UNIVERSITÉ PARIS 8

Prédiction de Diabète

Par Apprentissage

Automatique

Module : Techniques d'Apprentissage Artificiel

Par ZAID Ibtissam & ZHENG Caroline

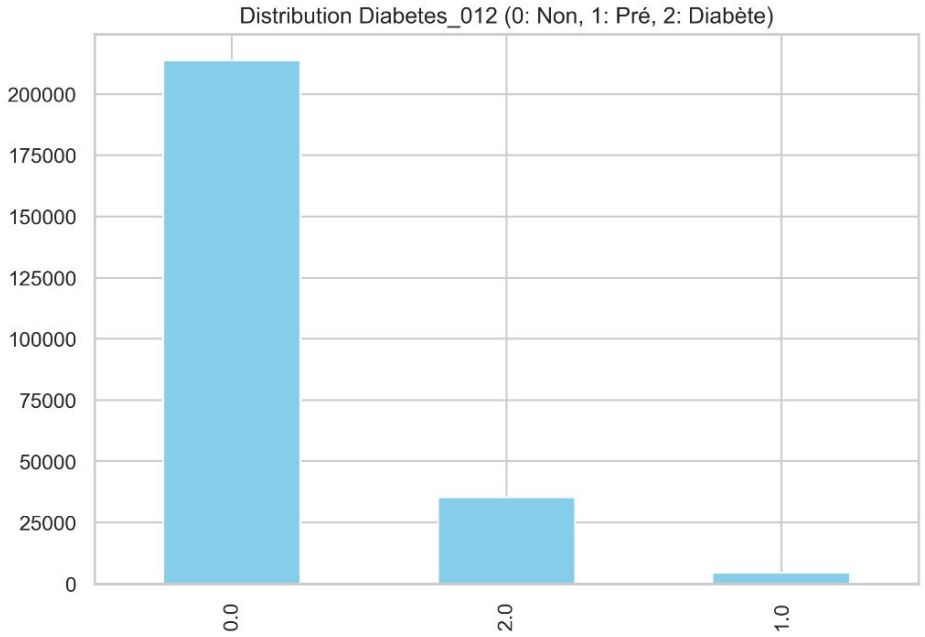
Master 1 Big Data

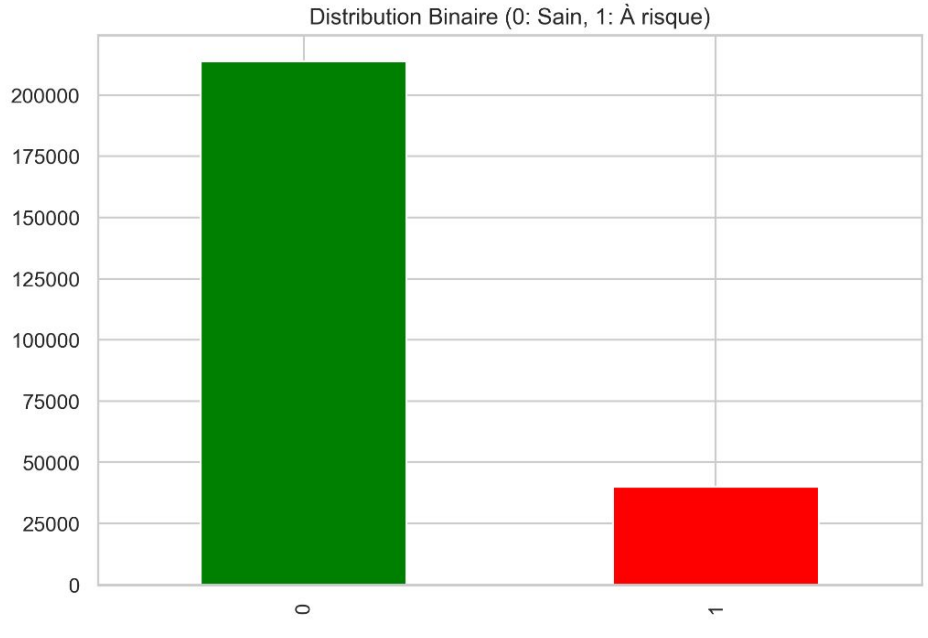
CONTEXTE & PROBLÉMATIQUE

> **"Notre priorité : Maximiser le Rappel (Ne pas rater de malades)."**

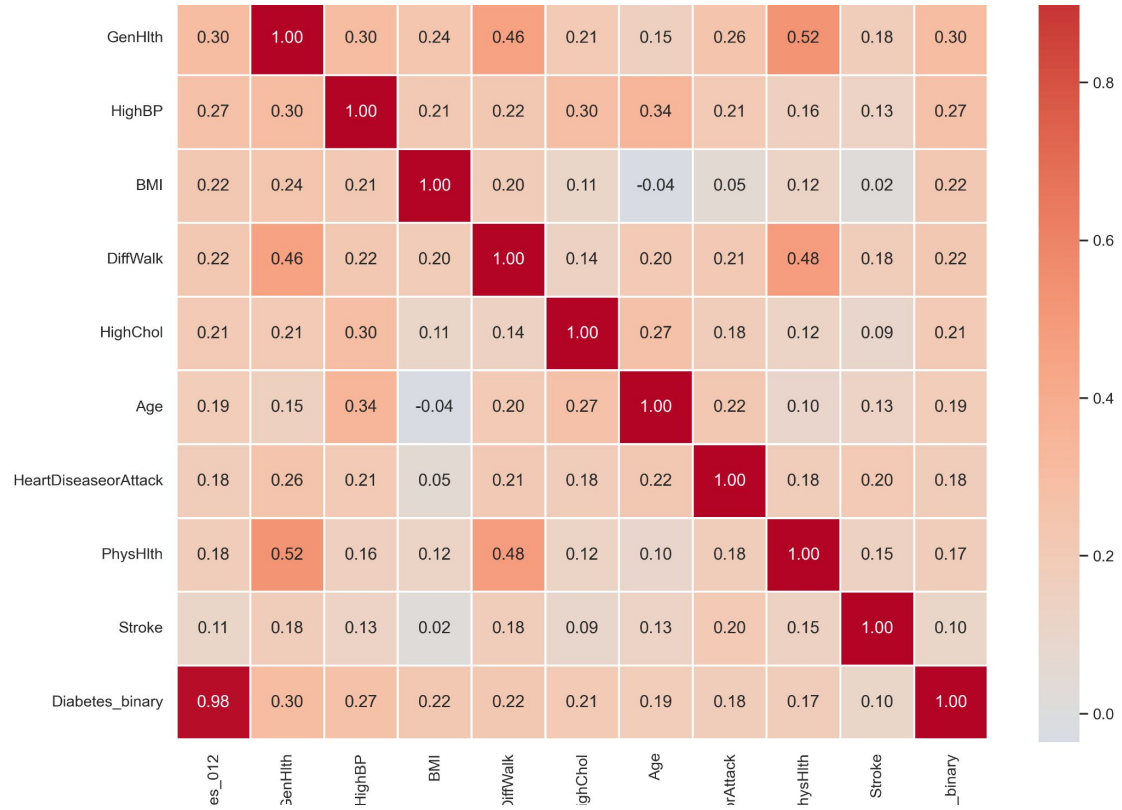
Le diabète est un enjeu de santé publique majeur. Une détection précoce est cruciale pour prévenir les complications.

- **Objectif :** Classification Binaire (Sain vs Diabétique).
- **Données :** Kaggle Health Indicators (~253k patients).
- **Défi Technique :** Fort déséquilibre des classes.





ANALYSE EXPLORATOIRE (EDA)



Facteurs de Risque Identifiés

L'analyse de corrélation révèle des liens logiques avec les données cliniques :

Hypertension (HighBP)
Corrélation de +0.27.

Santé Générale (GenHlth)
Corrélation de +0.30.

IMC (BMI)
Corrélation de +0.22.

PIPELINE DE PRÉPARATION

Une préparation rigoureuse est essentielle pour garantir la robustesse du modèle.

1. Split Stratifié

Train (60%) / Val (20%) / Test (20%).

2. Scaling

StandardScaler sur Xtrain uniquement.

