Analyse prédictive à l'aide de la régression et du KNN

Réalisé par Khalil El Achkhem et Ouiza SADI OUFELLA

Table of Contents

hexagone

Introduction

Objectif du projet

L’objectif principal de ce projet est d’utiliser différentes techniques de machine learning supervisé pour réaliser des prédictions à partir de deux jeux de données réels : l’un médical (sur les maladies cardiaques), l’autre sportif (statistiques de joueurs NBA).

Introduction

Problématique

Dans des domaines aussi variés que la santé ou le sport, la capacité à prédire une issue ou une performance à partir de données peut s'avérer cruciale.

Introduction

Problématique

Peut-on prédire si une personne est atteinte d'une maladie cardiaque simplement à partir de son âge ou d'autres données médicales ?

Peut-on estimer le totale de points marques par un joueur NBA à partir de ses statistiques de jeu ?

Introduction

Solution proposée

- **Régression linéaire simple pour prédire une valeur continue à partir d'une seule variable**
- **Régression linéaire multiple pour prédire une valeur continue à partir de plusieurs variables**
- **Régression logistique pour modéliser la probabilité d'un événement binaire**
- **K-Nearest Neighbors (KNN) pour classer les individus sur la base de la proximité avec des cas similaires**

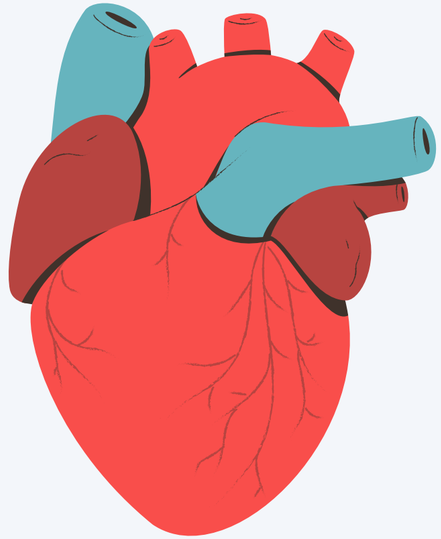
Chaque méthode a été testée sur un jeu de données spécifique pour évaluer sa pertinence et son efficacité

Jeux de données

Données sur les maladies cardiaques

Ce dataset médical contient des informations sur des patients, incluant :

- age : Âge du patient
- thalach : Fréquence cardiaque maximale atteinte
- target : Présence de maladie cardiaque (1 = oui, 0 = non)
- Et d’autres variables : cholestérol, pression artérielle, type de douleur thoracique, etc.



Jeux de données

Données NBA

Ce dataset statistique regroupe les performances de joueurs de NBA pour une saison donnée. Les colonnes incluent :

- PTS : Points par match (notre variable cible)
- MP : Minutes jouées par match
- 3PA : Tentatives de tirs à 3 points
- AST : Passes décisives
- FG : Field Goals réussis
- Age : l'âge du joueur
- Et bien d'autres statistiques de performance

Régression Linéaire Simple

Cette étude s'intéresse à la relation linéaire entre le nombre total de points marqués (PTS) et le nombre de tirs réussis (FG) dans les performances des joueurs de la NBA.
Pour cela, une régression linéaire simple est utilisée afin de modéliser cette relation et de prédire les FG à partir de PTS.

Hypothèses statistiques :

- **H_0 (hypothèse nulle) :** Il n'existe aucune relation linéaire significative entre PTS et FG.

- **H_1 (hypothèse alternative) :** Il existe une relation linéaire significative entre PTS et FG.

Avant de modéliser, une analyse de corrélation a été réalisée :

- **Coefficient de corrélation de Pearson : 0.991,** indiquant une corrélation très forte et positive entre les deux variables.
- **p-value : 0.000,** ce qui signifie que la corrélation est hautement significative (on rejette H_0).

Régression Linéaire Multiple

L’objectif est de prédire le nombre de goals Le nombre de points marqués en moyenne par match (PTS) par un joueur NBA, à partir de plusieurs statistiques de performance.

