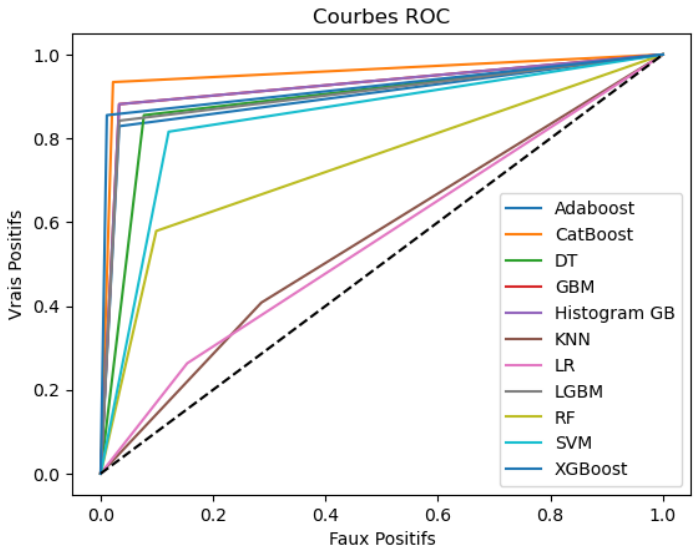

FIGURE 3.36 – Courbe ROC collective du dataset1 avec 2 HbA1c

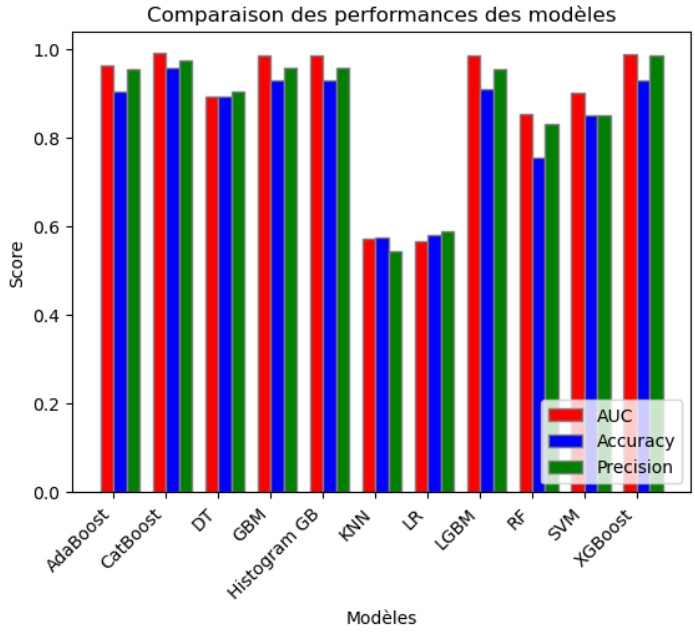

FIGURE 3.37 – Courbe des performances collectives du dataset1 avec 2 HbA1c

En utilisant deux valeurs de HbA1c, tous les modèles ont atteint des performances remarquables et nous constatons une différence non négligeable. Le meilleur modèle est CatBoost avec une valeur AUC de 0.989.

Les paramètres utilisés dans ce modèle sont affichés dans la figure 3.38 et Table 3.10.

```
cb=CatBoostClassifier(iterations=100,
                      learning_rate=0.1,
                      depth=5,
                      l2_leaf_reg=1,
                      random_strength= 0,
                      loss_function='Logloss',
                      logging_level='Silent',
                      eval_metric='AUC')
```

FIGURE 3.38 – Paramètres du meilleur modèle du dataset1 avec 2 HbA1c

TABLE 3.10 – Explication des paramètres du modèle CatBoost 2

Paramètre	**Explication**
Logging_level	Peut prendre plusieurs valeurs à savoir 'Silent', 'Verbose', 'Info' et 'Debug' il sert à définir le niveau de journalisation global pour l'ensemble du processus d'entraînement du modèle [75].

Les paramètres restants sont précédemment expliqués dans la Table 3.6.

Les courbes et les résultats du modèle CatBoost sont affichés dans les figures 3.39 ,3.40 , 3.41, 3.42 et 3.43 :

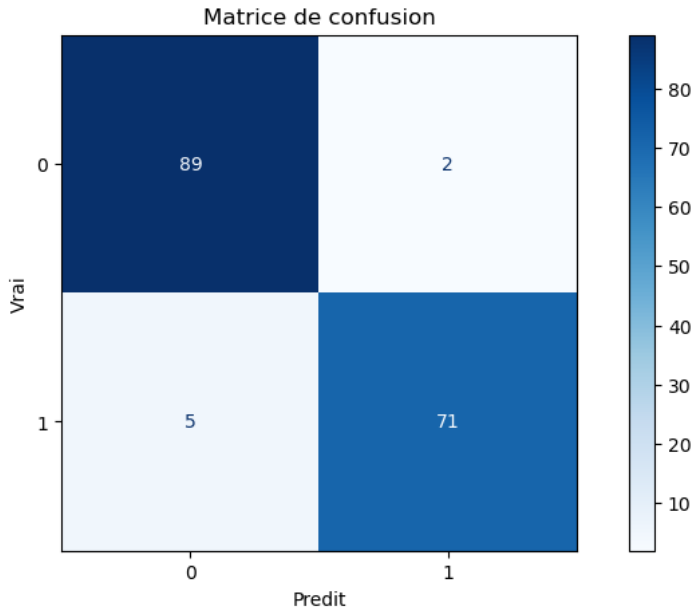

FIGURE 3.39 – Matrice de confusion meilleur modèle du dataset1 avec 2 HbA1c

	precision	recall	f1-score	support
0	0.95	0.98	0.96	91
1	0.97	0.93	0.95	76
accuracy			0.96	167
macro avg	0.96	0.96	0.96	167
weighted avg	0.96	0.96	0.96	167

FIGURE 3.40 – Rapport de classification du meilleur modèle du dataset1 avec 2 HbA1c

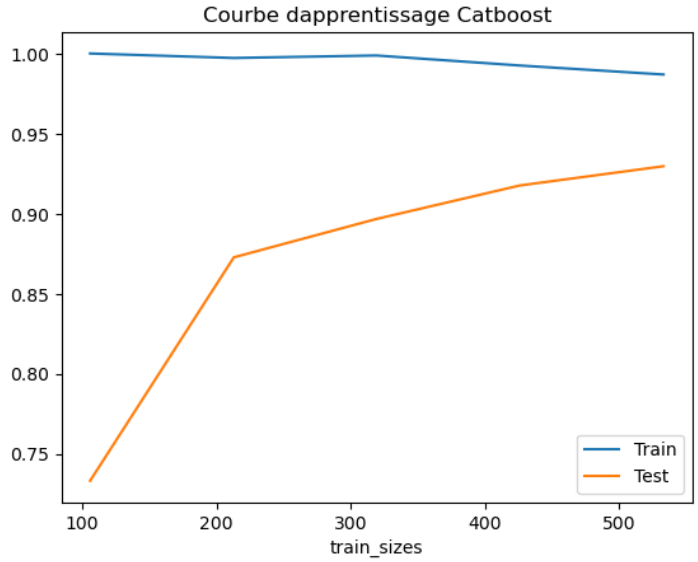

FIGURE 3.41 – Courbe d'apprentissage du meilleur modèle du dataset1 avec 2 HbA1c

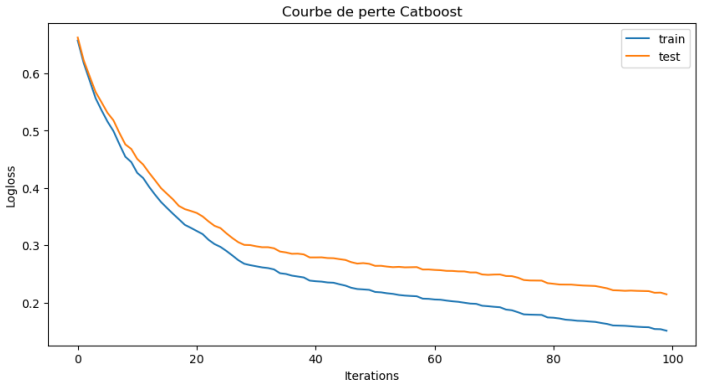

FIGURE 3.42 – Courbe perte du meilleur modèle du dataset1 avec 2 HbA1c

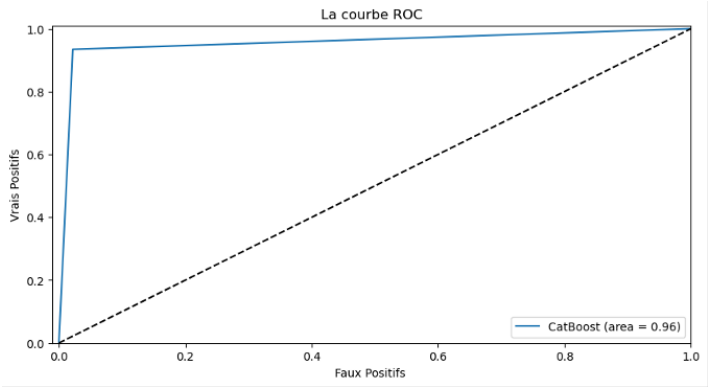

FIGURE 3.43 – Courbe ROC du meilleur modèle du dataset1 avec 2 HbA1c

B- Deuxième étape

Nous avons affiché les graphes d'importances des meilleurs modèles : Catboost, Xtreme Gradient Boosting et nous avons utilisé la méthode SHAP pour Histogram Gradient Boosting comme illustré dans les figures 3.44, 3.45 et 3.46. Ces graphes indiquent que les caractéristiques : "Gender" et "VLDL" sont les moins influentes. Par conséquent, elles ont été supprimées dans le meilleur modèle. Les résultats sont présentés dans la Table 3.28.

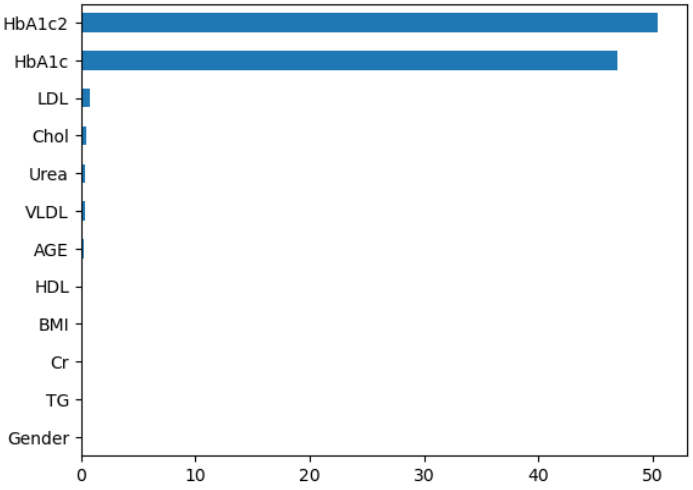

FIGURE 3.44 – Graphe d'importance du modèle CatBoost du dataset1 avec 2 HbA1c

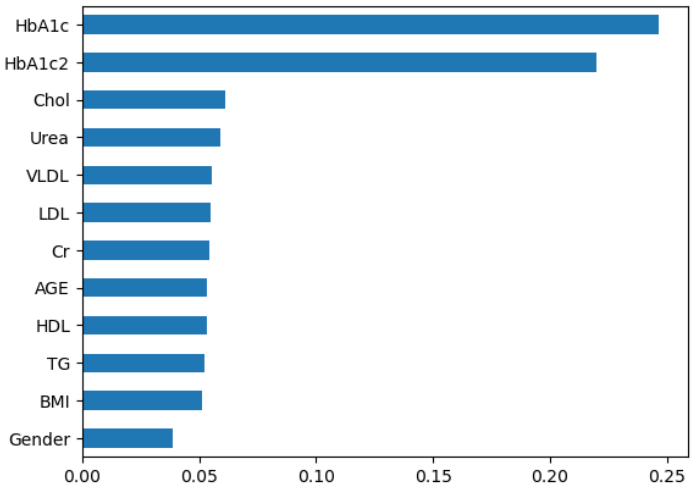

FIGURE 3.45 – Graphe d'importance du modèle XGBoost du dataset1 avec 2 HbA1c

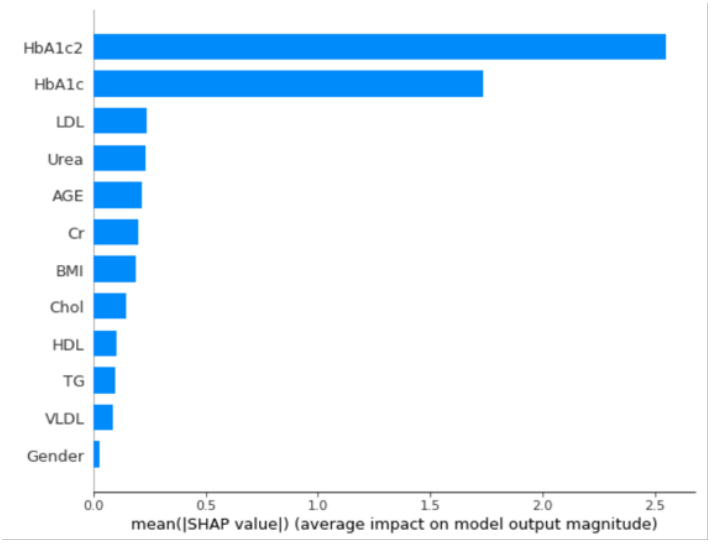

FIGURE 3.46 – Graphe SHAP du modèle HGB du dataset1 avec 2 HbA1c

TABLE 3.11 – Performances du modèle optimisé du dataset1 avec 2 HbA1c

Modèle optimisé	**AUC**	**Accuarcy**	**Précision**
CatBoost	0.99046	0.93413	0.97101

C- Troisième étape

De même qu'avec une seule valeur de HbA1c, nous entamons la construction du nouveau modèle en utilisant le meilleur modèle optimisé, les facteurs de risques et la technique du Bagging. Voici le modèle final et les résultats obtenus :

```
from sklearn.ensemble import BaggingClassifier
bagging_cb = BaggingClassifier(base_estimator=cb, n_estimators=10)
bagging_cb.fit(x, y)
y_pred = bagging_cb.predict(x_test)
accuracy = accuracy_score(y_test, y_pred)
y_prob_test=bagging_cb.predict_proba(x_test)[:,1]
auc_test= roc_auc_score(y_test,y_prob_test)
Test_precision = precision_score(y_test,y_pred)
print('Accuracy = ' + Style.BRIGHT + str(accuracy) + Style.RESET_ALL +
      '\nAUC = ' + Style.BRIGHT + str(Test_precision) + Style.RESET_ALL +
      '\nPrécision = ' + Style.BRIGHT + str(lgbm_precision) + Style.RESET_ALL+'\n')
```

