PRÉDICTION DES MALADIES CARDIAQUES

avec scikit-learn

GIULIA GOVERNATORI

LE POINT DE DÉPART

Pourquoi ce projet ?

Les maladies cardiovasculaires sont la première cause de décès dans le monde :

17,9 millions *de décès par an (OMS)*

Souvent évitables avec un diagnostic précoce...

La question :

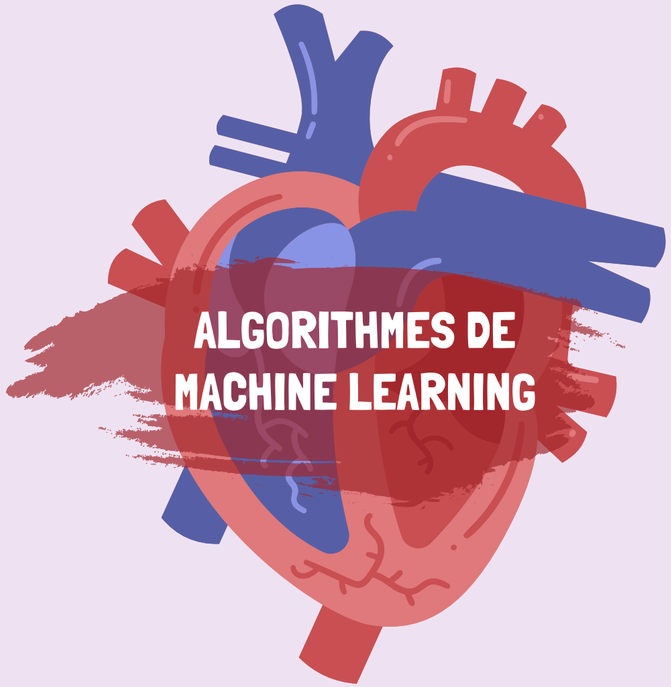
Peut-on prédire le risque en utilisant les données cliniques ?

NOTRE MISSION

Objectif du Projet

Créer un système intelligent qui aide à identifier les patients à risque :

DONNÉES CLINIQUES DU PATIENT



PROBABILITÉ DE MALADIE
+ NIVEAU DE RISQUE

LES DONNÉES

Le Dataset UCI Heart Disease

Qualité excellente : zéro valeur manquante !

303 patients de 4 hôpitaux internationaux : → 📋 13 variables cliniques par patient

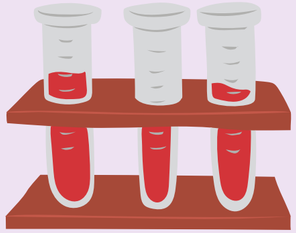
160 personnes saines | 137 avec maladie

QUE MESURONS-NOUS ?

Les 13 Variables Cliniques

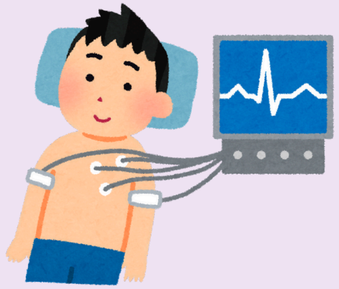
Données démographiques :

- Âge
- Sexe



Examens de base :

- Pression artérielle
- Cholestérol
- Glycémie à jeun



Tests avancés :

- ECG au repos
- Fréquence cardiaque maximale à l'effort
- Type de douleur thoracique
- Angine induite par l'effort
- Dépression du segment ST à l'effort (oldpeak)
- Pente du segment ST
- Nombre de vaisseaux colorés (angiographie)
- Thalassémie (anomalie sanguine)

LE DÉFI INITIAL

Un Problème à Résoudre

Le dataset avait 5 classes de maladie →

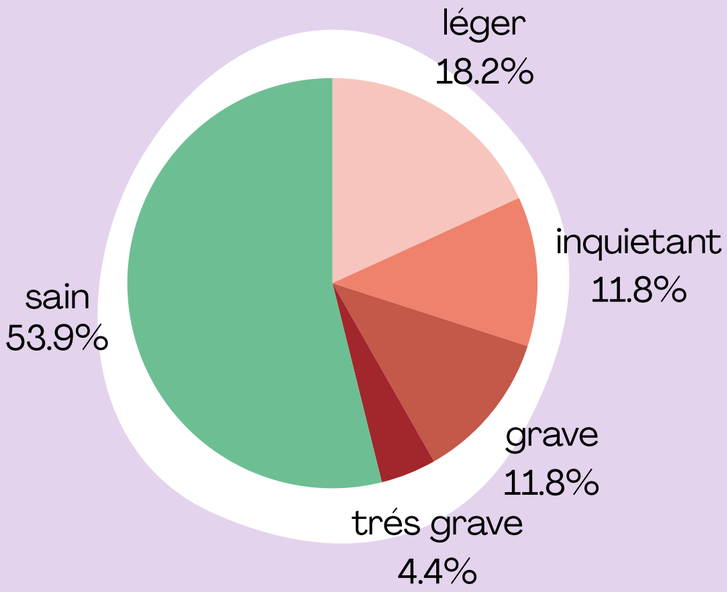
- Classe 0 : Sain (160 patients)
- Classes 1-4 : Gravité croissante