3. ADASYN

Rééquilibrage de la classe minoritaire.

4. Entraînement

Comparaison RF vs XGBoost.

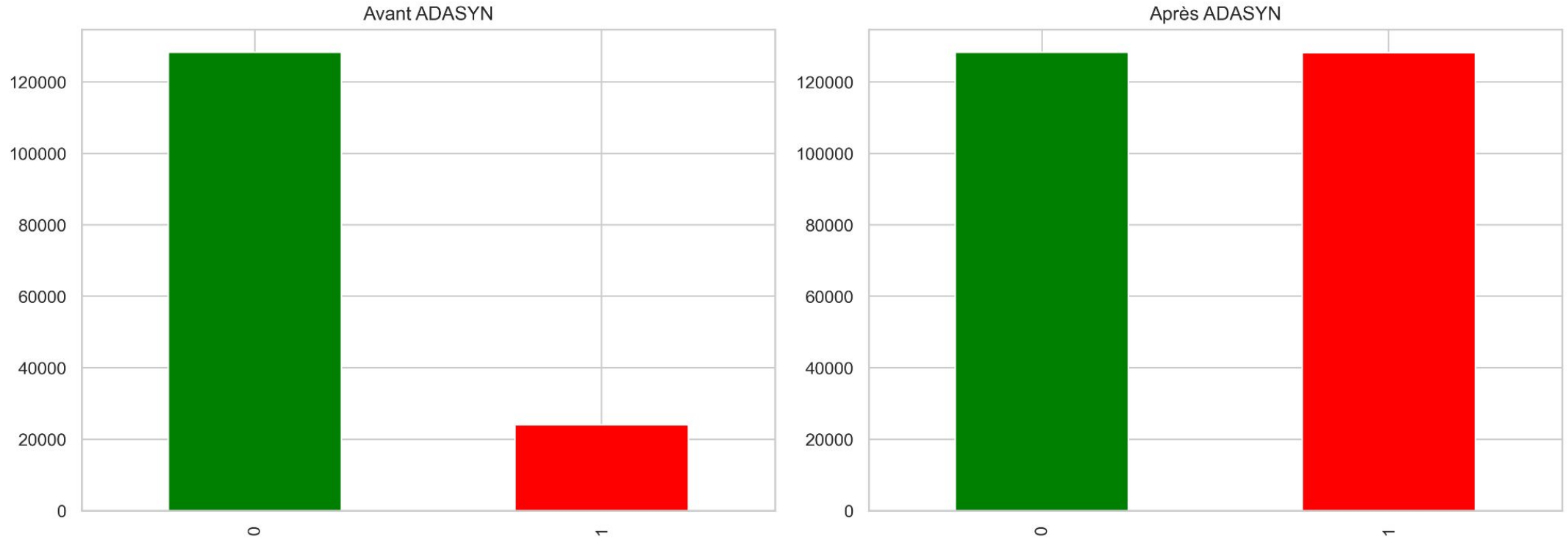
GESTION DU DÉSÉQUILIBRE : ADASYN

Le Problème Initial

Seulement environ 15% de cas positifs. Le modèle apprendrait principalement à prédire la majorité.

