

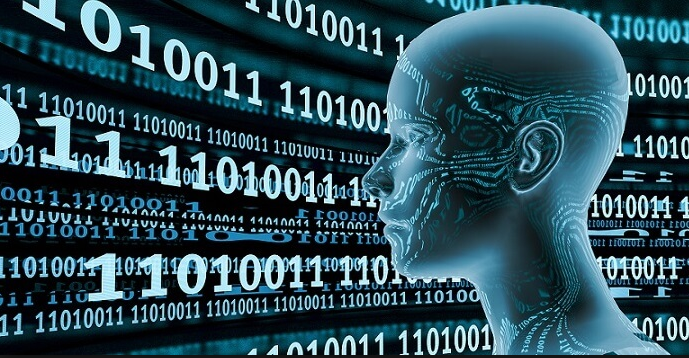
**2018**

Amaar Asaaf

Bahaa Gdiesat

12/24/2018





**Titanic: Machine Learning from Disaster**

Problem

* You are given a set of data about 890 passengers who were on the Titanic the moment of the sinking and you must find the best relationship between this dataset ("train.csv") and the survival. ( train your model with this data )

Pclass, Name Sex, Age, SibSp, Parch, Ticket, Fare, Cabin and Embarked => X

Survived => Y

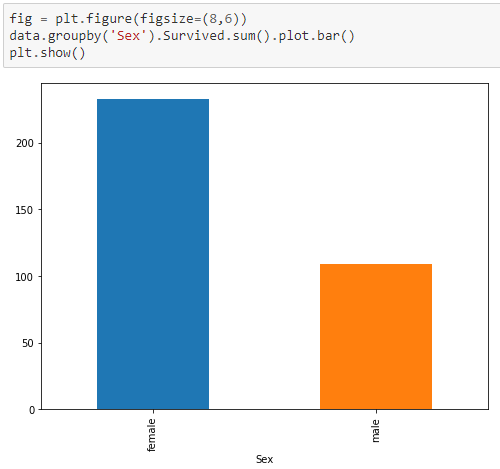
F(X) = Y

* use your trained model on the ("test.csv") data set and predict the survival of each passenger in the test file and add them to a column named ("survivedPre")
* compare your prediction with the result inside the ("gender\_submission") file (use the column " PassengerID" to uniquely identify each passenger).
* compute the accuracy of your prediction.
* use another way to train your model, predict the survival, compute the accuracy (submitted the code with best accuracy).
* play with the inputs and chose the inputs that gives the best accuracy.
* save your prediction of survival of each passenger into csv file with the name ("tst.csv").
* submit the code with best accuracy, the csv file and your report

**Calculations and Visual analysis used:**

1) sex:

* what is the probability of survival given the sex is "female"?
* what is the probability of survival given the sex is "male"?

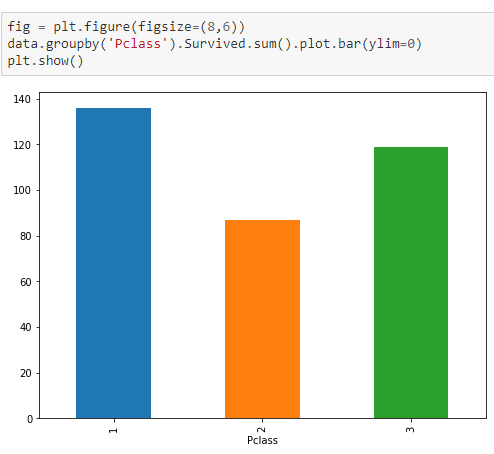


\*prob(suvived|female) > 2 x prob(suvived|male)

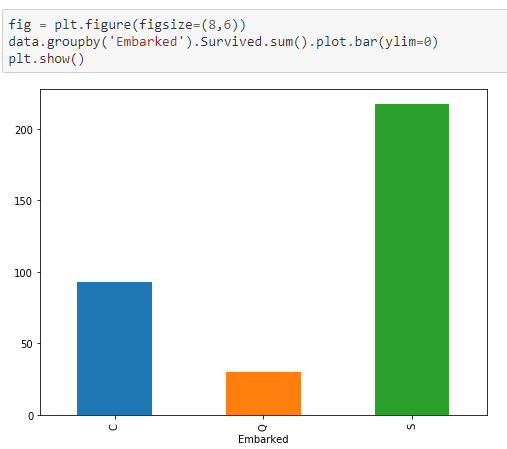
2) Pclass:

* what is the probability of survival given the Pclass is 1?
* what is the probability of survival given the Pclass is 2?
* what is the probability of survival given the Pclass is 3?