❌ **PROBLÈME :**

Classes trop déséquilibrées et seulement 303 observations

↓

Modèle qui n'apprend pas des petites classes



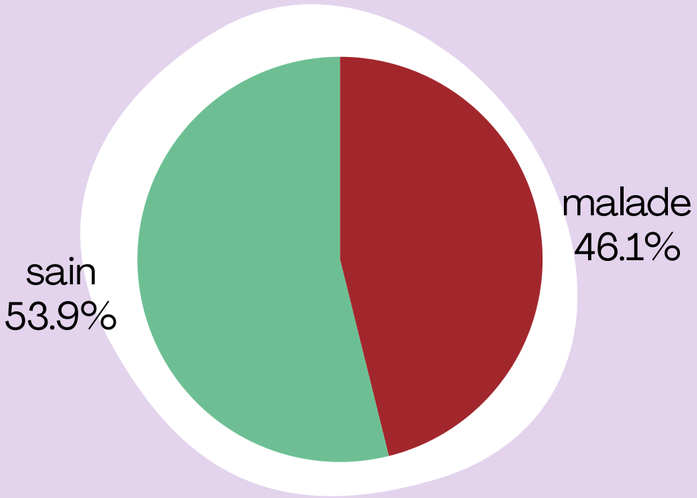
→

✅ **SOLUTION :**

Transformation en problème binaire

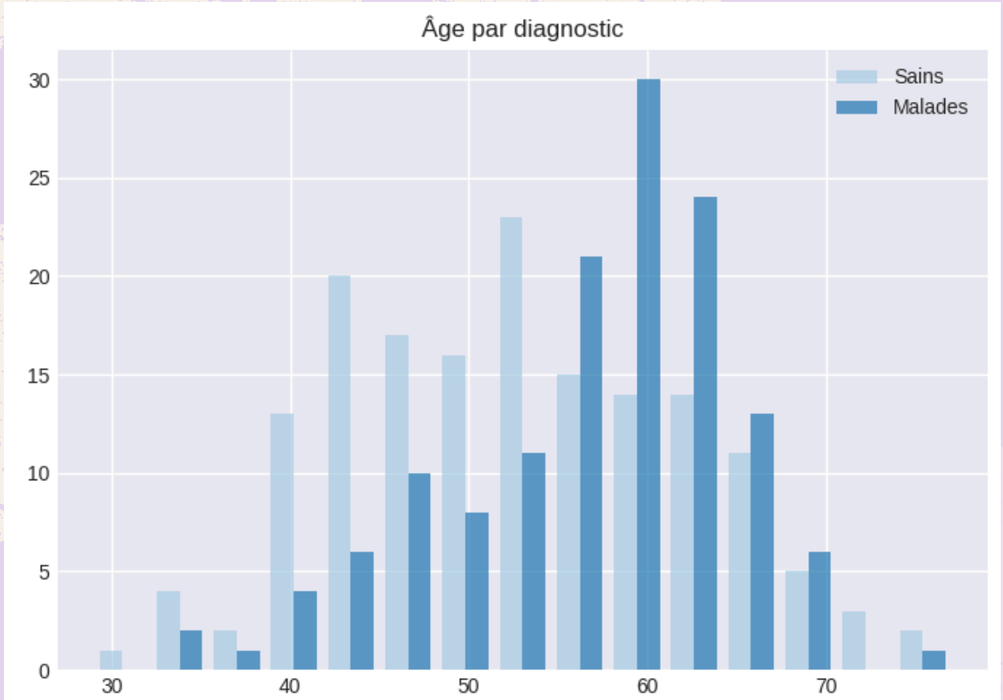
↓

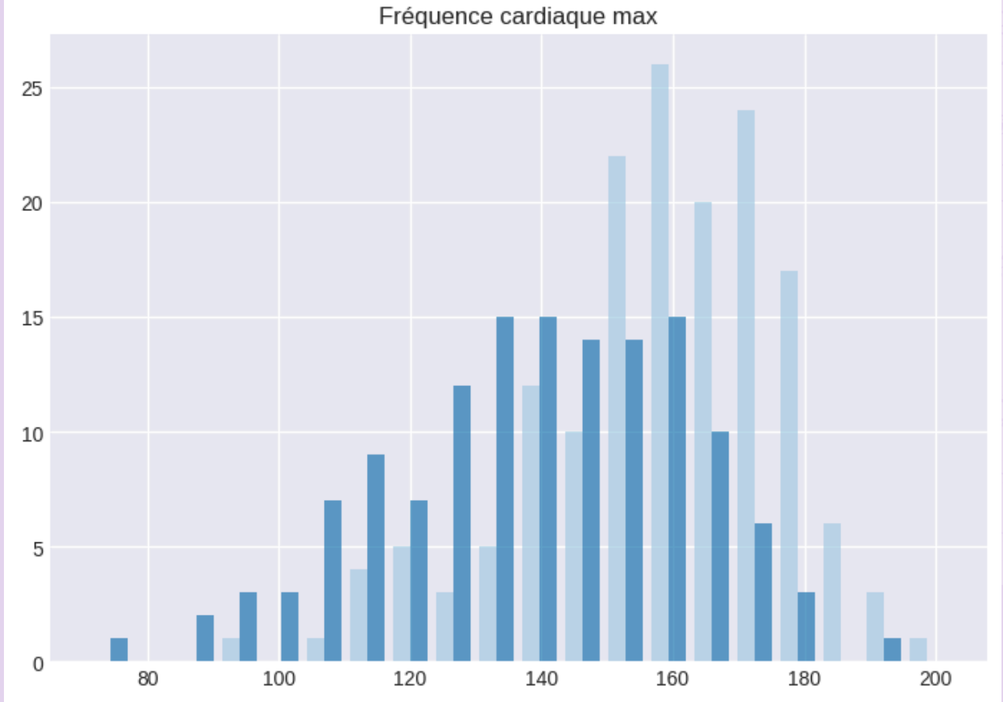
Sain vs Malade → classes équilibrées

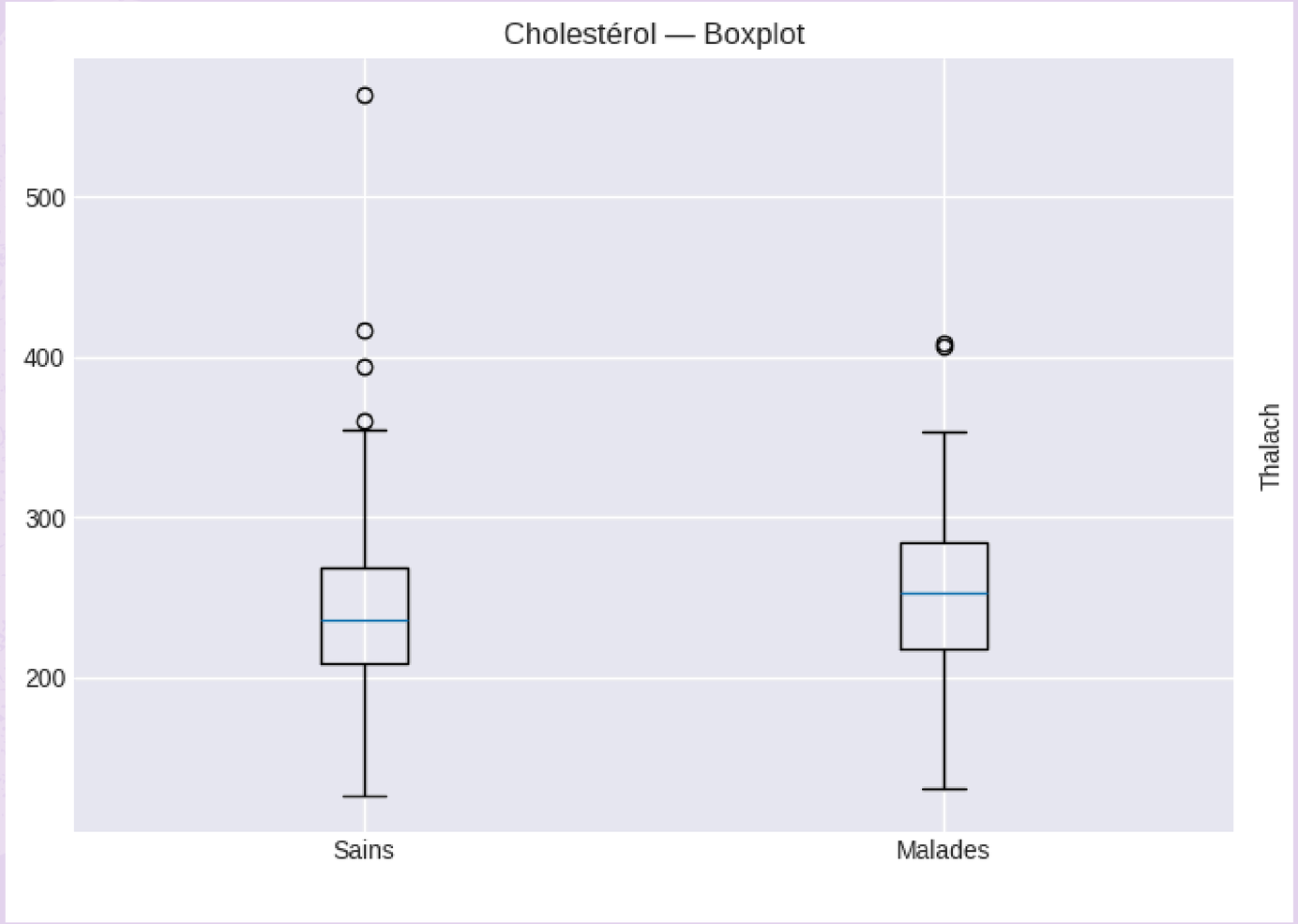


EXPLORER LES DONNÉES

Découvertes clés de l'analyse exploratoire







PRÉPARER LES DONNÉES

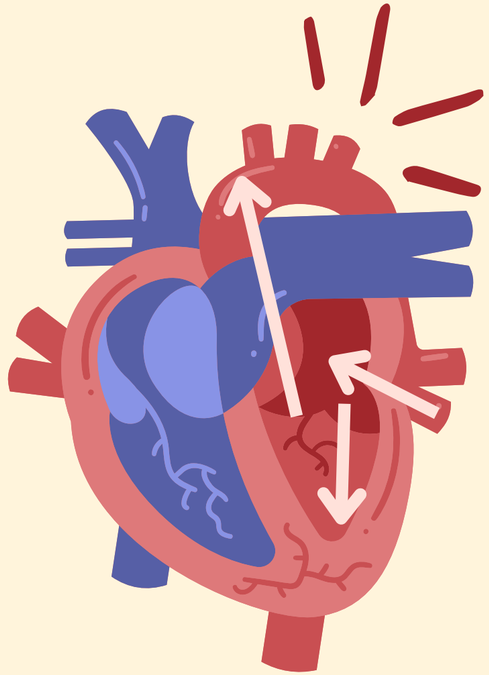
Le Preprocessing

STEP 1:

Train-Test Split + Cross-Validation

La structure à 3 niveaux :

📦 Dataset complet (303 patients)
- 🎓 **80% Training Set** (237 patients)
  - Divisé en **5** "**fold**" pour la ***Cross-Validation***
  - → Fold 1, 2, 3, 4, 5 alternent comme validation
  - → Utilisé pour choisir les meilleurs hyperparamètres
- 🧪 **20% Test Set** (60 patients)
  - → JAMAIS touché jusqu'à l'évaluation finale !



PRÉPARER LES DONNÉES

Le Preprocessing

STEP 1:

Train-Test Split + Cross-Validation

Comment fonctionne la validation croisée à 5 plis (5-Fold Cross-Validation) :

Training Set → divisé en 5 parties égales
Val → Training set usé comme Validation

Round 1: [**Val**] [**Train**] [**Train**] [**Train**] [**Train**]
Round 2: [**Train**] [**Val**] [**Train**] [**Train**] [**Train**]
Round 3: [**Train**] [**Train**] [**Val**] [**Train**] [**Train**]
