Compte rendu — Système d’aide au diagnostic des maladies cardiaques

Auteur : Anouar LAHLOU

Objectif du document : Version ergonomique et prête à déposer du rapport de projet — prédiction de la présence d’une maladie cardiaque à partir de variables cliniques.

---

1. Résumé exécutif

Ce projet vise à développer un outil d’aide au diagnostic capable de prédire la présence d’une maladie cardiaque à partir de variables cliniques. En s’appuyant sur un jeu de données de 270 patients (14 variables), un modèle XGBoost optimisé atteint une AUC de 0.87 et une précision globale de 80% après optimisation. Le principal point d’attention reste la présence de 6 faux négatifs — patients malades classés comme sains — qui motivent des recommandations concrètes pour réduire les risques cliniques.

---

2. Contexte et enjeu

- **Problème métier :** Les maladies cardiovasculaires nécessitent des diagnostics fiables ; les erreurs (en particulier les faux négatifs) peuvent entraîner des conséquences vitales.
- **Objectif :** Maximiser le **rappel** (recall) afin de réduire le nombre de patients malades non détectés, tout en conservant une performance de séparation élevée (AUC).

---

3. Données — Description synthétique

- **Taille :** 270 patients
- **Variables :** 14 variables cliniques (âge, sexe, cholestérol, tension au repos, thalach, cp, thal, slope, ca, oldpeak, …)
- **Répartition cible :** ~44.4% malades / 55.6% sains (jeu relativement équilibré)

Statistiques clés (après nettoyage) - Âge moyen : 54.5 ans (écart-type ≈ 8.9) - Tension moyenne : 131.5 mmHg - Cholestérol moyen : 250 mg/dL

---

4. Prétraitement & qualité des données

- **Valeurs manquantes :** Simulation de ~182 valeurs manquantes (≈5%). Imputation par la moyenne pour variables numériques.
- **Outliers :** Présence d’extrêmes en cholestérol et oldpeak; jugés cliniquement plausibles et conservés.
- **Justification :** L’imputation par la moyenne a préservé la distribution globale et maintenu des écarts-types raisonnables.

5. Analyse exploratoire (points saillants)

- Variables fortement corrélées à la cible : **thal**, **cp**, **slope**, **ca**.
- Distribution des variables : majorité d'hommes (68%).
- Cas cliniques « masqués » identifiés — patients aux valeurs proches de la normale mais malades.

---

6. Modélisation & performances

6.1 Modèle initial (XGBoost)

- **Accuracy :** 77.8%
- **AUC :** 0.86
- **Matrice de confusion (extrait) :** 24 TN | 6 FP — 6 FN | 18 TP

6.2 Optimisation (GridSearchCV)

- **Hyperparamètres retenus :** colsample_bytree=0.7, learning_rate=0.1, max_depth=3, n_estimators=50, subsample=0.9
- **Résultat :** Accuracy 80%, AUC 0.87
- **Remarque :** Passage à SMOTE n'a pas réduit les FN (classes déjà équilibrées).

---

7. Analyse des faux négatifs

- **Nombre :** 6 patients
- **Caractéristiques moyennes normalisées :** oldpeak plus bas que la moyenne, thalach légèrement inférieur, trestbps légèrement supérieur.
- **Interprétation :** Profils proches de la normale → « cas masqués » difficiles à détecter par le modèle. Risque clinique élevé si non corrigé.

---

8. Recommandations opérationnelles

1. **Ajuster le seuil de décision** (ex. 0.50 → 0.40) pour réduire les faux négatifs ; accepter une légère hausse des faux positifs (préférable cliniquement).
2. **Enrichir le jeu de données** : ajouter IMC, tabagisme, antécédents familiaux, pression systolique/diastolique séparée.
3. **Ingénierie de features** : construire variables dérivées (chol/trestbps, age/thalach, oldpeak×slope) pour exposer interactions non linéaires.
4. **Interprétabilité** : intégrer l'analyse SHAP pour expliquer les prédictions individuelles et aider le clinicien à vérifier les cas limites.
5. **Ensemble de modèles** : combiner XGBoost, Random Forest et SVM dans un Voting Classifier pour stabiliser les décisions.
6. **Validation clinique** : tester le modèle sur un jeu externe et impliquer des cardiologues pour revue des cas « masqués ».

---

9. Plan de déploiement (synthèse)

- Phase 1 — Prototype : modèle optimisé + seuil abaissé + tableau de bord SHAP pour revue clinicien.
- Phase 2 — Validation : tests rétrospectifs sur jeu externe, ajustement du seuil en concertation médicale.
- Phase 3 — Déploiement pilote : intégration dans le flux clinique (outil d'aide, non décisionnel) avec suivi des alertes et retour d'expérience.

---

10. Conclusion

Le modèle développé offre une base solide (AUC = 0.87) pour un outil d'aide au diagnostic. La priorité immédiate est la réduction des faux négatifs via ajustements de seuil, enrichissement des données et analyses explicatives. Avec ces mesures, l'outil peut devenir un support fiable pour améliorer la détection des maladies cardiaques en milieu clinique.

Annexes (à joindre si nécessaire)

- Matrices de confusion complètes
- Courbe ROC
- Table des hyperparamètres testés
- Exemples d'explications SHAP pour 3 cas (TP, FP, FN)

Fin du document.