# BRIEF PROJET

## INTRODUCTION AUX ALGORITHMES DE MACHINE LEARNING

## TABLE DES MATIERES



## INTRODUCTION

L'objectif de ce brief projet est de retracer scientifiquement l'histoire du naufrage du Titanic en utilisant les données disponibles sur le Kaggle.

Il s'agit d'un jeu de données public très facile d'accès et qui possède plusieurs vertus pédagogiques. Bien que l'analyse des données liées au naufrage du Titanic n'a aucun intérêt métier, les données sont riches pour pouvoir mettre en pratique les techniques et les modèles que nous avons abordés les dernières semaines.

Pour cela, il vous est demandé de mettre en œuvre une démarche complète d'exploitation de données allant de la compréhension du besoin jusqu'à l'évaluation des modèles élaborés en passant par une phase de préparation et d'analyse de données.

## I. MISE EN ŒUVRE

Dans ce qui suit, vous trouverez une trame à suivre pour mener à bien ce projet.

Cette trame a valeur d'exemple, il n'est pas obligatoire de la suivre scrupuleusement.

### 1. PREPARATION DES DONNEES

Les données sont à télécharger sur Kaggle Titanic. Vous disposez de deux jeux de données :

- **Train** : jeu d'apprentissage
- **Test** : jeu de test

### TRAVAIL DEMANDE

1. Importer les deux jeux de données.
2. Transformer les tables en data frame.
3. Analyser la signification de chaque variable.
4. Afficher le type de chacune des variables ainsi que le nombre de valeurs nulles par variables.
5. Supprimer les observations incomplètes.

## 2. EXPLORATION DES DONNEES

Avant d'aller plus loin, il est essentiel d'explorer les données pour les comprendre et les utiliser de manière efficiente.

### DATA VISUALISATION

1. Pour chacune des variables suivantes, créer un ou plusieurs diagrammes qui la résume au mieux :
    - Alive
    - Age
    - Sex
    - Class
    - Fare
2. Analyser les diagrammes et en tirer des conclusions et/ou des hypothèses

### TESTS D'HYPOTHESES

#### *LES FEMMES ET LES ENFANTS D'ABORD*

1. Faire un test d'hypothèse pour savoir si oui ou non, les enfants ont été privilégiés lors du naufrage.
2. Faire un test d'hypothèse pour vérifier si oui ou non, les femmes ont été privilégiées lors du naufrage.
3. Conclure.

#### *L'INFLUENCE DU PRIX DU BILLET SUR LA SURVIE DES PASSAGERS*

1. Faire un test d'hypothèse pour savoir si oui ou non, le prix du billet a une influence sur la survie d'un passager.
2. Conclure.

## 3. MACHINE LEARNING

L'objectif est de trouver un modèle qui nous permettra de prédire si un passager est mort ou vivant en se basant sur les données du dataset.

### REGRESSION LINEAIRE

1. Elaborer des modèles de régression linaire pour prédire la variable (target) *alive*. En sélectionnant comme variables explicatives (features) :
    a. *age*, *sex*, *class*, *fare* prises individuellement
    b. *age*, *sex*, *class*, *fare*
    c. différentes combinaisons de *age*, *sex*, *class*, *fare*
    d. ayant une forte corrélation avec la variable *alive*
2. Evaluer chaque modèle de régression.
3. Conclure.

### REGRESSION LOGISTIQUE

1. Sélectionner les variables explicatives en utilisant l'algorithme Recursive feature elimination.

**ATTENTION** : Cette méthode n'a pas encore été abordée en cours. Néanmoins, elle est simple à appliquer avec scikit-learn.
Vous pouvez bien évidement choisir une autre méthode de sélection de variables explicatives.

2. Elaborer un modèle de régression logistique pour prédire la variable *alive* en fonction des variables explicatives choisies.
3. Conclure.

## II. POUR ALLER PLUS LOIN

Les modèles élaborés précédemment sont basés sur un jeu de données incomplet.

Il est possible de reconstituer le jeu de données en remplaçant les données manquantes par des valeurs que vous devez choisir.

1. Refaire les modèles de régression en utilisant le jeu de données complet.
2. Evaluer les modèles.
3. Conclure.

## III. LIVRABLES

Au cours de ce brief projet, il vous sera demandé de nous transmettre les livrables suivants via un **dépôt git** :

- **Planning prévisionnel** : **lundi 06 avril à 13h**
- **DataViz** : **mardi 07 avril à 13h**
- **Modèle de régression linéaire** : **mercredi 08 avril à 13h**
- **Modèle de régression logistique** : **jeudi 09 avril à 13h**
- **Rapport du projet** : **jeudi 09 avril à 17h**

Les restitutions auront lieu le **vendredi 10 avril**.

Il est à noter que l'objectif de ce brief projet est de « faire parler » les données liées au naufrage du Titanic et raconter cette histoire mythique au travers d'une démarche scientifique.

On rappelle que le sujet de ce brief projet est purement pédagogique et n'a aucun intérêt métier.

Une attention particulière sera accordée à la démarche scientifique mise en place.

## IV. BIBLIOGRAPHIE


*Quelques conseils pour rédiger un rapport de projet* . Consulté le 04 05, 2020, sur https://www.irit.fr/~Armelle.Bonenfant/Enseignement/JavaSTRI/conseilsRapport.pdf

*Rédaction d'un rapport en informatique.* Récupéré sur univ-amu.fr: http://nicolas.durand.perso.luminy.univ-amu.fr/pub/projet3a/RedactionRapportInfo.pdf

*data science : fondamentaux et études de cas* . Eyrolles.

*Présentation générale de CRISP-DM*. Récupéré sur https://www.ibm.com/support/knowledgecenter/fr/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html

*Feature selection*. Récupéré sur scikit-learn: https://scikit-learn.org/stable/modules/feature_selection.html

*Titanic : Machine Learning from Disaster.* (s.d.). Récupéré sur kaggle: https://www.kaggle.com/c/titanic/overview/evaluation