Round 4: [**Train**] [**Train**] [**Train**] [**Val**] [**Train**]
Round 5: [**Train**] [**Train**] [**Train**] [**Train**] [**Val**]

Moyenne des 5 résultats = Performance estimée

Pourquoi cela évite le **DATA SNOOPING** :

✅ Test set = vu une seule fois à la fin
✅ Cross-Validation = “test” toujours sur le “training set”
✅ Décisions basées uniquement sur le Training Set
✅ Estimation honnête de la capacité de généralisation

🔒 Stratification : les proportions sains/malades (54 %/46 %) sont maintenues dans chaque Fold

PRÉPARER LES DONNÉES

Le Preprocessing

STEP 2:

Standardisation des nombres

→ Toutes les variables sur la même échelle

STEP 3:

Encodage des catégories

→ Transformer le texte (de classes) en nombres compréhensibles par la machine (0/1)

PREMIER MODÈLE – RANDOM FOREST

La Forêt qui Apprend

Qu'est-ce qu'un Random Forest ?

Imaginez 100 docteurs qui donnent leur opinion :

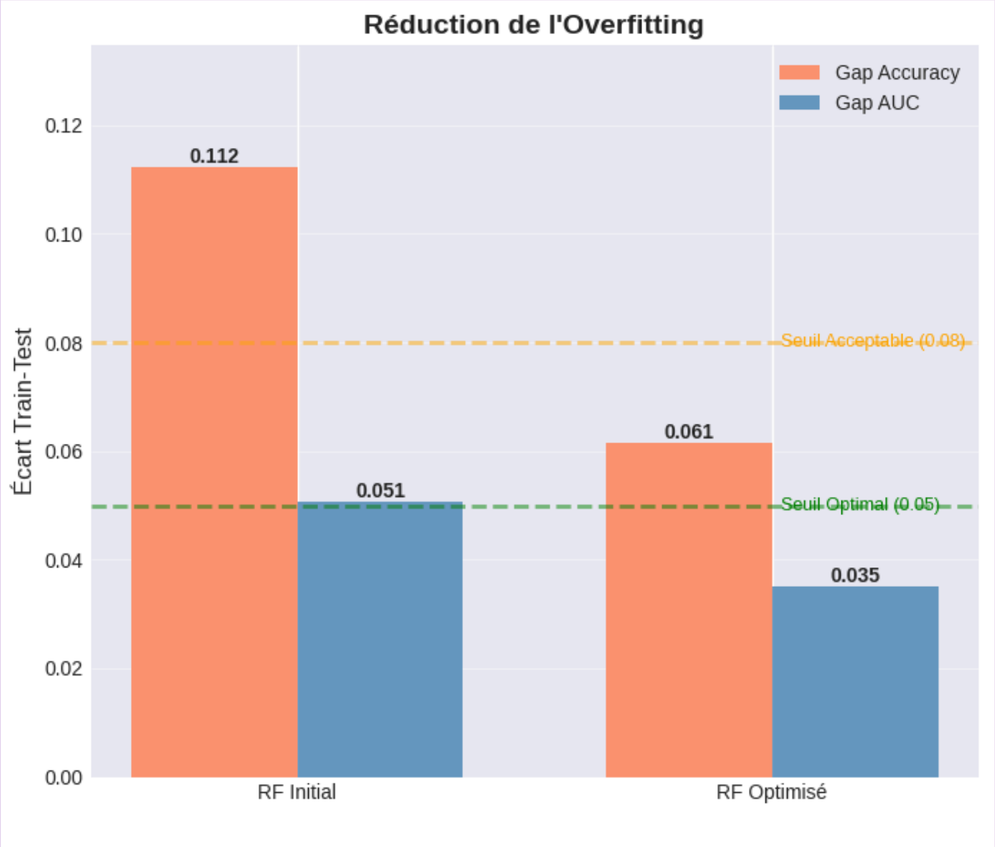
- Chaque docteur examine le patient sous un angle différent
- Puis ils votent ensemble pour le diagnostic final
- La majorité l'emporte !

Résultat initial :

- ✅ Précision test : 86,7%
- ❌ Mais... trop de "mémorisation" (***overfitting***)

OPTIMISATION

Overfitting du Random Forest : Un Problème à Résoudre



Le problème : Le modèle Random Forest apprend "par cœur" au lieu de comprendre.