FIGURE 3.47 – Code Bagging du dataset1 avec 2 HbA1c

TABLE 3.12 – Performances du modèle final du dataset1 avec 2 HbA1c

Modèle final	AUC	Accuarcy	Précision
CatBoost	1.0	1.0	0.95522

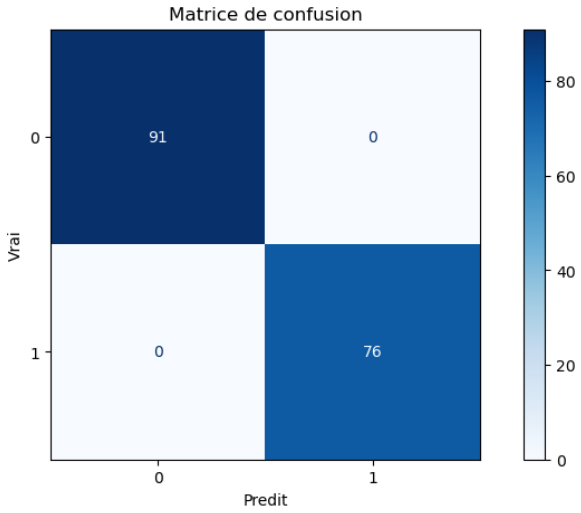

FIGURE 3.48 – Matrice de confusion du modèle final du dataset1 avec 2 HbA1c

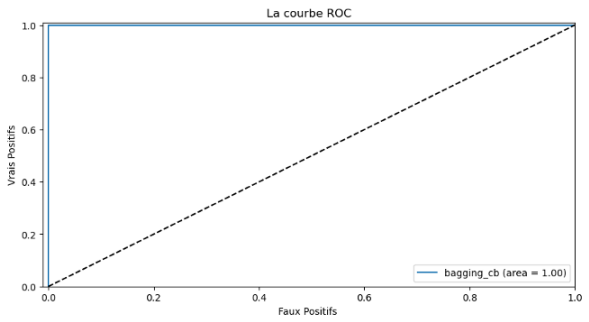

FIGURE 3.49 – Courbe ROC du modèle final du dataset1 avec 2 HbA1c

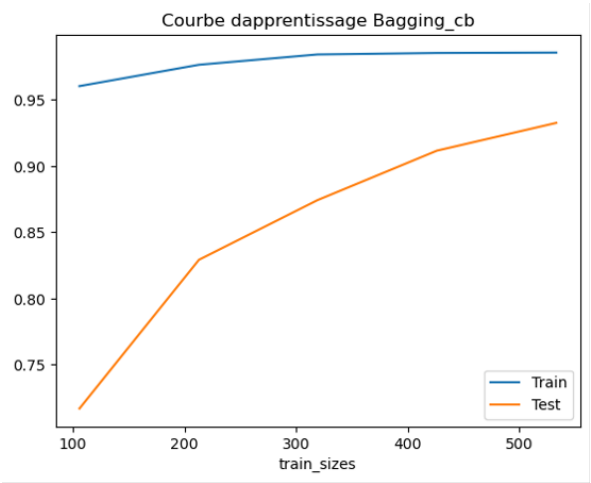

FIGURE 3.50 – Courbe d'apprentissage du modèle final du dataset1 avec 2 HbA1c

D- Quatrième étape

Dans ce qui suit deux exemples de validation :

	No_Pation	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	BMI	HbA1c2	Class
0	454316	57	4.1	70	9.3	5.3	3.3	1.0	1.4	29.0	7.0	0
1	4543	55	4.1	34	13.9	5.4	1.6	1.6	3.1	33.0	10.0	1
2	454316	55	3.1	39	8.5	5.0	2.5	1.9	2.9	27.0	7.0	0
3	454316	28	3.5	61	8.5	4.5	1.9	1.1	2.6	37.0	9.5	0
4	454316	69	10.3	185	7.7	4.9	1.9	1.2	3.0	37.0	11.0	1
5	454317	71	11.0	97	7.0	7.5	1.7	1.2	1.8	30.0	8.0	0
6	876534	31	3.0	60	12.3	4.1	2.2	0.7	2.4	37.2	8.6	1
7	87654	30	7.1	81	6.7	4.1	1.1	1.2	2.4	27.4	7.2	0
8	24004	38	5.8	59	6.7	5.3	2.0	1.6	2.9	40.5	8.5	0
9	24054	54	5.0	67	6.9	3.8	1.7	1.1	3.0	33.0	11.8	1

FIGURE 3.51 – Dataset de validation du dataset1 avec 2 HbA1c

Exemple 1

```
def patient1 (bagging_cb,AGE=57,Urea=4.1,Cr=70,HbA1c=9.3,Chol=5.3,TG=3.3,HDL=1.0,LDL=1.4,BMI=29.0,HbA1c2=7):
    x= np.array([AGE,Urea,Cr,HbA1c,Chol,TG,HDL,LDL,BMI,HbA1c2]).reshape(1,10)
    print(bagging_cb.predict(x))
patient1(bagging_cb)
```

```
[0]
```

FIGURE 3.52 – Exemple 1 de validation du dataset1 avec 2 HbA1c

Le résultat de validation est correct.

Exemple 2

```
def patient2 (bagging_cb,AGE=55,Urea=4.1,Cr=34,HbA1c=13.9,Chol=5.4,TG=1.6,HDL=1.6,LDL=3.1,BMI=33.0,HbA1c2=10):
    x= np.array([AGE,Urea,Cr,HbA1c,Chol,TG,HDL,LDL,BMI,HbA1c2]).reshape(1,10)
    print(bagging_cb.predict(x))
patient2(bagging_cb)
```

```
[1]
```

FIGURE 3.53 – Exemple 1 de validation du dataset1 avec 2 HbA1c

De même que le premier, le résultat de validation est correct.

Comparaison

Le tableau suivant résume les performances des modèles appliqués au dataset1 avec un et deux HbA1c.

TABLE 3.13 – Table de comparaison du dataset1 avec un et deux HbA1c

	Performance	**1 HbA1c**	**2 HbA1c**
Nom du modèle	/	CatBoost	CatBoost
Modèle de base	**AUC**	0.616	0.989
	ACC	0.580	0.958
	Précison	0.8	0.973
Modèle optimisé	**AUC**	0.624	0.990
	ACC	0.605	0.934
	Précison	0.917	0.971
Modèle final	**AUC**	0.857	1.0
	ACC	0.635	1.0
	Précision	0.8	0.95522

Nous remarquons une différence apparente entre les performances des modèles prédictifs en utilisant une seule valeur d'hémoglobine glyquée contre l'utilisation de deux valeurs. Ceci représente alors le premier pas vers la validation de notre approche.

3.5.2 Dataset 2

Dans cette partie, les mêmes étapes sont suivies en utilisant le deuxième dataset dans le but de comparer les performances des modèles prédictifs mais également d'étudier l'impact des différentes caractéristiques.

Avant ajout de HbA1c2

A-Première étape

Les résultats obtenus sont les suivants :

TABLE 3.14 – Les performances des 11 méthodes sur le dataset 2 avec 1 HbA1c

Modèle	Méthode d'amélioration	AUC	Accuracy	Précision
Adaboost	Grid search + réajustement des paramètres	0.755435	0.717949	0.692308
CatBoost	Manuelle	**0.861413**	0.794872	0.9
DT	Bayes search	0.824728	0.74359	0.875
GBM	Grid search	0.834239	0.794872	0.9
HGB	Gp_minimize	0.84375	0.717949	0.777778
KNN	Grid search + réajustement des paramètres	0.638587	0.666667	1
LGBM	Manuelle	0.853261	0.794872	0.9
LR	Manuelle	0.796196	0.692308	0.75
RF	Manuelle	0.841033	0.769231	0.888889
SVM	Manuelle	0.703804	0.589744	0
XGBoost	Manuelle	0.817935	0.717949	0.727273

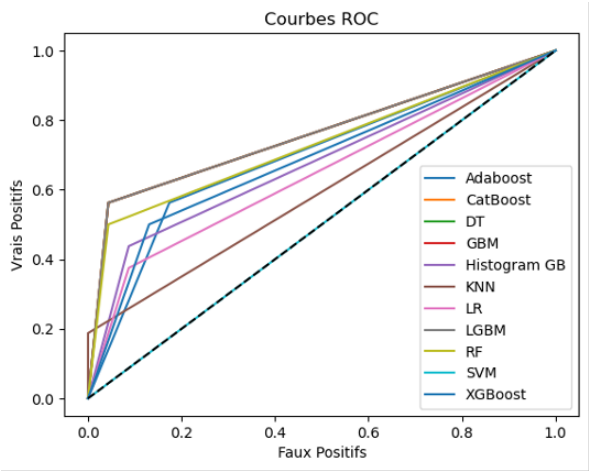

FIGURE 3.54 – Courbe ROC collective dataset2 avec 1 HbA1c

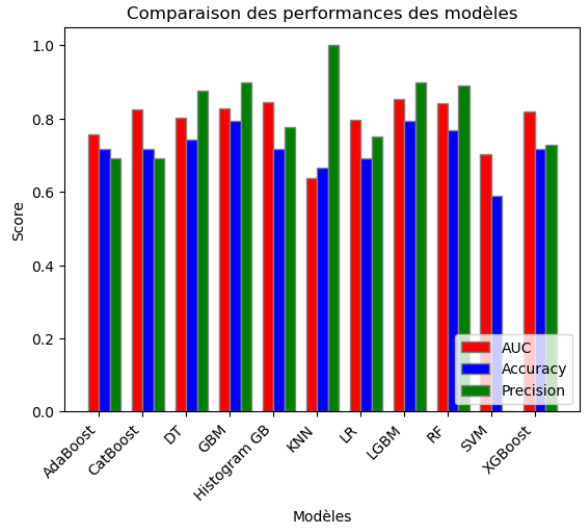

FIGURE 3.55 – Courbe des performances collectives du dataset2 avec 1HbA1c

Le meilleur modèle s'agit de CatBoost avec une AUC de 86.14%. La performance de ce modèle a été obtenue en utilisant les paramètres suivants :

```
cb=CatBoostClassifier(iterations=300,
                      learning_rate=0.1,
                      depth=5,
                      l2_leaf_reg=1,
                      random_strength= 0,
                      loss_function='Logloss',
                      logging_level='Silent',
                      eval_metric='AUC')
```

FIGURE 3.56 – Paramètres du meilleur modèle du dataset2 avec 1 HbA1c

Dans ce qui suit les courbes et les résultats du meilleur modèle :

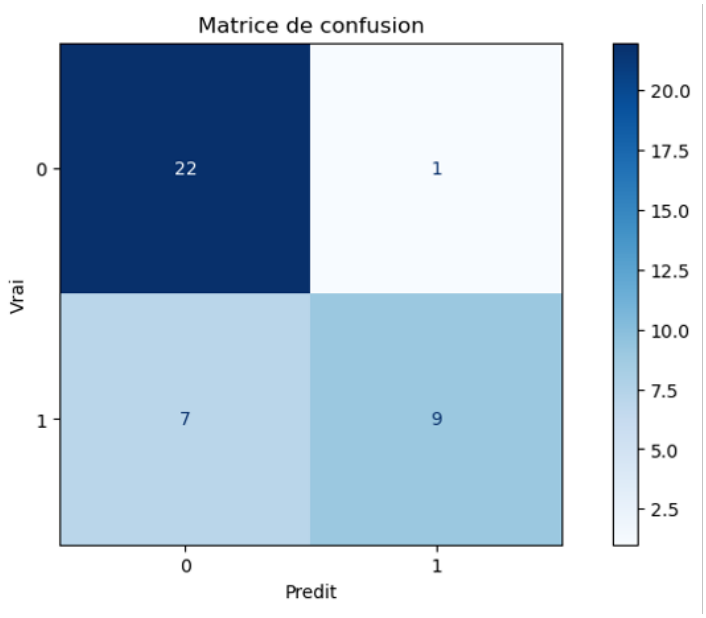

FIGURE 3.57 – Matrice de confusion du meilleur modèle du dataset2 avec 1 HbA1c

	precision	recall	f1-score	support
0	0.76	0.96	0.85	23
1	0.90	0.56	0.69	16
accuracy			0.79	39
macro avg	0.83	0.76	0.77	39
weighted avg	0.82	0.79	0.78	39

FIGURE 3.58 – Rapport de classification du meilleur modèle du dataset2 avec 1 HbA1c

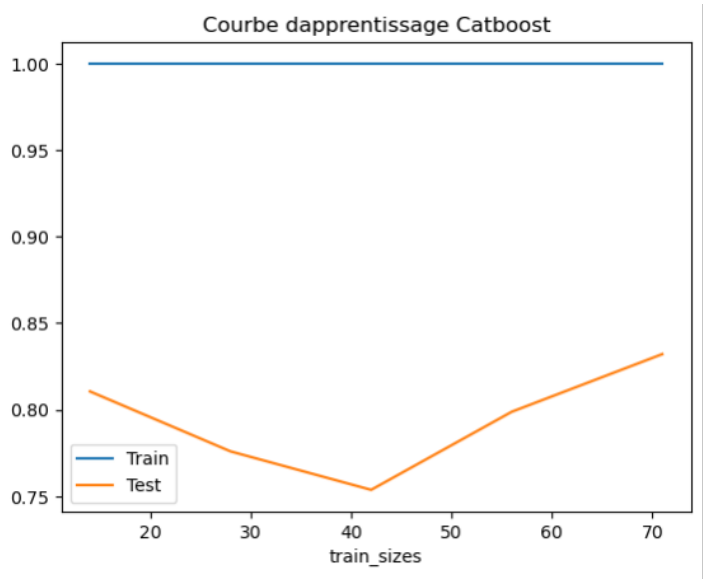

FIGURE 3.59 – Courbe d'apprentissage du meilleur modèle du dataset2 avec 1 HbA1c

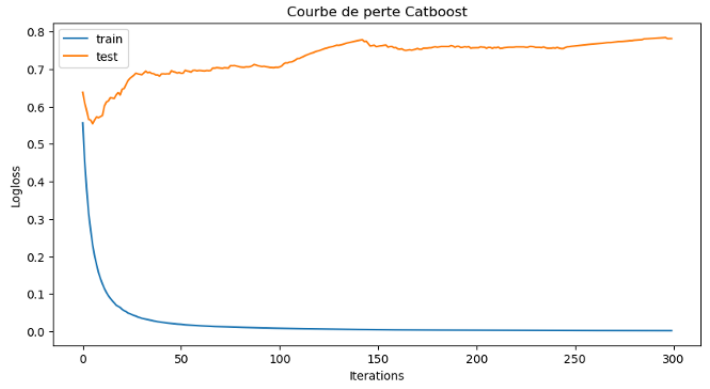

FIGURE 3.60 – Courbe perte du meilleur modèle du dataset2 avec 1 HbA1c

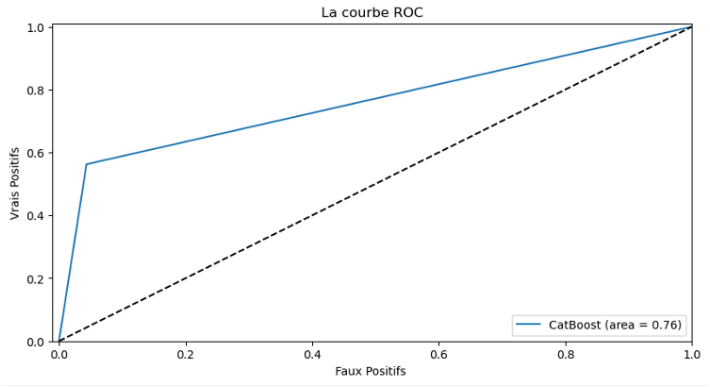

FIGURE 3.61 – Courbe ROC du meilleur modèle du dataset2 avec 1 HbA1c.

