

Université de Nouakchott  
Département mathématique et  
informatique

---

Projet:Machine Learning from  
Disaster

---



Fatimetou/limam — Fatimetou/Elhadj  
C16698 — C14683

June 4, 2024

# Contents

<b>Introduction</b>	<b>4</b>
<b>1 Charger et Prétraiter les Données</b>	<b>5</b>
1.1 Objectif . . . . .	5
1.2 Étapes . . . . .	5
1.2.1 Chargement des Données . . . . .	5
1.2.2 Exploration des Données . . . . .	5
1.2.3 Nettoyage des Données . . . . .	5
1.2.4 Conversion des Variables Catégorielles . . . .	6
<b>2 Créer les Modèles pour tester</b>	<b>7</b>
2.1 modèle de régression logistique . . . . .	7
2.1.1 objectif . . . . .	7
2.2 modèle de forêt aléatoire . . . . .	7
2.2.1 objectif . . . . .	7
2.3 modèle de K-nearest neighbors . . . . .	7
2.3.1 objectif . . . . .	7
2.4 modèle de Support Vector Machine . . . . .	8
2.4.1 objectif . . . . .	8
2.5 Étapes . . . . .	8
2.5.1 Séparation des Données . . . . .	8
2.5.2 Entraînement du Modèle . . . . .	8
2.5.3 Évaluation du Modèle . . . . .	8
<b>3 Développer la Fonction de Prédiction</b>	<b>9</b>

3.1	Objectif . . . . .	9
3.2	Étapes . . . . .	9
3.2.1	Définition de la Fonction . . . . .	9
3.2.2	Prétraitement des Données d'Entrée . . . . .	9
<b>4</b>	<b>Évaluer la Performance du Modèle</b>	<b>10</b>
4.1	Objectif . . . . .	10
4.2	Étapes . . . . .	10
4.2.1	Évaluation Continue . . . . .	10
<b>5</b>	<b>Interface Gradio</b>	<b>11</b>
5.1	Objectif . . . . .	11
5.2	Définir les Composants d'Entrée . . . . .	11
5.3	Définir les Composants de Sortie . . . . .	11
5.4	Tester et Déboguer l'Interface . . . . .	11
5.5	Déployer l'Interface Gradio . . . . .	12
	<b>Conclusion</b>	<b>13</b>

# Introduction

Objective de ce projet est de prédire la survie des passagers du Titanic en utilisant un modèle Random Forest. Pour ce faire, nous devons d'abord charger et prétraiter les données, puis entraîner et évaluer le modèle, et enfin, créer une interface utilisateur interactive avec Gradio pour permettre aux utilisateurs de faire des prédictions basées sur les caractéristiques des passagers.

# Chapter 1

## Charger et Prétraiter les Données

### 1.1 Objectif

L'objectif de cette étape est de charger les données à partir d'un fichier CSV, de nettoyer et prétraiter les données, de gérer les valeurs manquantes, et de convertir les variables catégorielles en variables numériques.

### 1.2 Étapes

#### 1.2.1 Chargement des Données

Charger les données du fichier CSV en utilisant une bibliothèque telle que pandas pour examiner la structure des données.

#### 1.2.2 Exploration des Données

Analyser les premières lignes des données pour comprendre leur structure et utiliser des méthodes statistiques pour obtenir des informations globales.

#### 1.2.3 Nettoyage des Données

Identifier et gérer les valeurs manquantes. Par exemple, remplacer les valeurs manquantes dans la colonne 'Age' par la médiane des âges.

#### 1.2.4 Conversion des Variables Catégorielles

Convertir les variables catégorielles en variables numériques par une méthode comme le one-hot encoding. Par exemple, transformer la colonne 'Sex' en deux colonnes binaires '*Sex<sub>m</sub>ale*' et '*Sex<sub>f</sub>emale*'

# Chapter 2

## Créer les Modèles pour tester

### 2.1 modèle de régression logistique

#### 2.1.1 objectif

La régression logistique est un modèle puissant et interprétable pour les tâches de classification binaire, offrant des prédictions probabilistes et une compréhension claire des facteurs influençant les résultats

### 2.2 modèle de forêt aléatoire

#### 2.2.1 objectif

Le modèle de forêt aléatoire est un outil puissant et flexible pour les tâches de classification et de régression, offrant une précision élevée, une robustesse au surapprentissage, et une capacité à traiter des données complexes et variées.

