

Master 1 Big Data

Rapport d'Apprentissage Automatique : Prédiction de Diabète de Bout en Bout

Module :
Techniques d'Apprentissage Artificiel

Réalisé par :
ZAID Ibtissam
ZHENG Caroline

Table des matières

1 Contexte et Analyse Exploratoire des Données (EDA)	2
1.1 Problématique et Jeu de Données	2
1.2 Analyse de Corrélation	2
2 Préparation des Données et Méthodologie	3
2.1 Division et Normalisation	3
2.2 Sur-échantillonnage ADASYN	3
3 Modélisation et Sélection du Modèle	3
3.1 Comparaison des Modèles	3
3.2 Analyse de la Courbe d'Apprentissage	4
4 Évaluation Finale et Optimisation du Seuil	4
4.1 Optimisation du Seuil de Décision	4
4.2 Résultats Finaux sur l'Ensemble de Test	5
4.3 Importance des Variables	5
5 Déploiement : Application Web (React)	6
5.1 Fonctionnalités du Dashboard	6
5.2 Prédiction en Temps Réel	6
6 Conclusion et Perspectives	6
6.1 Synthèse des Résultats	6
6.2 Perspectives	7

1 Contexte et Analyse Exploratoire des Données (EDA)

1.1 Problématique et Jeu de Données

Ce projet a pour objectif d'utiliser l'apprentissage automatique pour prédire le risque de diabète à partir d'indicateurs de santé et de mode de vie. Le jeu de données utilisé est le **Diabetes Health Indicators Dataset** de Kaggle.

- **Taille du Dataset :** 253 680 lignes.
- **Classe Cible :** La variable *Diabetes_012* est binarisée en *Diabetes_binary* (1 si l'individu est diabétique ou pré-diabétique, 0 sinon).
- **Déséquilibre de Classes :** Le dataset présente un déséquilibre significatif, avec un **taux de diabétiques** d'environ **15.76%** contre **84.24%** de non-diabétiques.

1.2 Analyse de Corrélation

L'analyse de corrélation avec la variable cible (*Diabetes_binary*) identifie les principaux facteurs de risque.

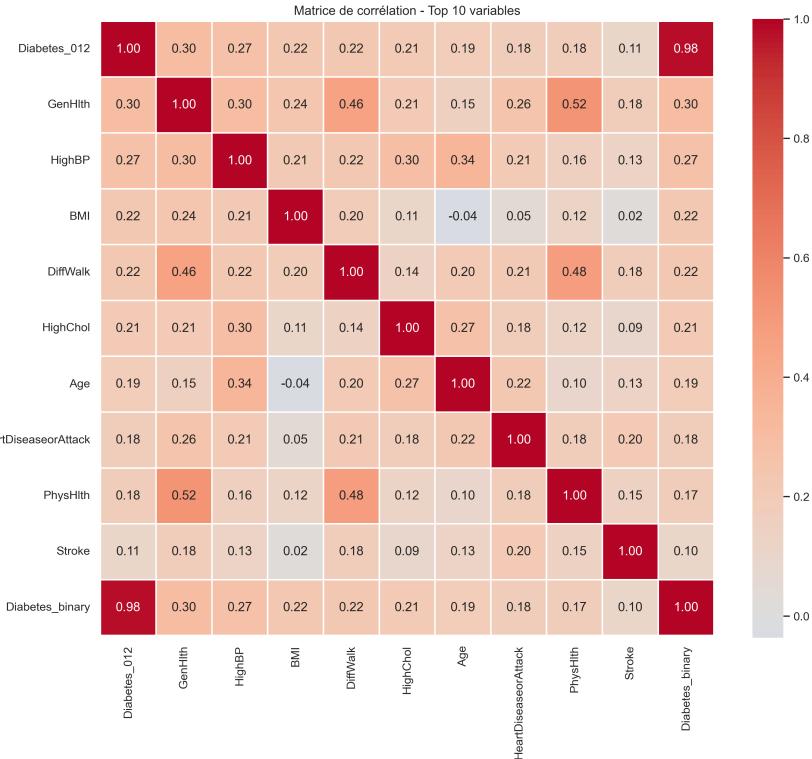


FIGURE 1 – Matrice de corrélation des 10 variables les plus corrélées au diabète.