La Solution : ADASYN

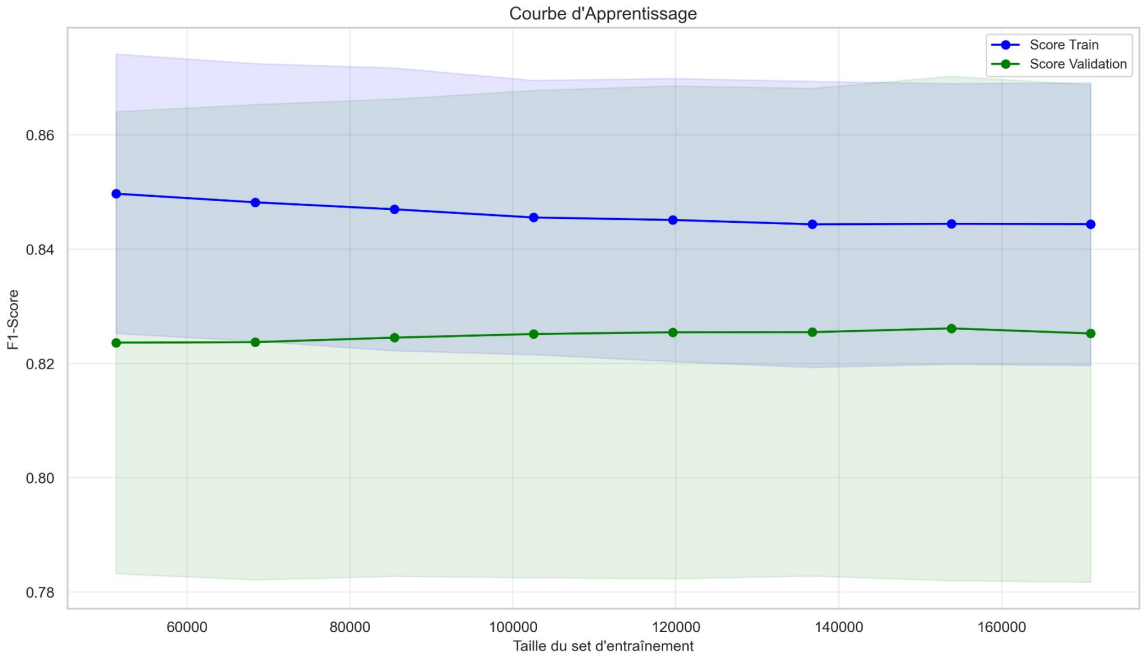
Génération de nouveaux exemples diabétiques artificiels, permettant un quasi-équilibre 50/50 pour un entraînement non biaisé.



MODÉLISATION & SÉLECTION

Nous avons comparé deux algorithmes d'ensemble robustes : **Random Forest** et **XGBoost**.

La métrique de décision est le **F1-Score** sur la validation croisée (CV=5), car l'Accuracy est trompeuse sur des données déséquilibrées.



Modèle	F1-Score (CV)	F1-Score (Val)
Random Forest	**0.867**	**0.460**
XGBoost	0.852	0.455

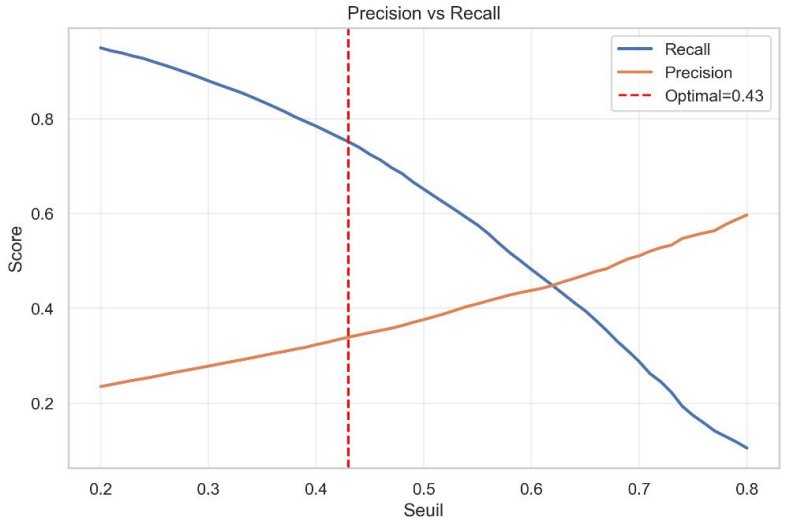
**Le Random Forest est retenu pour sa performance stable.*