B-Deuxième étape

Les graphes d'importances des trois meilleurs modèles sont les suivants :

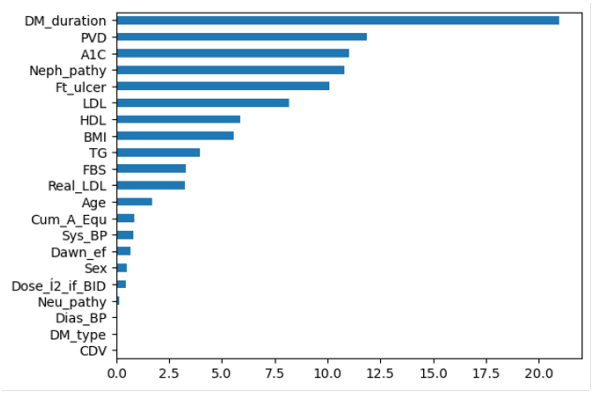

FIGURE 3.62 – Graphe d'importance du modèle CatBoost du dataset2 avec 1 HbA1c

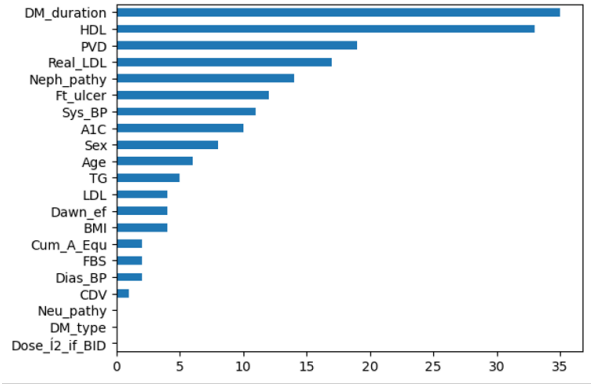

FIGURE 3.63 – Graphe d'importance du modèle LGBM du dataset2 avec 1 HbA1c

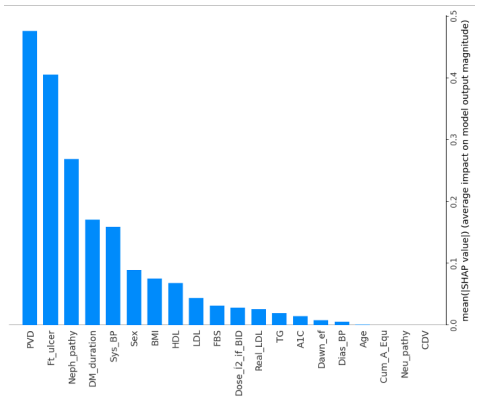

FIGURE 3.64 – Graphe SHAP du modèle HGB du dataset2 avec 1 HbA1c

Les caractéristiques "Dose_Í2_if_BID", "Dawn_ef", "DM_type", "Sys_BP", "Neu_pathy"

et “Cum_A_Equ” ont été supprimées car jugées les moins influentes par les trois modèles. Les résultats obtenus sont les suivants :

TABLE 3.15 – Performances du modèle optimisé du dataset2 avec 1 HbA1c

Modèle optimisé	AUC	Accuarcy	Précision
CatBoost	0.87340	0.85	0.86667

C-Troisième étape

En combinant notre modèle avec le Bagging nous avons obtenu les performances suivantes :

```
bagging_cb= BaggingClassifier(base_estimator=cb, n_estimators=10)
bagging_cb.fit(x, y)
y_pred = bagging_cb.predict(x_test)
accuracy = accuracy_score(y_test, y_pred)
y_prob_test=bagging_cb.predict_proba(x_test)[:,1]
auc_test= roc_auc_score(y_test,y_prob_test)
Test_precision = precision_score(y_test,y_pred)
print('Accuracy = ' + Style.BRIGHT + str(accuracy) + Style.RESET_ALL +
      '\nAUC = ' + Style.BRIGHT + str(Test_precision) + Style.RESET_ALL +
      '\nPrécision = ' + Style.BRIGHT + str(cb_precision) + Style.RESET_ALL+'\n')
```

FIGURE 3.65 – Code Bagging du dataset2 avec 1 HbA1c

TABLE 3.16 – Performances du modèle final du dataset2 avec 1 HbA1c

Modèle final	AUC	Accuarcy	Précision
CatBoost + Bagging	0.94118	0.95	0.86667

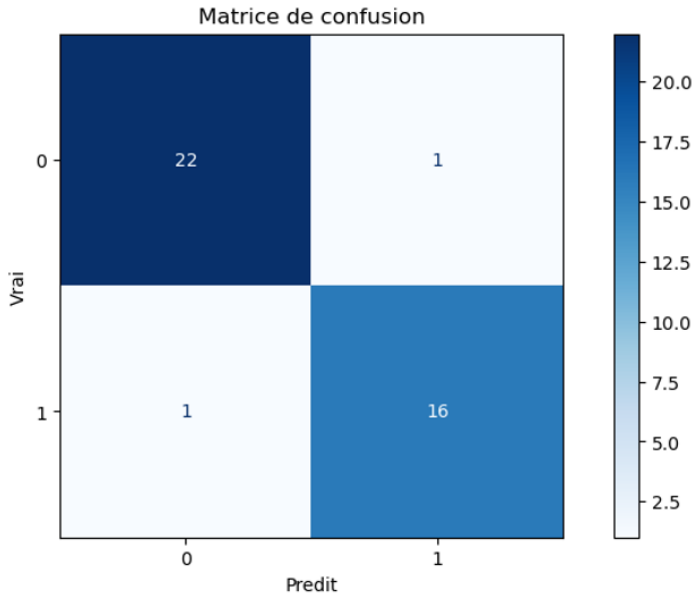

FIGURE 3.66 – Matrice de confusion du modèle final du dataset2 avec 1 HbA1c

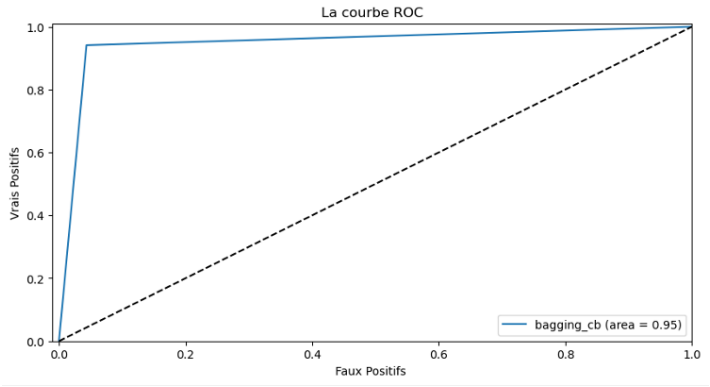

FIGURE 3.67 – Courbe ROC du modèle final du dataset2 avec 1 HbA1c

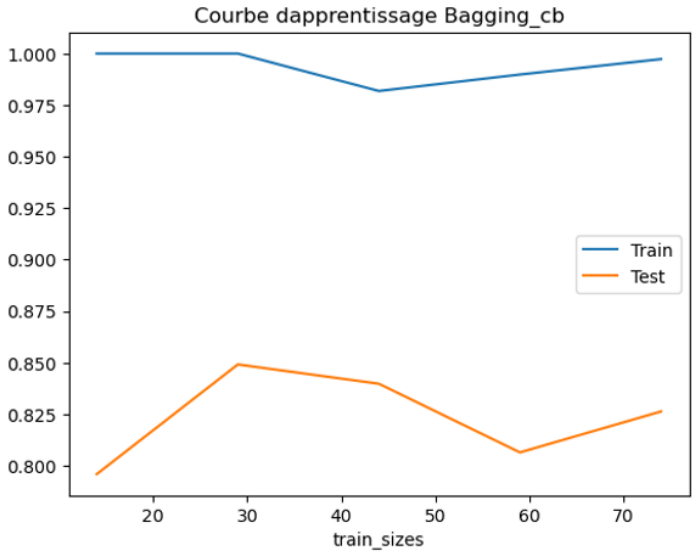

FIGURE 3.68 – Courbe d'apprentissage du modèle final du dataset2 avec 1 HbA1c

D- Quatrième étape

Voici deux exemples de validation :

Name	Sex	Age	BMI	DM_duration	FBS	A1C	LDL	HDL	TG	Neph_pathy	Ret_pathy	PVD	CDV	Ft_ulcer	Dias_BP	Real_LDL
129.0	0.0	52.0	34.5	3.0	138.0	7.33	113.0	33.0	217.0	0.0	0.0	0.0	1.0	0.0	75.0	154.81
130.0	1.0	56.0	22.0	3.0	156.0	7.16	80.0	43.0	179.0	0.0	0.0	0.0	0.0	0.0	90.0	109.60
131.0	0.0	65.0	24.0	3.0	300.0	9.67	58.0	20.0	96.0	1.0	1.0	1.0	1.0	1.0	80.0	89.90
132.0	0.0	56.0	33.0	5.0	213.0	8.74	107.0	34.0	756.0	1.0	1.0	1.0	1.0	1.0	110.0	165.85
133.0	0.0	48.0	31.0	2.0	159.0	7.45	141.0	39.0	197.0	0.0	0.0	0.0	0.0	0.0	80.0	141.00

FIGURE 3.69 – Dataset de validation du dataset2 avec 1 HbA1c

Exemple 1

```
def malade(bagging_cb, Sex=0, Age=52, BMI=34.5,  DM_duration=3.0,
           FBS=138, A1C=7.33, LDL=113, HDL=33, TG=217,
            Neph_pathy=0, PVD=0, CDV=1, Ft_ulcer=0,
            Dias_BP=75,  Real_LDL=154.81):

    x = np.array([Sex, Age, BMI,  DM_duration, FBS, A1C, LDL, HDL, TG,
                  Neph_pathy, PVD, CDV, Ft_ulcer,   Dias_BP, Real_LDL]).reshape(1, 15)
    print(bagging_cb.predict(x))
malade(bagging_cb)
```

```
[0]
```

FIGURE 3.70 – Exemple 1 de validation du dataset2 avec 1 HbA1c

La validation du premier exemple est correcte.

Exemple 2

```
def malade(bagging_cb, Sex=0, Age=65, BMI=24.0,  DM_duration=3.0,
           FBS=300, A1C=9.67, LDL=58, HDL=20, TG=96,
            Neph_pathy=1, PVD=1, CDV=1, Ft_ulcer=1,
            Dias_BP=80,  Real_LDL=89.9):

    x = np.array([Sex, Age, BMI,  DM_duration, FBS, A1C, LDL, HDL, TG,
                  Neph_pathy, PVD, CDV, Ft_ulcer,   Dias_BP, Real_LDL]).reshape(1, 15)
    print(bagging_cb.predict(x))
malade(bagging_cb)
```

```
[1]
```

FIGURE 3.71 – Exemple 2 de validation du dataset2 avec 1 HbA1c

La validation du deuxième exemple est également correcte.

Après ajout de HbA1c2

A- Première étape

Dans ce qui suit les résultats des performances des 11 méthodes :

TABLE 3.17 – Performances des 11 méthodes sur le dataset 2 avec 2 HbA1c

Modèle	Méthode d'amélioration	AUC	Accuracy	Précision
Adaboost	Grid search	0.86413	0.794872	0.9
CatBoost	Manuelle	0.870924	0.769231	1
DT	Manuelle	0.842391	0.74359	0.875
GBM	Grid search	0.86413	0.794872	0.9
HGB	Grid search + réajustement des paramètres	0.862772	0.717949	0.777778
KNN	Grid search + réajustement des paramètres	0.648098	0.666667	1
LR	Manuelle	0.804348	0.717949	0.777778
LGBM	Grid search + réajustement des paramètres	**0.875**	0.769231	0.818182
RF	Manuelle	0.842391	0.769231	1
SVM	Grid search	0.711957	0.692308	0.833333
XGBoost	Manuelle	0.794872	0.794872	0.833333

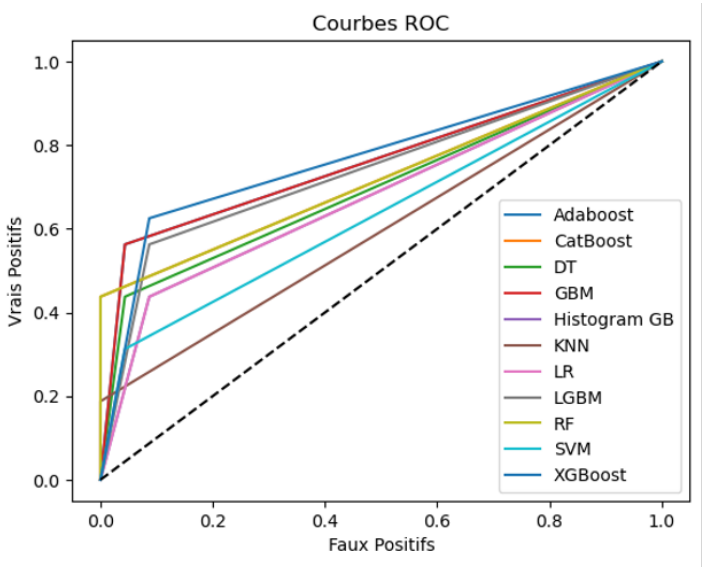

FIGURE 3.72 – Courbe ROC collective du dataset2 avec 2 HbA1c

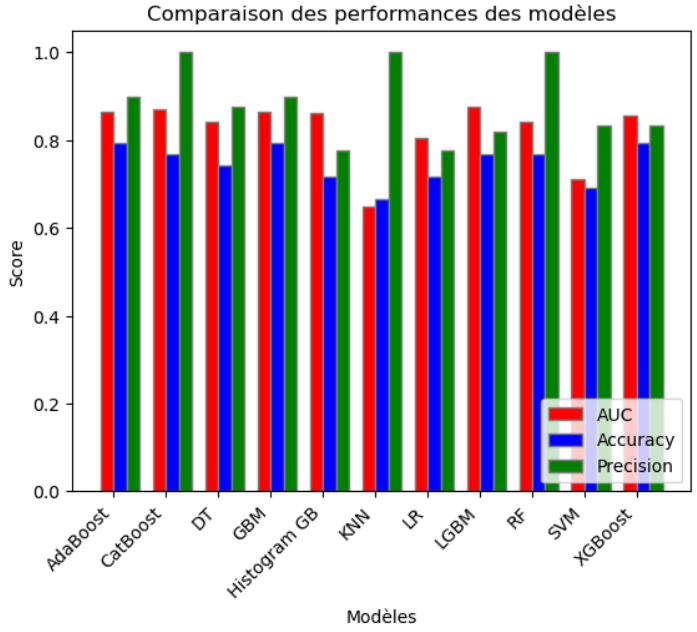

FIGURE 3.73 – Courbe des performances collectives du dataset2 avec 2 HbA1c

Les performances de tous les modèles ont augmenté et le meilleur modèle s'agit de Light Gradient Boosting. L'AUC obtenue avec ce dernier est de 87.5%. Les paramètres utilisés dans ce modèle sont :

```
lgbm=lgb.LGBMClassifier(objective='binary',
                        metric='AUC',
                        boosting_type='gbdt',
                         n_estimators=200,
                         max_bin=255,
                        num_leaves=31,
                         max_depth=4,
                        learning_rate=0.02,
                        feature_fraction=0.8,
                        bagging_fraction=0.5,
                        min_data_in_leaf= 10,
                        )
```

FIGURE 3.74 – Paramètres du meilleur modèle du dataset2 avec 2 HbA1c

TABLE 3.18 – Explication des paramètres du modèle LGBM

Paramètre	**Explication**
objective	Pour l'application de classification binaire il prend la valeur 'binary' [76]
metric	Sert à définir la métrique d'évaluation [76]
boosting_type	Sa valeur par défaut est 'gbdt ' qui est le diminutif de Gradient Boosting Decision Tree [76]
n_estimators	Nombre d'itérations de boosting [76]
max_bin	Nombre maximal de Bin pour chaque feature [76]
num_leaves	Représente nombre maximum de feuilles dans un arbre, c'est le paramètre principal pour contrôler la complexité du modèle d'arbre [76]
max_depth	La profondeur maximale de l'arbre [76]
learning_rate	Vitesse d'apprentissage [76]
feature_fraction	Peut être utilisé pour accélérer la formation faire face au surapprentissage [76]
bagging_fraction	Comme feature_fraction, il peut être utilisé pour accélérer la formation faire face au surapprentissage [76]
min_data_in_leaf	Il permet d'éviter un ajustement excessif dans un arbre feuille par feuille [76]

Dans ce qui suit les résultats et les courbe du modèle Light Gradient Boosting.