Les variables présentant la plus forte corrélation positive avec le diabète sont (par ordre décroissant) :

1. *HighBP* (Hypertension Artérielle)
2. *HighChol* (Taux de Cholestérol Élevé)
3. *GenHlth* (Perception de la Santé Générale)
4. *BMI* (Indice de Masse Corporelle)
5. *Age* (Catégorie d'âge)

2 Préparation des Données et Méthodologie

2.1 Division et Normalisation

Le jeu de données est divisé en trois ensembles stratifiés (entraînement 60%, validation 20%, test 20%). Une **StandardScaler** est appliquée pour normaliser les caractéristiques.

2.2 Sur-échantillonnage ADASYN

Afin de pallier le déséquilibre des classes, la technique **ADASYN** (*Adaptive Synthetic Sampling*) est appliquée sur l'ensemble d'entraînement uniquement. ADASYN génère des échantillons synthétiques pour la classe minoritaire, en se concentrant sur les exemples les plus difficiles à classer.

TABLE 1 – Distribution de la Classe Cible (Ensemble d'Entraînement)

Classe	Avant ADASYN	Après ADASYN
Non-diabétique (0)	≈ 128 221	≈ 128 221
Diabétique (1)	≈ 23 987	≈ 128 155 (Synthétiques)

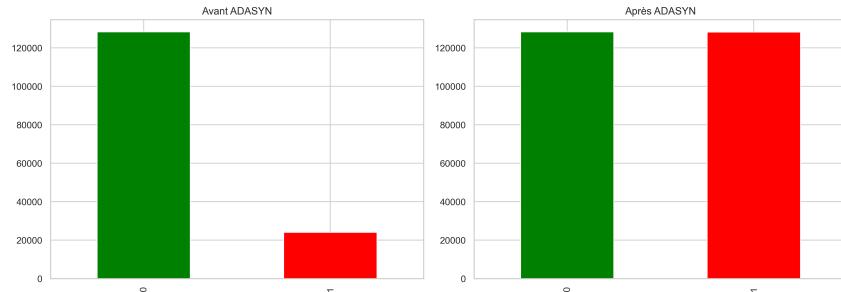


FIGURE 2 – Distribution binaire des classes avant et après l'application d'ADASYN.

3 Modélisation et Sélection du Modèle

3.1 Comparaison des Modèles

Deux modèles d'ensemble puissants ont été entraînés sur l'ensemble rééquilibré (X_{train}, Y_{train}) : **Random Forest** et **XGBoost**.

- **Mesure de Performance** : Le **F1-Score** est privilégié pour évaluer la performance, car il représente la moyenne harmonique de la *Precision* et du *Recall*.
- **Validation Croisée** : Une validation croisée ($CV = 5$) a été réalisée. Le Random Forest a obtenu un F1-Score CV moyen de 0.8240 sur les données synthétiques.

TABLE 2 – Comparaison des performances (F1-Score)

Modèle	CV Moyen (Train)	Validation (Réel)
RandomForest	0.824	0.481
XGBoost	0.792	0.371

Le modèle **Random Forest** a été retenu pour l'évaluation finale car il présentait une meilleure capacité de généralisation.

3.2 Analyse de la Courbe d'Apprentissage

La courbe d'apprentissage est un outil de diagnostic essentiel montrant la performance en fonction de la taille de l'ensemble d'entraînement.

