# **Rapport de Projet — Prédiction du Risque de Diabète par Analyse Clinique**

## **1. Introduction**

Dans le cadre de notre mission au sein d’un laboratoire biomédical, nous avons développé un système d’intelligence artificielle capable de prédire si un patient est à risque de développer un diabète de type 2, en se basant sur des données cliniques simples. Ce projet combine deux approches de l’apprentissage automatique : le **clustering** pour identifier des profils types, et la **classification supervisée** pour prédire le risque d’un nouvel individu.

## **2. Objectifs**

* Identifier des groupes de patients similaires (clustering).
* Prédire le risque de diabète sur la base de mesures cliniques (classification).
* Déployer un modèle IA robuste, interprétable, et reproductible.

## **3. Données**

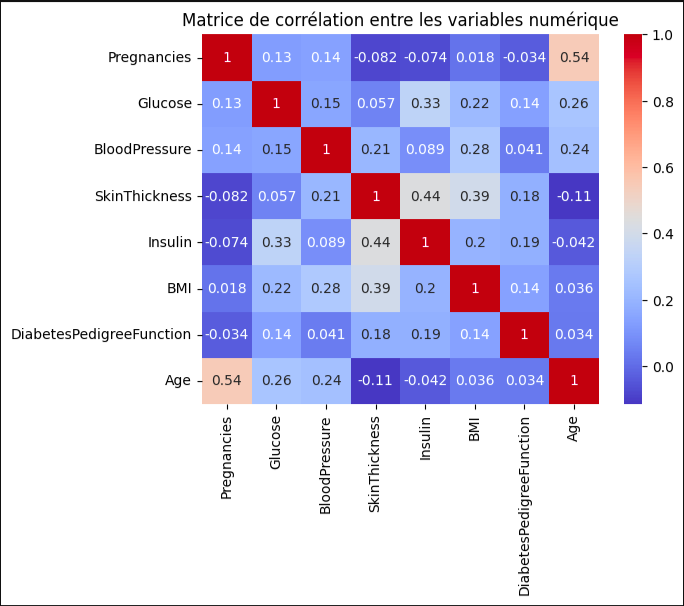
Nous avons utilisé un jeu de données issu d’examens médicaux de patients. Les variables disponibles sont :

* Glucose
* BloodPressure
* SkinThickness
* Insulin
* BMI
* DiabetesPedigreeFunction
* Age

Taille du jeu de données : 768 entrées.

## **4. Analyse exploratoire (EDA)**

* **Valeurs manquantes** : Aucune valeurs manquantes trouvées.
* **Outliers** : détectés avec les boxplots et la méthode IQR.
* **Matrice de corrélation**

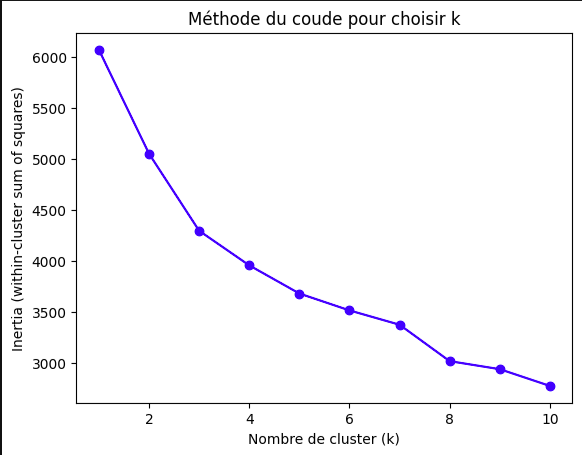


## **5. Prétraitement**

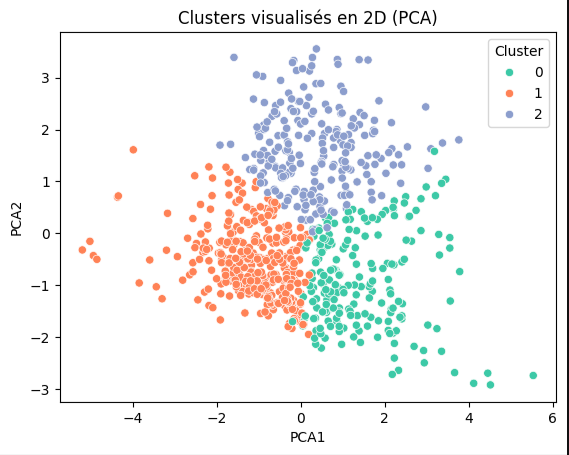
* Suppression des outliers.
* Suppression des valeurs manquantes.
* Normalisation avec StandardScaler.

## **6. Clustering non supervisé**

* Algorithme utilisé : **K-Means**.
* Nombre de clusters k=3 via la **méthode du coude**.



* Visualisation des clusters par **PCA (Analyse en Composantes Principales)**.



