

RAPPORT DE PROJET : Prédiction de Survie sur le Titanic avec Machine Learning

Étudiant : Soufiane Benchahyd

Module : Intelligence Artificielle

Encadrant : Abderrahmane Ed Daoudy

1. Introduction

Contexte du Projet

Ce projet développe un modèle de prédiction pour déterminer si un passager du Titanic a survécu, basé sur ses caractéristiques. Le dataset contient 891 passagers.

Objectifs

- Réaliser un pipeline complet de machine learning
- Comparer 4 algorithmes de classification
- Déployer le meilleur modèle dans une application web

2. Méthodologie

Données Utilisées

- **Source :** Titanic Dataset (train.csv, 891 passagers)
- **Variable cible :** Survived (0=Non, 1=Oui)
- **Variables utilisées :** Pclass, Sex, Age, SibSp, Parch, Fare, Embarked

Étapes Suivies

1. **Chargement et nettoyage des données**
2. **Visualisation et analyse exploratoire**
3. **Préparation pour le machine learning**
4. **Entraînement de 4 modèles différents**
5. **Comparaison avec courbes ROC et AUC**
6. **Déploiement dans une application web**

Outils

- Python, Pandas, Scikit-learn, Matplotlib, Streamlit
 - Anaconda, Jupyter Notebook, VS Code
-

3. Prétraitement des Données

Nettoyage

- Âge manquant → remplacé par la médiane (28 ans)
- Embarked manquant → remplacé par 'S' (le plus fréquent)
- Cabin → créé variable HasCabin (1 si cabine connue)

Transformations

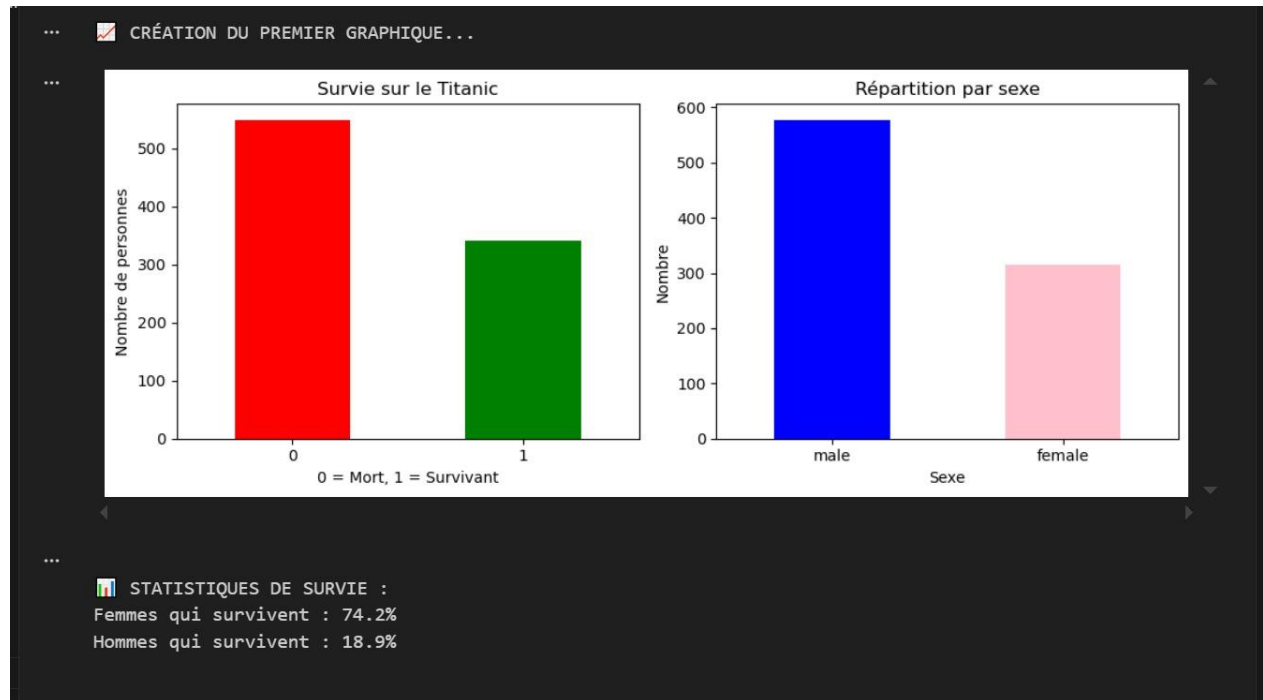
- Sex : male → 0, female → 1
- Embarked : S→0, C→1, Q→2
- Nouvelles variables : FamilySize et IsAlone