Les solutions appliquées :

1. 🌳Limiter la profondeur des arbres pour éviter des structures trop complexes.
2. 🌿Exiger plus d'observations dans chaque feuille pour empêcher l'apprentissage de cas isolés.
3. 🎲Introduire davantage d'aléatoire dans le choix des variables pour diversifier les arbres.
4. 🌲Augmenter le nombre total d'arbres afin de stabiliser les prédictions par moyennage.
5. 🧪Exiger un plus grand nombre d'exemples avant de diviser un nœud pour réduire les séparations inutiles.
6. ⚓Activer le rééquilibrage automatique des classes pour limiter l'influence d'une classe dominante.

Résultat après optimisation :

- **Précision :** 85% (similaire)
- **Overfitting :** réduit de plus 30% ! ✨

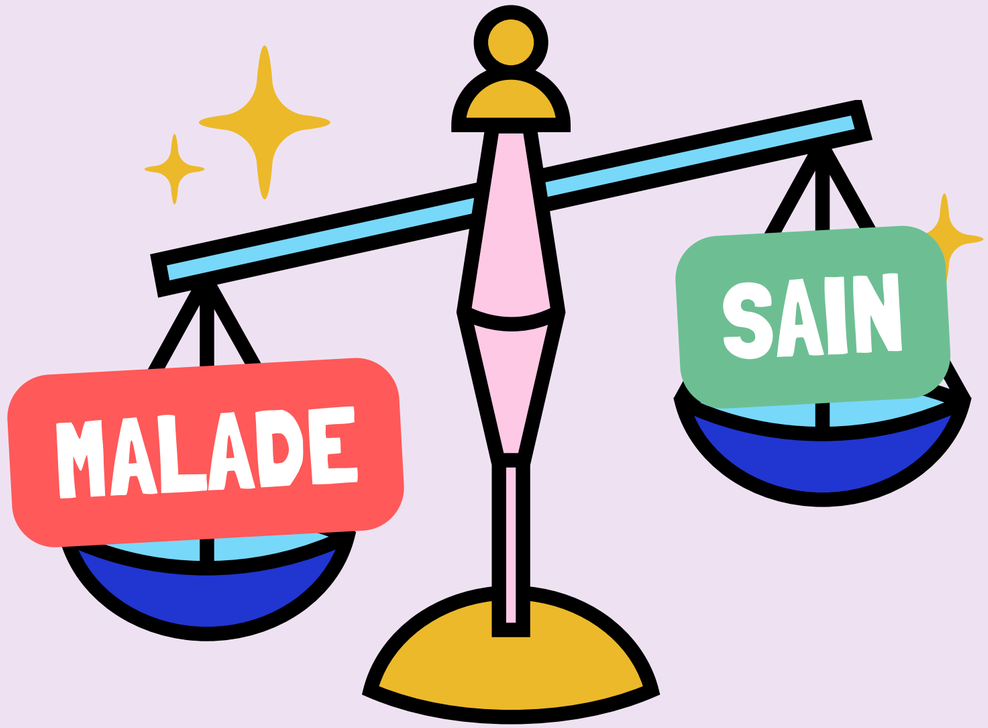
DEUXIÈME MODÈLE – RÉGRESSION LOGISTIQUE

L'Approche Linéaire

Un modèle plus simple mais puissant.

Pensez à une balance qui pèse l'importance de chaque symptôme :

- Douleur thoracique → +3 points de risque
- Âge élevé → +2 points
- Cholestérol bas → -1 point
- Total > seuil = maladie

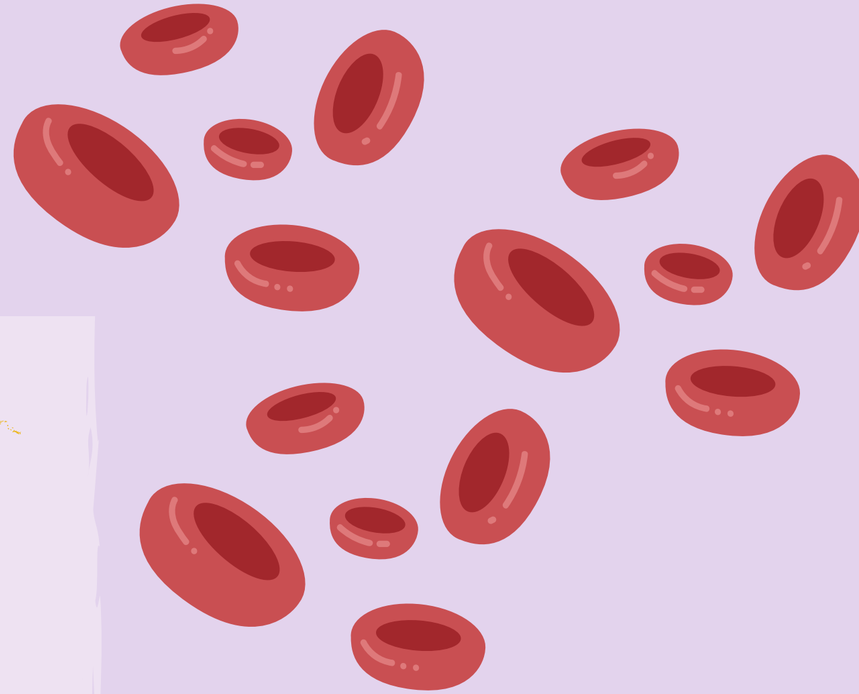


Résultat :

- ✅ Précision test : 86,7%
- ✅ Zéro overfitting !
- ✅ Facilement interprétable

LE MEILLEUR DES DEUX MONDES

Stacking Ensemble : combiner les deux modèles !



Niveau 1 : Random Forest fait des prédictions + Régression Logistique fait des prédictions

Niveau 2 : Un troisième modèle (Stacking) apprend des 2 et décide

C'est comme demander conseil à :

- Un expert intuitif (Random Forest)
- Un analyste méthodique (Régression Logistique)
- Un directeur qui synthétise les deux (Méta-modèle)

LA COMPARAISON FINALE

Qui Gagne ?

Gagnant pour ce projet : **Stacking**
→ Meilleure discrimination globale (**AUC = 0,953**)

Critère	Random Forest	Logistic Regression	Stacking	Gagnant
AUC Test	0.946	0.951	**0.953**	🏆 Stacking
Accuracy	0.867	**0.867**	0.833	🏆 RF/LR
Precision Malade	0.88	0.92	0.88	🏆 LR
Recall Malade	0.82	0.79	0.75	🏆 RF
Overfitting	Modéré	✅ Aucun	Aucun	🏆 LR/Stack
Interprétabilité	Faible	✅ Élevée	Moyenne	🏆 LR

QU'EST-CE QUI COMPTE VRAIMENT ?

