Rapport de Machine Learning : Prédiction de la mortalité liée à l'insuffisance cardiaque

RANDRIANJAFY Voahanginiaina Roberte

December 18, 2025

Contents

1 Introduction

L'objectif de ce projet est de développer un modèle pour predire la mortalité causée par l'insuffisance cardiaque. Nous utilisons pour cela un jeu de données contenant des variables cliniques (age,anaemia,creatinine phosphokinase,diabetes,ejection fraction, high blood pressure,platelets,serum creatinine,serum sodium,sex,smoking,time,DEATH EVENT).

2 Méthodologie

2.1 Prétraitement des données

Le pipeline de données comprend :

- **Imputation :** Remplacement des valeurs manquantes par la médiane (numérique) et le mode (catégoriel).
- **Standardisation :** Mise à l'échelle des variables numériques pour équilibrer leur influence.
- **Gestion du déséquilibre :** Utilisation de la méthode SMOTE et de la pondération des classes (*class weight*) pour compenser le faible nombre de décès observés.

2.2 Modèles testés

Nous avons évalué 8 modèles issus de différentes familles algorithmiques afin de couvrir un large spectre de logiques décisionnelles :

2.2.1 Modèles Linéaires et de Voisinage

- **Régression Logistique :** Modèle statistique qui calcule la probabilité d'appartenance à une classe (décès/survie) en utilisant une fonction sigmoïde. C'est le modèle de référence pour sa simplicité et son interprétabilité.
- **K-Nearest Neighbors (KNN) :** Algorithme basé sur la proximité. Il classifie un patient en regardant les k patients les plus proches (similaires) dans le jeu de données.
- **Support Vector Machine (SVC) :** Cet algorithme cherche à tracer une frontière (appelée hyperplan) qui sépare les survivants des décès avec la plus grande marge de sécurité possible. Pour ce projet, nous utilisons un **noyau RBF** (*Radial Basis Function*). Ce mécanisme permet de projeter les données dans un espace à plus haute dimension pour séparer des groupes qui ne le seraient pas par une simple ligne droite. Concrètement, il permet au modèle de capturer des zones de risque complexes et "courbes" dans les données cliniques.

2.2.2 Modèles basés sur les Forêts (Bagging)

- **Random Forest :** Combine plusieurs arbres de décision entraînés sur des sous-ensembles de données. La décision finale est prise par vote majoritaire, ce qui réduit le risque de sur-apprentissage (*overfitting*).

- **Extra Trees :** Similaire au Random Forest, mais utilise des seuils de division aléatoires pour chaque caractéristique, ce qui peut encore améliorer la robustesse face au bruit.

2.2.3 Modèles de Boosting (Séquentiels)

- **Gradient Boosting :** Construit des arbres de manière séquentielle, où chaque nouvel arbre tente de corriger les erreurs commises par les précédents.
- **XGBoost (Extreme Gradient Boosting) :** Une version optimisée et très rapide du Gradient Boosting, intégrant des régularisations avancées pour une meilleure performance.
- **LightGBM :** Un framework de boosting développé par Microsoft, qui utilise une croissance des arbres par feuilles (*leaf-wise*) plutôt que par niveau, le rendant extrêmement efficace sur les petits et grands jeux de données.

3 Résultats et Analyse

Les 3 meilleurs modèles qui sont RandomForest,LogisticRegression et GradBoost ont été optimisés via une recherche d'hyperparamètres (*RandomizedSearchCV*) et évalués sur un ensemble de validation.

3.1 Tableau comparatif

Voici les performances obtenues sur l'ensemble de validation :

Modèle	Accuracy	F1-Score	Recall	Precision	ROC AUC
RandomForest	0.833	0.687	0.579	0.846	**0.881**
LogisticReg.	0.833	**0.722**	**0.684**	0.765	0.879
GradBoost	0.817	0.686	0.632	0.750	0.863

Table 1: Comparaison des performances des 3 meilleurs modèles

3.2 Interprétation des métriques

- **ROC AUC** (Le gagnant : RandomForest - 0.880)

  C'est la métrique la plus importante ici. Elle mesure la capacité du modèle à séparer les patients qui vont décéder de ceux qui vont survivre.

  0.88 est un excellent score (proche de 1.0). Cela signifie que dans 88 % des cas, le modèle classera un patient décédé avec un risque plus élevé qu'un patient survivant.RandomForest gagne ici d'un cheveu face à la LogisticRegression.