Sélection des variables explicatives (features) :

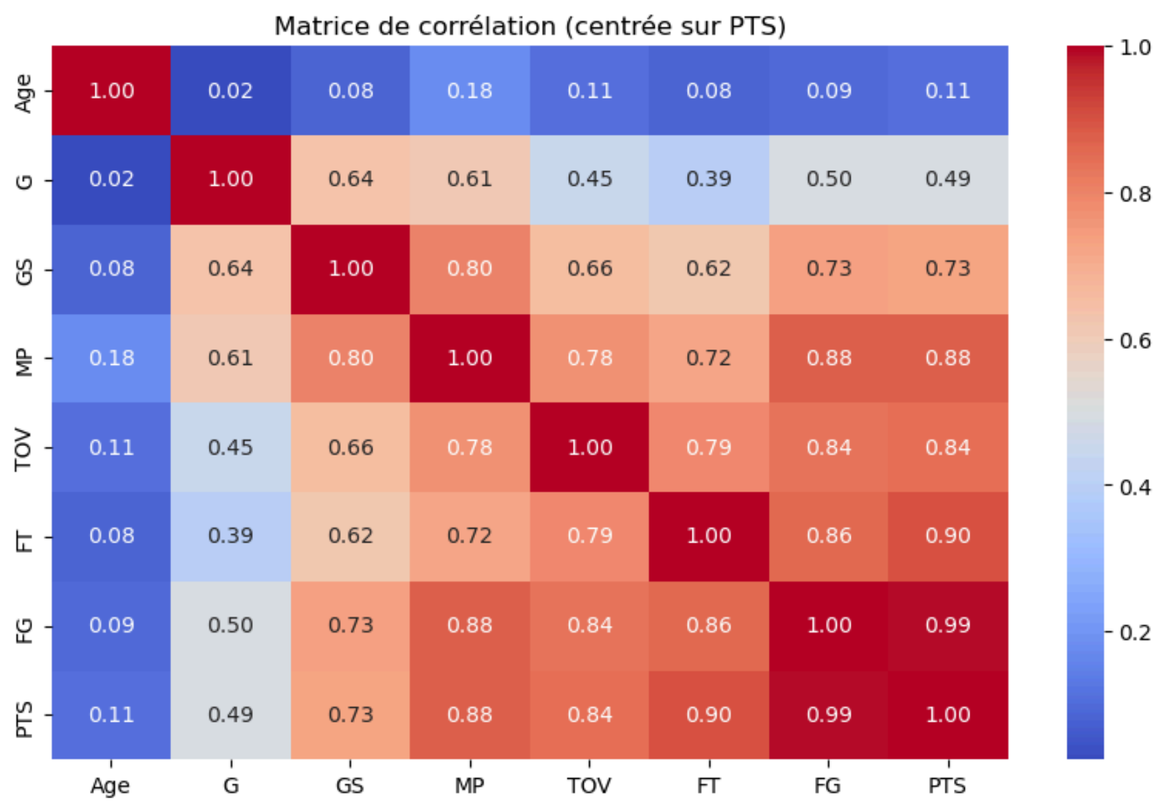
Le choix des variables explicatives a été fait en deux étapes :

