Titanic survival evaluation

# ****Introduction/Business Problem.****

The problem is simple: to use machine learning to create a model that predicts which passengers survived the Titanic shipwreck based on the information provided about the passengers. This is not a secret that first class passengers were the ones who mostly survived, but the idea is to find more factors (other than the class of the ticket) that could influence the survival of a given passenger.

# ****Data.****

The data is gathered in a csv file with 891 entry. Each entry represents a passenger and consists of his/her name, age, marital status, ticket class, sex, age in years, no. of siblings/spouses aboard the Titanic, no. of parents/children aboard the Titanic, ticket number, passenger fare, cabin number and the port of embarkation. And of course, either a person survived or not.

The first five rows of the file are these:

|  |
| --- |
| PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked |
| 1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S |
| 2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C |
| 3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S |
| 4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,113803,53.1,C123,S |

It is already understandable from these four entries that the data is incomplete somewhere. So, the whole dataset should be checked for consistency, the second column (Survived) should be moved to become the last column and several columns (such as Name, Ticket, Fare, Cabin, Embarked) removed from the trained model.