Polish companies bankruptcy Data Set

Mehdi CHIKH – Artus Chapelain de La Villeguérin - DIA2

Data Set Information

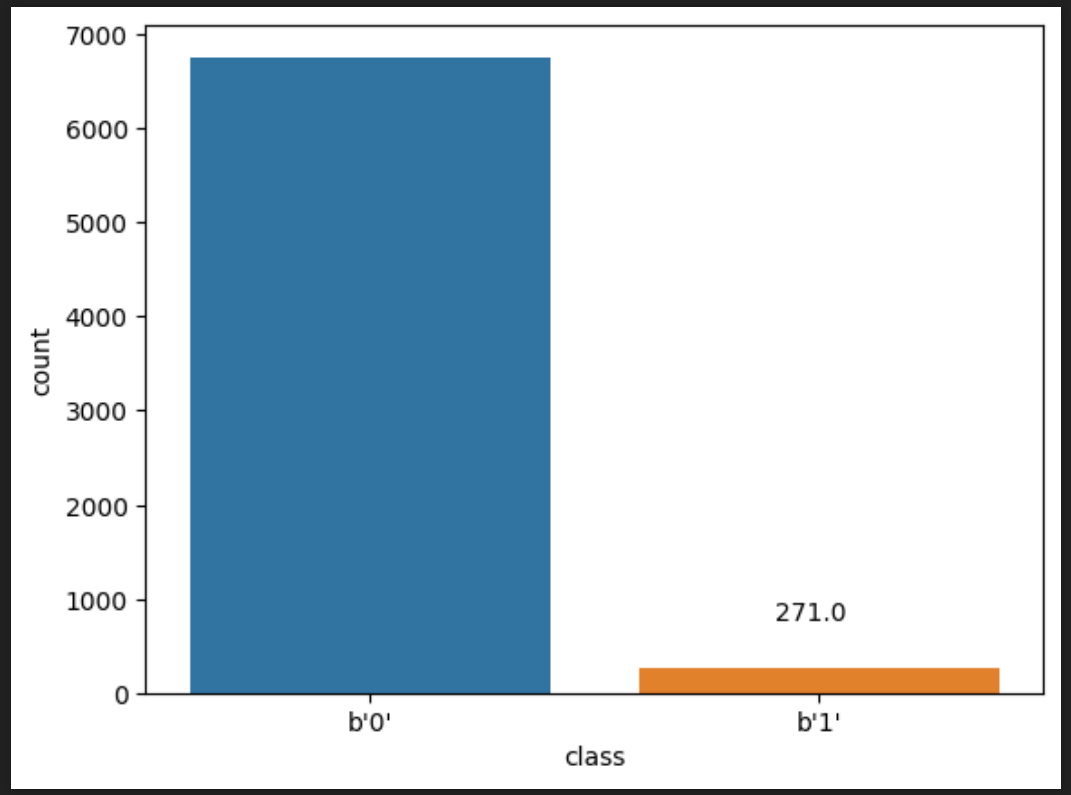
- The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service, which is a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012
- We have a total of 43 405 financial statements over 5 dataframes
- Each dataframe is composed of 64 features. We need to know the best features explaining the bankruptcy of a company

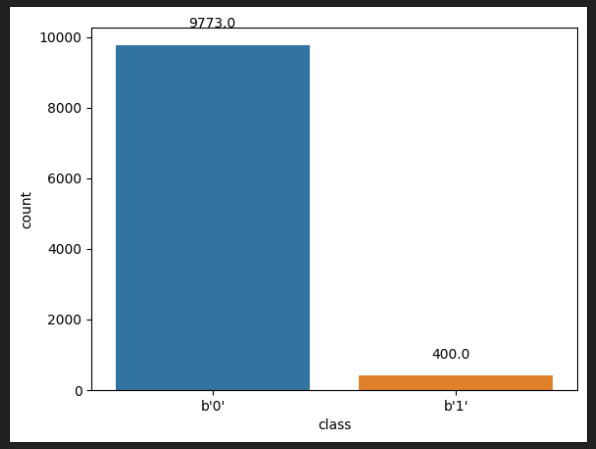
What are the main factors explaining the bankruptcy of Polish companies ?