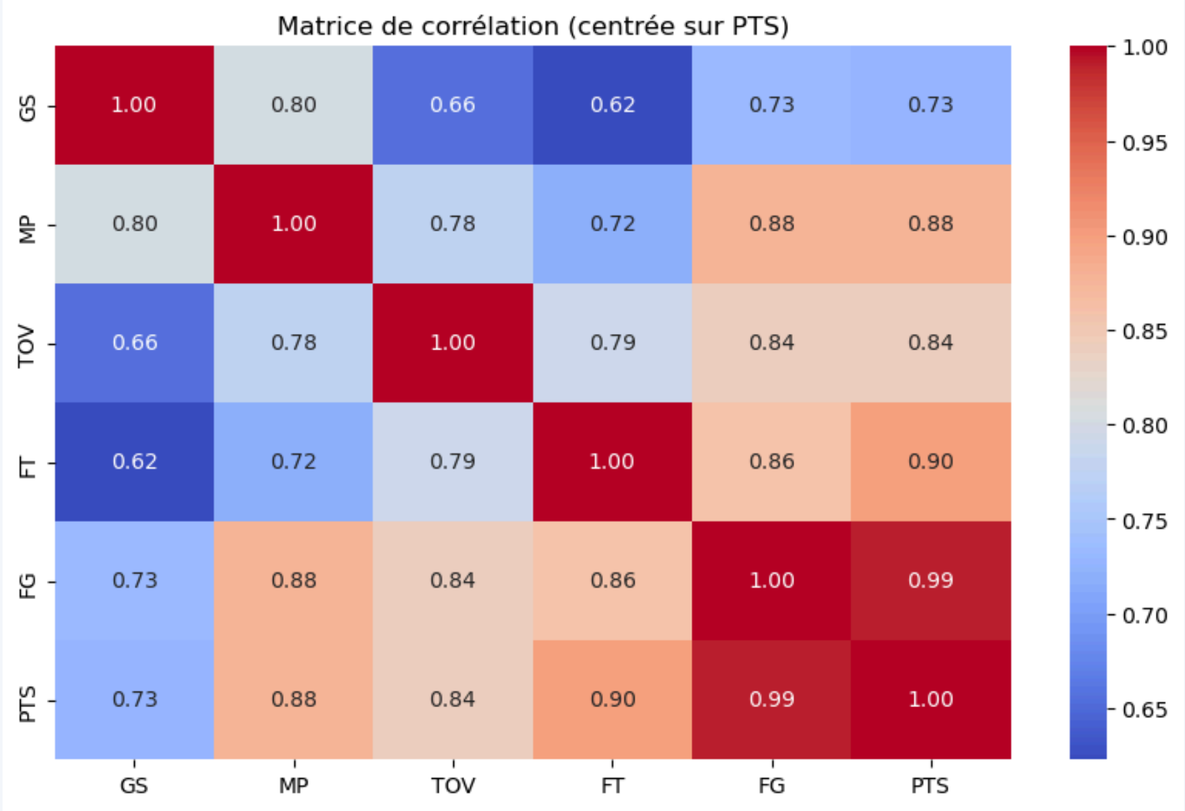
1. Analyse de corrélation : On a choisi des variables ayant une bonne corrélation linéaire avec la variable cible PTS. **Exemple :** MP (minutes jouées), FG (tentatives de tir).
2. Analyse du VIF (Variance Inflation Factor) :

Pour éviter les problèmes de multicolinéarité (variables fortement corrélées entre elles), on a calculé le VIF de chaque variable.

- Seules les variables avec un VIF < 10 ont été conservées
- Cela garantit une meilleure stabilité du modèle et une interprétation plus fiable des coefficients.

Régression Linéaire Multiple





Régression Linéaire Multiple

hexagone

```
Variance Inflation Factors (VIF):

  Feature       VIF
0   const  7.386440
1      GS  2.844373
2      MP  5.933031
3     TOV  3.748191
4      FT  4.235352
5      FG  8.811105
```

- GS Le nombre de matchs commencés
- MP Le nombre moyen de minutes jouées par match.
- FG Le nombre de tirs réussis au total
- FT Le nombre de lancers francs réussis
- ORB Le nombre de rebonds offensifs
- TOV Le nombre de ballons perdu
- PTS Le nombre de points marqués en moyenne par match

Régression Logistique

Prédire la probabilité qu'un patient soit atteint d'une maladie cardiaque à partir de ses caractéristiques cliniques

Variable cible : target où 1 indique la présence de la maladie, et 0 son absence.

Sélection des variables explicatives (features) :
Le modèle s'appuie sur plusieurs variables explicatives clés : l'âge, le sexe, la fréquence cardiaque maximale atteinte (thalach), ainsi que des indicateurs liés aux anomalies cardiaques comme la dépression du segment ST (oldpeak), sa pente (slope), ou encore le nombre de vaisseaux détectés (ca). D'autres informations comme le type de douleur thoracique (cp) et les résultats du test thalium (thal) enrichissent l'analyse.

Type de prédiction : classification binaire

K-Nearest Neighbors (KNN)

L’objectif de cette étude est de prédire la présence ou l’absence d’une maladie cardiaque à l’aide d’un algorithme de classification K-Nearest Neighbors (KNN).
Le modèle repose sur le principe selon lequel un individu est classé en fonction des catégories majoritaires parmi ses k voisins les plus proches.

Le jeu de données utilisé comprend plusieurs variables cliniques telles que :

- **l’âge (**age**)**
- **le type de douleur thoracique (**cp**)**
- **la fréquence cardiaque maximale atteinte (**thalach**)**
- **le cholestérol (**chol**)**
- **le nombre de vaisseaux colorés (**ca**)**

Une normalisation des données a été effectuée avant l’apprentissage, et le modèle a été entraîné avec k = 2-8.