\*prob(suvived|1) > prob(suvived|3) > prob(suvived|1)



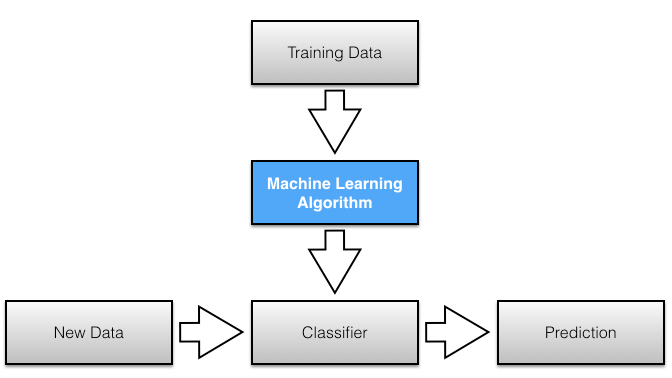
3)Embarked:



after a lots of analysis we find that the biggest reliable attributes are :

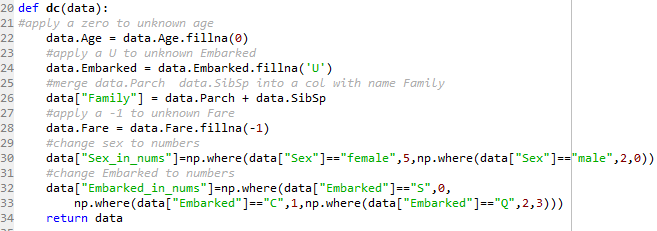
* Sex, Age, Embarked.

next for the code using the best method we find:  

1) clean the data:

* fill the null values with a unique values (may affect prediction).
* change Sex, Embarked values to numbers.
* for solving the assumption of independence among features we combine (merge) columns together in a new column which gives us a huge number of state space inputs ( may affect accuracy ).

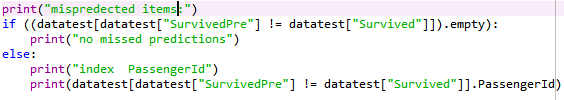


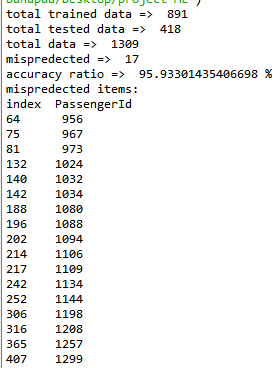


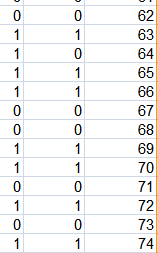
2)train your model, predict survival, save to pandas.



3)print miss predicted items and accuracy.  



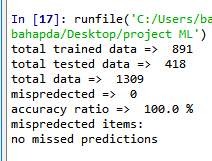


* A => index , B => Actual Survival , C => my prediction
* Accuracy = 96% .
* as
* as items tested more the 400 items.

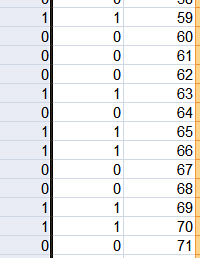
how can we improve code accuracy?

final step: play with the inputs used  
from calculations and visual analysis we find using the Sex, Age, Embarked will give better accuracy.





* wow a 100% accuracy with no miss predictions?
* will lets check it in the "tst.csv" file where we got index 64 miss predicted remember?



we see that for every row the survival prediction is equal to the actual survival.

ملاحظات:

هذا الكود لا يستخدم القرائة بفنجان القهوة او التنجيم او السحر الاسود.

ولكن من الممكن ان تقل دقة البرنامج عند تطبيقه على داتا مختلفة.