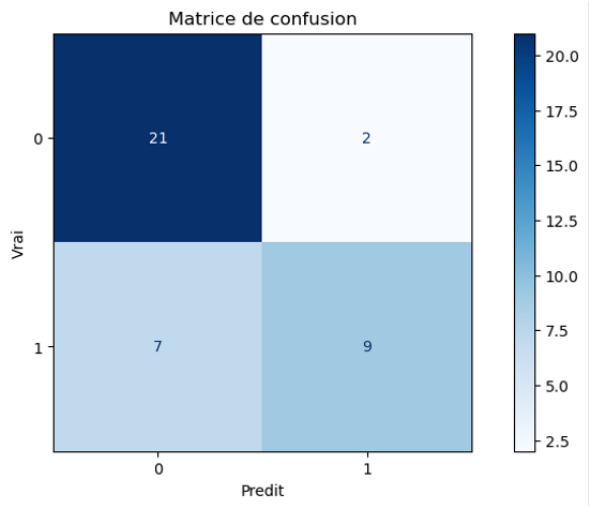

FIGURE 3.75 – Matrice de confusion du meilleur modèle du dataset2 avec 2 HbA1c

	precision	recall	f1-score	support
0	0.75	0.91	0.82	23
1	0.82	0.56	0.67	16
accuracy			0.77	39
macro avg	0.78	0.74	0.75	39
weighted avg	0.78	0.77	0.76	39

FIGURE 3.76 – Rapport de classification du meilleur modèle du dataset2 avec 2 HbA1c

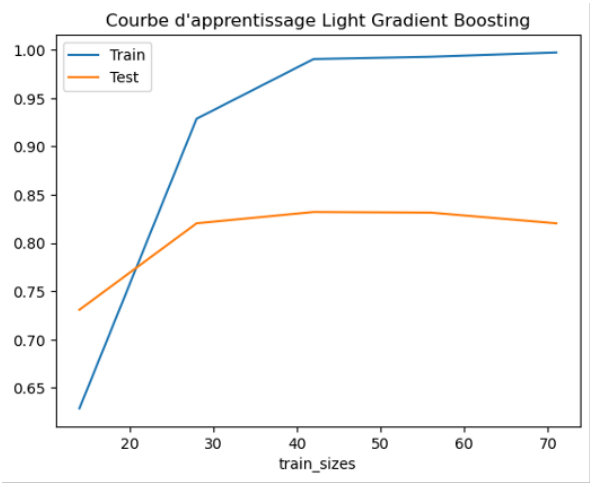

FIGURE 3.77 – Courbe d'apprentissage du meilleur modèle du dataset2 avec 2 HbA1c

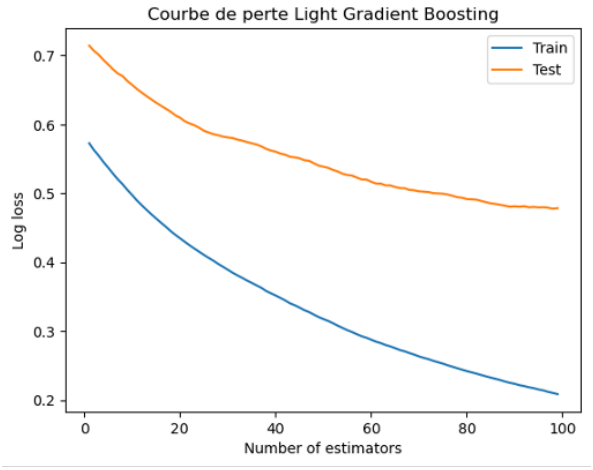

FIGURE 3.78 – Courbe perte du meilleur modèle du dataset2 avec 2 HbA1c

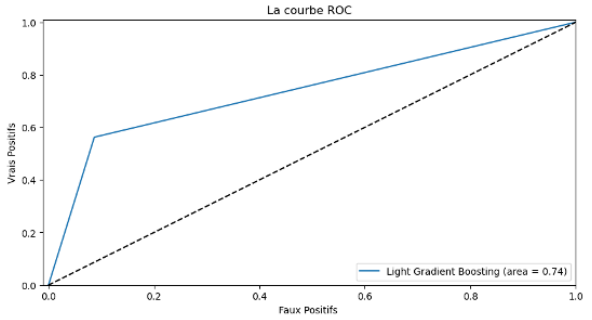

FIGURE 3.79 – Courbe ROC du meilleur modèle du dataset2 avec 2 HbA1c

B- Deuxième étape

D'après les graphes d'importances de Light Gradient Boosting, AdaBoost et CatBoost, les caractéristiques à éliminer sont : "Dose_Í2_if_BID", "DM_type, Cum_A_Equ", "CDV", "Sex","Neu_pathy" et "Real_LDL". Les résultats sont affichés dans les figures et le tableau suivants.

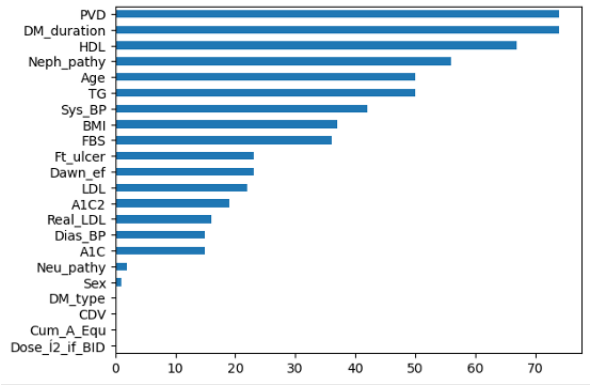

FIGURE 3.80 – Graphe d'importance du modèle LGBM du dataset2 avec 2 HbA1c

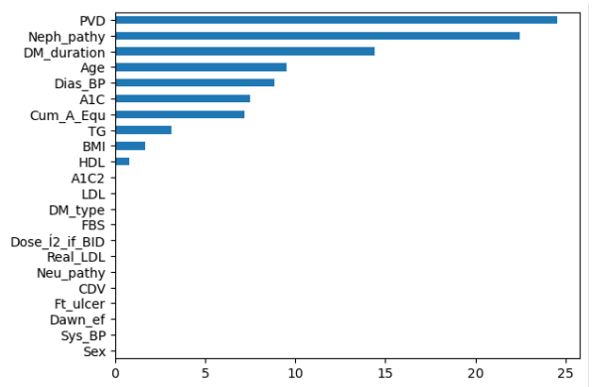

FIGURE 3.81 – Graphe d'importance du modèle CatBoost du dataset2 avec 2 HbA1c

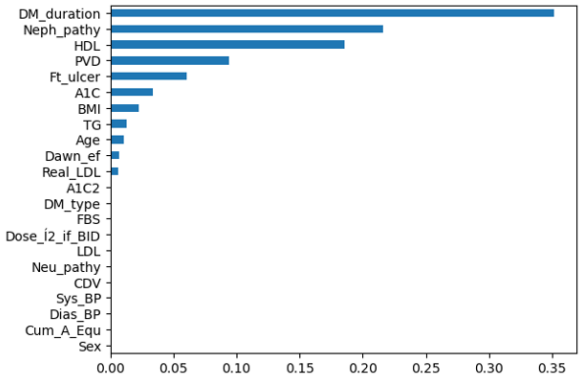

FIGURE 3.82 – Graphe d'importance du modèle AdaBoost du dataset2 avec 2 HbA1c

TABLE 3.19 – Performances du modèle optimisé du dataset2 avec 2 HbA1c

Modèle optimisé	**AUC**	**Accuarcy**	**Précision**
Light Gradient Boosting	0.88587	0.76923	0.76923

C- Troisième étape

Après la combinaison du modèle LGBM étant le meilleur et la technique du Bagging, nos résultats sont les suivants :

```
bagging_lgbm = BaggingClassifier(base_estimator=lgbm, n_estimators=10)
bagging_lgbm.fit(x, y)
y_pred = bagging_lgbm.predict(x_test)
accuracy = accuracy_score(y_test, y_pred)
y_prob_test=bagging_lgbm.predict_proba(x_test)[:,1]
auc_test= roc_auc_score(y_test,y_prob_test)
Test_precision = precision_score(y_test,y_pred)
print('Accuracy = ' + Style.BRIGHT + str(accuracy) + Style.RESET_ALL +
      '\nAUC = ' + Style.BRIGHT + str(Test_precision) + Style.RESET_ALL +
      '\nPrécision = ' + Style.BRIGHT + str(lgbm_precision) + Style.RESET_ALL+'\n')
```

FIGURE 3.83 – Code Bagging du dataset2 avec 2 HbA1c

TABLE 3.20 – Performances du modèle final du dataset2 avec 2 HbA1c

Modèle final	**AUC**	**Accuarcy**	**Précision**
Light Gradient Boosting + Bagging	1.0	0.89744	0.76923

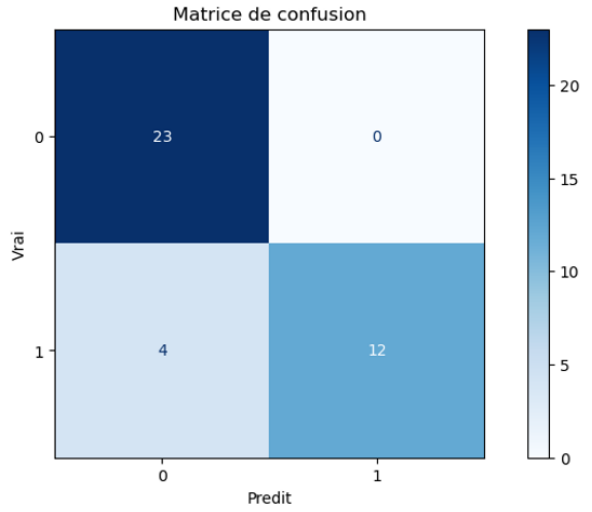

FIGURE 3.84 – Matrice de confusion du modèle final du dataset2 avec 2 HbA1c

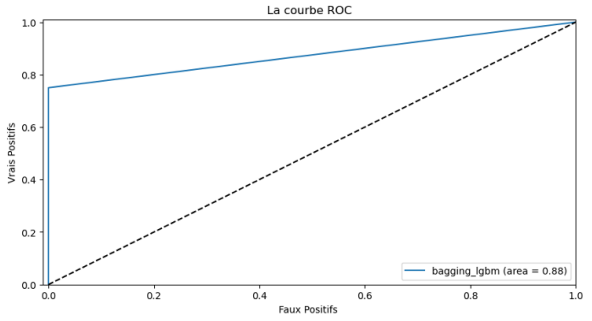

FIGURE 3.85 – Courbe ROC du modèle final du dataset2 avec 2 HbA1c

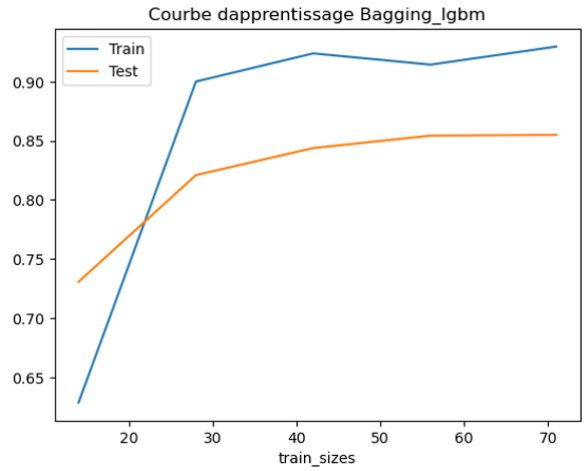

FIGURE 3.86 – Courbe d'apprentissage du modèle final du dataset2 avec 2 HbA1c

D- Quatrième étape

Name	Age	BMI	DM_duration	FBS	A1C	LDL	HDL	TG	Neph_pathy	Ret_pathy	PVD	Ft_ulcer	Dawn_ef	Sys_BP	Dias_BP	Real_LDL	A1C2
129	52	34.5	3	138	7.33	113	33	217	0	0	0	0	1	115	75	154.81	8.063
130	56	22.0	3	156	7.16	80	43	179	0	0	0	0	1	130	90	109.60	7.017
131	65	24.0	3	300	9.67	58	20	96	1	1	1	1	0	120	80	89.90	8.509
132	56	33.0	5	213	8.74	107	34	756	1	1	1	1	1	155	110	165.85	8.827
133	48	31.0	2	159	7.45	141	39	197	0	0	0	0	0	120	80	141.00	7.748

FIGURE 3.87 – Dataset de validation sur le dataset 2 avec 2HbA1c

Exemple 1

```
def malade(bagging_lgbm,  Age=52, BMI=34.0,  DM_duration=3.0,
           FBS=138, A1C=7.33, LDL=113, HDL=33, TG=217,
            Neph_pathy=0, PVD=0,  Ft_ulcer=0, Dawn_ef=1,
           Sys_BP=115, Dias_BP=75,A1C2=8.063):

    x = np.array([ Age, BMI,  DM_duration, FBS, A1C, LDL, HDL, TG,
                  Neph_pathy, PVD, Ft_ulcer, Dawn_ef, Sys_BP, Dias_BP ,A1C2]).reshape(1, 15)
    print(bagging_lgbm.predict(x))
malade(bagging_lgbm)
```

```
[0]
```

FIGURE 3.88 – Exemple 1 de validation du dataset2 avec 2 HbA1c

Le patient du premier exemple a été correctement validé.
Exemple 2

```
def malade(bagging_lgbm,  Age=56, BMI=33.0,  DM_duration=5.0,
           FBS=213, A1C=8.74, LDL=107, HDL=34, TG=756,
            Neph_pathy=1, PVD=1,  Ft_ulcer=1, Dawn_ef=1,
           Sys_BP=155, Dias_BP=80,  A1C2=7.748):

    x = np.array([ Age, BMI,  DM_duration, FBS, A1C, LDL, HDL, TG,
                  Neph_pathy, PVD, Ft_ulcer, Dawn_ef, Sys_BP, Dias_BP ,A1C2]).reshape(1, 15)
    print(bagging_lgbm.predict(x))
malade(bagging_lgbm)
```

```
[1]
```

FIGURE 3.89 – Exemple 1 de validation du dataset2 avec 2 HbA1c

La validation du deuxième patient est également correcte.

Comparaison

Le tableau suivant résume les performances des modèles appliqués au dataset2 avec un et deux HbA1c.

TABLE 3.21 – Table de comparaison du dataset2 avec un et deux HbA1c

	Performance	**1 HbA1c**	**2 HbA1c**
Nom du modèle	/	CatBoost	Light Gradient Boosting
Modèle de base	**AUC**	0.861	0.875
	ACC	0.795	0.769
	Précison	0.9	0.82
Modèle optimisé	**AUC**	0.873	0.886
	ACC	0.85	0.769
	Précison	0.867	0.769
Modèle final	**AUC**	0.94118	1.0
	ACC	0.95	0.89744
	Précision	0.86666	0.76923

3.5.3 Dataset 3

Avant ajout de HbA1c2

A-Première étape

Les résultats obtenus sont les suivants :

TABLE 3.22 – Les performances des 11 méthodes sur le dataset 3 avec 1 HbA1c

Modèle	**Méthode d'amélioration**	**AUC**	**Accuracy**	**Précision**
Adaboost	Grid search + réajustement des paramètres	0.8125	0.75	0.571429
CatBoost	Manuelle	0.84375	0.666667	0.5
DT	Grid search	0.734375	0.75	0.6
GBM	Grid search + réajustement des paramètres	0.84375	0.666667	0.5
HGB	Manuelle	**0.90625**	0.666667	0.5
KNN	Grid search	0.625	0.666667	0.5
LR	Grid search + réajustement des paramètres	0.84375	0.75	0.571429
LGBM	Grid search	0.875	0.75	0.571429
RF	Grid search	0.875	0.75	0.571429
SVM	Grid search + réajustement des paramètres	0.8125	0.583333	0.444444
XGBoost	Grid search + réajustement des paramètres	0.875	0.75	0.571429

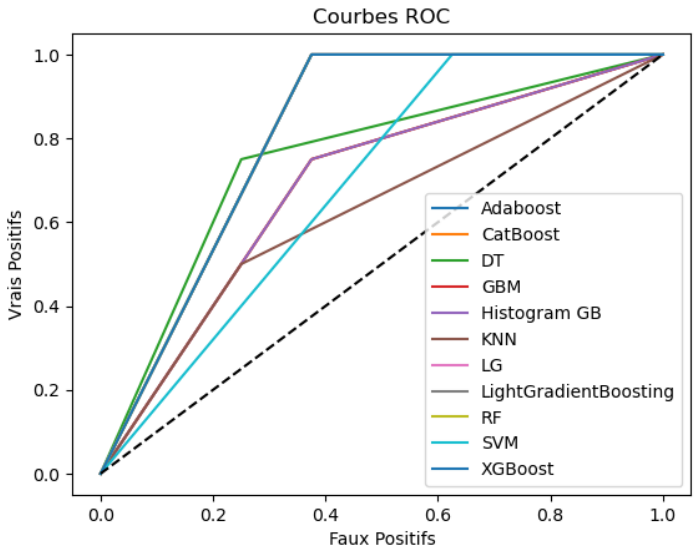

FIGURE 3.90 – Courbe ROC collective dataset3 avec 1 HbA1c

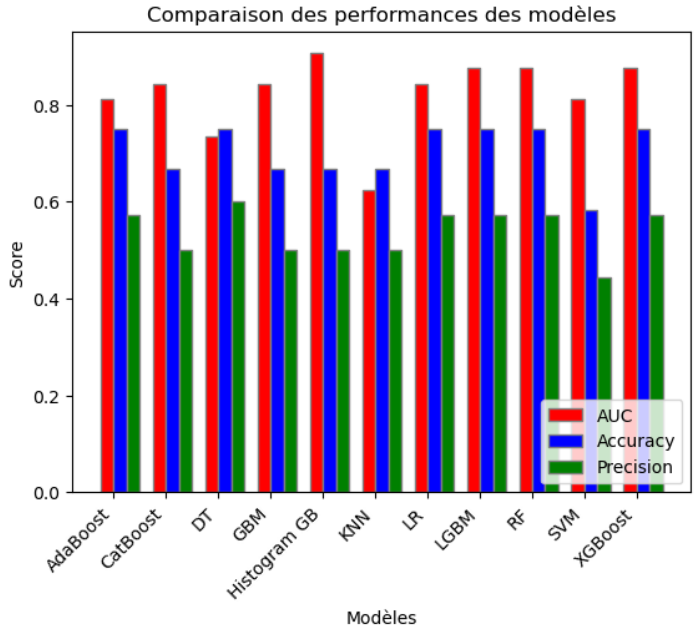

FIGURE 3.91 – Courbe des performances collectives du dataset3 avec 1HbA1c

Le meilleur modèle s'agit de Histogram Gradient Boosting avec une AUC de 90.625%. La performance de ce modèle a été obtenue en utilisant la combinaison des paramètres suivants. Voir figure 3.92 et la Table 3.23.

```
hgb = HistGradientBoostingClassifier(
    learning_rate=0.1,
    max_iter=200,
    max_depth=5,
    min_samples_leaf=8,
    l2_regularization=0.1,
    verbose=1
)
```

