Projet : Machine Learning avec Python

Les voyageurs à bord du Titanic

Objectifs :

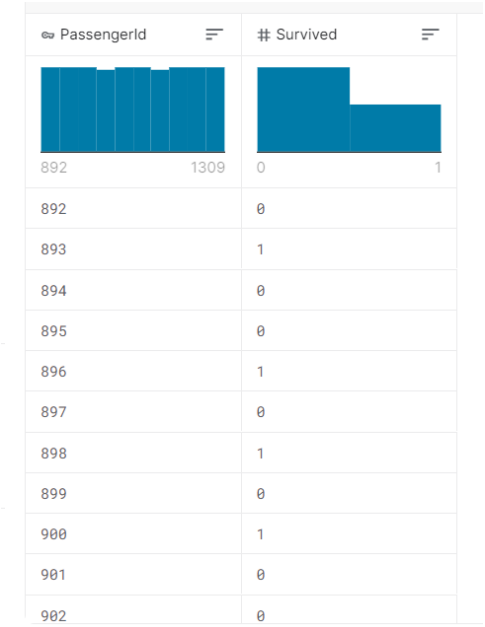
L'objectif de ce projet est de prédire si les voyageurs à bord du titanic ont survécu lors du naufrage.

Les données :

Il s'agit d'un projet issu du site Kaggle. Vous disposez d'un échantillon train et test. Vous devrez à l'issu du projet prédire les modalités de la variable 'Survival ' pour l'échantillon test.

Si vous souhaitez durant le projet vous pourrez effectuer des soumissions sur kaggle afin de voir le score de prédiction associé à votre algo sur l'échantillon test :
https://www.kaggle.com/c/titanic/submit

Le fichier de soumission doit être composé de cette façon :



PassengerId	Survived
892 – 1309	0 – 1
892	0
893	1
894	0
895	0
896	1
897	0
898	1
899	0
900	1
901	0
902	0

Il s'agit des prédictions associées à chaque passenger_id (issus du fichier test)

Les variables :

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Pour avoir une vision plus détaillée des données : https://www.kaggle.com/c/titanic/data

Les attendus :

Vous devrez rendre pour le 15 janvier au plus tard :

- Votre fichier de soumission final
- Un jupyter Notebook qui détaille toute votre démarche pour arriver à l'algorithme le plus performant (avec toutes les cellules compilées) :
    - Importation des données (merci de créer une variable 'path' qui contiendra le chemin associé aux datasets, pour que ce soit plus simple, et de l'appeler dans votre importation de données)
    - Analyse du jeu de données (analyse des variables, identification valeurs manquantes, outliers)
    - Nettoyage du jeu de données (gestion outliers, valeurs manquantes)
    - Rééchantillonnage si besoin
    - Création de modèle
    - Optimisation modèles
    - Interprétation du meilleur modèle

Les critères de notation :

- Réalisation de chaque étape
- Les explications de chaque étape (le but est de voir que vous comprenez ce que vous exécutez : pourquoi on le fait ? explication des paramètres, quels scores utilisés pour performance du modèle ? quelle validation croisée ? comparaison des différents scores (f1, accuracy etc …) etc …)
- Rédaction (propreté du notebook, graphiques tous avec titres, légendes etc ..)
- La recherche (n'hésitez pas à laisser des algos qui sont moins performants, mais expliquez pourquoi ils ne sont pas ceux que vous retiendrez)
- La performance de votre algo final et son interprétation (variables explicatives les plus importantes du modèle, matrice de confusion etc …)