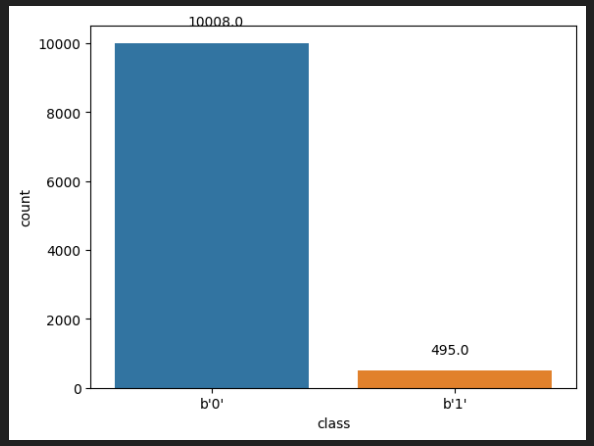
Data pre-processing

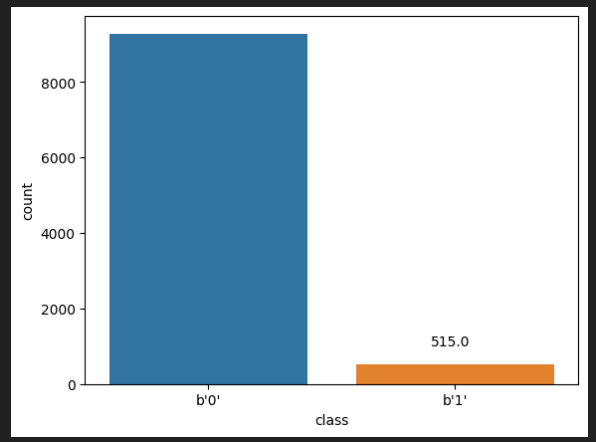
Class distribution

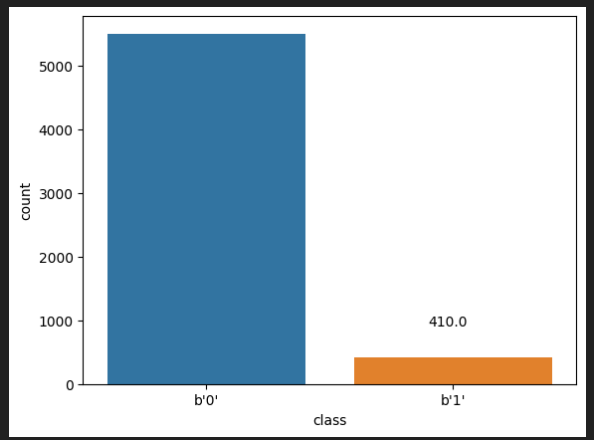
The b0 class is the majority. We are going to add in the final dataset 500 objects of classes b1 so that the trained algorithm does not lean too much in favor of b0.