Importance des Variables

Top 5 facteurs prédictifs (62% du pouvoir prédictif) :

🫀 **Type de douleur thoracique**

15,8%

🩸 **Thalassémie**

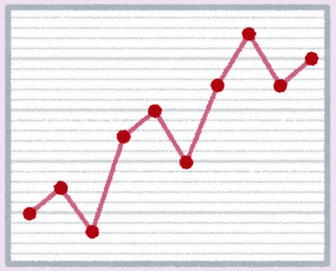
15,0%

💓 **Fréquence cardiaque max sous effort**

11,1%

🔬 **Vaisseaux sanguins colorés**

10,4%



📉 **Dépression ST**

9,8%

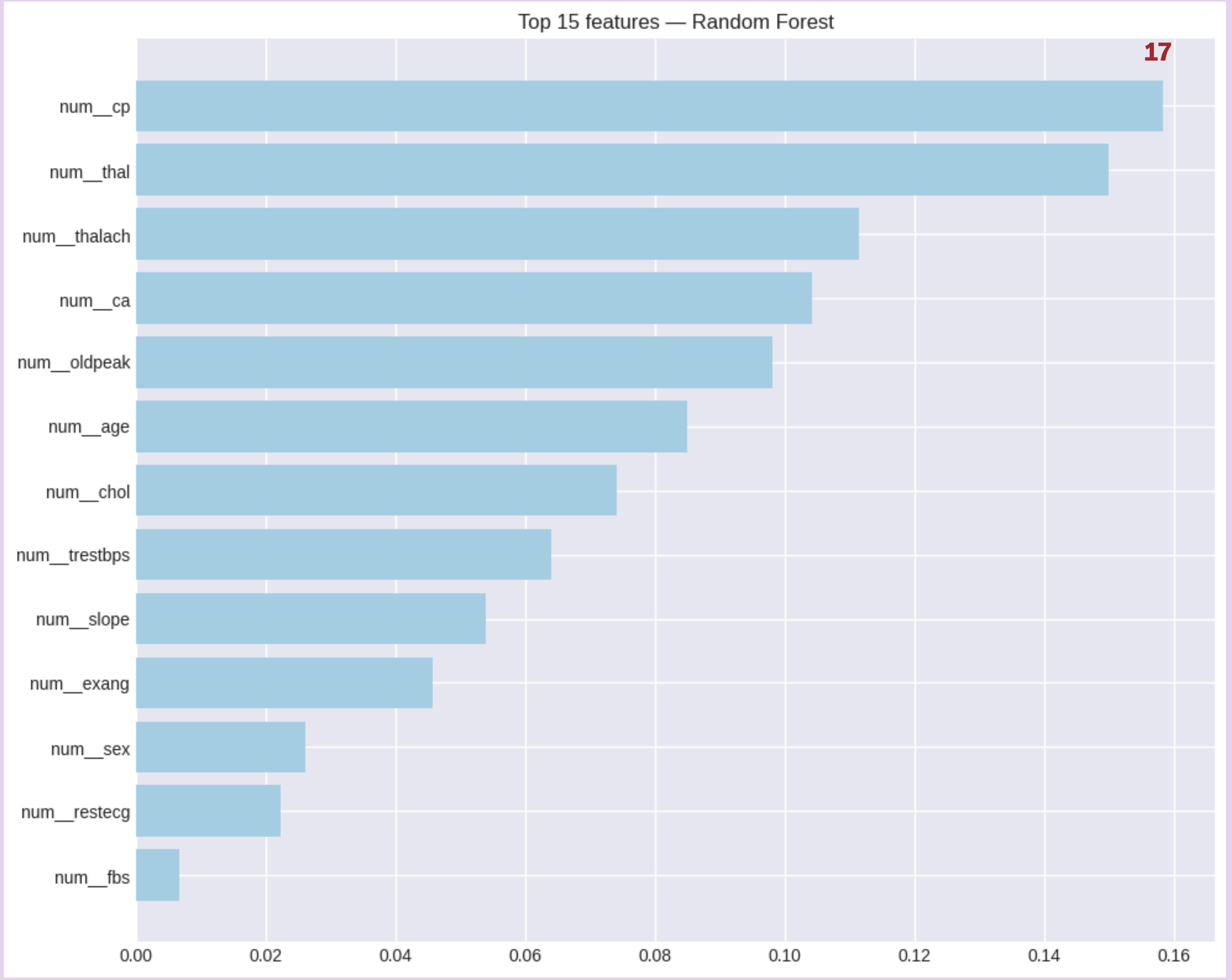
L'analyse de l'importance des variables dans le modèle Random Forest révèle que cinq caractéristiques dominent la prédiction du risque cardiovasculaire.

QU'EST-CE QUI COMPTE VRAIMENT ?

Importance des Variables

Variables les Plus Prédictives

Variable	Importance (%)
`cp` (type de douleur thoracique)	15,8%
`thal` (thalassémie)	15,0%
`thalach` (fréquence cardiaque maximale)	11,1%
`ca` (nombre de vaisseaux colorés)	10,4%
`oldpeak` (dépression ST)	9,8%



DU MODÈLE À L'APPLICATION

Le Dashboard What-If

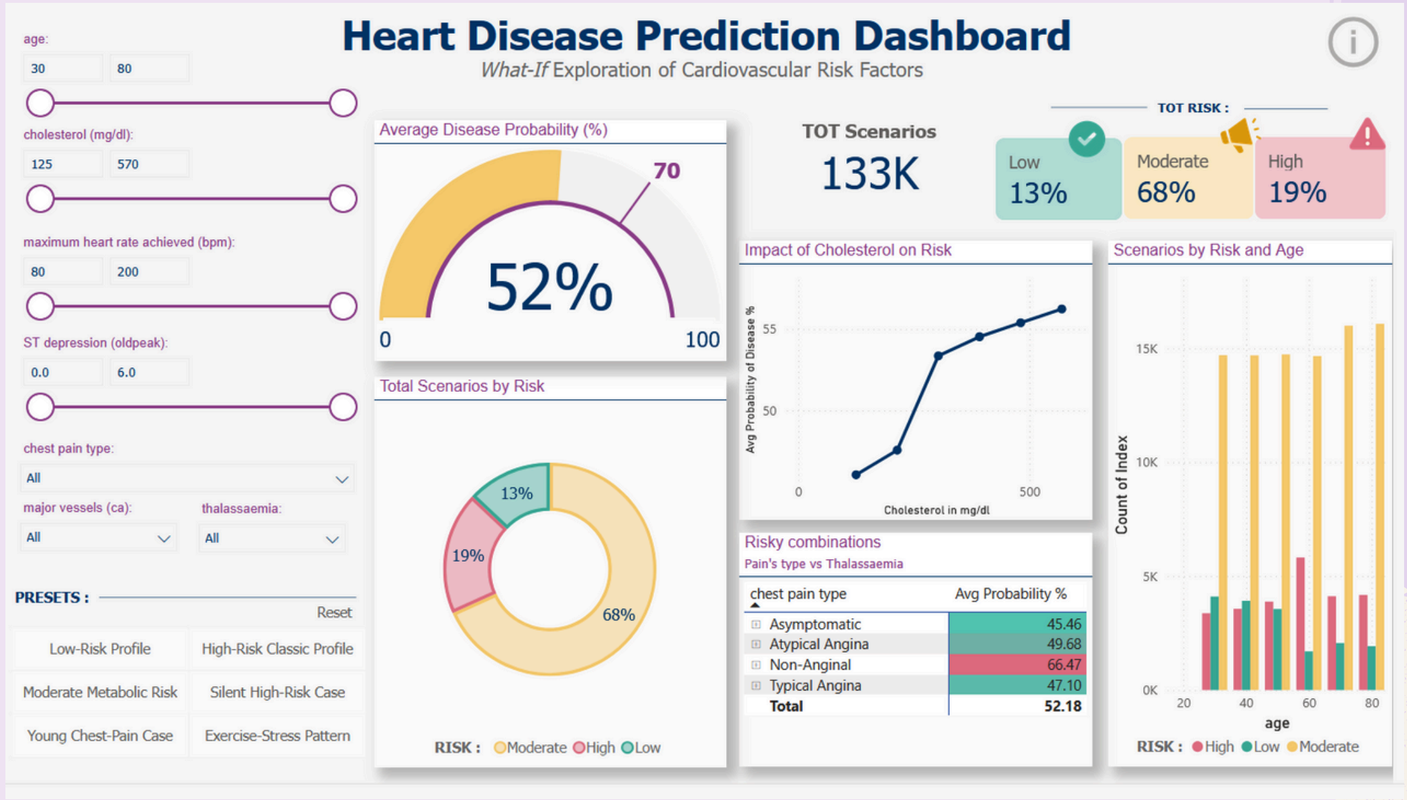
Nous avons créé 133 056 scénarios simulés !

