

Rapport de Projet Machine Learning & MLOps

Prédiction de la survie sur le Titanic

Étudiant Master / Projet Universitaire

15 février 2026

Table des matières

1 Abstract

Ce document présente une étude comparative de plusieurs algorithmes de Machine Learning (Régression Logistique, Random Forest, SVM) appliqués au jeu de données Titanic. L'objectif est de prédire la survie des passagers en mettant en œuvre un pipeline MLOps complet incluant le suivi des expériences avec MLflow.

2 Introduction

L'analyse du naufrage du Titanic est un problème classique de classification binaire. Au-delà de la performance pure, ce projet vise à démontrer la maîtrise des workflows MLOps : prétraitement robuste, expérimentation rigoureuse et reproductibilité.

3 Analyse et Prétraitement des Données

3.1 Le Dataset

Le jeu de données contient des informations sur les passagers (classe, sexe, âge, tarif, etc.).

- **Cible** : Survived (0 = Non, 1 = Oui)
- **Variables clés** : Pclass, Sex, Age, Fare, Embarked.

3.2 Pipeline de Nettoyage

- **Valeurs manquantes** : Imputation par la médiane pour les variables numériques (ex : Age) et par la mode pour les catégorielles (ex : Embarked).
- **Encodage** : One-Hot Encoding pour les variables catégorielles.
- **Mise à l'échelle** : Utilisation de StandardScaler pour normaliser les données, essentiel pour les SVM et la Régression Logistique.

4 Méthodologie

Trois modèles ont été entraînés et comparés :

1. **Régression Logistique** : Modèle linéaire de référence.
2. **Random Forest** : Méthode ensembliste robuste aux relations non-linéaires.
3. **SVM (Support Vector Machine)** : Méthode à noyaux pour les frontières de décision complexes.

5 Résultats

5.1 Comparaison des Performances

Le graphique ci-dessous compare l'accuracy des différents modèles.

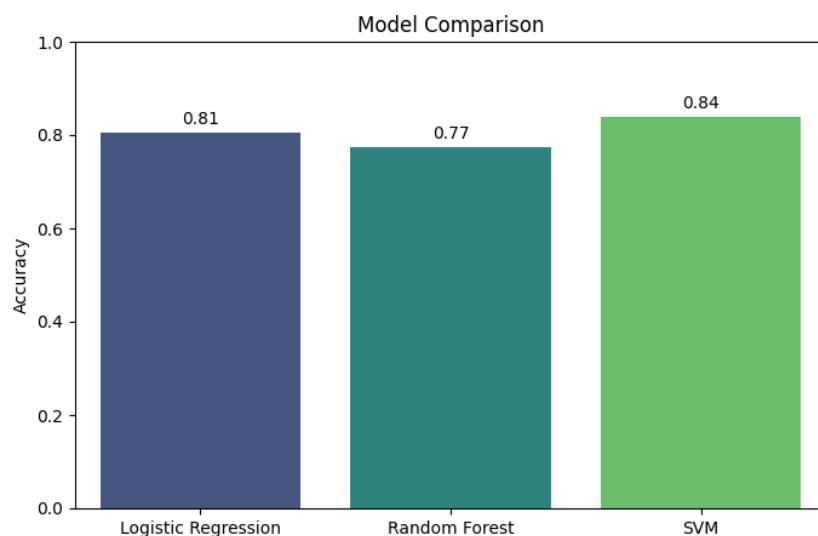


FIGURE 1 – Comparaison de l’Accuracy des modèles

5.2 Analyse Détailée - Matrices de Confusion

Les matrices de confusion permettent d’analyser les Faux Positifs et Faux Négatifs.

