# Rapport du Projet Machine Learning

## Titre : Prédiction du diabète à l'aide d'algorithmes de classification

**Nom :** Loubna Laâkouk

**Filière :** CISI3 – SupMTI Oujda

**Année universitaire :** 2024 / 2025

## 1. Introduction

Dans ce projet de machine learning, l'objectif était de concevoir un modèle capable de prédire si une personne est atteinte de diabète à partir de données médicales simples. L'idée principale est d'automatiser une prédiction basée sur les antécédents et mesures cliniques d'un patient, ce qui pourrait aider à la détection précoce de cette maladie.

Ce projet m'a permis de mettre en pratique plusieurs notions vues en cours : la manipulation de données avec Python, le nettoyage, l'analyse exploratoire, et surtout l'entraînement et l'évaluation de modèles d'apprentissage automatique.

## 2. Description du dataset

J'ai utilisé un dataset public très connu : **Pima Indians Diabetes Database**, disponible sur Kaggle. Il contient des informations médicales de 768 femmes d'origine amérindienne âgées de 21 ans ou plus.

Le fichier contient 9 colonnes :

- **Pregnancies** : nombre de grossesses
- **Glucose** : taux de glucose dans le sang
- **BloodPressure** : pression artérielle
- **SkinThickness** : épaisseur du pli cutané
- **Insulin** : taux d'insuline
- **BMI** : indice de masse corporelle
- **DiabetesPedigreeFunction** : facteur héréditaire

- **Age** : âge de la patiente
- **Outcome** : 0 = non-diabétique, 1 = diabétique

## 3. Prétraitement des données

Après avoir chargé les données, j'ai remarqué que certaines valeurs étaient anormalement égales à zéro dans des colonnes comme le glucose, la pression artérielle ou l'insuline. Or, une valeur de 0 pour ces indicateurs n'est pas réaliste d'un point de vue médical.

J'ai donc remplacé ces zéros par la **médiane** de chaque colonne correspondante.

Ensuite, j'ai séparé les données en deux ensembles :

- 80 % pour l'entraînement (614 lignes)
- 20 % pour le test (154 lignes)

## 4. Analyse exploratoire

J'ai réalisé une première visualisation avec un **countplot**, qui montre qu'il y a un léger déséquilibre entre les classes : un peu plus de patientes non-diabétiques (0) que diabétiques (1).

Une **matrice de corrélation (heatmap)** a également été générée. Elle m'a permis d'identifier que les variables **Glucose**, **BMI** et **Age** sont les plus corrélées avec le fait d'être diabétique ou non.

## 5. Entraînement des modèles

Trois modèles de classification ont été testés :

- **Logistic Regression**
- **Decision Tree**
- **K-Nearest Neighbors (KNN)**

**Résultats obtenus :**

| Modèle | Accuracy | F1-score (diabétiques) | Recall (diabétiques) |
|---|---|---|---|
| Logistic Regression | 76 % | 0.65 | 0.64 |
| Decision Tree | 73 % | 0.64 | **0.69** |
| KNN (k=5) | 66 % | 0.56 | 0.62 |

## 6. Interprétation des résultats

Le modèle de **régression logistique** a obtenu la meilleure précision globale.

Cependant, le modèle **Decision Tree** a eu un **meilleur rappel** pour les patients diabétiques, ce qui est important si on veut éviter les faux négatifs.

KNN a été le moins performant dans ce cas, peut-être à cause du déséquilibre des classes et du fait qu'il est très sensible à l'échelle des données.

## 7. Conclusion

Ce mini-projet m'a permis de pratiquer concrètement l'ensemble des étapes d'un projet de machine learning, de l'analyse des données jusqu'à la comparaison de plusieurs modèles.

J'ai choisi de retenir le modèle **Logistic Regression**, car il est simple, efficace et donne de bons résultats.

À l'avenir, j'aimerais tester des modèles plus puissants comme les **Random Forests** ou encore des méthodes d'optimisation comme la **validation croisée** ou le **tuning d'hyperparamètres**.