Technologies utilisées

Python

Scikit-learn

Matplotlib, Seaborn

Pandas, NumPy

Résultats et Discussion

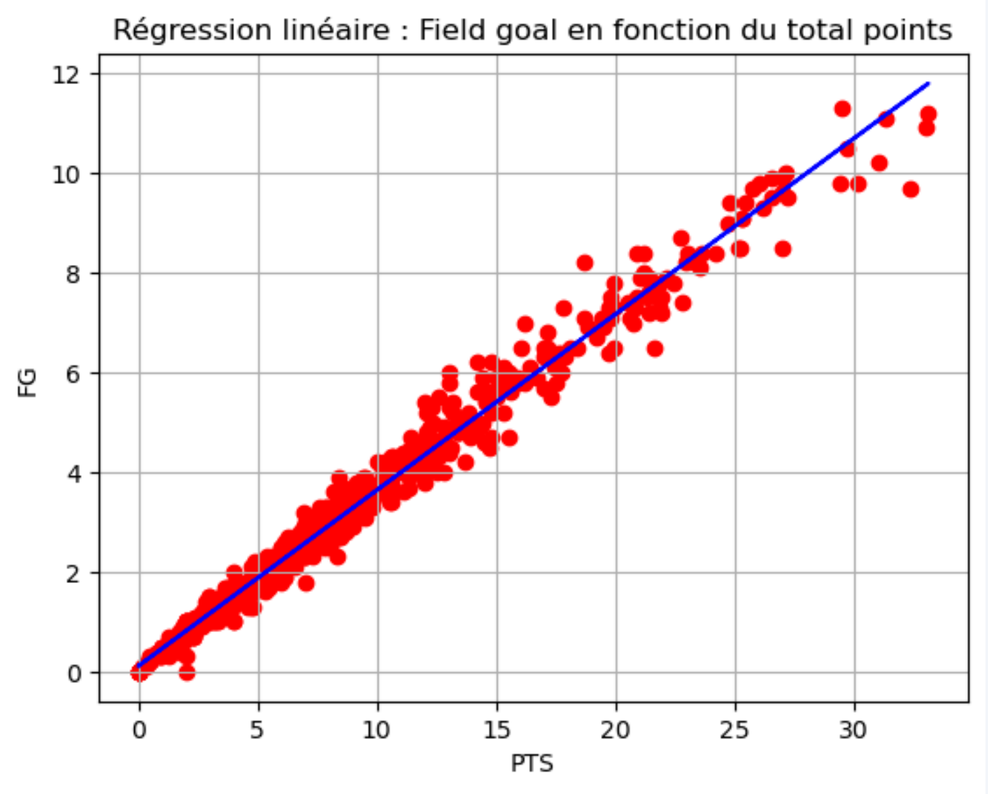
Régression Linéaire Simple

Le modèle de régression linéaire a été entraîné sur un échantillon aléatoire de 80 % des données, et testé sur les 20 % restants.

Évaluation du modèle sur le jeu de test :

- **R^2 Score : 0.978 →** le modèle explique 97.8 % de la variance des tirs réussis (FG) par les points (PTS)
- **MAE : 0.227 →** en moyenne, le modèle se trompe de 0.227 tirs
- **RMSE : 0.332 →** faible erreur quadratique, bon ajustement global, les prédictions sont à 0.33 tirs près des valeurs réelles.

La visualisation des prédictions montre une droite de régression bien alignée avec les données, confirmant la bonne qualité du modèle.



Le modèle révèle une relation presque linéaire parfaite entre PTS et FG, ce qui est logique dans un contexte sportif. Néanmoins, des variables complémentaires (comme les minutes jouées ou les matchs joués) pourraient enrichir l'analyse dans une version multiple du modèle.

Résultats et Discussion

hexagone

Régression Multiple

Le modèle présente des performances impressionnantes, avec un R^2 de 0.99, ce qui signifie qu'il explique presque toute la variation des points marqués (PTS). Les variables les plus influentes sont les minutes jouées (MP) et les lancers francs (FT). Le modèle montre également que plus de pertes de balle (TOV) réduisent les chances de marquer, tandis que plus de tirs réussis (FG) augmentent considérablement le nombre de points.