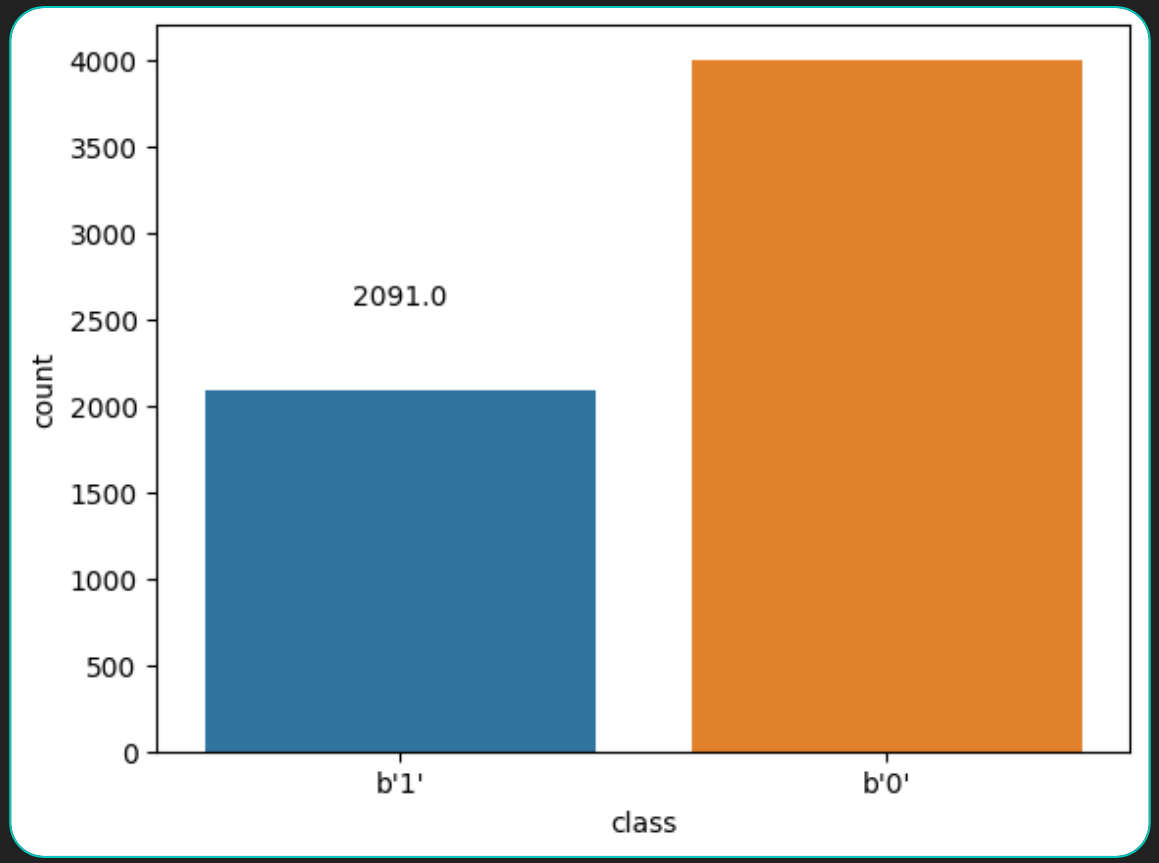








After adding the b1 classes, we get this dataframe, which is much more balanced :