Comment ?
En combinant toutes les possibilités :

- *Âge : 30-80 ans*
- *Cholestérol : 125-570 mg/dl*
- *Fréquence cardiaque : 80-200 bpm*
  - *4 autres variables cliniques*

Résultat :

Un simulateur interactif dans ***Power BI*** *où les utilisateurs peuvent explorer des scenarios "what if..."*



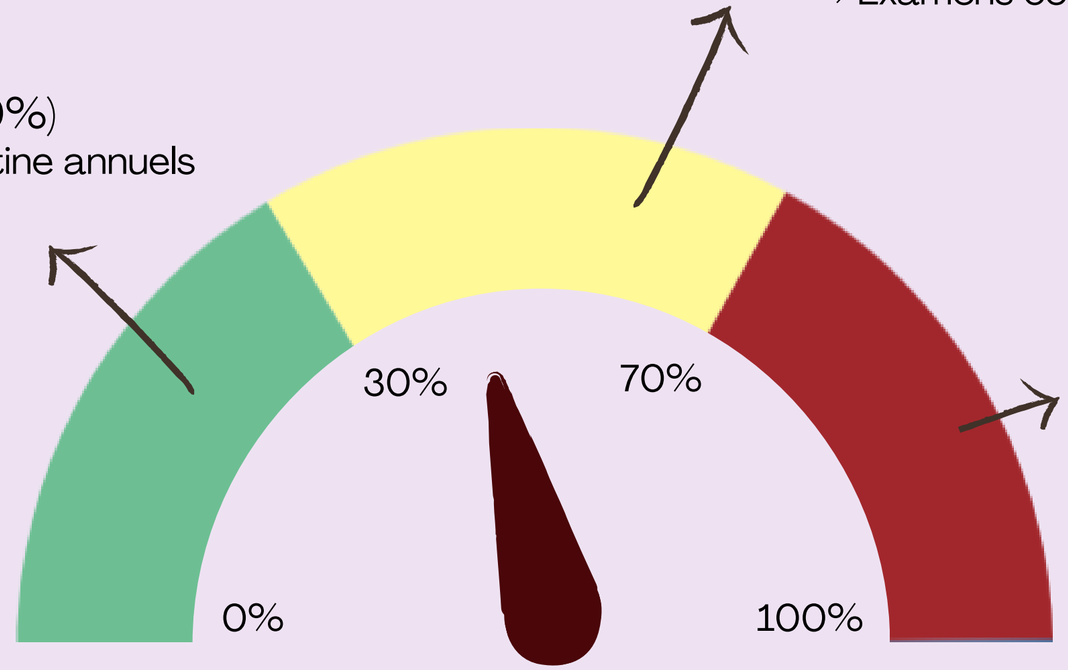
CATÉGORIES DE RISQUE

Communiquer de Façon Claire

Nous avons traduit les probabilités en actions concrètes :

Risque MODÉRÉ (30-70%)
→ Examens complémentaires conseillés

Risque BAS (< 30%)
→ Contrôles de routine annuels



Risque ÉLEVÉ (≥ 70%)
→ Intervention cardiologique immédiate

CAS PRATIQUE

Exemple d'Utilisation

Patient type :

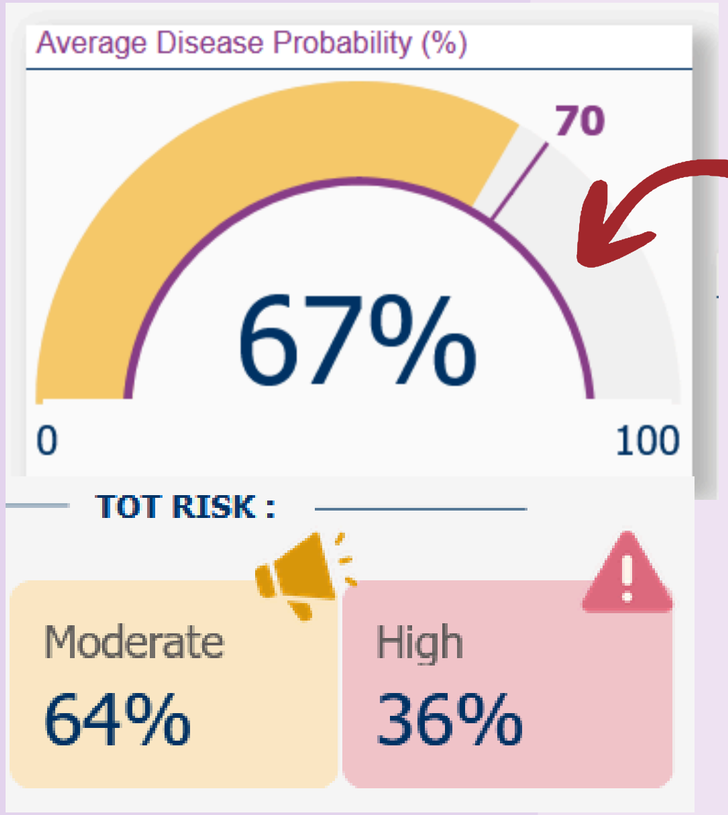
- 👤 Homme, **55-65** ans
- 🩺 Cholestérol **240-450** mg/dl
- ❤️‍🔥 Fréquence max **80-120** bpm
- 😣 Douleur thoracique: **typique anginal**
- 🔬 Thalassémie **defeaut fixe**