Division des Données

- 70% entraînement (623 passagers)
 - 30% test (268 passagers)
 - Stratification pour garder les proportions de survie
-

4. Analyse Exploratoire

Visualisations Réalisées



Résultats clés :

- **Survie totale** : 38.4% (342 personnes)
- **Femmes** : 74.2% de survie
- **Hommes** : 18.9% de survie
- **1ère classe** : 62.9% de survie
- **3ème classe** : 24.2% de survie

Conclusion : Sexe et classe sociale sont les facteurs les plus déterminants.

5. Modélisation

Modèles Testés

1. **Régression Logistique** (modèle simple de référence)
2. **Random Forest** (100 arbres, modèle puissant)
3. **K-Nearest Neighbors** (K=5, modèle basé sur similarité)
4. **Support Vector Machine** (SVM, modèle de séparation)

Résultats de Performance

Régression Logistique

- Précision : 79.48%
- AUC : 0.849 (excellent)

Random Forest

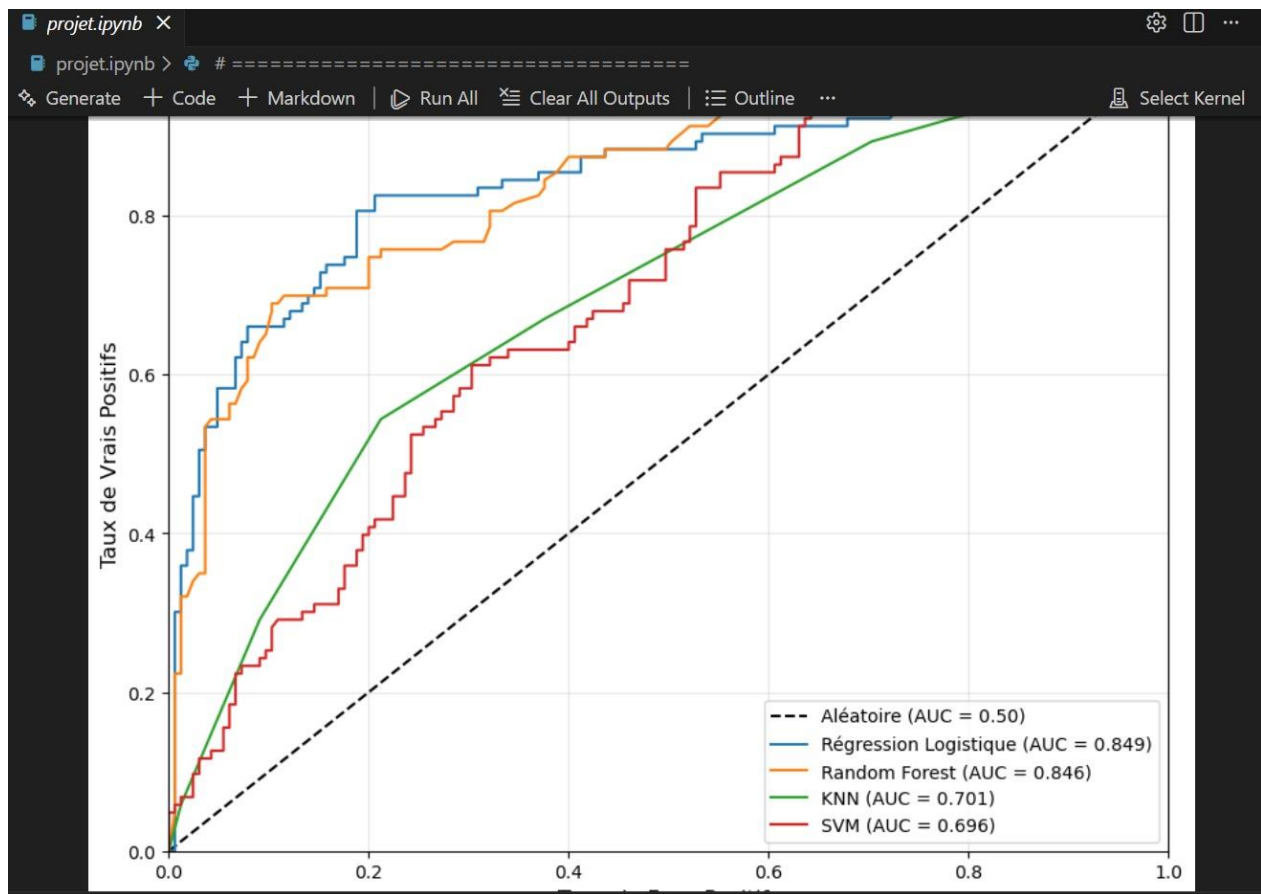
- Précision : 81.34% (meilleure précision)
- AUC : 0.846