FIGURE 3.92 – Paramètres du meilleur modèle du dataset3 avec 1 HbA1c

TABLE 3.23 – Explication des paramètres du modèle HGB

Paramètre	**Explication**
learning_rate	Le taux d'apprentissage, il est considéré comme facteur multiplicatif pour les valeurs des feuilles [77]
max_iter	Représente le nombre maximal d'arbre. Par défaut sa valeur est 100 [77]
max_depth	Il s'agit de la profondeur maximal des arbres [110] [77]
min_samples_leaf	C'est le nombre minimum d'échantillons par feuille [77]
l2_regularization	Le paramètre de régularisation L2 [77]
verbose	Par défaut sa valeur est 0, Représente le niveau de verbosité [77]

Dans ce qui suit toutes les courbes et les résultats affichées pour le meilleur modèle :

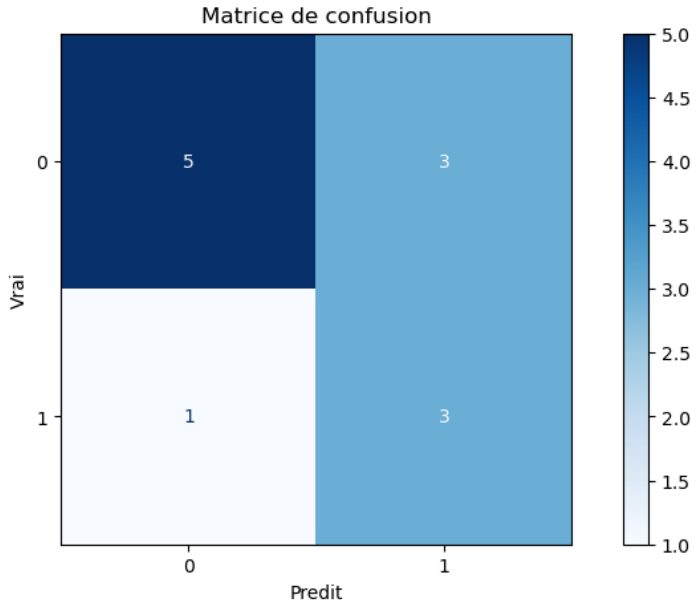

FIGURE 3.93 – Matrice de confusion du meilleur modèle du dataset3 avec 1 HbA1c

	precision	recall	f1-score	support
0	0.83	0.62	0.71	8
1	0.50	0.75	0.60	4
accuracy			0.67	12
macro avg	0.67	0.69	0.66	12
weighted avg	0.72	0.67	0.68	12

FIGURE 3.94 – Rapport de classification du meilleur modèle du dataset3 avec 1 HbA1c

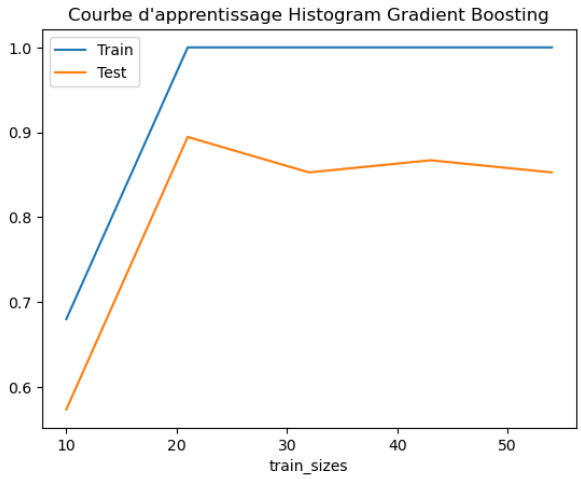

FIGURE 3.95 – Courbe d'apprentissage du meilleur modèle du dataset3 avec 1 HbA1c

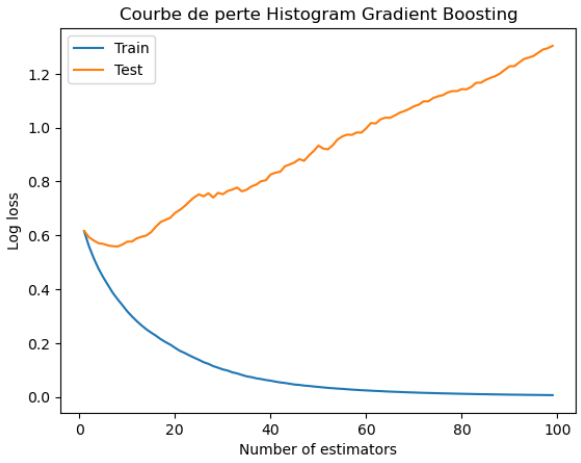

FIGURE 3.96 – Courbe perte du meilleur modèle du dataset3 avec 1 HbA1c

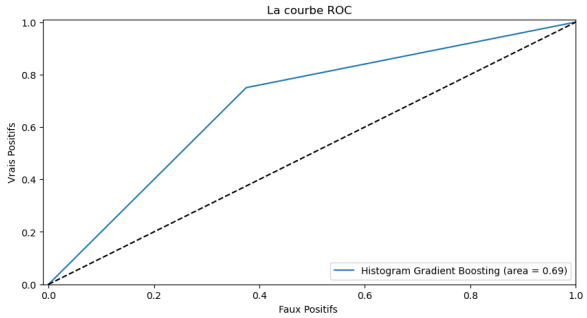

FIGURE 3.97 – Courbe ROC du meilleur modèle du dataset3 avec 1 HbA1c.

B-Deuxième étape

Afin d'identifier les caractéristiques les moins influentes, nous nous sommes basés sur les trois meilleurs modèles à savoir : Histogram Gradient Boosting, Light Gradient Boosting et Xtreme Gradient Boosting dont les graphes d'importances sont les suivants :

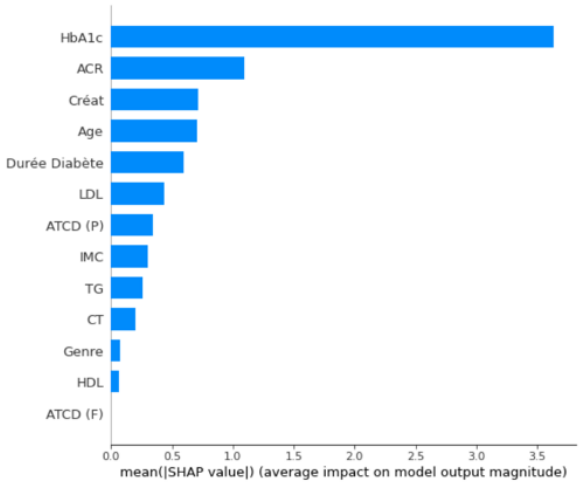

FIGURE 3.98 – Graphe SHAP du modèle HGB du dataset3 avec 1 HbA1c

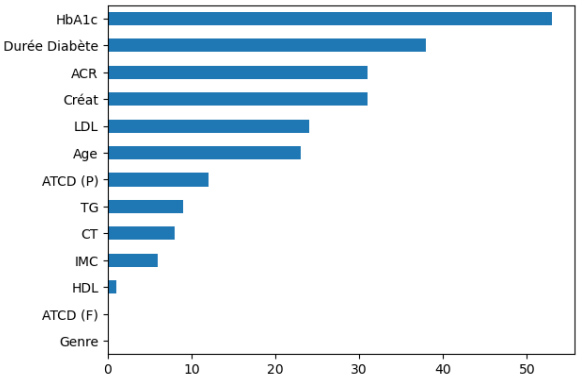

FIGURE 3.99 – Graphe d'importance du modèle LGB du dataset3 avec 1 HbA1c

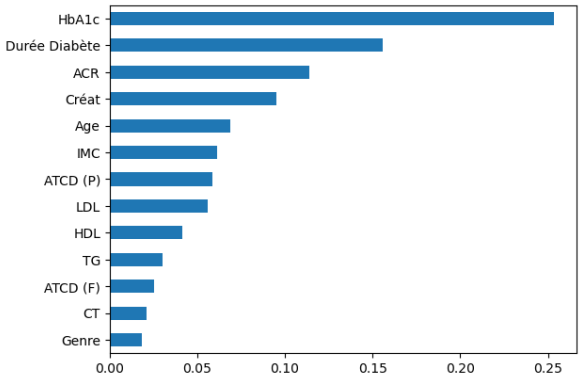

FIGURE 3.100 – Graphe d'importance du modèle XGBoost du dataset3 avec 1 HbA1c

Nous remarquons que les caractéristiques : "Genre" et "ATCD (F)" sont celles qui se répètent le plus comme caractéristiques non influentes dans les trois meilleurs modèles. Par conséquent, elles ont été supprimées dans le meilleur modèle c'est à dire Histogram Gradient Boosting. La table suivante montre les résultats obtenus.

TABLE 3.24 – Performances du modèle optimisé du dataset3 avec 1 HbA1c

Modèle optimisé	AUC	Accuarcy	Précision
HGB	0.90625	0.666667	0.5

On ne remarque aucun changement et cela est dû à la cohérence des caractéristiques de ce dataset.

C-Troisième étape

Notre modèle est désormais optimisé et les facteurs de risques sont identifiés. Nous passons alors à la construction d'un nouveau modèle en le combinant avec la technique du Bagging comme suit :

```
bagging_hgb = BaggingClassifier(base_estimator=hgb1, n_estimators=10)
bagging_hgb.fit(x, y)
y_pred = bagging_hgb.predict(x_test)
accuracy = accuracy_score(y_test, y_pred)
y_prob_test=bagging_hgb.predict_proba(x_test)[:,1]
auc_test= roc_auc_score(y_test,y_prob_test)
Test_precision = precision_score(y_test,y_pred)
print('Accuracy = ' + Style.BRIGHT + str(accuracy) + Style.RESET_ALL +
      '\nAUC = ' + Style.BRIGHT + str(Test_precision) + Style.RESET_ALL +
      '\nPrécision = ' + Style.BRIGHT + str(cb_precision) + Style.RESET_ALL+'\n')
```

FIGURE 3.101 – Code Bagging du dataset3 avec 1 HbA1c

Voici les nouvelles performances de notre modèle :

TABLE 3.25 – Performances du modèle final du dataset3 avec 1 HbA1c

Modèle final	AUC	Accuarcy	Précision
HGB + Bagging	1.0	1.0	0.5

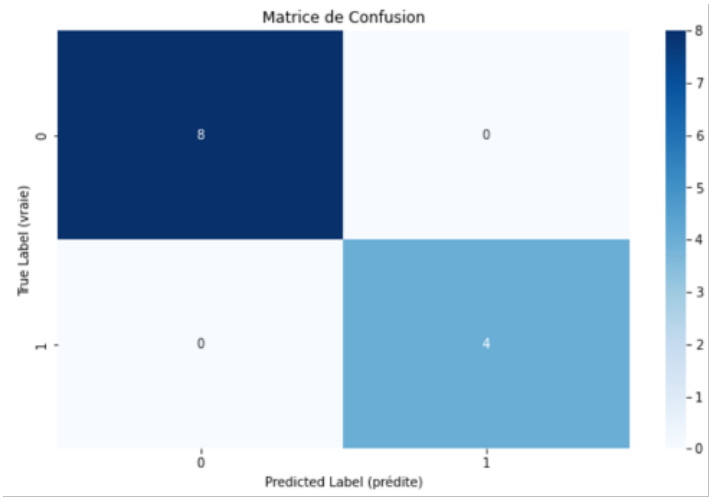

FIGURE 3.102 – Matrice de confusion du modèle final du dataset3 avec 1 HbA1c

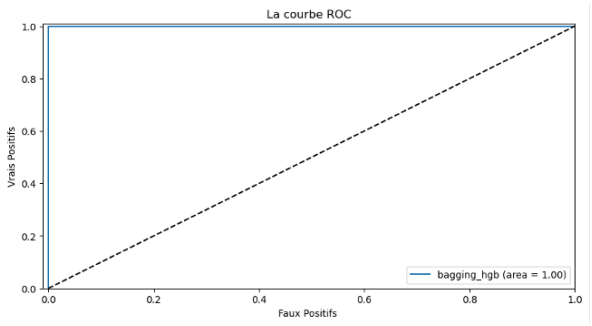

FIGURE 3.103 – Courbe ROC du modèle final du dataset3 avec 1 HbA1c

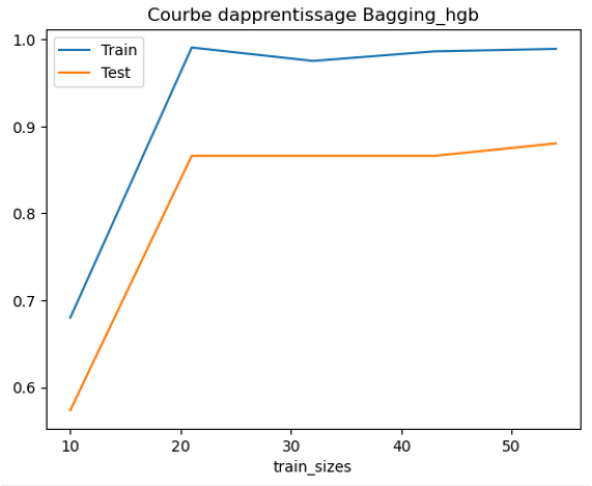

FIGURE 3.104 – Courbe d'apprentissage du modèle final du dataset3 avec 1 HbA1c

D- Quatrième étape

	Age	IMC	Durée	ATCD_P	HbA1c	CT	HDL	LDL	TG	Créat	ACR	Rétinopathie
0	56	35.42	11	1	6.9	1.41	0.28	0.77	1.81	8.30	20.00	1
1	71	20.96	28	0	9.1	1.74	0.47	1.10	0.86	10.49	31.00	1
2	38	23.18	1	0	5.9	1.86	0.26	1.28	1.60	10.38	10.00	0
3	59	22.27	6	0	6.7	1.63	0.42	1.04	0.85	8.79	4.39	0

FIGURE 3.105 – Dataset de validation du dataset3 avec 1 HbA1c

Exemple 1

```
def malade (bagging_hgb,Age=71,IMC=20.96,Durée_Diabète=28,ATCD_P=0,HbA1c=9.1,CT=1.74,HDL=0.47,LDL=1.10,TG=0.86,
           Créat=10.49,ACR=31.00):
    x= np.array([Age,IMC,Durée_Diabète,ATCD_P,HbA1c,CT,HDL,LDL,TG,Créat,ACR]).reshape(1,11)
    print(bagging_hgb.predict(x))
malade(bagging_hgb)
```

```
[1]
```

FIGURE 3.106 – Exemple 1 de validation du dataset3 avec 1 HbA1c

Le résultat de validation est correct.

