Report about the titanic data

# Link to the Gitlab page :

https://gitlab.com/python7963908/cassie\_doguet\_iris\_dataset\_dia/-

# Description of the algorithm used for each of the solution

With the titanic dataset we have to predict if the passenger of titanic will survive or not. I decided to apply the logistic regression model to this dataset. Here are the 5 first rows of the dataset and its details:

Une image contenant texte, Police, nombre, ligne

Description générée automatiquement

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Now we have to clean the data, and first plot where do we have a lack of information:

Une image contenant texte, capture d’écran, Rectangle

Description générée automatiquement

Une image contenant texte, capture d’écran, Rectangle

Description générée automatiquement

There are several things to do to clean the data :

* Encode the sex column so that it’s no longer “male” and “female” but “0” and “1”
* Encode the Embarked column
* Encode the ticket column
* Drop the columns Name and Cabin
* Replace the NaN values of Age by the mean value of age

Une image contenant texte, capture d’écran, affichage, Rectangle

Description générée automatiquement

Une image contenant texte, capture d’écran, affichage, Rectangle

Description générée automatiquement

I chose to drop the name’s and cabin’s columns. The names are not use full as we have the parch and we do not have enough values in the cabin’s column.

We now want to see if we have outliers values.

Une image contenant texte, capture d’écran, diagramme, affichage

Description générée automatiquement

Une image contenant texte, capture d’écran, diagramme, affichage

Description générée automatiquement

I do the same with the testing dataset.

Then we have to define the training and testing data :

Une image contenant texte, Police, capture d’écran, ligne

Description générée automatiquement

# How the model performance was improved

First, we have to see how Logistic regression works. Logistic regression is used to predict the probability based on features. The prediction is always between 0 and 1. It’s very suitable for our data set because the closest we are from 1, the more chance we have of surviving. As we can see on the picture below it calculated the probability of the event using the data :

Une image contenant ligne, diagramme, Tracé, pente

Description générée automatiquement

I first tried to improve the model performance by changing the solver hyperparameter. I could then see the accuracy of each model using different solvers :

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Then I decided to use the Grid Search exactly like I did with the iris data set but for the hyperparameters of a logistical regression model.

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

I found an accuracy almost like the one where I was only changing the solver.