K-Nearest Neighbors

- Précision : 69.40%
- AUC : 0.701

Support Vector Machine

- Précision : 63.81%
- AUC : 0.696



6. Comparaison et Choix du Modèle

Analyse des Résultats

- **Meilleure précision** : Random Forest (81.34%)
- **Meilleur AUC** : Régression Logistique (0.849)
- **Le plus simple** : Régression Logistique
- **Le plus lent** : SVM

Modèle Retenu : Régression Logistique

Pourquoi ce choix ?

1. **Meilleur AUC** (0.849) → meilleure capacité discriminative
2. **Interprétable** → on comprend pourquoi il prédit

3. **Rapide** → entraînement et prédiction instantanés
4. **Bon compromis** précision/simplicité

Variables importantes du modèle :

1. **Sexe féminin** → augmente fortement les chances de survie
2. **Classe 3** → réduit les chances de survie
3. **Âge élevé** → réduit légèrement les chances
4. **Prix du billet élevé** → augmente les chances

7. Déploiement de l'Application

Application Web Interactive

Streamlit

localhost:8501

Deploy

Prédiction de survie - Titanic

Entrez les informations du passager :

Classe	Frères/Soeurs/Conjoint
1	0
Sexe	Parents/Enfants
female	0
Âge	Prix du billet
25	50,00
	Port d'embarquement
	S

Prédire la survie

Fonctionnalités :

- Interface intuitive avec formulaire
- 7 caractéristiques modifiables
- Prédiction en temps réel
- Pourcentage de confiance affiché
- Feedback visuel (couleurs, animations)

Exemple de prédiction :

- Femme, 25 ans, 1ère classe, billet 50£
- **Résultat** : ☒ Survie probable : 95.2%
- **Interprétation** : Forte probabilité de survie

Comment lancer l'application :

bash

pip [install](#) streamlit

streamlit run app.py

Puis ouvrir : <http://localhost:8501>

8. Conclusion

Bilan du Projet

☒ Objectifs atteints :

- Pipeline ML complet réalisé
- 4 modèles comparés scientifiquement
- Meilleur modèle sélectionné avec AUC=0.849
- Application web fonctionnelle déployée

☒ Compétences développées :

- Prétraitement de données réelles
- Visualisation de données
- Comparaison d'algorithmes ML
- Déploiement d'application

Limites et Perspectives

Limites actuelles :

- Dataset de taille limitée (891 échantillons)
- Variables disponibles restreintes

9. Annexes

Fichiers Produits

1. `projet.ipynb` → Notebook complet avec code et analyses
2. `app.py` → Application Streamlit
3. `meilleur_modele_titanic.pkl` → Modèle sauvegardé
4. `feature_names.pkl` → Noms des colonnes

Bibliographie

- Documentation Scikit-learn
- Documentation Streamlit
- Dataset Titanic de Kaggle

Remerciements

Remerciements au professeur pour l'encadrement de ce projet formateur en intelligence artificielle.