Exemple 2

```
def malade (bagging_hgb,Age=38,IMC=23.18,Durée_Diabète=1,ATCD_P=0,HbA1c=5.9,CT=1.86,HDL=0.26,LDL=1.28,TG=1.60,
           Créat=10.38,ACR=10.00):
    x= np.array([Age,IMC,Durée_Diabète,ATCD_P,HbA1c,CT,HDL,LDL,TG,Créat,ACR]).reshape(1,11)
    print(bagging_hgb.predict(x))
malade(bagging_hgb)
```

```
[0]
```

FIGURE 3.107 – Exemple 2 de validation du dataset3 avec 1 HbA1c

La validation est également correcte.

Après ajout de HbA1c2

A- Première étape

Le tableau suivant montre les résultats obtenus par les 11 méthodes :

TABLE 3.26 – Performances des 11 méthodes sur le dataset 3 avec 2 HbA1c

Modèle	Méthode d'amélioration	AUC	Accuracy	Précision
Adaboost	Grid search + réajustement des paramètres	**0.9375**	0.833333	0.666667
CatBoost	Manuelle	0.90625	0.666667	0.5
DT	Bayes search	0.75	0.75	0.6
GBM	Grid search	0.9375	0.75	0.571429
HGB	Gp_minimize	0.90625	0.75	0.6
KNN	Bayes search + réajustement des paramètres	0.8125	0.833333	1
LR	Grid search + réajustement des paramètres	0.875	0.833333	0.666667
LGBM	Grid search + réajustement des paramètres	0.90625	0.833333	0.666667
RF	Bayes search + réajustement des paramètres	0.9375	0.75	0.571429
SVM	Grid search + réajustement des paramètres	0.8125	0.666667	0.5
XGBoost	Grid search + réajustement des paramètres	0.90625	0.75	0.571429

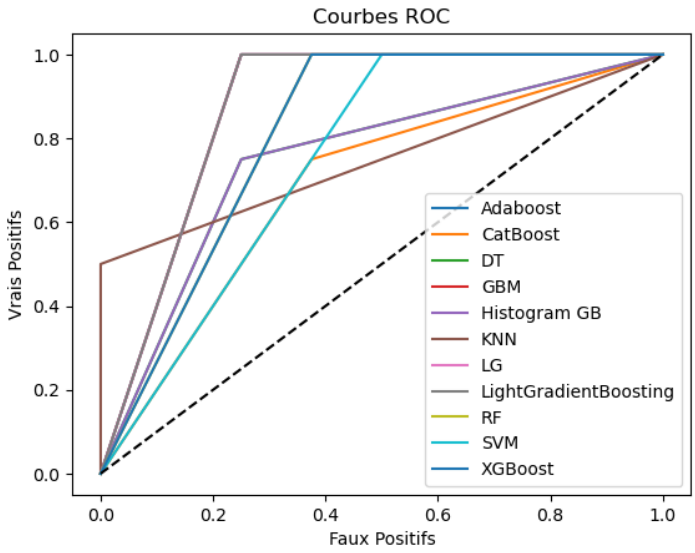

FIGURE 3.108 – Courbe ROC collective du dataset3 avec 2 HbA1c

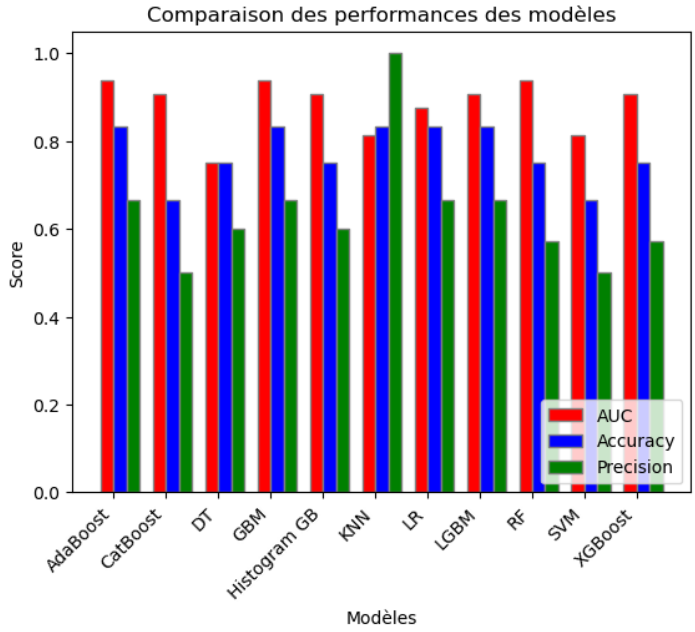

FIGURE 3.109 – Courbe des performances collectives du dataset3 avec 2 HbA1c

En utilisant deux valeurs de HbA1c, on remarque que les performances de tous les modèles se sont améliorées. Le meilleur modèle est Adaboost avec une valeur AUC de 93.75%. Les paramètres utilisés dans ce modèle sont illustrées dans la figure 3.110 et Table 3.27.

```
ab=AdaBoostClassifier(base_estimator=None,
                      n_estimators=100,
                      learning_rate=1,
                      algorithm='SAMME.R',
                      random_state=42)
```

FIGURE 3.110 – Paramètres du meilleur modèle du dataset3 avec 2 HbA1c

TABLE 3.27 – Explication des paramètres du modèle Adaboost

Paramètre	**Explication**
base_estimator	L'estimateur de base à partir duquel l'ensemble est construit. Par defaut il prend la valeur None [78].
n_estimators	L'estimateur de base à partir duquel l'ensemble est construit. Par defaut il prend la valeur None [78]
learning_rate	Traduit comme le poids appliqué pour le classifieur à chaque itération du boosting [78]
algorithm	Il sert à spécifier l'algorithme utilisé pour l'entrainement il peut prendre la valeur ' SAMME' ou 'SAMME.R' [78]
random_state	Détermine une valeur aléatoire attribuée aux estimateurs à chaque itération [78]

Les courbes et les résultats du modèle Adaboost sont affichés dans ce qui suit :

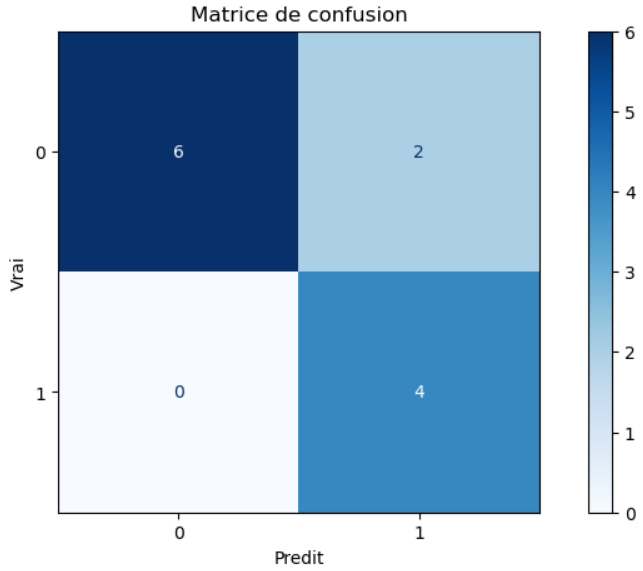

FIGURE 3.111 – Matrice de confusion meilleur modèle du dataset3 avec 2 HbA1c

	precision	recall	f1-score	support
0	1.00	0.75	0.86	8
1	0.67	1.00	0.80	4
accuracy			0.83	12
macro avg	0.83	0.88	0.83	12
weighted avg	0.89	0.83	0.84	12

FIGURE 3.112 – Rapport de classification du meilleur modèle du dataset3 avec 2 HbA1c

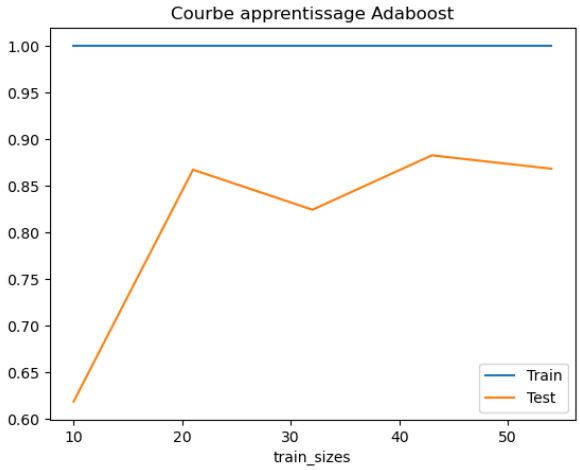

FIGURE 3.113 – Courbe d'apprentissage du meilleur modèle du dataset3 avec 2 HbA1c

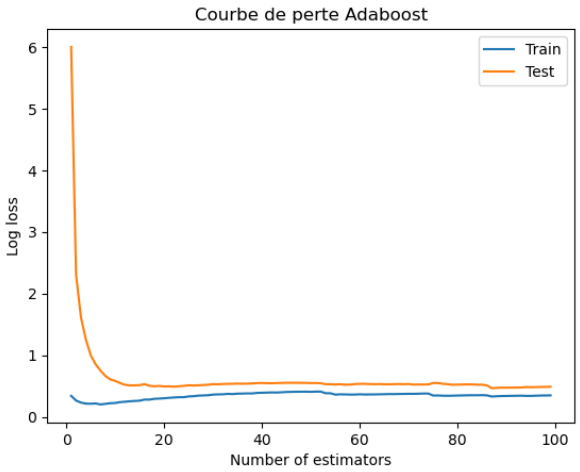

FIGURE 3.114 – Courbe perte du meilleur modèle du dataset3 avec 2 HbA1c

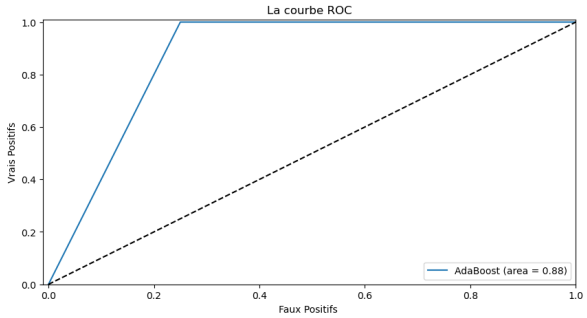

FIGURE 3.115 – Courbe ROC du meilleur modèle du dataset3 avec 2 HbA1c

B- Deuxième étape

Nous avons affiché les graphes d'importances des meilleurs modèles : Adaboost, Gradient boosting et Random Forest comme illustré dans les figures 3.116, 3.117 et 3.118. Ces graphes indiquent que les caractéristiques : Genre et ATCD (P) sont les moins influentes. Par conséquent, elles ont été supprimées dans le meilleur modèle. Les résultats sont présentés dans la table 3.28.

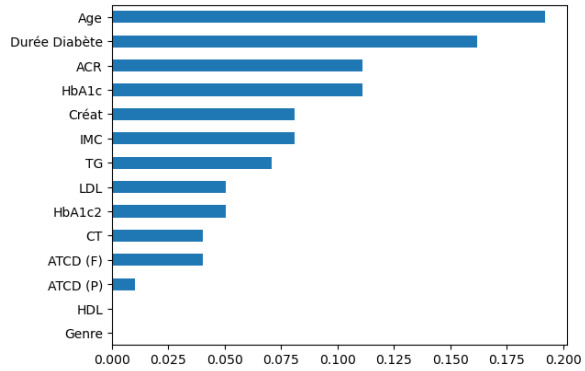

FIGURE 3.116 – Graphe d'importance du modèle Adaboost du dataset3 avec 2 HbA1c

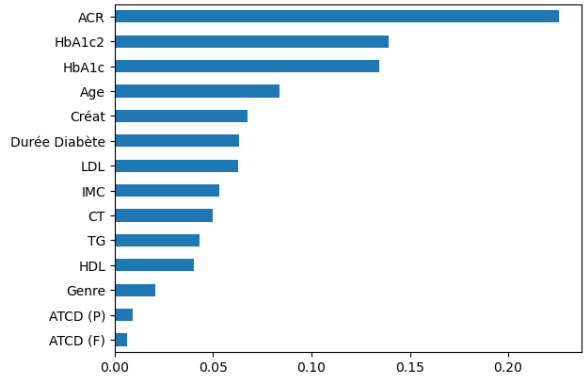

FIGURE 3.117 – Graphe d'importance du modèle GBM du dataset3 avec 2 HbA1c

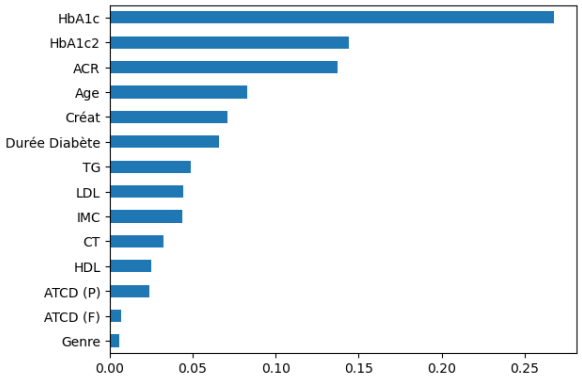

FIGURE 3.118 – Graphe d'importance du modèle RF du dataset3 avec 2 HbA1c

TABLE 3.28 – Performances du modèle optimisé du dataset1 avec 2 HbA1c

Modèle optimisé	AUC	Accuarcy	Précision
Adaboost	0.9375	0.83334	0.66667

C- Troisième étape

De même qu'avec une seule valeur de HbA1c, nous entamons la construction du nouveau modèle en utilisant le meilleur modèle optimisé, les facteurs de risques et la technique du Bagging. Voici le modèle final et les résultats obtenus :

```
bagging_ab = BaggingClassifier(base_estimator=ab1, n_estimators=10)
bagging_ab.fit(x, y)
y_pred = bagging_ab.predict(x_test)
accuracy = accuracy_score(y_test, y_pred)
y_prob_test=bagging_ab.predict_proba(x_test)[:,1]
auc_test= roc_auc_score(y_test,y_prob_test)
Test_precision = precision_score(y_test,y_pred)
print('Accuracy = ' + Style.BRIGHT + str(accuracy) + Style.RESET_ALL +
      '\nAUC = ' + Style.BRIGHT + str(Test_precision) + Style.RESET_ALL +
      '\nPrécision = ' + Style.BRIGHT + str(cb_precision) + Style.RESET_ALL+'\n')
```

FIGURE 3.119 – Code Bagging du dataset3 avec 2 HbA1c

TABLE 3.29 – Performances du modèle final du dataset1 avec 2 HbA1c

Modèle final	AUC	Accuarcy	Précision
CatBoost	1.0	1.0	0.6

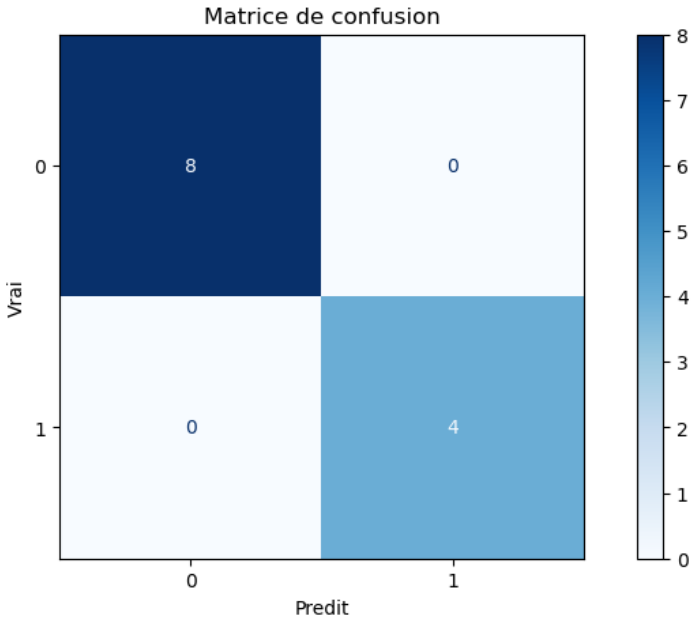

FIGURE 3.120 – Matrice de confusion du modèle final du dataset3 avec 2 HbA1c

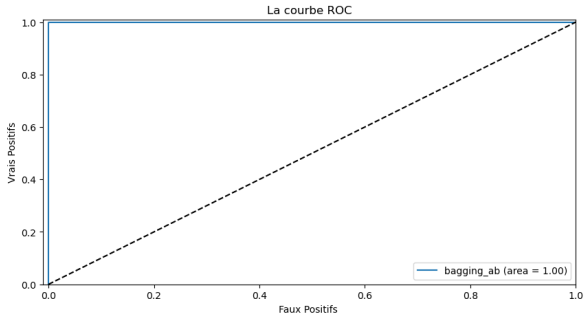

FIGURE 3.121 – Courbe ROC du modèle final du dataset3 avec 2 HbA1c

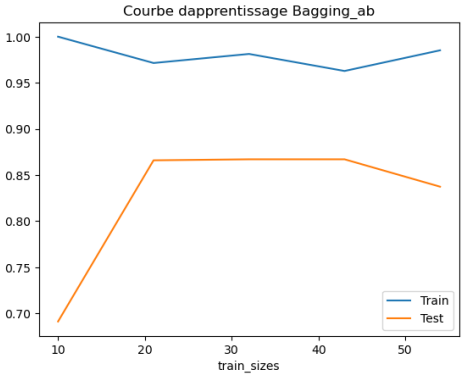

FIGURE 3.122 – Courbe d'apprentissage du modèle final du dataset3 avec 2 HbA1c

D- Quatrième étape

Dans ce qui suit deux exemples de validation :