Avec une erreur moyenne d'environ 0.44 point par joueur, le modèle est très précis dans ses prédictions des PTS, bien qu'il y ait toujours une petite marge d'erreur.

```
# Exemple avec une observation
valeurs = [77, 32, 10, 3, 8]
#[ 'GS', 'MP','TOV','FT', 'FG']
prediction = predict_y(valeurs)

print("Prédiction de PTS :", prediction)
```

✓ 0.0s

```
Prédiction de PTS : 19.717337038682174
```

$$Y = \text{Intercept} + \beta_1 \cdot X1 + \beta_2 \cdot X2 + \beta_3 \cdot X3$$

```
Model Coefficients: [-0.00811325  0.05760015 -0.14114568  0.84347501  2.21365376]
Model Intercept: -0.329345708709031
R² Score (Accuracy): 0.9906751732628261
Mean Squared Error (MSE): 0.37818012890136965
Root Mean Squared Error (RMSE): 0.6149635183499665
Mean Absolute Error (MAE): 0.43690629473862147
```

Résultats et Discussion

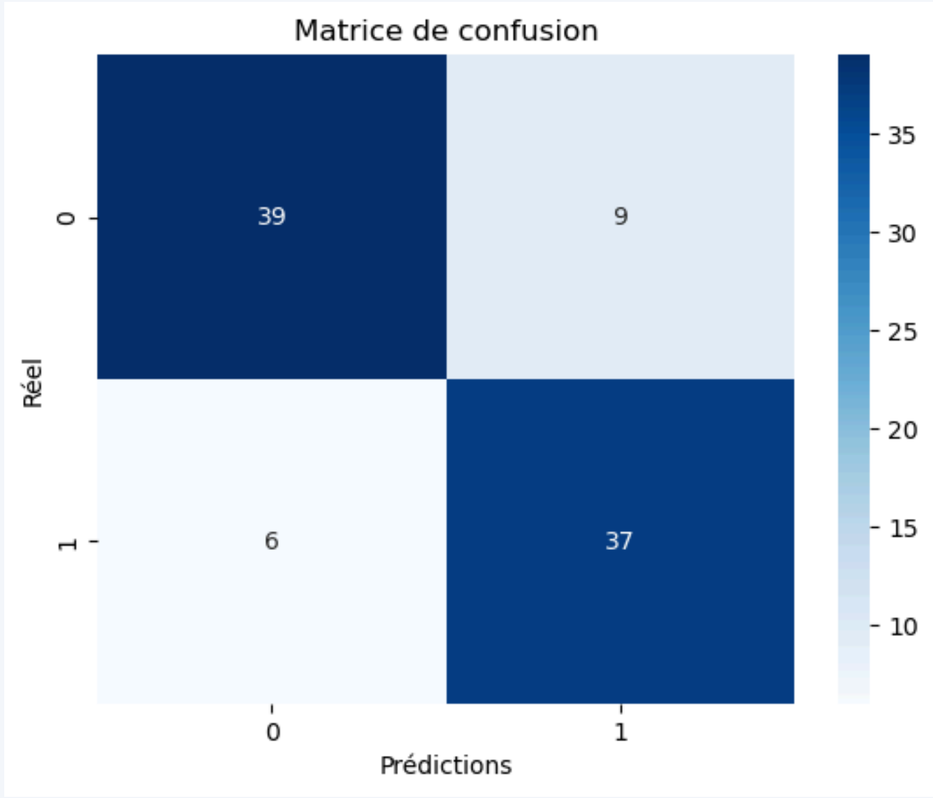
Régression Logistique

Le modèle a été évalué sur un échantillon test de 91 observations. Voici les résultats détaillés :

Précision globale (Accuracy) : 84.0 %

Matrice de confusion :

- 39 vrais négatifs (classe 0 bien prédite)
- 37 vrais positifs (classe 1 bien prédite)
- 9 faux positifs (prédit 1 alors que c'est 0)
- 6 faux négatifs (prédit 0 alors que c'est 1)



Résultats et Discussion

hexagone

Régression Logistique

Classe	Précision	Recall	F1-Score	Support
0	0.87	0.81	0.84	48
1	0.80	0.86	0.83	43

```
Profil 1: Probabilité prédite = 21.36%
Profil 2: Probabilité prédite = 37.86%
Profil 3: Probabilité prédite = 94.70%
```

Moyennes :

- **Macro avg : Précision = 0.84, Recall = 0.84, F1 = 0.84**

Discussion :

Le modèle présente un bon équilibre entre précision et rappel, ce qui est essentiel en contexte médical.

