Prédiction de la maladie cardiaque à partir de données cliniques

Projet de machine learning supervisé

Contexte & Problématique

- Les maladies cardiovasculaires sont l'une des premières causes de mortalité mondiale.
- Identifier précocement les patients à risque permet de prévenir des complications graves.
- **Question posée :** Peut-on prédire la présence d'une maladie cardiaque à partir de simples données cliniques (âge, tension, cholestérol, ECG, etc.) ?

FR

Objectifs du projet

- Construire un **modèle prédictif fiable et interprétable**.
- Comparer plusieurs algorithmes de classification (Logistic Regression, Random Forest, Gradient Boosting).
- Identifier les **facteurs cliniques clés** associés au risque.
- Fournir une **aide à la décision médicale** simple à exploiter.

Données utilisées

- **Source :** *Heart Disease Dataset – UCI / Kaggle*
- 1025 patients, 14 variables cliniques, dont :age, sex, trestbps, chol, thalach, oldpeak, etc.
- **Variable cible** : target (1 = malade, 0 = sain)
- Données équilibrées, peu de valeurs manquantes.

Présentation des données

	age	Sex	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	1	1	106	0	1.9	1	3	2	0
5	58	0	0	100	0	0	122	0	1.0	1	0	2	1
6	58	1	0	114	0	2	140	0	4.4	0	3	1	0
7	55	1	0	160	0	0	145	1	0.8	1	1	3	0
8	46	1	0	120	0	0	144	0	0.8	2	0	3	0
9	54	1	0	122	0	0	116	1	3.2	1	2	2	0

Méthodologie

- **Préparation des données :**
  - Standardisation des variables numériques
  - Encodage OneHot des variables catégorielles
- **Modélisation :**
  - Régression Logistique
  - Random Forest (tuning via GridSearchCV)
  - Gradient Boosting
- **Évaluation :**
  - Accuracy, F1, Recall, ROC-AUC
  - Matrices de confusion
- **Split 70/30** (train/test, stratifié)

Résultats

Modèle	Accuracy	F1	ROC-AUC
Régression Logistique	0.87	0.88	0.94
Gradient Boosting	0.96	0.97	0.98
Random Forest (Best)	0.98	0.98	0.995