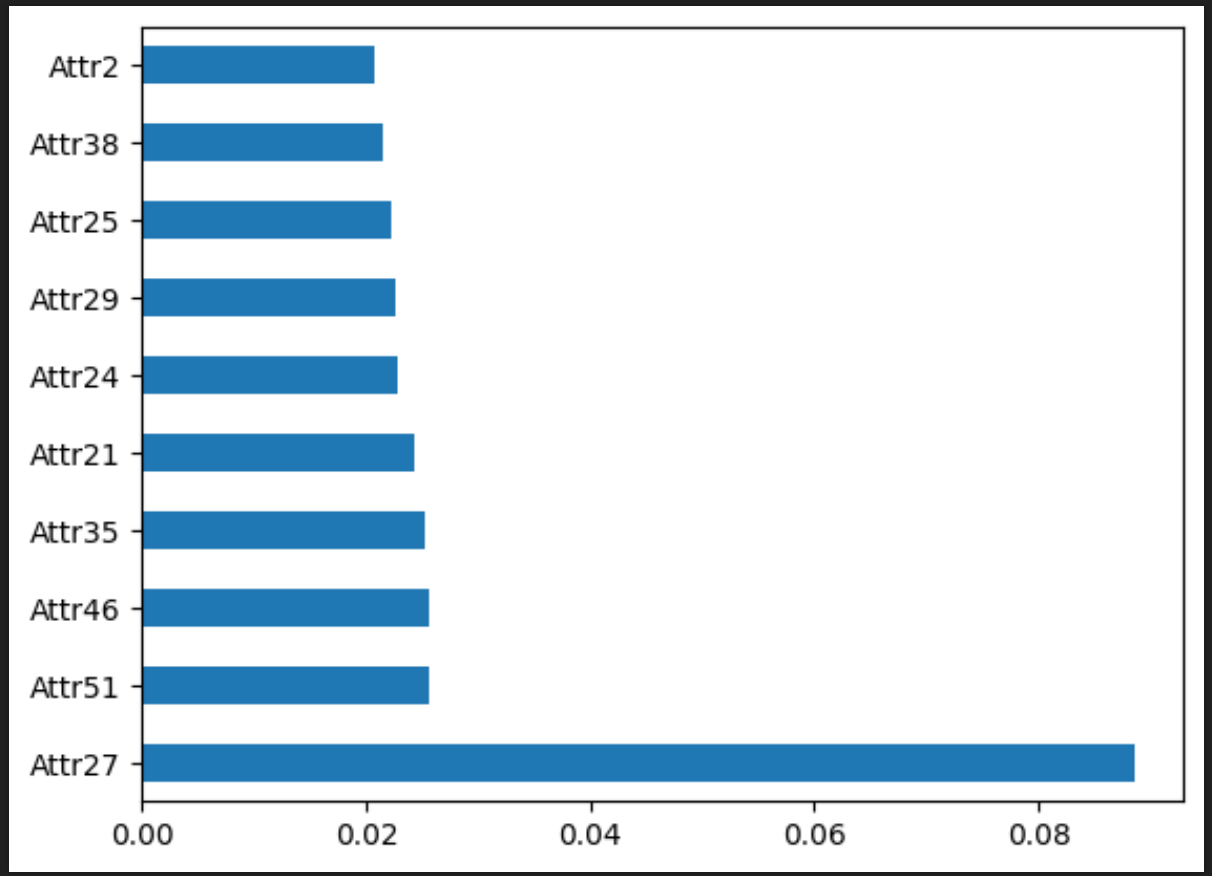
Checking for NaN values and duplicates

Since the number of NaN are reasonable, we're gonna replace

Attr1	2
Attr2	2
Attr3	2
Attr4	19
Attr5	10
Attr6	2
Attr7	2
Attr8	13
Attr9	2
Attr10	2
Attr11	38
Attr12	19
Attr13	21
Attr14	2

Attr15	2
Attr16	13
Attr17	13
Attr18	2
Attr19	21
Attr20	21
Attr21	1103
Attr22	2
Attr23	21
Attr24	99
Attr25	2
...	
Attr63	19
Attr64	140

We're gonna use 2 features selection methods and keep the features with the most importance



After running the algorithms several times, here are the features that come up most often :

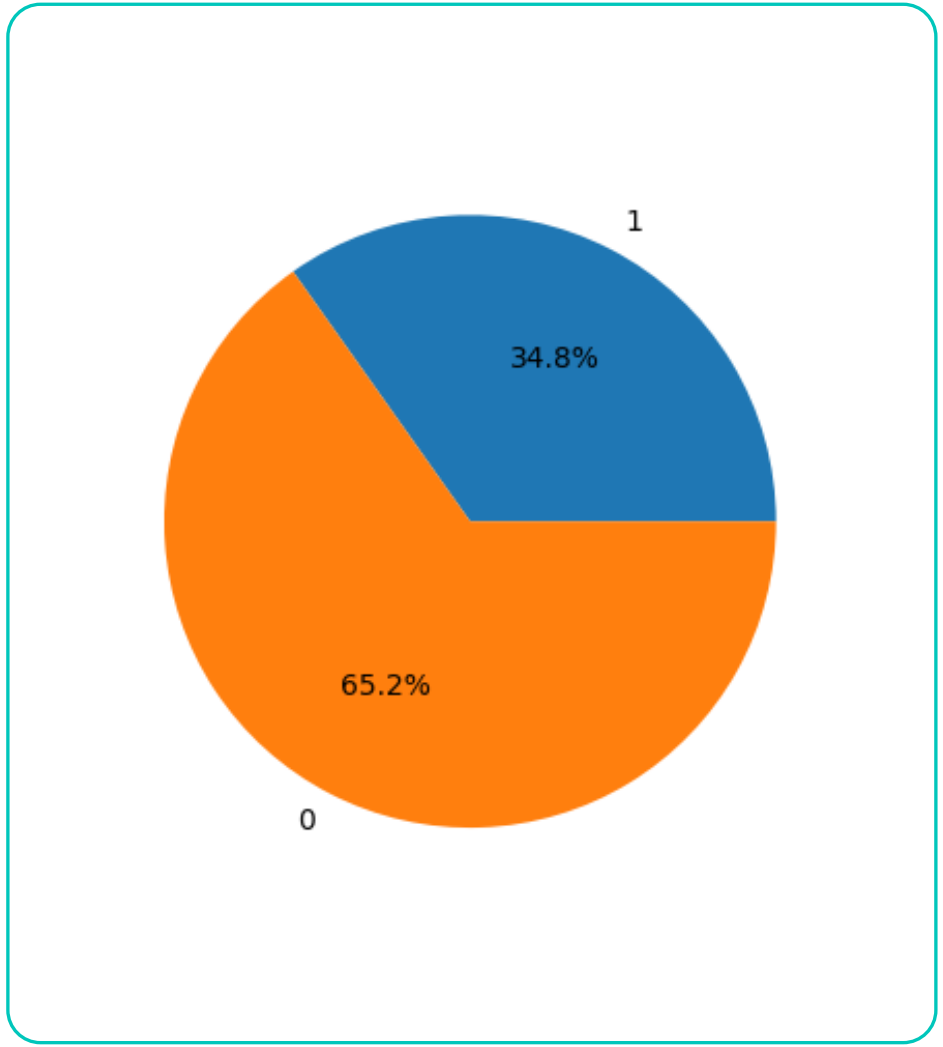
- 27 : profit on operating activities / financial expenses
- 56 : (sales - cost of products sold) / sales
- 1 : net profit / total assets
- 2 : total liabilities / total assets
- 35 : profit on sales / total assets

We wil keep these features to train the model because these features are the best to undestand how a polish company can go bankrupt.