* Interprétation :

Les moyennes des variables Glucose > 130 et BMI > 30 Diabetes Pedigree Function >0,5 donc le Clustur 0 est à haut risque de diabète.

Les moyennes des variables Glucose < 130 et BMI < 30 Diabetes Pedigree Function <0,5 donc le Clustur 1 est à faible risque de diabète.

Les moyennes des variables Glucose < 130 et BMI > 30 Diabetes Pedigree Function <0,5 donc le Clustur 2 est à risque moyen de diabète.

* Ajout d’une colonne risk\_category : 0 = Haut Risque , 1 = Faible Risque, 2 = Moyen Risque.

## **7. Classification supervisée**

### **7.1 Préparation des données**

* X : Glucose, BMI, Age, DiabetesPedigreeFunction
* y : colonne Cluster isssue du clustering
* Division : train\_test\_split(test\_size=0.2)
* Rééquilibrage des classes avec RandomOverSampler:

Avant équilibrage Counter({1: 330, 2:225, 0:204})

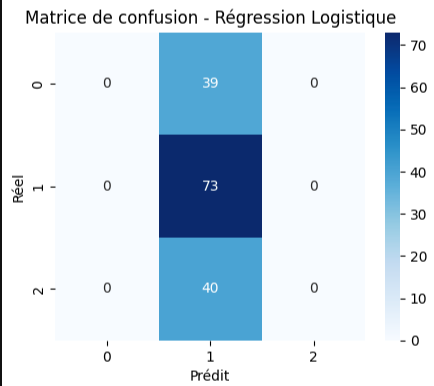
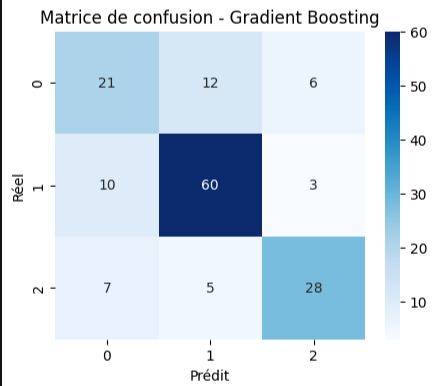
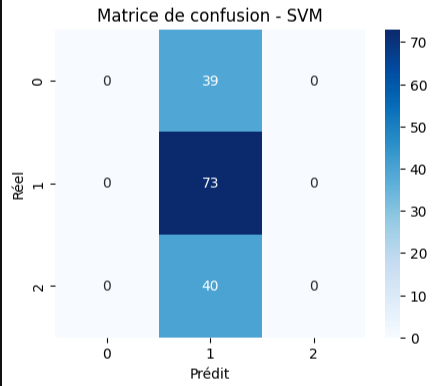
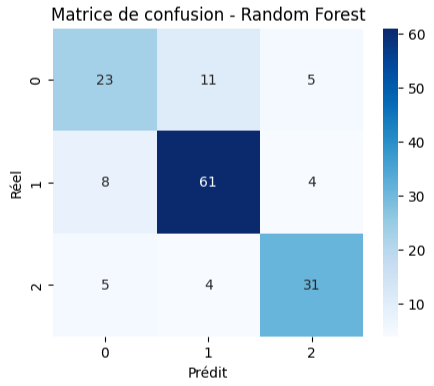
Avant équilibrage Counter({1: 330, 2:330, 0:330})

### **7.2 Modèles testés**

* Logistic Regression
* Random Forest
* SVM
* Gradient Boosting

### **7.3 Évaluation**

* Métriques : Accuracy, Precision, Recall, F1-score, Confusion Matrix



* Meilleur modèle : Gradient Boosting Classifier avec F1-score de 71%
* Validation croisée cv=5 pour robustesse

### **7.4 Optimisation**

* GridSearchCV pour affiner les hyperparamètres (n\_estimators, max\_depth...)

## **8. Sauvegarde du modèle**

* Modèle final sauvegardé dans models/best\_model.pkl
* Peut être chargé pour faire des prédictions sur de nouveaux patients

## **9. Documentation**

* Le projet est documenté dans un notebook Jupyter.
* README fourni pour exécution locale.
* Code commenté à chaque étape pour reproductibilité.

## **10. Conclusion**

Ce projet a permis de combiner l’analyse exploratoire, l’apprentissage non supervisé et la classification supervisée pour offrir un outil d’aide à la décision médicale. Il constitue une base solide pour un déploiement en milieu hospitalier ou dans une application de suivi patient.

## **11. Perspectives**

* Intégration dans une interface Web (Flask ou Streamlit)
* Déploiement sur le cloud (API REST)
* Ajout de nouvelles variables cliniques (historique familial, mode de vie, etc.)
* Explication des prédictions avec des outils comme SHAP ou LIME

**Auteur** : M.G.

**Rôle** : Développeur IA Junior

**Contact** : [[votre@email.com](mailto:votre@email.com)]