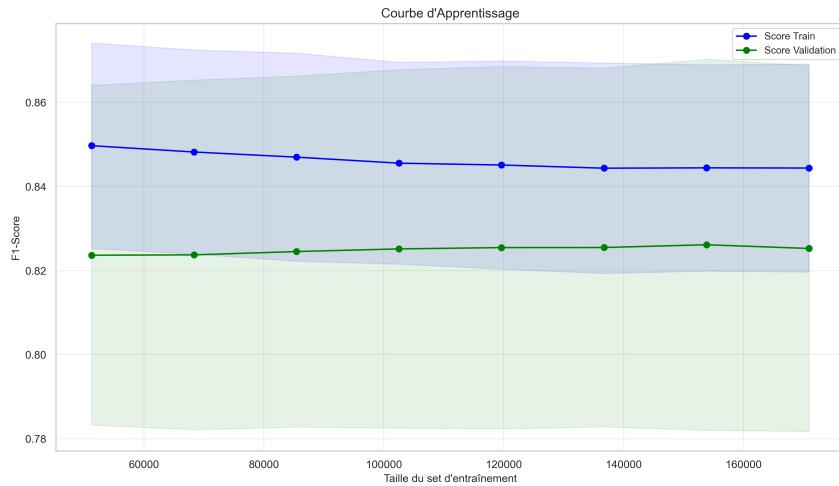


FIGURE 3 – Courbe d'apprentissage. La convergence indique une stabilité du modèle.

4 Évaluation Finale et Optimisation du Seuil

4.1 Optimisation du Seuil de Décision

Dans un contexte de santé, le coût d'un **FauxNégatif** (FN, un diabétique non détecté) est supérieur à celui d'un **FauxPositif** (FP). L'objectif est d'optimiser le seuil de probabilité ($P(\text{Diabète} = 1) \geq \text{Seuil}$) pour maximiser le *Recall* (Sensibilité).

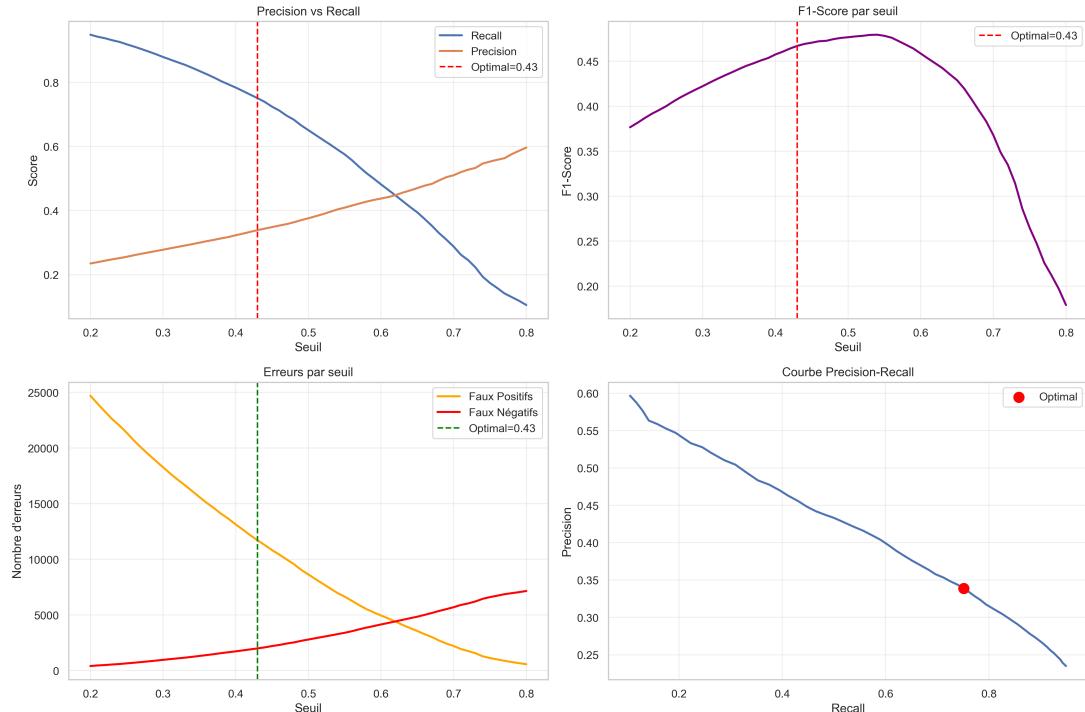


FIGURE 4 – Analyse de l'évolution des métriques (Précision, Recall, F1) en fonction du seuil.

- **Seuil Optimal identifié : 0.430**
- **Objectif atteint : Recall de 75.1%**

4.2 Résultats Finaux sur l'Ensemble de Test

Le modèle est évalué avec le seuil optimal de 0.430.

