

# Predicting Survival on the Titanic Wreck

Nicolas Drizard, Eloi Zablocki

March, 1st 2015

# Introduction

In the report we present our approach to the *Kaggle* competition<sup>1</sup>. The goal is to predict the passengers that survived at the shipwreck of Titanic, given some characteristics such as the genre, the age, etc. First, we will explain the preprocessing task that has be made in order to clean, extract and create the features. Then we will present the algorithms tries and used for the classification task.

## Première partie

### *Data pre-processing and Feature Engineering*

#### 1 Data pre-processing

use of pandas (handle floats and strings) tres pratique

#### 2 Feature Engineering

**Creation of features** extraction du title, et regroupement en 4 classes de title. (master c'est pour les enfants)  
extraction du deck,  
creation de familysize,  
creation de fareperperson,  
creation de age\*Pclass,

**Filling missing values** remplacer les missing values de l'age par l'age moyen des mecs qui ont le même title.

**Dimensionality reduction** (on en a pas fait, donc a voir si on en fait ou pas)

---

1. More infos to be found at <https://www.kaggle.com/c/titanic-gettingStarted/>

## Deuxième partie

# Classification algorithm

## 1 Data visualization

insister sur le fait qu'on a fait beaucoup de data visualization, sur weka ou avec pandas (quite a Ã©crire 3 lignes de code en plus)

## 2 Seeking the best classification algorithm

**weka** explication rapide du logiciel, ce qu'il peut faire

**finding a good algo** test de pleins d'algo, on a choisit "bagging"

**Explication de bagging** il accepte les missing values, il accepte les attributs Ã  valeurs nominales ce qu'il fait, son implÃ©mentation

## Troisième partie

# What has not worked

ne pas hésiter à dire qu'on a essayé plein de trucs qui n'ont pas forcément fonctionné. bien expliquer que trouver une bonne solution consiste à approcher pas à pas.

## Conclusion