Data visualization

Pie chart representing the share of classes 0 and 1 in the dataframe

- We can see that we have greatly improved the class 1 share in the dataset. In each other dataset taken individually, we had a share of only 4% of class 1. Here, we have 34.8%.

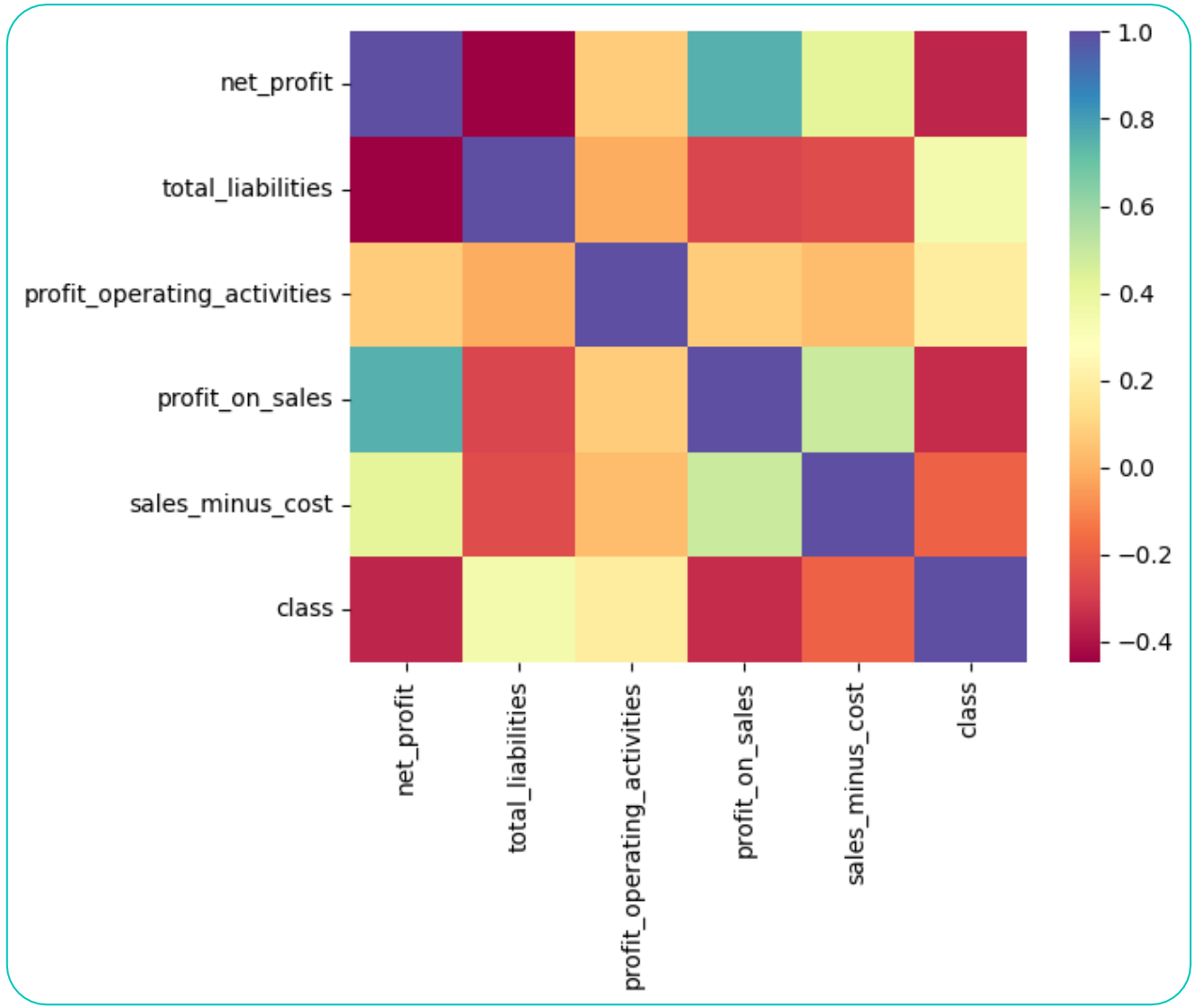


Heatmap of all the variables and class

In this heatmap, we observe two interesting correlations with the target class :