### 2.3 modèle de K-nearest neighbors

#### 2.3.1 objectif

En résumé, le modèle kNN est une méthode efficace et intuitive pour la classification et la régression, mais il est crucial de choisir correctement le nombre de voisins et de normaliser les données pour obtenir des performances optimales.

## **2.4 modèle de Support Vector Machine**

### **2.4.1 objectif**

Le modèle de Support Vector Machine (SVM) est un puissant algorithme d'apprentissage supervisé utilisé principalement pour la classification, bien qu'il puisse également être appliqué à la régression. Il est particulièrement efficace pour les problèmes de haute dimension et est connu pour sa capacité à maximiser la marge de séparation entre les classes.

## **2.5 Étapes**

### **2.5.1 Séparation des Données**

Diviser les données en ensembles d'entraînement et de test pour permettre une évaluation fiable du modèle.

### **2.5.2 Entraînement du Modèle**

Utiliser une bibliothèque de machine learning pour créer et entraîner un modèle choisi, avec les données d'entraînement.

### **2.5.3 Évaluation du Modèle**

Évaluer la performance du modèle en utilisant des métriques comme l'accuracy, la précision, le rappel et le F1-score. Ces métriques permettent de comprendre la qualité des prédictions du modèle.



## Chapter 3

# Développer la Fonction de Prédiction

### 3.1 Objectif

Développer une fonction de prédiction qui prend les caractéristiques des passagers en entrée et retourne la prédiction de survie.

### 3.2 Étapes

#### 3.2.1 Définition de la Fonction

Créer une fonction qui prend en entrée les caractéristiques d'un passager (par exemple, classe, âge, sexe) et utilise le modèle entraîné pour prédire la survie.

#### 3.2.2 Prétraitement des Données d'Entrée

Appliquer les mêmes transformations que celles utilisées lors du prétraitement des données d'entraînement pour assurer la cohérence des entrées du modèle.

# Chapter 4

## Évaluer la Performance du Modèle

### 4.1 Objectif

Évaluer la performance du modèle avec des métriques appropriées et améliorer la précision si nécessaire.

### 4.2 Étapes

#### 4.2.1 Évaluation Continue

Évaluer la performance du modèle régulièrement avec des nouvelles données pour assurer sa robustesse et sa précision.

# Chapter 5

## Interface Gradio

### 5.1 Objectif

Créer les composants d'entrée pour l'interface Gradio, tels que les Dropdowns, Radios et Numbers pour capturer les caractéristiques des passagers.

### 5.2 Définir les Composants d'Entrée

Créer les composants d'entrée pour l'interface Gradio, tels que les Dropdowns, Radios et Numbers pour capturer les caractéristiques des passagers. S'assurer que les entrées capturent correctement les informations requises pour faire une prédiction.

### 5.3 Définir les Composants de Sortie

Créer le composant de sortie pour afficher la prédiction de survie. Assurez-vous que le résultat soit un message clair et précis. Rédiger un message clair et précis indiquant si le passager a des chances de survivre ou non, basé sur la prédiction du modèle.

### 5.4 Tester et Déboguer l'Interface

Tester l'interface Gradio avec divers scénarios pour s'assurer qu'elle fonctionne correctement et déboguer toute erreur. Identifier et cor-

riger les erreurs éventuelles en utilisant des techniques de débogage et des messages de journalisation.

## **5.5 Déployer l'Interface Gradio**

Lancer l'interface Gradio dans le navigateur et s'assurer qu'elle est accessible et opérationnelle pour les utilisateurs finaux.

# Conclusion

Le projet de prédiction de survie sur le Titanic a permis d'appliquer des techniques de machine learning pour résoudre un problème historique bien connu. En utilisant des étapes méthodiques de prétraitement des données, d'entraînement et d'évaluation du modèle, ainsi que la création d'une interface utilisateur intuitive avec Gradio, nous avons pu développer un outil prédictif performant. La documentation détaillée garantit que le projet est reproductible et compréhensible pour d'autres utilisateurs.