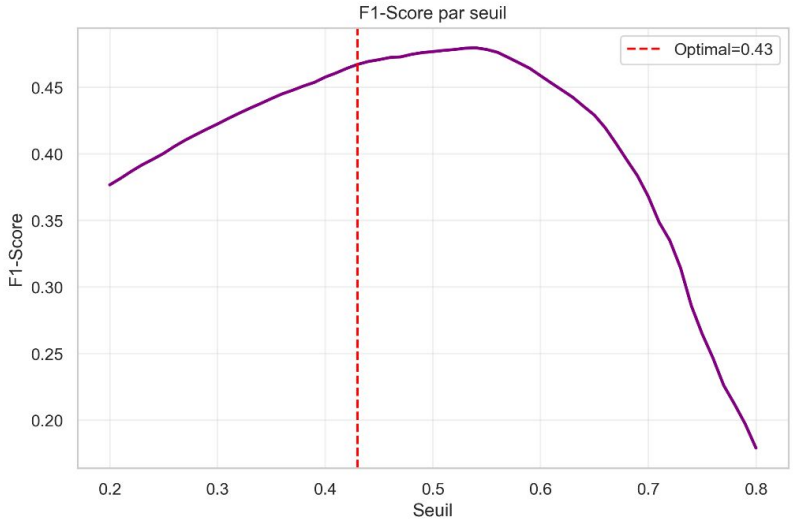
La courbe d'apprentissage montre une faible variance : le modèle est robuste.

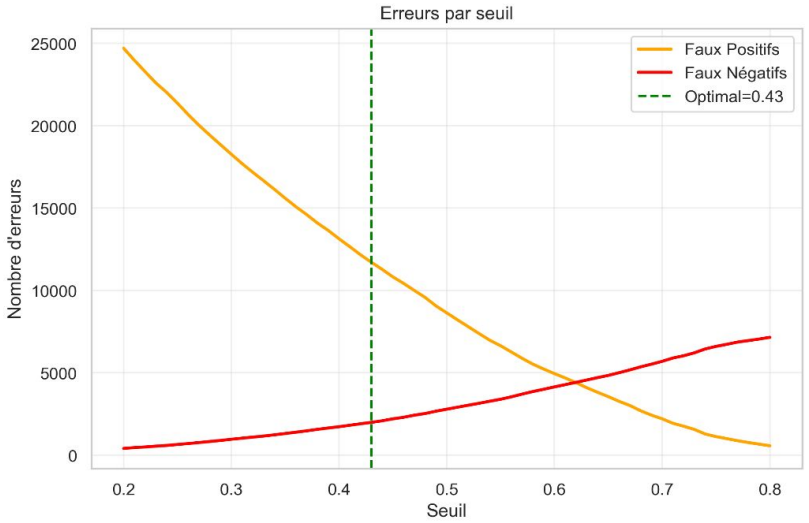
OPTIMISATION DU SEUIL

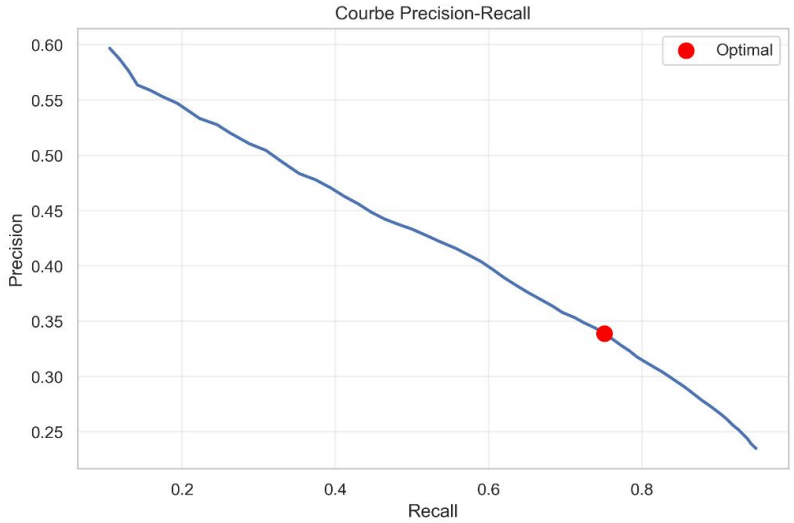
Seuil Optimal Final

0.430









Contexte et Stratégie

Le seuil par défaut (0.5) entraîne trop de Faux Négatifs (malades ratés).

Nous recherchons le F1-Score maximal tout en garantissant un Recall élevé.

RÉSULTATS FINAUX (TEST SET)

Évaluation avec le seuil optimisé (0.430).

