Rapport du Projet de Machine Learning

Nom : Amina Nagaz

Sujet : Prédiction du diabète

1. Objectif :

Ce projet a pour objectif de construire un modèle supervisé de Machine Learning permettant de prédire si une personne est atteinte de diabète en se basant sur ses données médicales. J'ai utilisé deux modèles: la régression logistique et l'algorithme KNN. J'ai comparé leurs performances et choisi le modèle le plus efficace.

2. Source des données :

- La Source : Kaggle

Variable cible : Outcome (0 = non diabétique, 1 = diabétique)

3. Analyse exploratoire des données :

-Plusieurs colonnes contiennent des zéros (valeurs manquantes cachées)

-Répartition équilibrée : 50% non diabétiques / 50% diabétiques

-Glucose est la variable la plus corrélée avec la présence de diabète

4. Prétraitement :

-Remplacement des zéros par NaN.

-Imputation des valeurs manquantes avec la médiane (SimpleImputer).

-Normalisation des variables numériques (StandardScaler).

-Utilisation d'un Pipeline pour automatiser le traitement.

5. 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

6. Résultats obtenus :

Évaluation via validation croisée (StratifiedKFold, 5 plis) :

Modèle	Précision moyenne
Régression logistique	74.88 %
KNN (k = 5)	69.45 %

Matrice de confusion et rapport de classification (test set) :

- Régression logistique : meilleure précision globale
- KNN : plus de faux négatifs

7. 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