

Projet ML

18 pts pour la 1ère partie

Projet : Prédiction de survie des passagers du Titanic

Introduction

Le Titanic est l'un des naufrages les plus célèbres de l'histoire. En 1912, lors de son voyage inaugural, le paquebot RMS Titanic a heurté un iceberg et a coulé, entraînant la mort de la plupart de ses passagers et membres d'équipage. Dans ce projet, nous utiliserons les données sur la survie des passagers du Titanic pour concevoir un système capable de prédire si quelqu'un a survécu ou non au naufrage.

Objectif

L'objectif de cet exercice est de trouver le meilleur modèle pour prédire la survie des passagers du Titanic en fonction de différentes caractéristiques telles que la classe socio-économique, le sexe, l'âge, le nombre de frères et sœurs / conjoints à bord, etc.

Données

Les données sur la survie des passagers du Titanic sont disponibles sur Kaggle à l'adresse suivante : [Titanic - Machine Learning from Disaster](#). Voici une description des principales caractéristiques présentes dans l'ensemble de données :

- **Survived:** Résultat de la survie (0 = Non; 1 = Oui)

- **Pclass**: Classe socio-économique (1 = Classe supérieure; 2 = Classe moyenne; 3 = Classe inférieure)
- **Name**: Nom du passager
- **Sex**: Sexe du passager
- **Age**: Âge du passager (certaines entrées contiennent NaN)
- **SibSp**: Nombre de frères et sœurs / conjoints du passager à bord
- **Parch**: Nombre de parents / enfants du passager à bord
- **Ticket**: Numéro de billet du passager
- **Fare**: Prix payé par le passager
- **Cabin**: Numéro de cabine du passager (certaines entrées contiennent NaN)
- **Embarked**: Port d'embarquement du passager (C = Cherbourg; Q = Queenstown; S = Southampton)

Nous allons utiliser ces données pour entraîner plusieurs modèles de machine learning et déterminer celui qui prédit le mieux la survie des passagers du Titanic.

Pour le passage du text au numérique, utilisez Onehotencoding (si vous êtes curieux, trouvez comment ça marche)

Étapes du projet

1. Importation des données
2. Exploration des données (EDA)
3. Prétraitement des données
4. Division des données en ensembles d'entraînement et de test
5. Entraînement de différents modèles
6. Évaluation des performances des modèles
7. Sélection du meilleur modèle
8. Optimisation du modèle sélectionné (le cas échéant)

9. Déploiement du modèle sous streamlit