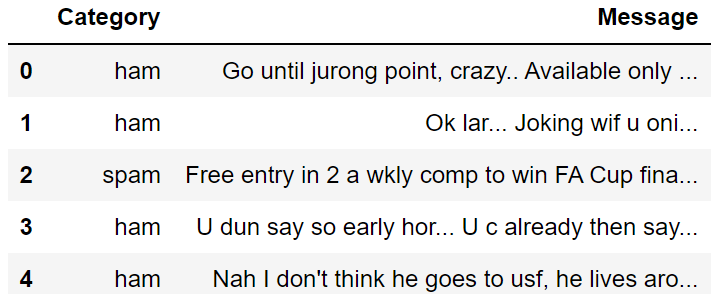
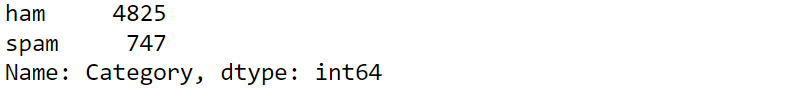
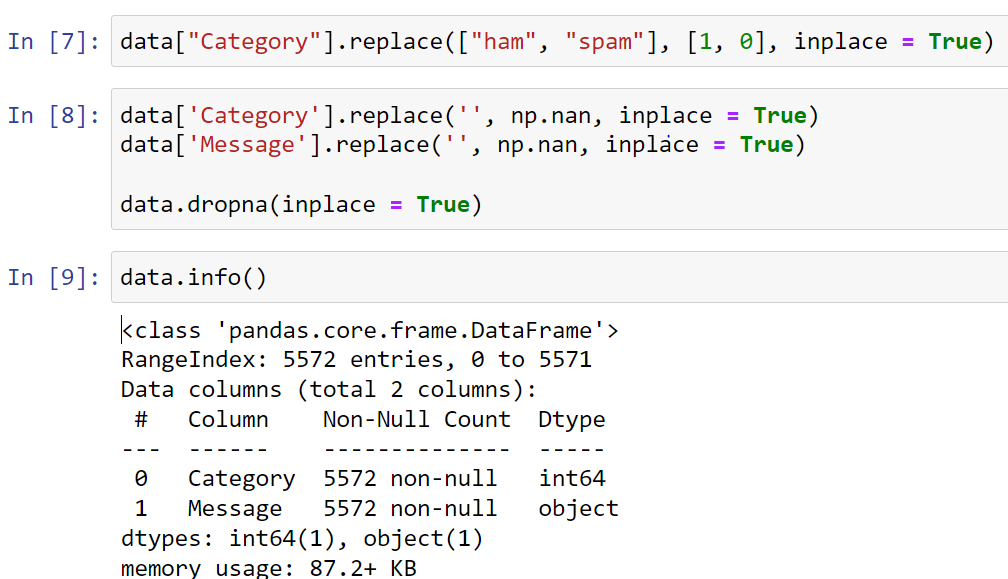
**Spam/phishing mail detection**

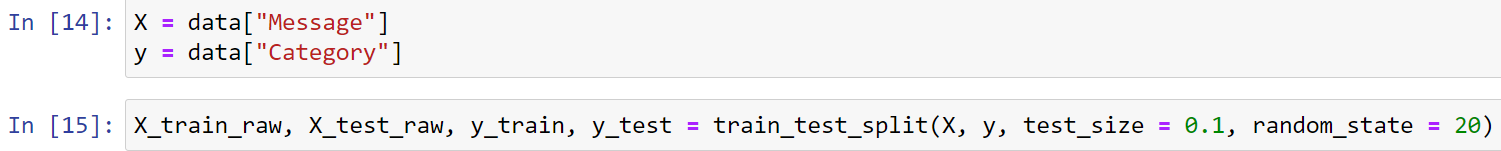
Dataset used:

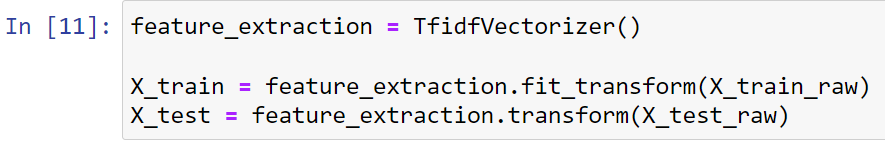
mail\_data.csv (5572 entries):



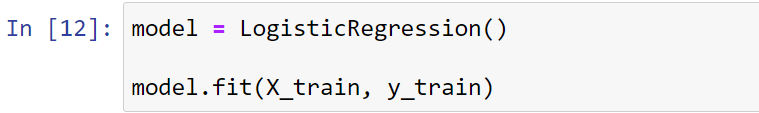
Firstly we preprocess the data, converting spam and ham(legit) to 0 and 1 and getting rid of any row that contains empty values.

Then, we save the message as X and category as y and we split the data into training and test sets.

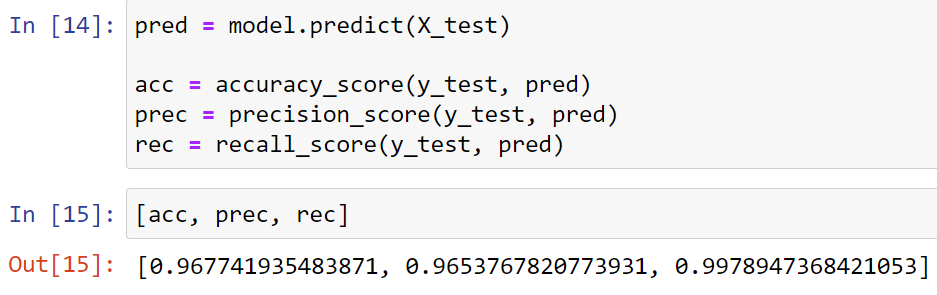
Because the message is a string, we will need to use TfidfVectorizer() from sklearn library to convert the text into a matrix of features before classifying.



After this, we can create the model and start training it. For the algorithm, I chose logistic regression because it’s one of the most reliable methods of classification, especially when the number of classes is only 2.



Here are the results:



Finally, I included an example on how to test the model manually.