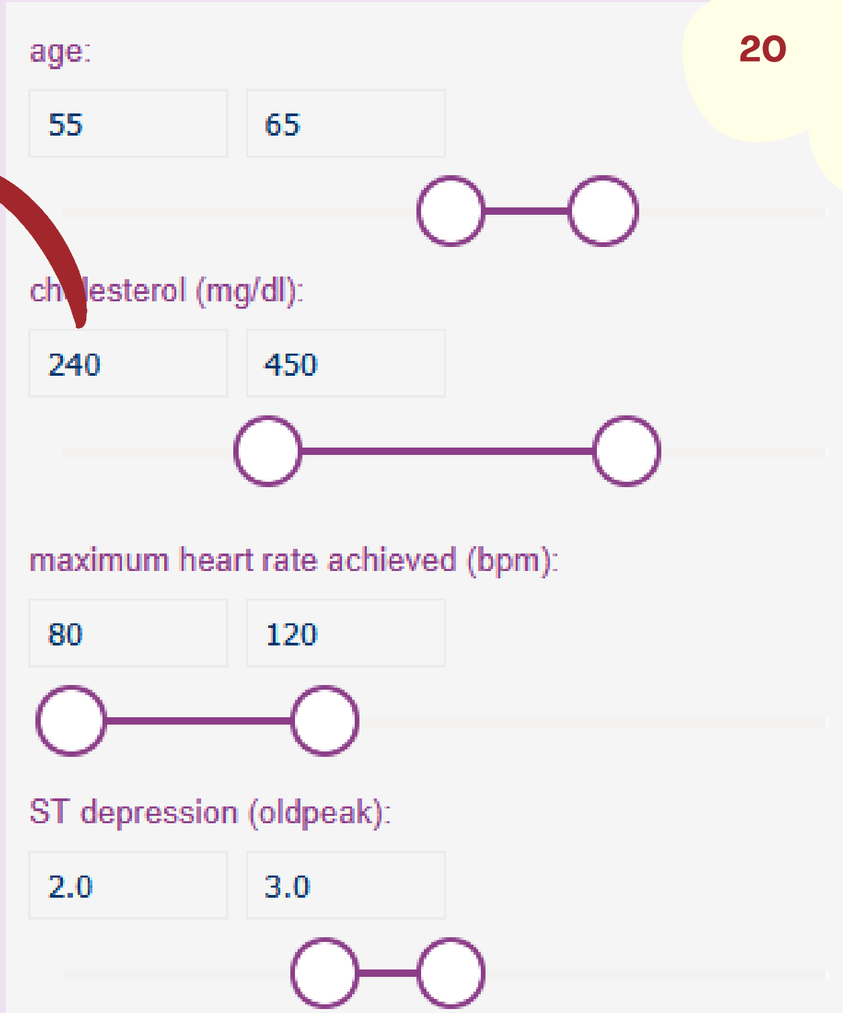
Prédiction du modèle :

→ **Probabilité maladie : 67%**

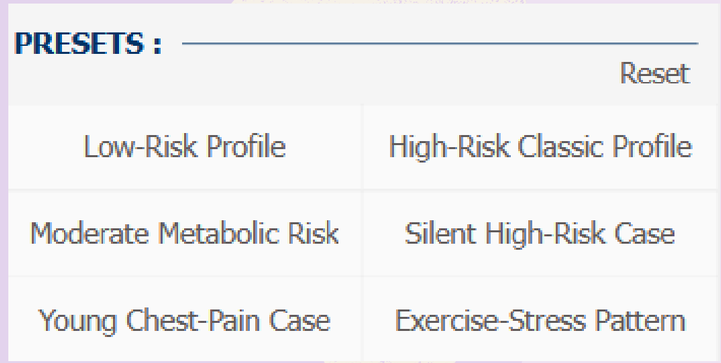
→ Catégorie : **RISQUE MODÉRÉ** 🟡 pour le **64%**

→ Action : **Échocardiogramme + test d'effort**





J'ai également prévu des préréglages pour faciliter l'exploration des données.



CE QUE NOUS AVONS APPRIS

Les Succès du Projet

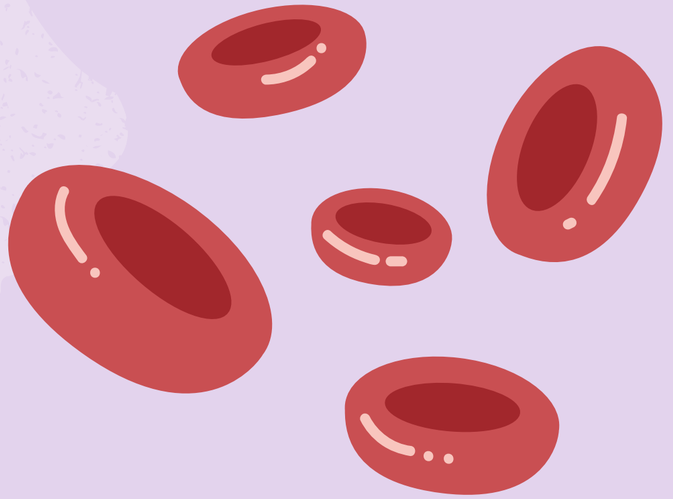
✅ **Performance excellente** : AUC 0,953 (presque parfait)

✅ **Modèle robuste** : Zéro overfitting dans le modèle final

✅ **Interprétabilité** : Nous savons quelles variables comptent

✅ **Applicabilité** : Dashboard prêt pour usage clinique

✅ **Méthodologie** : Approche rigoureuse et réplicable



LES DÉFIS ET LES LIMITES

Être Honnêtes

⚠️ *LIMITATIONS À CONSIDÉRER :*

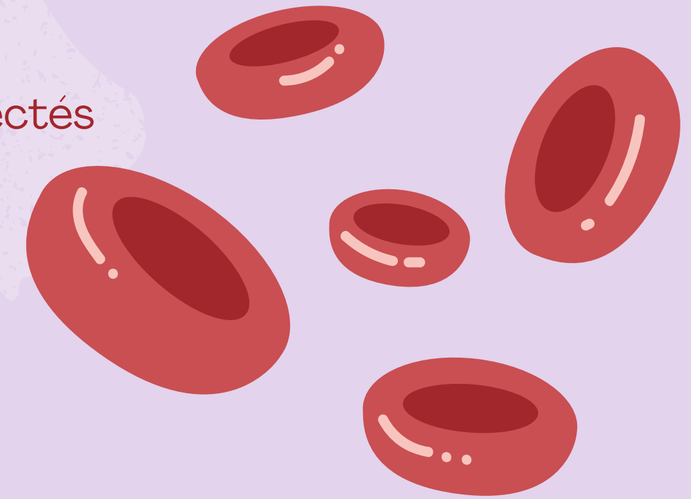
📊 **Dataset petit** : 303 patients → besoin de plus de données

🌍 **Contexte géographique** : Seulement 4 hôpitaux

⚡ **Recall classe Malades** : 75-82% → certains malades non détectés

🔄 **Validation** : Besoin de test sur nouveaux patients réels

Ces limites n'invalident pas le projet, mais indiquent où s'améliorer !



LE FUTUR DU PROJET

Prochaines Étapes

Le modèle actuel ne peut pas encore être utilisé sur des patients réels.

Ce projet nécessiterait quelques améliorations.

Que pouvons-nous faire mieux ?