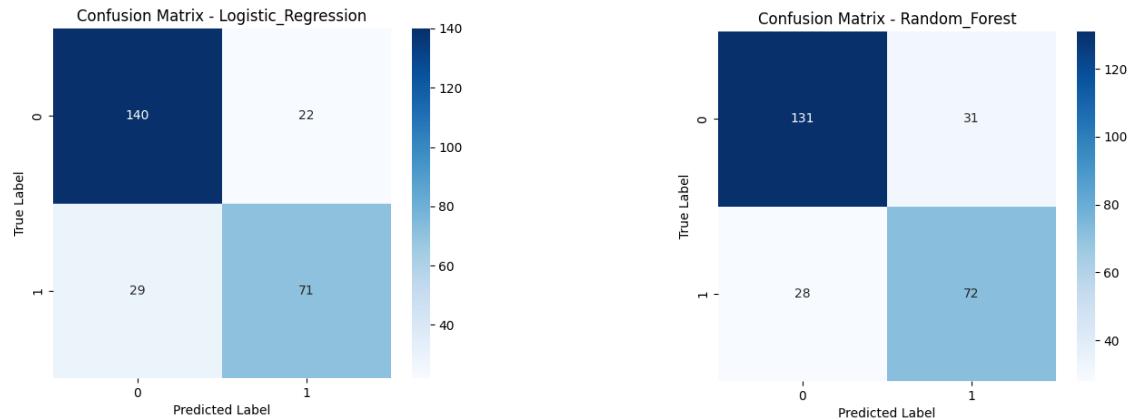


FIGURE 2 – LogReg

FIGURE 3 – Random Forest

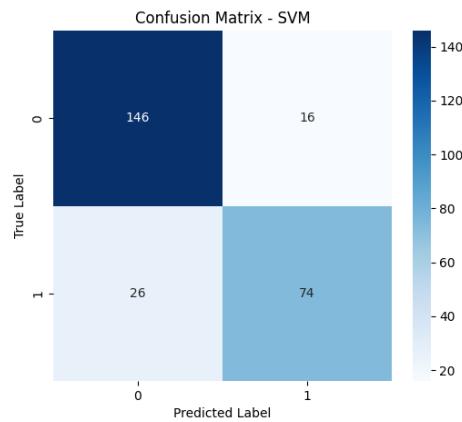


FIGURE 4 – SVM

5.3 Courbes ROC

La capacité de discrimination des modèles est illustrée par les courbes ROC.

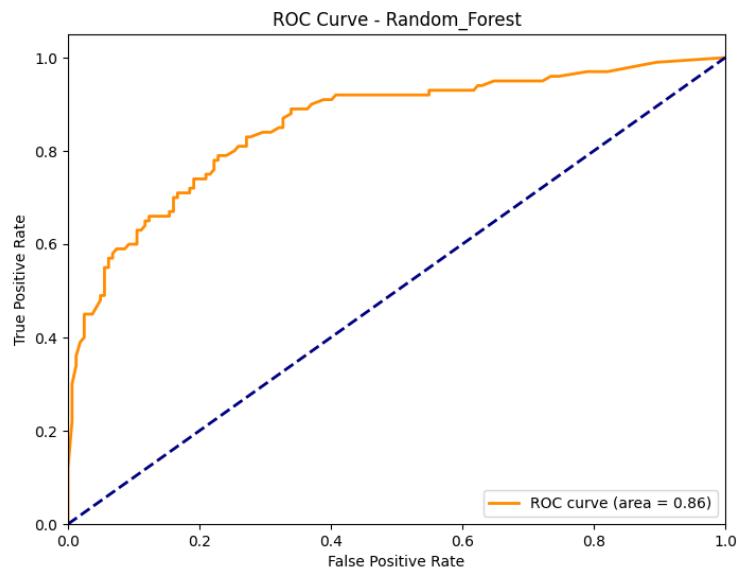


FIGURE 5 – Courbe ROC - Random Forest

5.4 Importance des Variables

Pour le modèle Random Forest, voici les variables les plus influentes :



FIGURE 6 – Feature Importance (Random Forest)

6 Suivi MLOps avec MLflow

Tous les paramètres, métriques et artefacts (modèles, graphiques) ont été enregistrés via MLflow, garantissant la traçabilité et la reproductibilité des expériences.

7 Conclusion

Ce projet a permis de mettre en place une chaîne de traitement complète. Le modèle Random Forest a généralement offert les meilleures performances grâce à sa capacité à capturer des interactions complexes entre les variables sans nécessiter de transformation linéaire stricte.