75.1%

Rappel (Sensibilité)

3 diabétiques sur 4 détectés

0.814

AUC-ROC

Très bonne capacité de discrimination

0.463

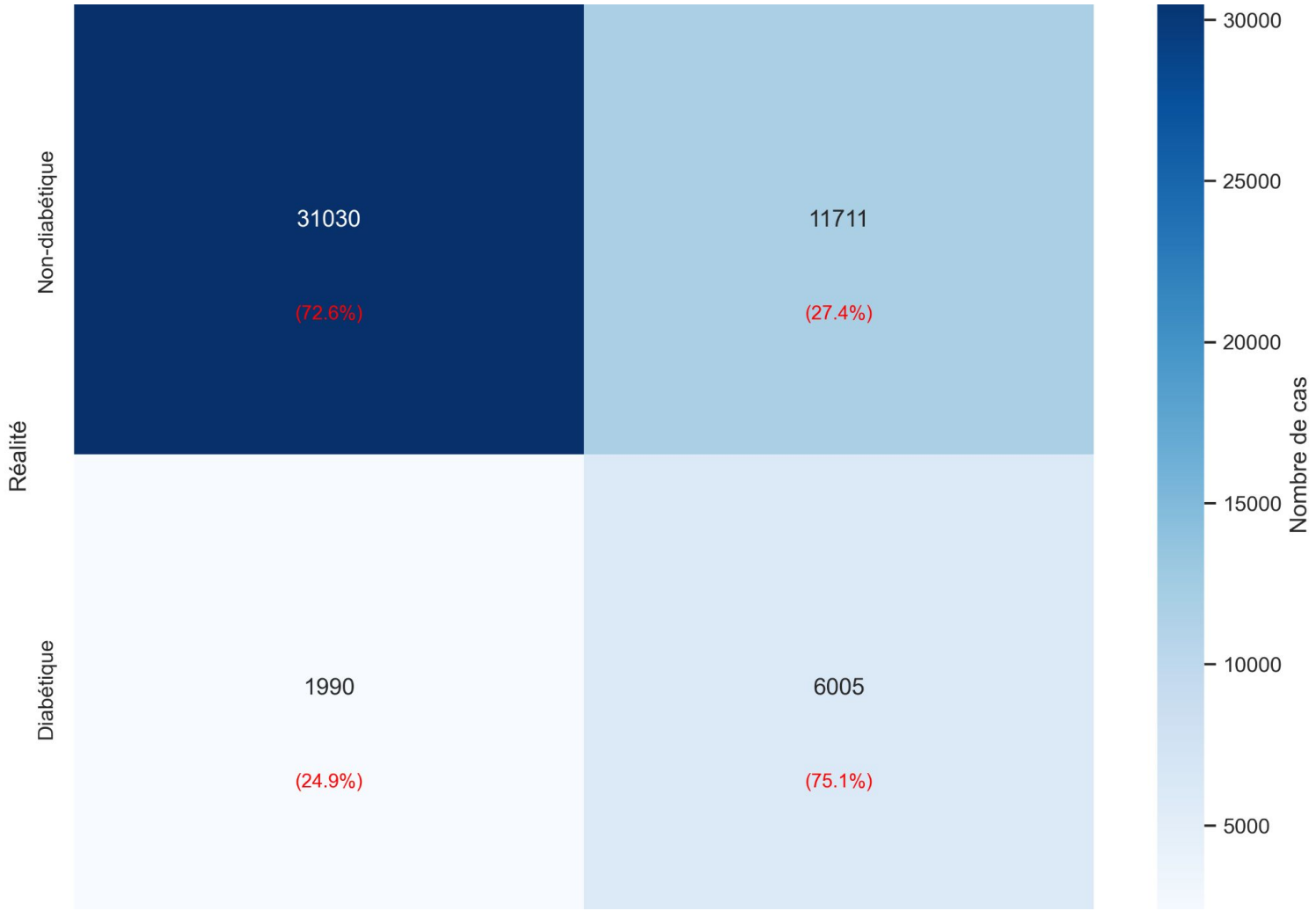
F1-Score

Meilleur compromis atteint à 0.43

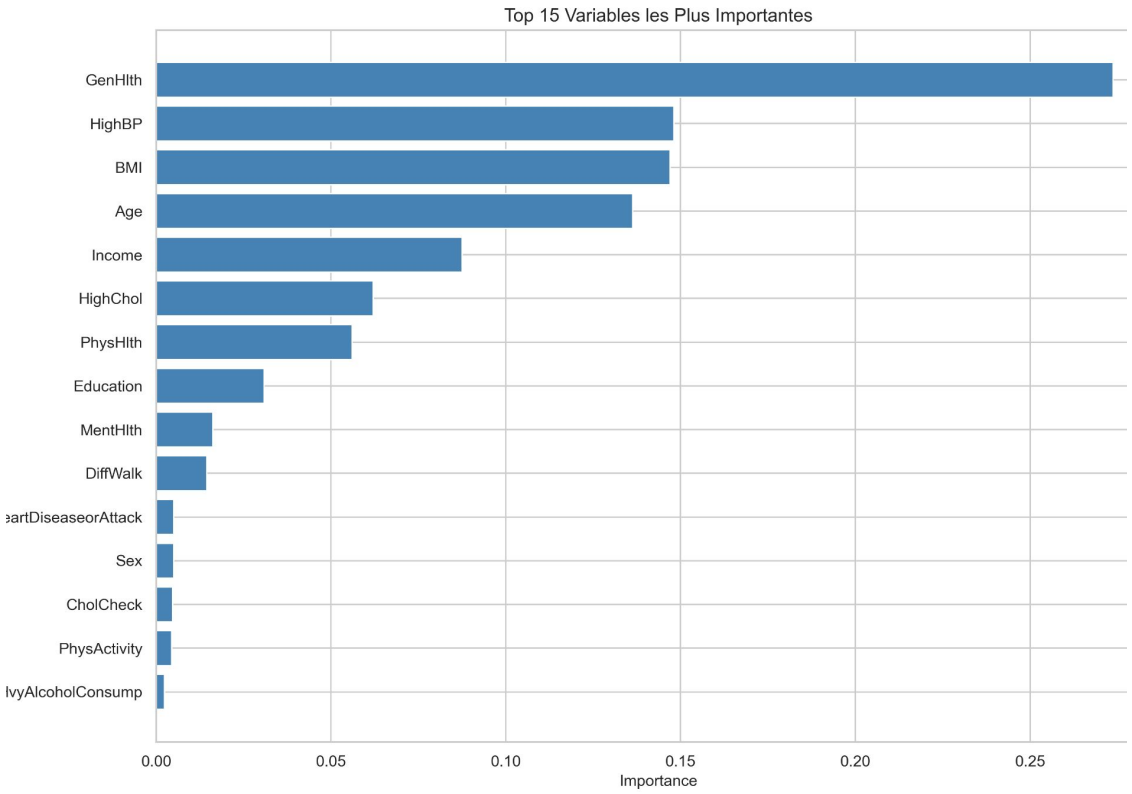
Analyse des Erreurs

- **Vrais Positifs (6005) :** Succès de la détection.
- **Faux Positifs (11711) :** C'est le prix du Recall élevé, acceptable en phase de dépistage.
- **Faux Négatifs (1990) :** Réduits au minimum (env. 4x moins que les FP).

L'analyse montre que les FP ont une probabilité prédite > 0.43, tandis que les FN sont largement sous le seuil optimal.



INTERPRÉTABILITÉ & PERSPECTIVES



Variables Décisives

Le modèle donne une importance maximale à GenHlth (Santé générale auto-déclarée) devant HighBP et BMI.

Bilan et Futur

Pipeline complet et robuste.

Objectif de Recall (>75%) atteint.

