Rapport du Projet de Machine Learning

Nom : Amina Nagaz

Sujet : Prédiction du diabète

1. Objectif :

Construire un modèle de machine learning supervisé pour prédire si une personne est diabétique ou non, à partir de ses données médicales.

2. Données utilisées

- Dataset : Pima Indians Diabetes (Kaggle)
- 768 lignes, 9 colonnes
- Variable cible : Outcome (0 = non diabétique, 1 = diabétique)

3. Traitement des données

- Remplacement des zéros invalides par des valeurs manquantes (NaN)
- Imputation avec la médiane
- Normalisation des données

4. Algorithmes utilisés :

a. Régression logistique

-Modèle de classification binaire

-Simple, rapide, efficace

b. KNN (K-Nearest Neighbors)

- Basé sur la distance entre les individus
- Sensible aux échelles, nécessite une normalisation
- K fixé à 5

5. Résultats obtenus :

Évaluation via validation croisée (StratifiedKFold, 5 plis) :

Modèle	Précision moyenne
Régression logistique	0.5914
KNN (k = 5)	0.5576

6. Conclusion :

La régression logistique a donné les meilleurs résultats pour ce problème de classification binaire. Le modèle est plus stable et plus fiable que KNN selon les métriques obtenues