TABLE 3 – Rapport de Classification Final (Seuil 0.430)

Classe	Precision	Recall	F1-Score	Support
Non-diabétique	0.94	0.73	0.82	42741
Diabétique	0.34	0.75	0.47	7995
AUC-ROC : 0.8138 Accuracy Globale : 73.00%				

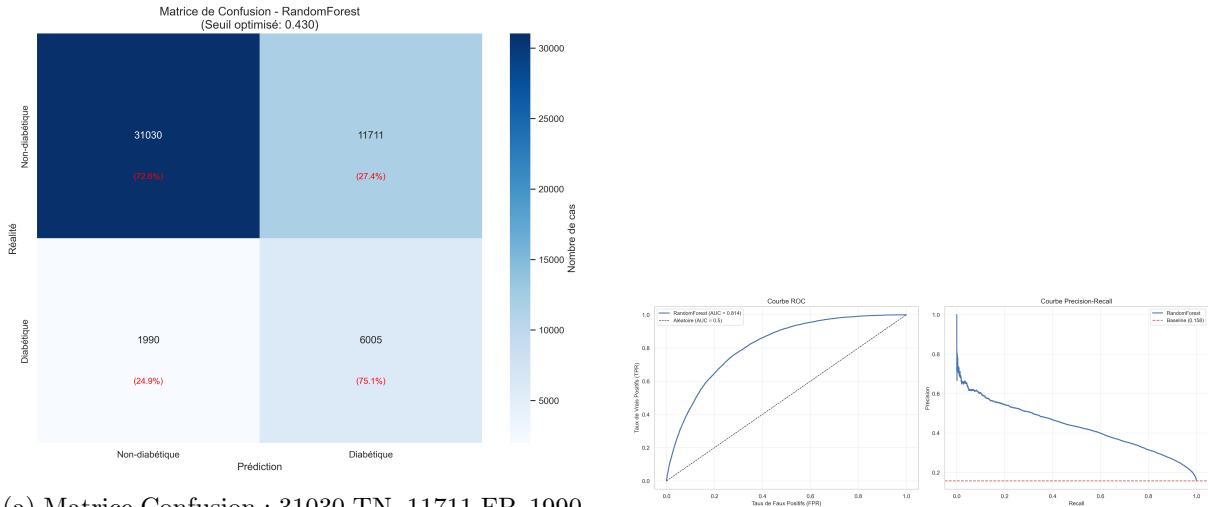


FIGURE 5 – Visualisation des métriques. Le Recall élevé permet de détecter plus de 6000 cas positifs réels, au prix d'une augmentation des Faux Positifs.

4.3 Importance des Variables

L'analyse des *Feature Importances* confirme le rôle central de l'état de santé général (GenHlth), de la tension (HighBP) et de l'IMC (BMI).

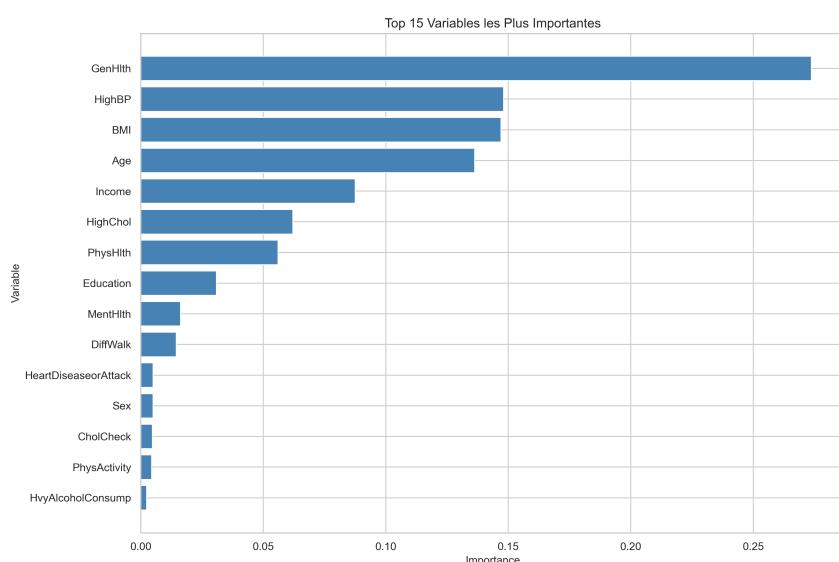


FIGURE 6 – Top 15 des variables expliquant la prédiction.