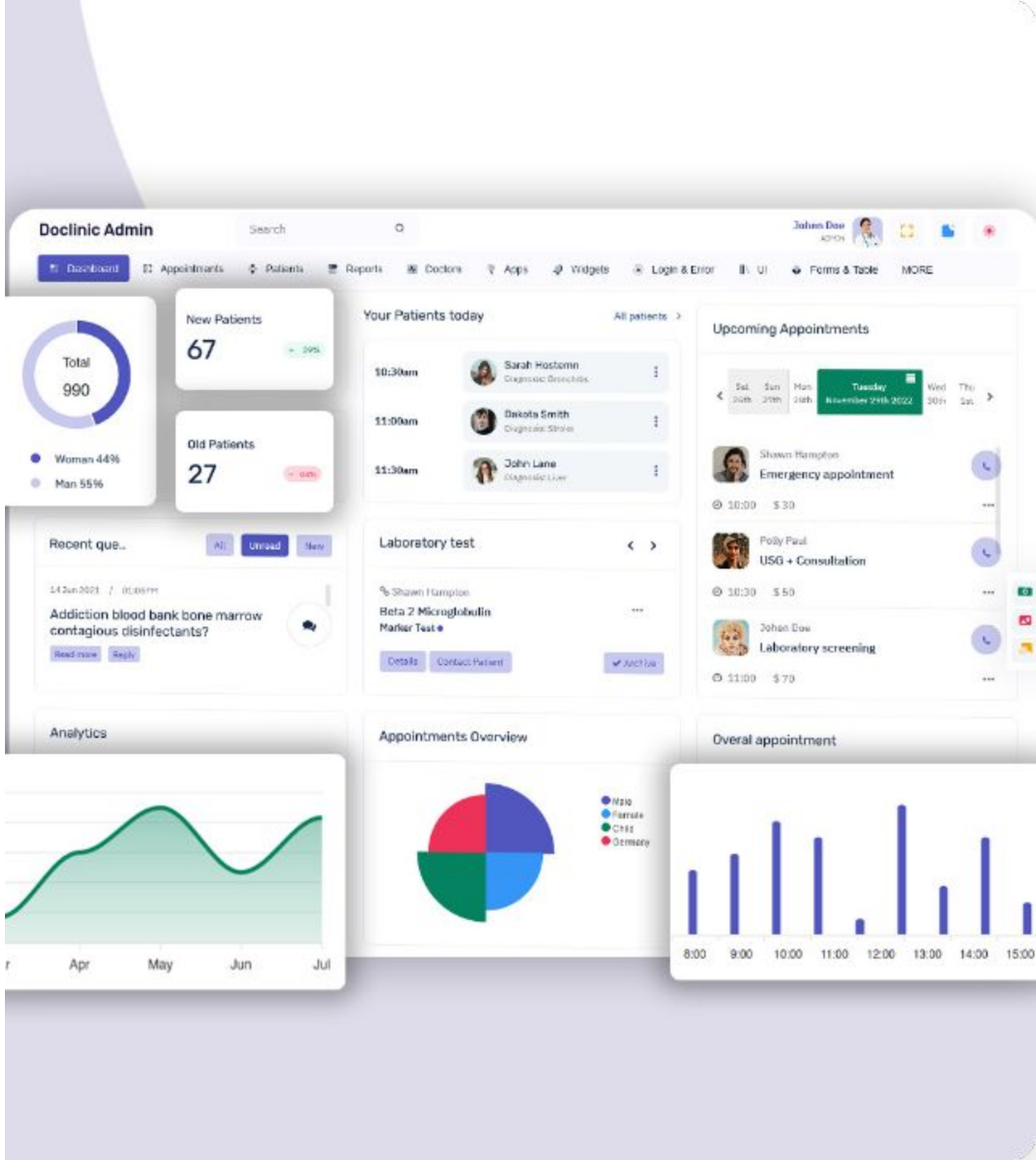
Perspectives : Intégrer l'analyse SHAP pour l'interprétabilité locale et explorer LightGBM pour la vitesse.

Déploiement Web

Application React.js

Développement d'une interface utilisateur moderne pour rendre le modèle accessible.

- **Simulateur :** Formulaire interactif pour les 21 paramètres.
- **Temps Réel :** Inférence immédiate avec le seuil de 0.43.
- **Aide à la décision :** Affichage clair du niveau de risque.



Merci de votre attention

Caasez vous ?

ZAID Ibtissam & ZHENG Caroline