	Age	IMC	Durée	ATCD_F	HbA1c	HbA1c2	CT	HDL	LDL	TG	Créat	ACR	Rétinopathie
0	56	35.42	11	1	6.9	6.7	1.41	0.28	0.77	1.81	8.30	20.00	1
1	71	20.96	28	1	9.1	8.9	1.74	0.47	1.10	0.86	10.49	31.00	1
2	38	23.18	1	0	5.9	6.3	1.86	0.26	1.28	1.60	10.38	10.00	0
3	59	22.27	6	0	6.7	6.2	1.63	0.42	1.04	0.85	8.79	4.39	0

FIGURE 3.123 – Dataset de validation du dataset3 avec 2 HbA1c

Exemple 1

```
def malade(bagging_ab, Age=38,IMC=23.18,Durée=1, ATCD_F=0, HbA1c=5.9,HbA1c2=6.3,CT=1.86,
           HDL=0.26,LDL=1.28,TG=1.6,Créat=10.38,ACR=0):
    x = np.array([Age,IMC,Durée, ATCD_F, HbA1c,HbA1c2,CT,HDL,LDL,TG,Créat,ACR]).reshape(1, 12)
    print(bagging_ab.predict(x))
malade(bagging_ab)
```

```
[0]
```

FIGURE 3.124 – Exemple 1 de validation du dataset3 avec 2 HbA1c

Cet exemple a été validé correctement.
Exemple 2

```
def malade(bagging_ab, Age=71,IMC=20.96,Durée=28, ATCD_F=1, HbA1c=9.1,HbA1c2=8.9,CT=1.74,
           HDL=0.47,LDL=1.10,TG=0.86,Créat=10.49,ACR=31):
    x = np.array([Age,IMC,Durée, ATCD_F, HbA1c,HbA1c2,CT,HDL,LDL,TG,Créat,ACR]).reshape(1, 12)
    print(bagging_ab.predict(x))
malade(bagging_ab)
```

```
[1]
```

FIGURE 3.125 – Exemple 2 de validation du dataset3 avec 2 HbA1c

De même que le premier, cet exemple a été également validé correctement.

Comparaison

Le tableau suivant résume les performances des modèles appliqués au dataset3 avec un et deux HbA1c.

TABLE 3.30 – Table de comparaison du dataset3 avec un et deux HbA1c

	Performance	**1 HbA1c**	**2 HbA1c**
Nom du modèle	/	Histogram Gradient Boosting	AdaBoost
Modèle de base	**AUC**	0.906	0.937
	ACC	0.667	0.834
	Précison	0.5	0.667
Modèle optimisé	**AUC**	0.906	0.937
	ACC	0.667	0.834
	Précison	0.5	0.667
Modèle final	**AUC**	1.0	1.0
	ACC	1.0	1.0
	Précision	0.5	0.6

3.5.4 Comparaison des trois datasets avec 1 seul HbA1c

En ce qui concerne le meilleur modèle prédictif, dans cette étude, les résultats de comparaison ont montré que le modèle CatBoost appliqué au dataset1 a obtenu la valeur la plus élevée de l'AUC (AUC=0,615), dépassant ainsi les autres méthodes.

De même pour le dataset2, CatBoost reste le plus performant (AUC=0.855) suivi de très près par le modèle LGBM avec une AUC de 0.853.

Quant au dataset3, Histogram Gradient Boosting est le meilleur modèle avec une AUC de 0.906.

3.5.5 Comparaison des trois datasets avec 2 HbA1c

Après l'ajout de la deuxième valeur de HbA1c, les résultats ont remarquablement augmenté.

Le meilleur modèle du dataset1 est toujours Catboost mais cette fois avec une **AUC** de **0.989**.

Quant au dataset2 CatBoost a été dépassé de très près par LGBM avec **AUC =0.875**.

En ce qui est du dataset3, AdaBoost a atteint une valeur de **AUC** de **0.937**, **accuracy** de **0.833** et une **précision** de **0.667** faisant de lui le meilleur modèle de ce dataset, suivi par RF et GBM avec la même AUC mais des valeurs d'accuracy et de précision moins élevée.

3.5.6 Comparaison de nos modèles finaux avec les travaux antérieurs

Le tableau suivant présente une comparaison des performances de nos modèles par rapport aux études antérieures.

Table 3.31 – Table de comparaison de nos modèles finaux avec les études antérieures

Étude		**Modèle**	**AUC**	**ACC**
Hsin-Yi Tsao et al. 2018 [55]		SVM	0.795	0.839
Mo et al. 2020 [56]		MLR	0.700	0.796
Zun Shen et al. 2021 [57]		Sel-Stacking	/	0.839
Wanyue Li et al. 2021 [58]		XGBoost	0.90	0.90
Yazan Jian et al. 2021 [59]		XGBoost	/	0.872
Li, He-Yan et al. 2022 [62]		KNN	0.98	/
Hong Pan et al. 2023 [69]		MLR	0.703	0.796
Zhu et al. 2023 [70]		LR	0.912	/
Zong et al. 2023 [71]		XGBoost Model 2	0.82	0.884
Dataset1	1 HbA1c	CatBoost	**0.857**	**0.635**
	2 HbA1c	CatBoost	**1.0**	**1.0**
Dataset2	1 HbA1c	CatBoost	**0.941**	**0.95**
	2 HbA1c	LGBM	**1.0**	**0.897**
Dataset3	1 HbA1c	HGB	**1.0**	**1.0**
	2 HbA1c	AdaBoost	**1.0**	**1.0**

Nous remarquons que les performances du modèle développé ont dépassé la plupart des études précédentes notamment avec l'utilisation de deux valeurs de HbA1c comme la table 3.31 le montre, ce qui confirme notre hypothèse initiale et ouvre une porte vers d'autres accomplissements dans le contexte de la prédiction de la RD.

3.6 Expérimentation

Comme précédemment constaté, un grand nombre de nos valeurs tendent vers 1. Cette observation peut s'expliquer par la taille réduite et le déséquilibre de nos datasets.

Pour remédier à cela, nous avons combiné les trois datasets en utilisant leurs caractéristiques communes afin d'obtenir un ensemble de données unifié et plus volumineux. Cela est illustré dans la figure suivante.

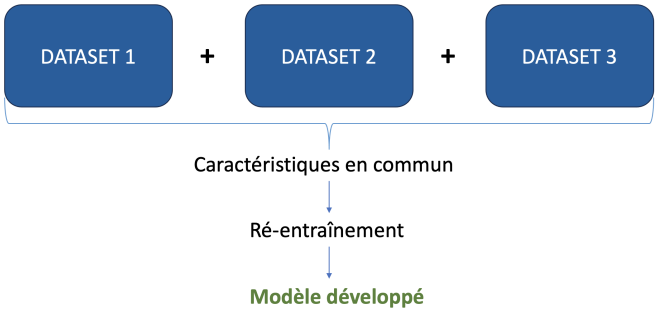

FIGURE 3.126 – Méthode d'unification

L'entraînement de notre modèle final avec ce nouveau dataset a donné les résultats suivants.

TABLE 3.32 – Performances du modèle avec le dataset unifié

Modèle	**AUC**	**Accuarcy**	**Précision**
Modèle développé	1	0.995	0.955

Les figures 3.127 ,3.129 , 3.129, 3.130 et 3.131 suivantes représentent les courbes et les résultats obtenus par ce dernier modèle.

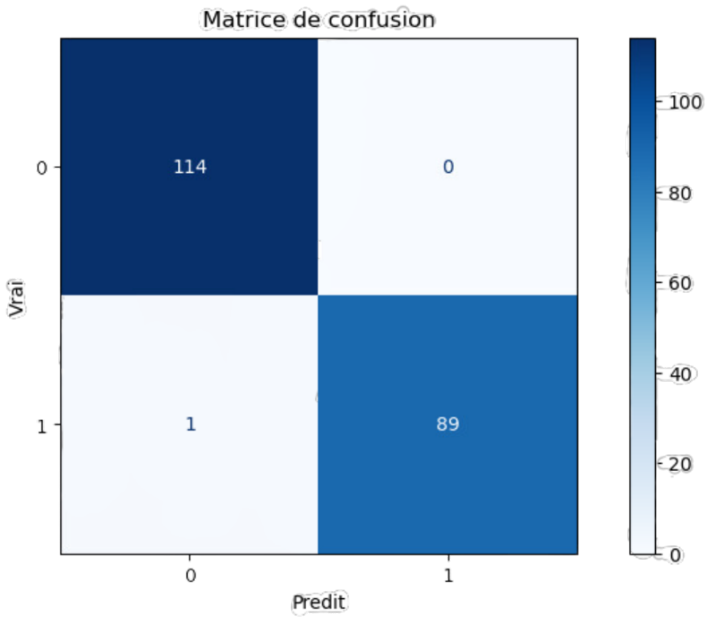

FIGURE 3.127 – Matrice de confusion du modèle avec le dataset unifié

	precision	recall	f1-score	support
0	0.98	1.00	0.99	114
1	1.00	0.98	0.99	90
accuracy			0.99	204
macro avg	0.99	0.99	0.99	204
weighted avg	0.99	0.99	0.99	204

FIGURE 3.128 – Rapport de classification du modèle avec le dataset unifié

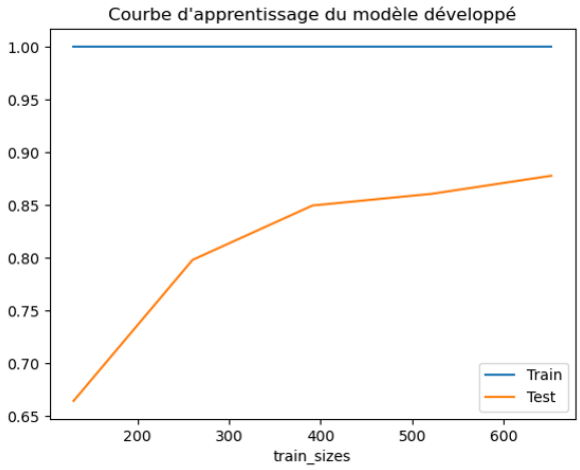

FIGURE 3.129 – Courbe d'apprentissage du modèle avec le dataset unifié

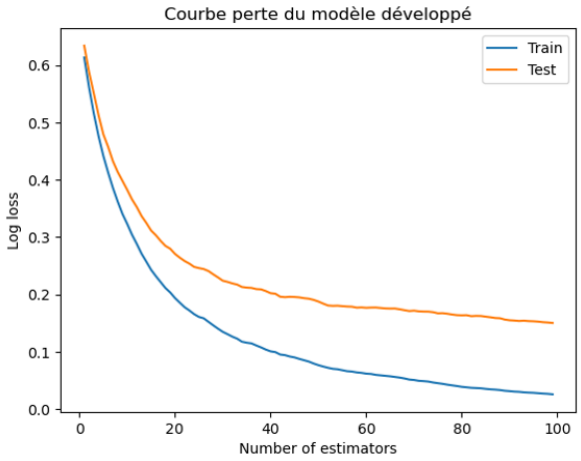

FIGURE 3.130 – Courbe perte du modèle avec le dataset unifié

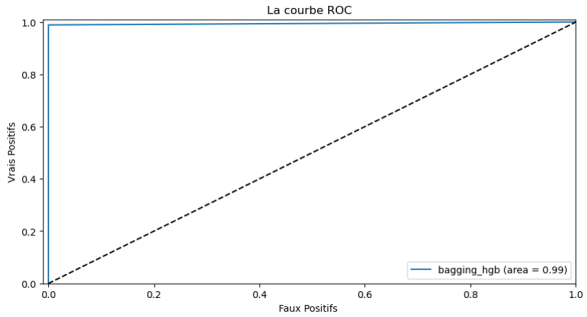

FIGURE 3.131 – Courbe ROC du modèle avec le dataset unifié

Conclusion

Dans ce chapitre, nous avons présenté la méthodologie adoptée dans le but de la construction de notre modèle permettant une prédiction précoce et précise de la RD identifiant ainsi l'impact des caractéristiques dans cette prédiction.

Les résultats obtenus sont très intéressants et ont démontré que les caractéristiques utilisées jouent un rôle prépondérant dans la performance du modèle prédictif notamment l'hémoglobine glyquée qui figure comme caractéristique importante dans pratiquement tous les modèles prédictifs, nous poussant ainsi à encore plus nous intéresser à sa variation qui s'est avérée être un facteur de risque majeur du développement de la RD.

Conclusion générale

La rétinopathie diabétique (RD) est une complication grave et fréquente chez les personnes âgées atteintes de diabète de type 2 entraînant une perte de vision significative voir la cécité. Cette maladie est souvent asymptomatique aux premiers stades soulignant l'importance et le rôle crucial d'une prédiction précoce. De ce fait, plusieurs recherches ont été faites notamment dans le domaine de l'intelligence artificielle dans le but de la détection, la classification ou encore la prédiction de la RD.

Nous avons proposé un nouveau modèle performant pour la prédiction de la RD et etudié l'impact des caractéristiques sur les performances du modèle et plus particulièrement l'influence de la variation du HbA1c.

Ce modèle est obtenu après avoir :
- Trouvé la meilleure méthode parmi 11 méthodes de machine learning.
- Optimisé une première fois le modèle.
- Déterminé les caractéristiques les plus influentes.
- entraîné les trois meilleurs modèles choisis après suppression des caractéristiques les moins pertinentes et optimisation des paramètres.
- Et enfin, appliqué le Bagging au modèle optimisé pour obtenir le modèle final avec les meilleures performances possibles.

Cette étude a été menée en utilisant trois datasets différents en nombre de caractéristiques et en nombre de patients. Chaque dataset a subi la même expérience une fois avec une seule valeur de HbA1c et une deuxième fois avec deux valeurs de HbA1c. Cela nous a permis de :

- Mettre en évidence que le choix du modèle de prédiction du risque de la RD peut être considérablement influencé par les caractéristiques utilisées. Ce travail a fait l'objet d'une communication dans une conférence internationale [10].
- Valider notre approche et de confirmer notre hypothèse initiale basée sur des études médicales antérieures.

En effet, les résultats obtenus ont démontré une nette amélioration des performances après l'ajout de la deuxième valeur de HbA1c. Ces résultats prometteurs indiquent que l'inclusion de l'historique de Hb1Ac dans les modèles de prédiction peut considérablement améliorer leur performance. Cette constatation ouvre de nouvelles perspectives quant à l'utilisation de ces caractéristiques comme élément crucial dans le processus de prédiction de la rétinopathie diabétique.

L'étude de l'impact d'un historique du Hb1AC sur la prédiction de la rétinopathie diabétique ouvre des perspectives prometteuses pour améliorer les méthodes de prédiction et la prise en charge clinique de cette maladie.