5 Déploiement : Application Web (React)

Pour répondre à l'exigence d'un projet de bout en bout, nous avons développé une interface utilisateur moderne utilisant la librairie **React.js**.

5.1 Fonctionnalités du Dashboard

L'application web permet de rendre les résultats du modèle accessibles aux utilisateurs non-techniques (médecins, patients) :

- **Visualisation des Performances** : Intégration des graphiques (Courbe ROC, Matrice de Confusion) via la librairie *Recharts* pour une lecture interactive des performances du modèle.
- **Simulateur de Risque (Inférence)** : Un formulaire interactif permet de saisir les 21 paramètres cliniques d'un nouveau patient.

5.2 Prédiction en Temps Réel

L'application implémente la logique du modèle entraîné (Random Forest) et applique le **seuil optimal de 0.430** défini précédemment.

1. L'utilisateur entre les données (Âge, IMC, Cholestérol...).
2. L'algorithme calcule la probabilité d'appartenance à la classe "Diabétique".
3. Si la probabilité > 0.43 , une alerte "Risque Élevé" est affichée avec des recommandations.

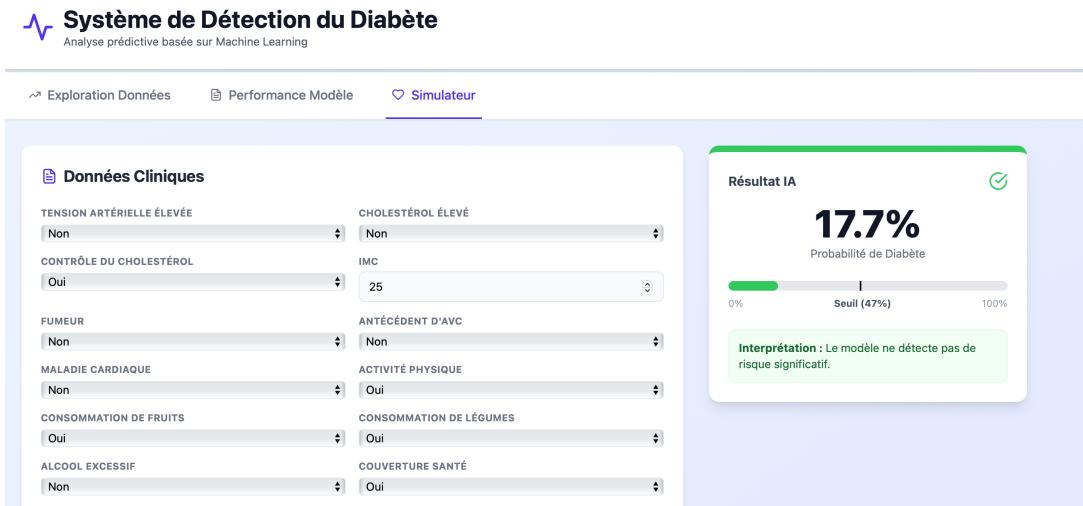


FIGURE 7 – Interface de l'application React : Formulaire de saisie et résultat de prédiction probabiliste.

6 Conclusion et Perspectives

6.1 Synthèse des Résultats

Ce projet a permis de construire une chaîne complète de traitement de données : du nettoyage à la visualisation web, en passant par l'entraînement d'un modèle **Random Forest**. L'utilisation d'**ADASYN** et l'optimisation du seuil de décision (fixé à 0.430) ont permis d'atteindre une sensibilité de **75.1%**, cruciale pour le dépistage médical, bien que cela entraîne une précision plus faible (34%) due aux faux positifs.

6.2 Perspectives

L'application React constitue une base solide pour un outil d'aide à la décision. Les développements futurs pourraient inclure une API Python (Flask/FastAPI) pour charger le modèle .pkl directement au lieu de recoder la logique en JavaScript, ainsi que l'ajout d'explications locales (SHAP values) pour chaque prédiction individuelle.