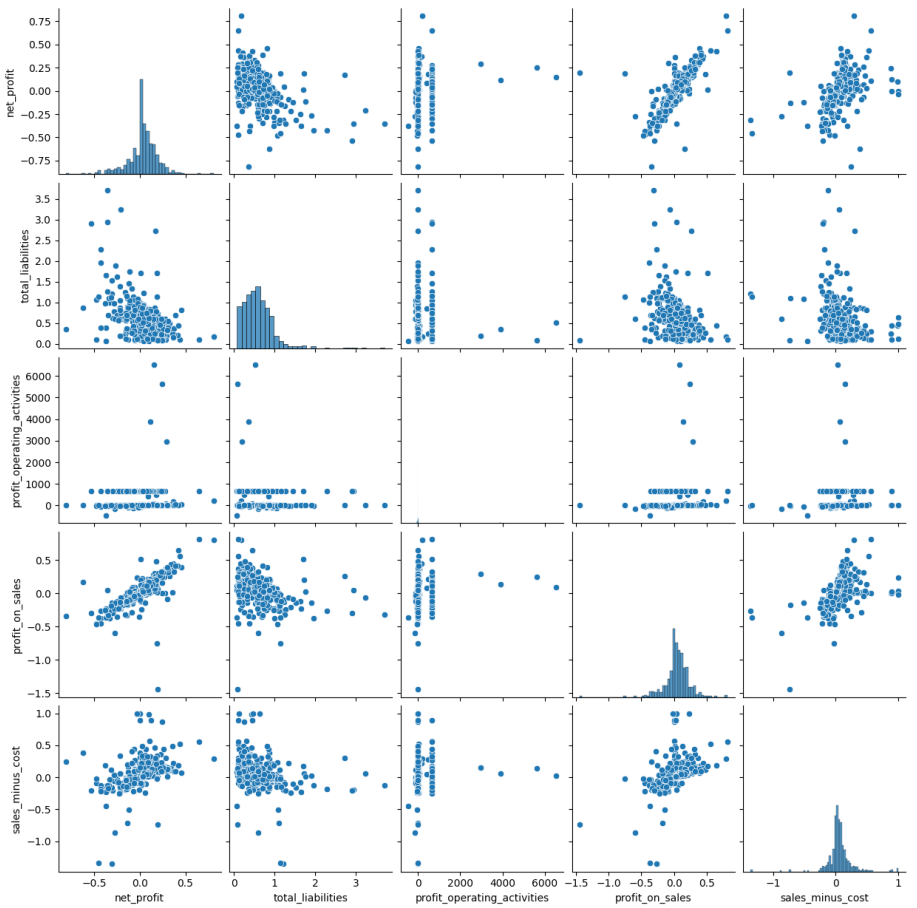
- Profit_on_sales seems to not be correlated to the target class
- Net_profit also seems to not be correlated to the target class

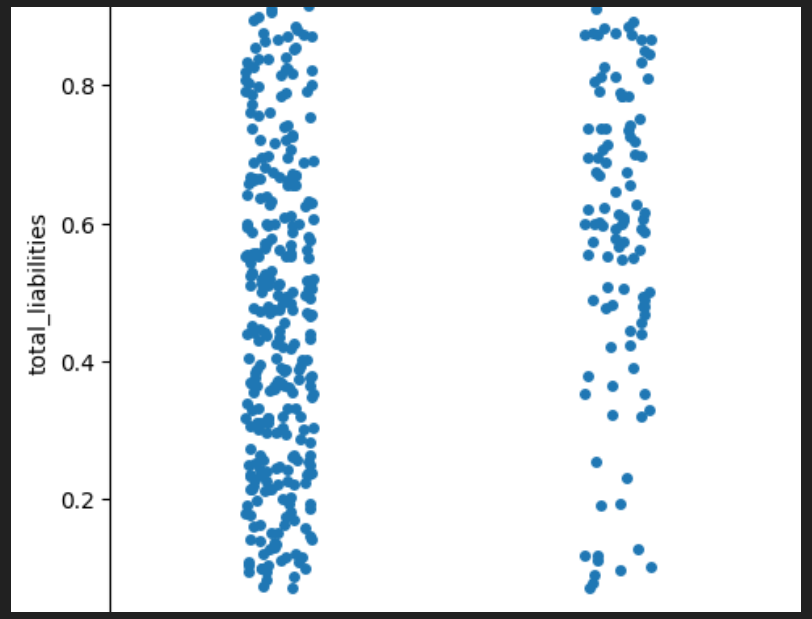
We can also see that profit_on_sales and net_profit are strongly correlated