Comme perspectives nous proposons :

1. Une expérimentation sur un large Dataset du type Dataset3 afin d'assurer des résultats plus cohérents et plus fiables.
2. Une validation externe des résultats : Il est important de reproduire ces résultats sur des ensembles de données plus larges et variés, pour évaluer la robustesse de l'impact des caractéristiques sur la prédiction de la RD.
3. Une intégration dans la pratique clinique : Si les résultats se généralisent, il pourrait être envisagé d'intégrer le modèle de prédiction de la RD basé sur un historique des Hb1Ac dans la pratique. Cela pourrait permettre une détection précoce de la maladie et une prise en charge adaptée pour réduire les risques de complications.

Bibliographie

[1] All About Vision. Anatomie de la rétine, symptômes de la rétinopathie diabétique : La façon dont le diabète affecte les yeux. `https://www.allaboutvision.com/fr-ca/anatomie-oeil/retine/`, Consulté le 14 Mai 2023.

[2] Qingqing Xu, Liye Wang, and Sujit S Sansgiry. A systematic literature review of predicting diabetic retinopathy, nephropathy and neuropathy in patients with type 1 diabetes using machine learning. *J. Med. Artif. Intell*, 3(6), 2020.

[3] Ramon Mayor Martins and Christiane Gresse Von Wangenheim. Findings on teaching machine learning in high school : A ten-year systematic literature review. *Informatics in Education*, 2022.

[4] Wolff B, Baudouin P, Girmens J, F, Quentel G, Sahel J, A, and Massin P. La rétinopathie diabétique non proliférante. *Rétine et vitré 2018*, Consulté le 01 Décembre 2022.

[5] Fettouma Mazari and Karim Ait Idir. Rétinopathie diabétique entre le diagnostic, la classification et le rythme de surveillance. *Med Sci*, 4(1) :5–9, 2017.

[6] Yogesh K Dwivedi, Laurie Hughes, Elvira Ismagilova, Gert Aarts, Crispin Coombs, Tom Crick, Yanqing Duan, Rohita Dwivedi, John Edwards, Aled Eirug, et al. Artificial intelligence (ai) : Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57 :101994, 2021.

[7] Alice Larroumet. Réduction majeure d'hba1c et rétinopathie dans le diabète de type 2. *Sciences du Vivant [q-bio]*, 2020.

[8] Catherine Gorst, Chun Shing Kwok, Saadia Aslam, Iain Buchan, Evangelos Kontopantelis, Phyo K Myint, Grant Heatlie, Yoon Loke, Martin K Rutter, and Mamas A Mamas. Long-term glycemic variability and risk of adverse outcomes : a systematic review and meta-analysis. *Diabetes care*, 38(12) :2354–2369, 2015.

[9] L Nalysnyk, M Hernandez-Medina, and G Krishnarajah. Glycaemic variability and complications in patients with diabetes mellitus : evidence from a systematic review of the literature. *Diabetes, Obesity and Metabolism*, 12(4) :288–298, 2010.

[10] Bessaa Tinhinane, Bellal Ferroudja, Ait Kaci Azzou Samira, Boukredera Djamila, and Tafoukt Rafik. Analyse de l'influence des caractéristiques sur la performance du modèle prédictif de la rétinopathie diabétique. *sciencesconf.org :coc23 :482083*, 2023.

[11] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3) :685–695, 2021.

[12] Iqbal H Sarker. Machine learning : Algorithms, real-world applications and research directions. *SN computer science*, 2(3) :160, 2021.

[13] Zhi-Hua Zhou. *Machine learning*. Springer Nature, 2021.

[14] J Han. M. kamber in j. pei, data mining : Concepts and techniques : Concepts and techniques, 3. izd, 2011.

[15] Andrew McCallum. Information extraction : Distilling structured data from unstructured text. *Queue*, 3(9) :48–57, 2005.

[16] Machine Learnia. Comment fonctionne le machine learning ? `https://machinelearnia.com/comment-fonctionne-machine-learning/`, Consulté le 23 Mai 2023.

[17] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. *Machine learning techniques for multimedia : case studies on organization and retrieval*, pages 21–49, 2008.

[18] Shichao Zhang, Debo Cheng, Zhenyun Deng, Ming Zong, and Xuelian Deng. A novel knn algorithm with data-driven k parameter computation. *Pattern Recognition Letters*, 109 :44–54, 2018.

[19] Wenchao Xing and Yilin Bei. Medical health big data classification based on knn classification algorithm. *IEEE Access*, 8 :28808–28819, 2019.

[20] Kaitlin Kirasich, Trace Smith, and Bivin Sadler. Random forest vs logistic regression : binary classification for heterogeneous datasets. *SMU Data Science Review*, 1(3) :9, 2018.

[21] Tianlong Xiao, Ping Zhang, Yuren Zhang, Dongxu Li, and Jinsong Shen. A research on the application of college students' physique data mining based on logistic regression algorithm. *ASP Transactions on Computers*, 1(2) :12–18, 2021.

[22] Charlotte Bonte and Frederik Vercauteren. Privacy-preserving logistic regression training. *BMC medical genomics*, 11 :13–21, 2018.

[23] Shujun Huang, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang, and Wayne Xu. Applications of support vector machine (svm) learning in cancer genomics. *Cancer genomics & proteomics*, 15(1) :41–51, 2018.

[24] Nianyin Zeng, Hong Qiu, Zidong Wang, Weibo Liu, Hong Zhang, and Yurong Li. A new switching-delayed-pso-based optimized svm algorithm for diagnosis of alzheimer's disease. *Neurocomputing*, 320 :195–202, 2018.

[25] Bahzad Charbuty and Adnan Abdulazeez. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01) :20–28, 2021.

[26] Harsh H Patel and Purvi Prajapati. Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, 6(10) :74–78, 2018.

[27] Belaidi Nada. L'apprentissage supervisé : définition et exemples. `https://blent.ai/blog/a/apprentissage-supervise-definition`, Consulté le 26 Novembre 2022.

[28] Iqbal H Sarker, ASM Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters, and Alex Ng. Cybersecurity data science : an overview from machine learning perspective. *Journal of Big data*, 7 :1–29, 2020.

[29] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2) :373–440, 2020.

[30] Mohssen Mohammed, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. *Machine learning : algorithms and applications*. Crc Press, 2016.

[31] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning : A survey. *Journal of artificial intelligence research*, 4 :237–285, 1996.

[32] Yagang Zhang. *New advances in machine learning*. BoD–Books on Demand, 2010.

[33] Jie Dou, Ali P Yunus, Dieu Tien Bui, Abdelaziz Merghadi, Mehebub Sahana, Zhongfan Zhu, Chi-Wen Chen, Zheng Han, and Binh Thai Pham. Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, japan. *Landslides*, 17 :641–658, 2020.

[34] Jason Brownlee. *Ensemble learning algorithms with Python : Make better predictions with bagging, boosting, and stacking*. Machine Learning Mastery, 2021.

[35] Li Wen and Michael Hughes. Coastal wetland mapping using ensemble learning algorithms : A comparative study of bagging, boosting and stacking techniques. *Remote Sensing*, 12(10) :1683, 2020.

[36] Matheus Henrique Dal Molin Ribeiro and Leandro dos Santos Coelho. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. *Applied soft computing*, 86 :105837, 2020.

[37] Yanqiu Zhang, Ming Ni, Chengwu Zhang, Shuang Liang, Sheng Fang, Ruijie Li, and Zhouyu Tan. Research and application of adaboost algorithm based on svm. In *2019 IEEE 8th joint international information technology and artificial intelligence conference (ITAIC)*, pages 662–666. IEEE, 2019.

[38] V Kishore Ayyadevara. Pro machine learning algorithms. *Apress : Berkeley, CA, USA*, 2018.

[39] Samir Touzani, Jessica Granderson, and Samuel Fernandes. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, 158 :1533–1543, 2018.

[40] Junliang Fan, Xin Ma, Lifeng Wu, Fucang Zhang, Xiang Yu, and Wenzhi Zeng. Light gradient boosting machine : An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural water management*, 225 :105758, 2019.

[41] Pedro Carmona, Francisco Climent, and Alexandre Momparler. Predicting failure in the us banking sector : An extreme gradient boosting approach. *International Review of Economics & Finance*, 61 :304–323, 2019.

[42] Brownlee Jason. Histogram-based gradient boosting ensembles in python. `https://machinelearningmastery.com/histogram-based-gradient-boosting-ensembles/`, Consulté le 07 Avril 2023.

[43] Abdullahi A Ibrahim, Raheem L Ridwan, Muhammed M Muhammed, Rabiat O Abdulaziz, and Ganiyu A Saheed. Comparison of the catboost classifier with other machine learning methods. *International Journal of Advanced Computer Science and Applications*, 11(11), 2020.

[44] R Sanjeetha, Anant Raj, Kolli Saivenu, Mumtaz Irteqa Ahmed, B Sathvik, and Anita Kanavalli. Detection and mitigation of botnet based ddos attacks using catboost machine learning algorithm in sdn environment. *International Journal of Advanced Technology and Engineering Exploration*, 8(76) :445, 2021.

[45] Qu'est-ce que c'est que le diabète ? `https://www.federationdesdiabetiques.org/information/diabete`, Consulté le 26 Novembre 2022.

[46] Diabète sucré. https://fr.wikipedia.org/wiki/DiabÃÍte_sucrÃľ, Consulté le 28 Novembre 2022.

[47] T Petitclerc. Syndrome polyuropolydipsique. *EMC - Néphrologie*, 1(2) :35–43, 2004.

[48] Œil humain. https://fr.wikipedia.org/wiki/%C5%92il_humain, Consulté le 28 Novembre 2022.

[49] Dr Leininger. Anatomie de l'oeil. l'œil : l'organe de la vision. https://dr-leininger.fr/loeil-et-la-vision/anatomie-de-loeil, Consulté le 20 Juin 2023.

[50] David A Atchison. *Optics of the human eye*. CRC Press, 2023.

[51] RAISS Abderrahmane. Chirurgien ophtalmologue. https://www.youtube.com/@drabderrahmaneraiss-chirur6436, Consulté le 30 Novembre 2022.

[52] Fettouma Mazari and Karim Ait Idir. Rétinopathie diabétique entre le diagnostic, la classification et le rythme de surveillance. *Med Sci*, 4(1) :5–9, 2017.

[53] Low Vision Aids. Rétinopathie diabétique. http://www.lowvisionaids.org/diabetic-retinopathy/, Consulté le 14 Mai 2023.

[54] Daniel Shu Wei Ting, Gemmy Chui Ming Cheung, and Tien Yin Wong. Diabetic retinopathy : global prevalence, major risk factors, screening practices and public health challenges : a review. *Clinical & experimental ophthalmology*, 44(4) :260–277, 2016.

[55] Hsin-Yi Tsao, Pei-Ying Chan, and Emily Chia-Yu Su. Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. *BMC bioinformatics*, 19 :111–121, 2018.

[56] Ruohui Mo, Rong Shi, Yuhong Hu, and Fan Hu. Nomogram-based prediction of the risk of diabetic retinopathy : a retrospective study. *Journal of Diabetes Research*, 2020, 2020.

[57] Zun Shen, Qingfeng Wu, Zhi Wang, Guoyi Chen, and Bin Lin. Diabetic retinopathy prediction by ensemble learning based on biochemical and physical data. *Sensors*, 21(11) :3663, 2021.

[58] Wanyue Li, Yanan Song, Kang Chen, Jun Ying, Zhong Zheng, Shen Qiao, Ming Yang, Maonian Zhang, and Ying Zhang. Predictive model and risk analysis for diabetic retinopathy using machine learning : a retrospective cohort study in china. *BMJ open*, 11(11) :e050989, 2021.

[59] Yazan Jian, Michel Pasquier, Assim Sagahyroon, and Fadi Aloul. A machine learning approach to predicting diabetes complications. In *Healthcare*, volume 9, page 1712. MDPI, 2021.

[60] SKMCA. Rashid centre for diabetes research. https://www.skmca.ae/rashid-centre-for-diabetes-research/, Consulté le 21 Janvier 2021.

[61] Yuedong Zhao, Xinyu Li, Shen Li, Mengxing Dong, Han Yu, Mengxian Zhang, Weidao Chen, Peihua Li, Qing Yu, Xuhan Liu, et al. Using machine learning techniques to develop risk prediction models for the risk of incident diabetic retinopathy among patients with type 2 diabetes mellitus : a cohort study. *Frontiers in Endocrinology*, page 885, 2022.

[62] He-Yan Li, Li Dong, Wen-Da Zhou, Hao-Tian Wu, Rui-Heng Zhang, Yi-Tong Li, Chu-Yao Yu, and Wen-Bin Wei. Development and validation of medical record-based logistic regression and machine learning models to diagnose diabetic retinopathy. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 261(3) :681–689, 2023.

[63] https://www.cdc.gov/nchs/nhanes/.

[64] Statistics for soil survey - book 2 : Spatial analysis of soil properties. http://ncss-tech.github.io/stats_for_soil_survey/book2/.

[65] Tawfiq Hasanin, Taghi M Khoshgoftaar, Joffrey L Leevy, and Naeem Seliya. Examining characteristics of predictive models with imbalanced big data. *Journal of Big Data*, 6(1) :1–21, 2019.

[66] Vincenzo Lagani, Franco Chiarugi, Shona Thomson, Jo Fursse, Edin Lakasing, Russell W Jones, and Ioannis Tsamardinos. Development and validation of risk assessment models for diabetes-related complications based on the dcct/edic data. *Journal of Diabetes and its Complications*, 29(4) :479–487, 2015.

[67] Thor Aspelund, Ó Þórisdóttir, Elin Olafsdottir, Arna Gudmundsdottir, AB Einarsdóttir, Jesper Mehlsen, S Einarsson, O Pálsson, G Einarsson, T Bek, et al. Individual risk assessment and information technology to optimise screening frequency for diabetic retinopathy. *Diabetologia*, 54 :2525–2532, 2011.

[68] Marios Skevofilakas, Konstantia Zarkogianni, Basil G Karamanos, and Konstantina S Nikita. A hybrid decision support system for the risk assessment of retinopathy development as a long term complication of type 1 diabetes mellitus. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 6713–6716. IEEE, 2010.

[69] Hong Pan, Jijia Sun, Xin Luo, Heling Ai, Jing Zeng, Rong Shi, and An Zhang. A risk prediction model for type 2 diabetes mellitus complicated with retinopathy based on machine learning and its application in health management. *Frontiers in Medicine*, 10 :1136653, 2023.

[70] Chengjun Zhu, Jiaxi Zhu, Lei Wang, Shizheng Xiong, Yijian Zou, Jing Huang, Huimin Xie, Wenye Zhang, Huiqun Wu, and Yun Liu. Development and validation of a risk prediction model for diabetic retinopathy in type 2 diabetic patients. *Scientific Reports*, 13(1) :5034, 2023.

[71] Guo-Wei Zong, Wan-Ying Wang, Jun Zheng, Wei Zhang, Wei-Ming Luo, Zhong-Ze Fang, Qiang Zhang, et al. A metabolism-based interpretable machine learning prediction model for diabetic retinopathy risk : A cross-sectional study in chinese patients with type 2 diabetes. *Journal of Diabetes Research*, 2023, 2023.

[72] Ahlam Rashid. Diabetes dataset. *Mendeley Data*, 1, 2020.

[73] Mendeley Data. Diagnosing and predicting clinical and para-clinical cutoffs for diabetes complications in lur and lak populations of iran : A roc curve analysis to design a regional guideline.

[74] Tafoukt R. Médecin spécialiste en endocrinologie - diabétologie. *Rue Guifri Ali Immeuble Ouadfel Laid, Cité Remla, Béjaia, Algérie.* rafiktafoukt6@gmail.com Phone : +213 5 53 120 865.

[75] Amazon SageMaker. Hyperparamètres de catboost. https://docs.aws.amazon.com/fr_fr/sagemaker/latest/dg/catboost-hyperparameters.html, Consulté le 12 Juin 2023.

[76] LightGBM documentation. Parameters. https://lightgbm.readthedocs.io/en/latest/Parameters.html, Consulté le 12 Juin 2023.

[77] Sklearn.ensemble. Histgradientboostingclassifier. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html#sklearn.ensemble.HistGradientBoostingClassifier`, Consulté le 12 Juin 2023.

[78] Sklearn.ensemble. Adaboostclassifier. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html#sklearn.ensemble.AdaBoostClassifier`, Consulté le 12 Juin 2023.