🔬 Plus de données : Objectif 1000+ patients

🌍 Diversité géographique : Inclure plus de populations

🧬 Nouvelles variables : Histoire familiale, mode de vie, génétique

🤖 Deep Learning : Si nous avons assez de données (plus de 300)

🔄 Étude prospective : Valider sur nouveaux patients dans le temps

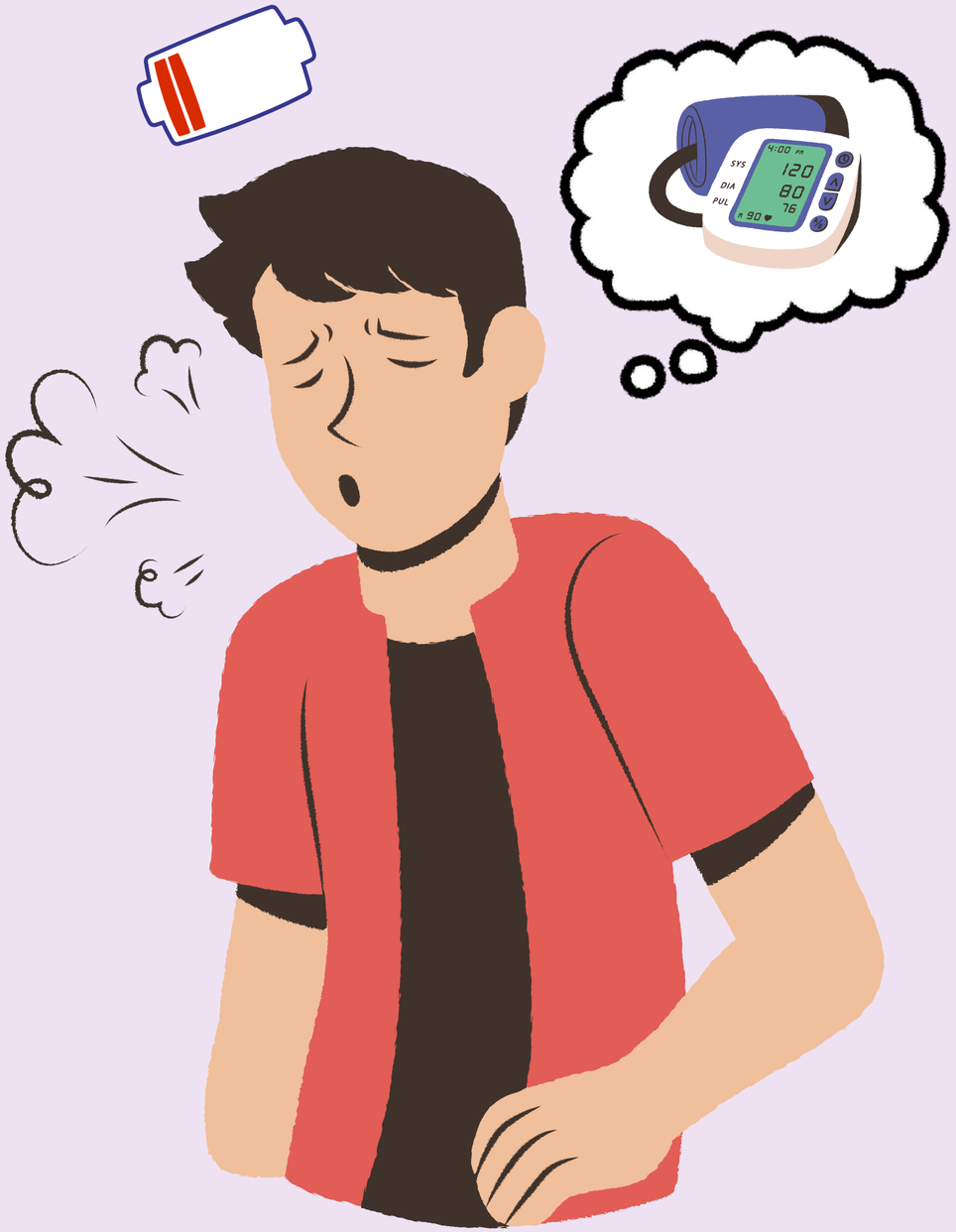
📱 App mobile : Dashboard accessible partout

IMPACT DU PROJET ACTUEL

À Quoi Sert Ce Projet ?

Avec le modèle actuel, les utilisateurs peuvent utiliser mon tableau de bord pour :

1. 📱 **Prendre conscience de leur propre risque**
2. 💪 **Se motiver à changer de mode de vie**
3. 🚴 **Voir l'impact du cholestérol, exercice, etc.**



TECHNOLOGIES UTILISÉES

Les Outils du Métier

Langage : Python

Bibliothèques clés :

- scikit-learn → Modèles ML et métriques
- pandas → Gestion des données
- matplotlib/seaborn → Visualisations
- joblib → Sauvegarde des modèles

Techniques appliquées :

- GridSearchCV (optimisation)
- Cross-validation (validation)
- Ensemble methods (stacking)

Output : Dashboard Power BI

LEÇONS APPRISES

Au-delà des Chiffres

Leçons techniques :

- L'AUC est plus informative que la précision pour les problèmes médicaux
- La régularisation est fondamentale contre l'overfitting
- Les methods *ensemble* améliorent la robustesse

Leçons pratiques :

- Équilibrer performance et interprétabilité
- Penser à l'utilisateur final (médecins/patients)
- Être transparent sur les limites

Leçon principale :

L'IA ne remplace pas le médecin, elle le soutient dans ses décisions

CONCLUSION

Le Message Final

Nous sommes partis d'une question :

"***Peut-on prédire les maladies cardiaques avec les données ?***"

Nous sommes arrivés à une réponse :

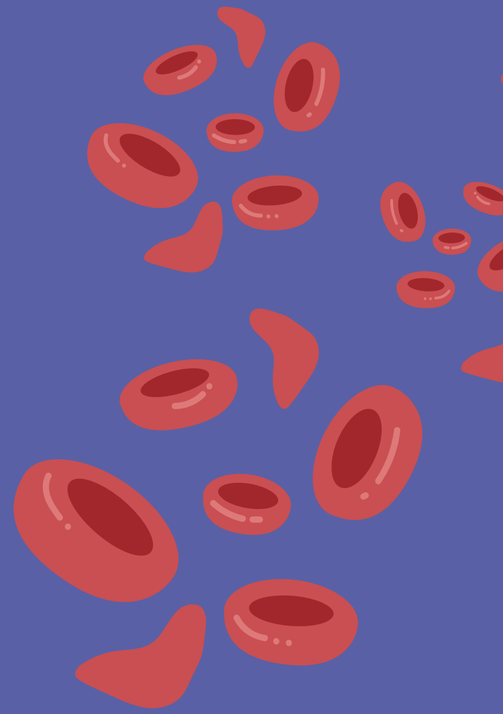
"***Oui, avec une précision de 95% et un outil utilisable***"

L'avenir de la médecine est la collaboration humain-machine

GLOSSAIRE

Termes Clés Expliqués

- **AUC** (**Area Under Curve**) : De 0 à 1, mesure à quel point le modèle distingue bien sains et malades. 0,95+ = excellent
- **Overfitting** : Quand le modèle apprend "par cœur" les données au lieu de comprendre les patterns généraux et il n'est plus capable de généraliser sur de nouvelles données.
- **Cross-validation** : Tester le modèle sur différentes parties des données pour vérifier la robustesse
- **Stacking** : Combiner les prédictions de plusieurs modèles pour en obtenir un meilleur
- **Recall** : % de malades réellement identifiés par le modèle



REMERCIEMENTS & CONTACT

Merci pour votre attention !

📊 Projet : **Prédiction Maladies Cardiaques**
📅 Date : **9 Novembre 2025**

👤 Autrice : **Giulia Governatori**
✉️ Email : **giuliagovernatori@hotmail.com**
💥 portfolio : **Giulia-Governatori.alwaysdata.net**
🌐 Linkedin : **@giuliagovernatori-bi-analyst/**

🔗 Dataset : **UCI Machine Learning Repository**
🙏 Crédits : Hungarian Institute of Cardiology, University Hospital Zurich, University Hospital Basel, V.A. Medical Center