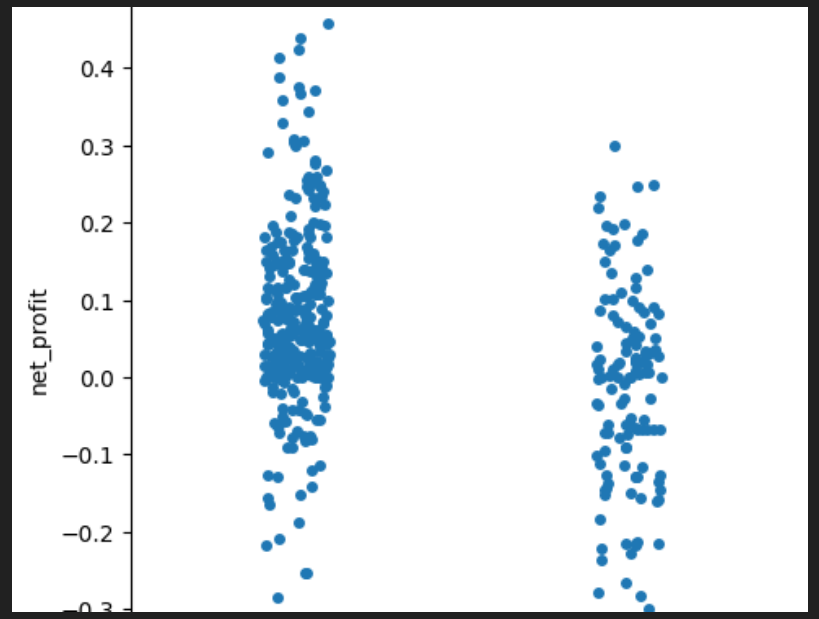


Using the pairplot function from seaborn, we can see the relations between all variables we have kept

- This will be useful to know which relation will be the most interesting to analyze