Il est légèrement plus performant pour détecter les cas positifs (classe 1) que négatifs, ce qui peut être favorable si l'objectif est de limiter les cas non détectés de maladies.

Donc :

L'âge (> 40) et le sexe influencent globalement le risque, mais le cœur ne ment pas sur ses symptômes : ce sont les signes cliniques mesurables (comme la douleur thoracique, la fréquence cardiaque, la réponse à l'effort...) qui ont un poids beaucoup plus fort dans la détection immédiate d'un problème cardiaque.

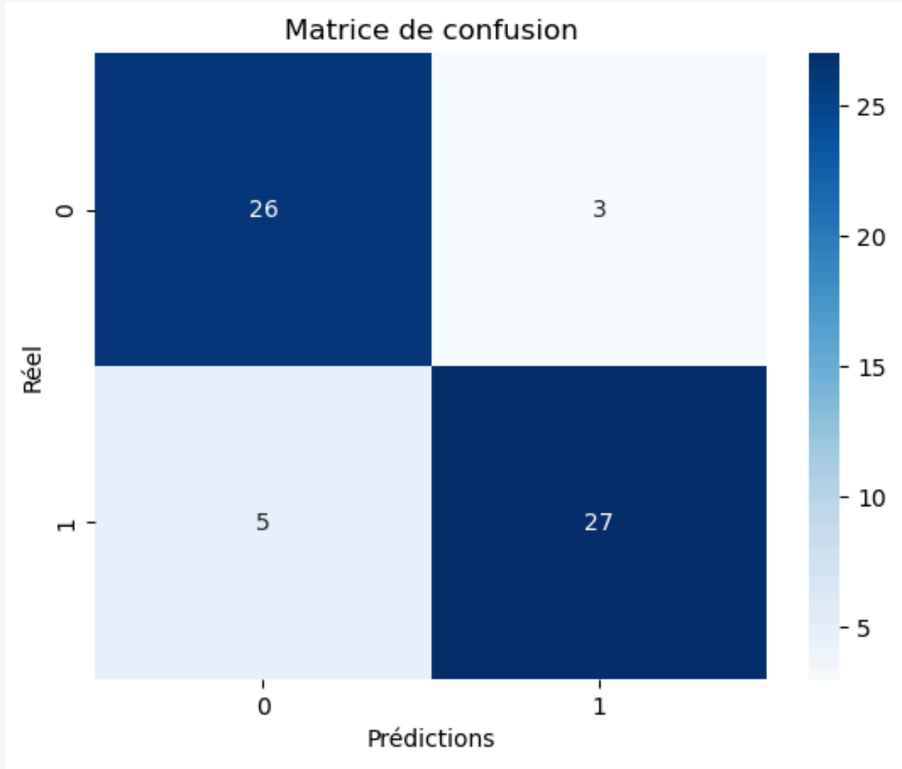
Résultats et Discussion

KNN

Le modèle KNN a été évalué sur un jeu de test avec k = 8. Voici les résultats détaillés :
Précision globale (Accuracy) : 86.89 %

Matrice de confusion :

- 26 vrais négatifs (classe 0 bien prédite)
- 27 vrais positifs (classe 1 bien prédite)
- 3 faux positifs (prédit 1 alors que c'est 0)
- 5 faux négatifs (prédit 0 alors que c'est 1)



Résultats et Discussion

KNN

Classe	Précision	Recall	F1-Score	Support
0	0.84	0.90	0.87	29
1	0.90	0.84	0.87	32

Moyennes :

- **Macro avg : Précision = 0.87, Recall = 0.87, F1 = 0.87**

Discussion :

Le modèle est équilibré entre les deux classes, avec une légère asymétrie : il est meilleur pour détecter les cas positifs (classe 1) que négatifs.

L'ajout de métriques comme le recall permet de vérifier qu'on ne "rate" pas trop de vrais cas positifs (important en médecine).

Conclusion

Chaque méthode a sa propre utilité :

- **La régression linéaire pour les prédictions continues simples**
- **La logistique pour les classifications binaires**
- **Le KNN pour les prédictions multivariées et intuitives**

Le choix du modèle dépend fortement du type de données et de la nature de la variable cible.

Passons à la démonstration