- **Recall / Sensibilité** (Le gagnant : LogisticRegression - 0.684)

  C'est la métrique la plus critique en médecine. Elle répond à la question : "Sur 100 personnes qui vont réellement décéder, combien le modèle a-t-il réussi à détecter ?"

  LogisticRegression (0.68) est meilleure que RandomForest (0.57).

C'est à dire la forêt aléatoire rate plus de patients à risque (elle a plus de "faux négatifs") que la régression logistique.

- **Precision** (Le gagnant : RandomForest - 0.846)

  Elle répond à : "Quand le modèle prédit un décès, à quel point est-il fiable ?"

  RandomForest (84%) est très fiable : s'il lance une alerte, il y a de fortes chances que ce soit vrai.

  LogisticRegression (76%) fait plus de "fausses alertes".

- **F1-Score** (Le gagnant : LogisticRegression - 0.722)

  C'est la moyenne entre la Precision et le Recall. Elle donne une idée de l'équilibre général.

  Ici, la LogisticRegression l'emporte car elle offre un meilleur compromis pour un usage clinique (elle détecte mieux les cas graves sans faire trop de fausses alertes).

L'analyse clinique nous impose de privilégier le **Recall** (Sensibilité). Dans ce contexte, la **Régression Logistique** s'avère être le modèle le plus prometteur malgré un score AUC légèrement inférieur à la Random Forest :

1. **Recall (0.684) :** Elle identifie correctement plus de 68% des patients à risque, contre seulement 58% pour la Random Forest.
2. **ROC AUC (0.879) :** Le modèle montre une excellente capacité de distinction entre les classes de survie et de décès.

4 Visualisations

Les figures suivantes présentent la performance diagnostique du modèle sélectionné.

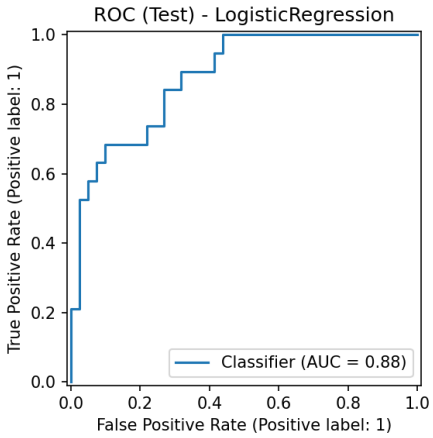

Figure 1: Courbe ROC

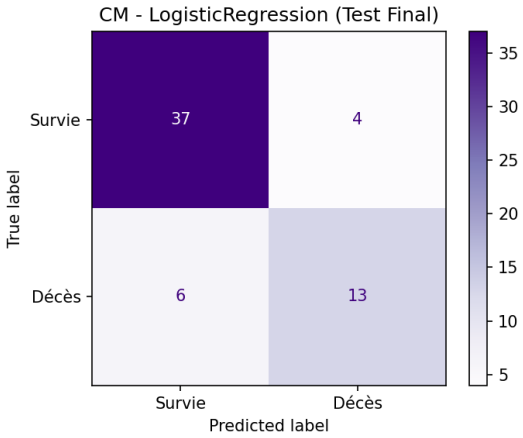

Figure 2: Matrice de Confusion

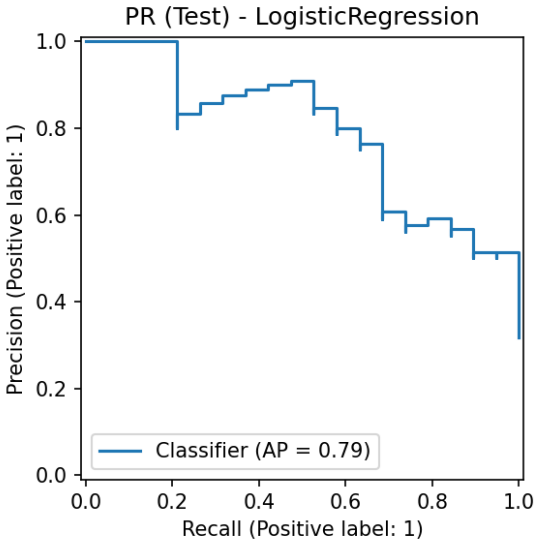

Figure 3: Courbe Precision-Recall - Équilibre justesse et détection

5 Conclusion

Le modèle final permet une détection robuste des patients à risque. Bien que la Random Forest soit plus précise (moins de fausses alertes), la Régression Logistique est préférable pour une application médicale car elle minimise le nombre de patients à risque non détectés.