

Prédire le diabète 5 ans avant le diagnostic avec 8 variables et un algorithme

XGBoost, valeurs SHAP et machine learning clinique pour la prévention du diabète de type 2

Pinel Baudelaire T. DAHOUI

Data Scientist & IA | Ingénieur en Travaux Statistiques
tpineldahoui@gmail.com | github.com/Baudelaire12

3 mai 2026

Série LinkedIn STEM au Quotidien — Jour 32/40

Plan

- 1 Mise en contexte
- 2 Développement mathématique
- 3 Cas pratique
- 4 Références

L'épidémie silencieuse : 537 millions de diabétiques

Chiffre clé

537 millions d'adultes vivent avec le diabète dans le monde un chiffre qui atteindra **783 millions** en 2045 (IDF, 2021). En Afrique subsaharienne, **24 millions** de personnes sont diabétiques, mais plus de la **moitié l'ignore**. Le dépistage précoce par algorithme ML est une révolution accessible.

- Une intervention simple au stade du pré-diabète (nutrition, activité physique, perte de poids de 5–7%) réduit la progression vers le diabète de **58%** (DPP Research Group, *NEJM*, 2002).
- L'enquête STEPS de l'OMS (Bénin, 2015) estimait la prévalence de l'hyperglycémie à **8,4%** chez les adultes de 25 à 64 ans à Cotonou.
- Un modèle XGBoost atteint une AUC de **0,87** pour la prédiction du diabète à 5 ans.

Variables prédictives cliniques clés

Variable	Importance SHAP	Seuil d'alerte
Glycémie à jeun (FPG)	28%	≥ 100 mg/dL
HbA1c	22%	$\geq 5.7\%$
Tour de taille	15%	> 94 cm (H), > 80 cm (F)
IMC	12%	≥ 25 kg/m ²
Âge	10%	≥ 45 ans
Antécédents familiaux	8%	Parents diabétiques
Pression artérielle	5%	$\geq 130/80$ mmHg

XGBoost et interprétabilité SHAP

XGBoost — gradient boosting sur arbres :

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot h_t(\mathbf{x}_i)$$

Valeur SHAP pour la variable j sur la prédiction de l'individu i :

$$\phi_j(i) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{j\}}(\mathbf{x}) - f_S(\mathbf{x}) \right]$$

Interprétabilité clinique

Les valeurs SHAP transforment un modèle « boîte noire » en outil cliniquement interprétable :
« Ce patient est à risque élevé **principalement** à cause de sa glycémie à jeun de 112 mg/dL et de son tour de taille de 98 cm ».

Proposition opérationnelle

Un score de risque diabète calculé automatiquement à chaque consultation de médecine générale au Bénin, à partir de 7 variables mesurées en routine (glycémie capillaire, IMC, tour de taille, tension, âge, hérédité, HbA1c), permettrait d'identifier les patients pré-diabétiques avant l'installation de la maladie.

Pipeline de déploiement envisageable :

- 1 Constitution d'une base de données béninoise (cohorte de 5 000 patients, CNHU et centres de santé)
- 2 Entraînement d'un modèle XGBoost sur données locales
- 3 Intégration dans OpenMRS pour le calcul automatique du score à chaque visite

- ④ Alerte au médecin si le score est supérieur à 0,30
- ⑤ Suivi de l'efficacité via la réduction de l'incidence du diabète sur 5 ans

« Prévenir le diabète coûte 10 fois moins cher que le traiter. La data peut faire ce calcul pour chaque patient. »

Références I

-  International Diabetes Federation. (2021). *IDF Diabetes Atlas* (10th ed.). IDF.
-  Diabetes Prevention Program Research Group. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *NEJM*, 346(6), 393–403.
-  Chen, T., & Guestrin, C. (2016). XGBoost : A scalable tree boosting system. *KDD 2016*, 785–794.
-  Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NeurIPS 2017*, 4768–4777.
-  WHO / Ministère de la Santé du Bénin. (2015). *Enquête STEPS sur les facteurs de risque des maladies non transmissibles au Bénin*. OMS.

Merci pour votre attention

Pinel Baudelaire T. DAHOUI | Ingénieur en Travaux Statistiques
tpineldahoui@gmail.com | github.com/Baudelaire12