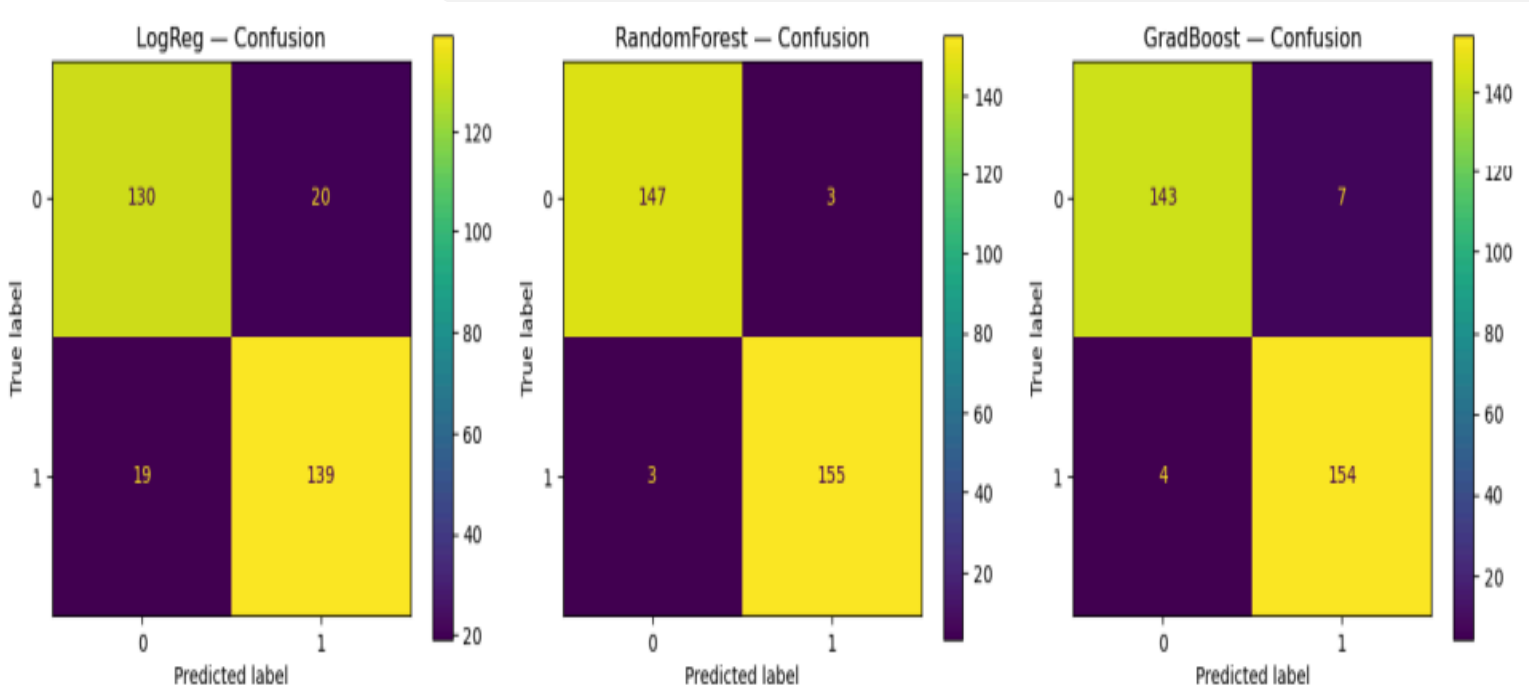
Interprétation

- **Variables les plus influentes :**
- thalach (fréquence cardiaque max)
- oldpeak (anomalies ECG)
- cp (type de douleur thoracique)
- age et ca (vaisseaux obstrués)

Ces facteurs sont cohérents avec la littérature médicale sur les risques cardiovasculaires.

Visualisation & Confusion

- Trois matrices de confusion côte à côte (LogReg, RF, GB).
- Commente brièvement :
  - peu de faux négatifs avec Random Forest → bon rappel clinique.
  - modèle fiable pour le dépistage préventif.



Recommandations & perspectives

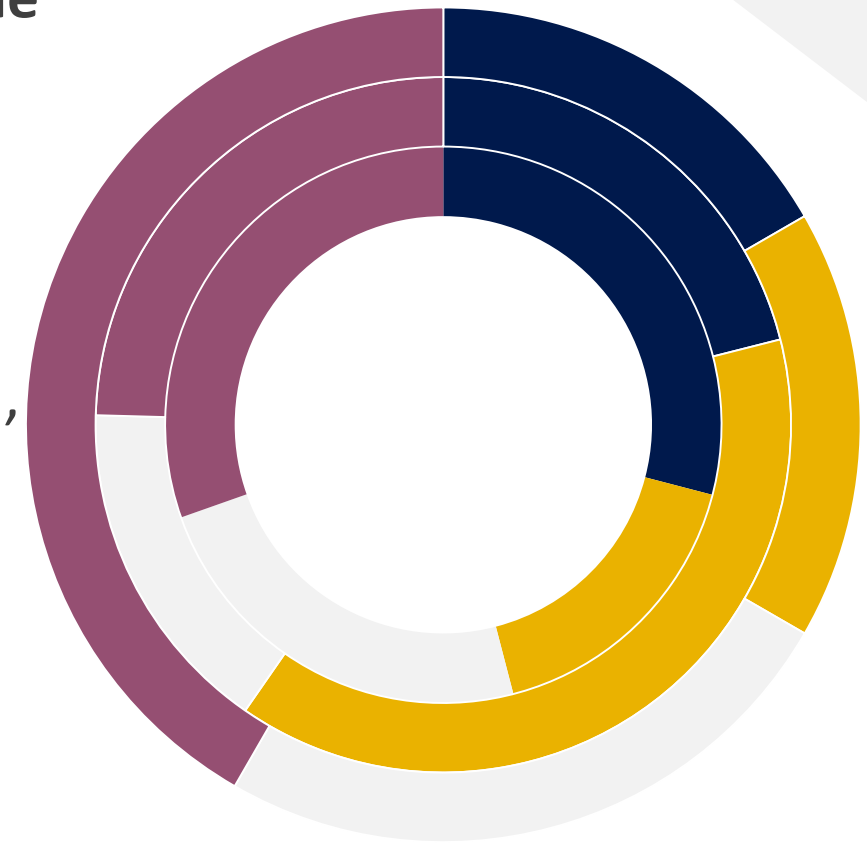
- Intégrer le modèle dans un **outil d'aide à la décision** hospitalier.
- Adapter le **seuil de décision** pour privilégier le *recall* (mieux vaut prévenir qu'omettre un malade).
- Étendre à des données réelles locales (Haïti, structures de santé publique).
- Utiliser **SHAP/LIME** pour interpréter patient par patient.

Conclusion

Le projet montre qu'un **modèle supervisé simple** peut prédire la maladie cardiaque avec une précision proche de 98 %.

Les **facteurs clés identifiés** sont cliniquement plausibles.

Approche **éthique et explicable** : l'IA en soutien, pas en remplacement du diagnostic médical.



Merci.

Amee Hashley JEUDY

+ 509 4846-4262

ameehashleyjeudy@gmail.com

https://www.linkedin.com/in/amee-hashley-jeudy-460449325