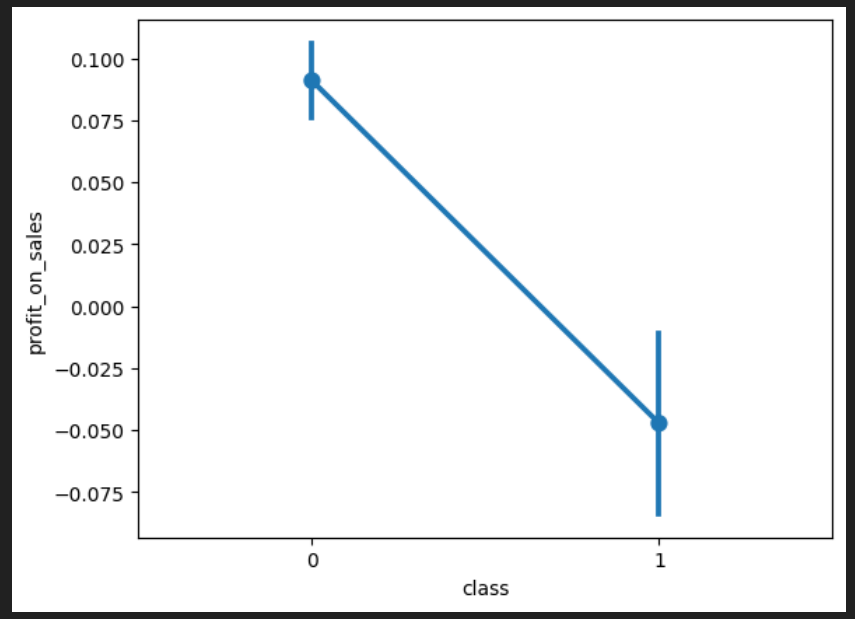


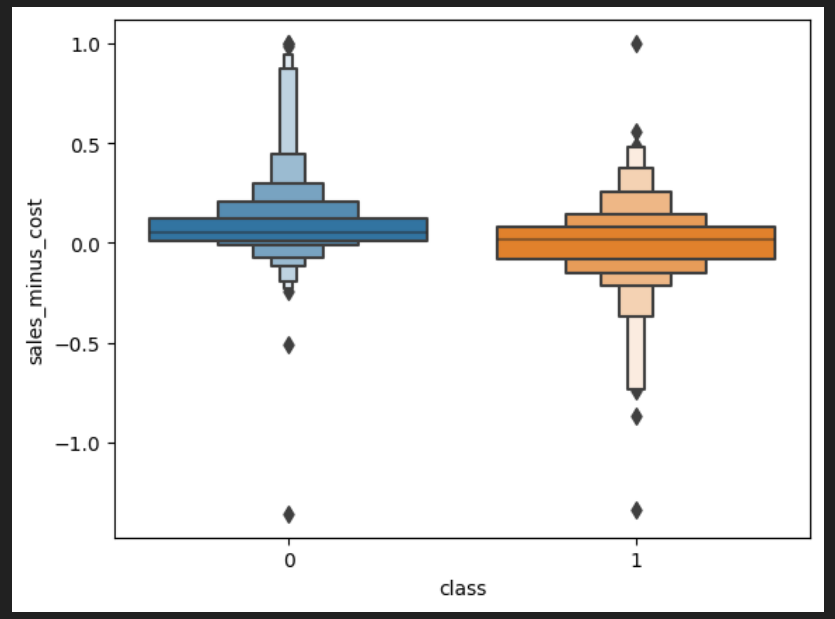




Catplot of class and net_profit/total_liabilities

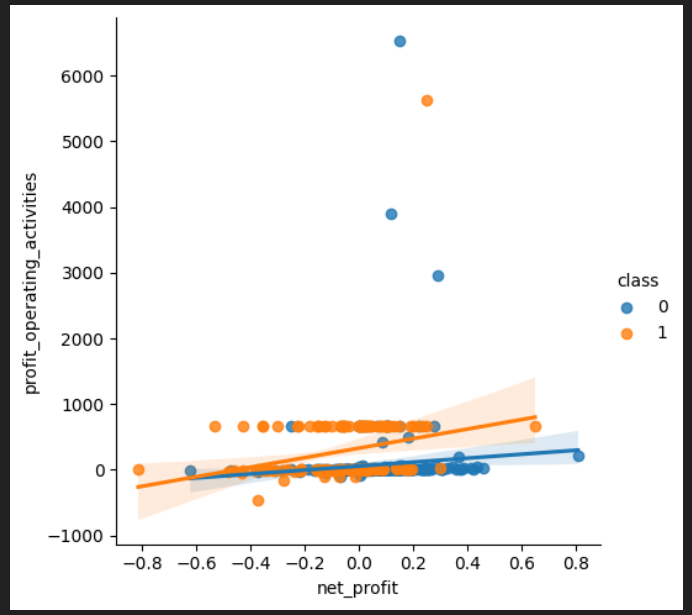
- In both of the graphs, we can see that the average of the values is different according to the class. This indicates that these variables are useful and necessary in predicting whether a company is bankrupt or not.

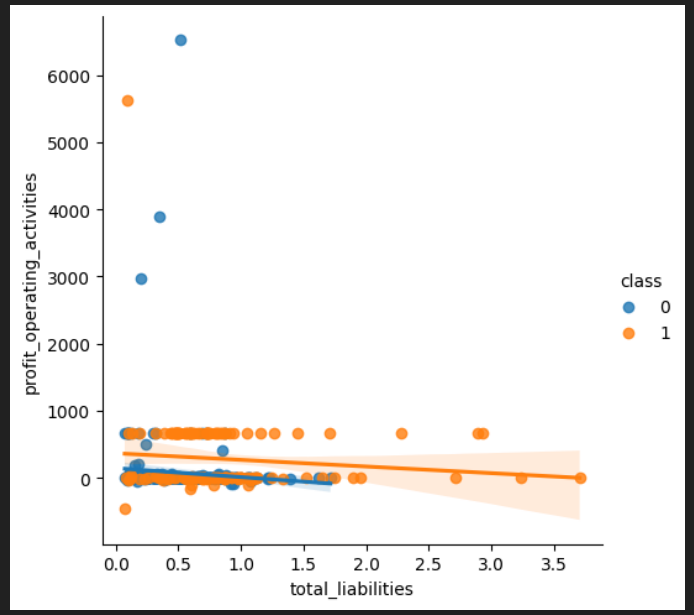


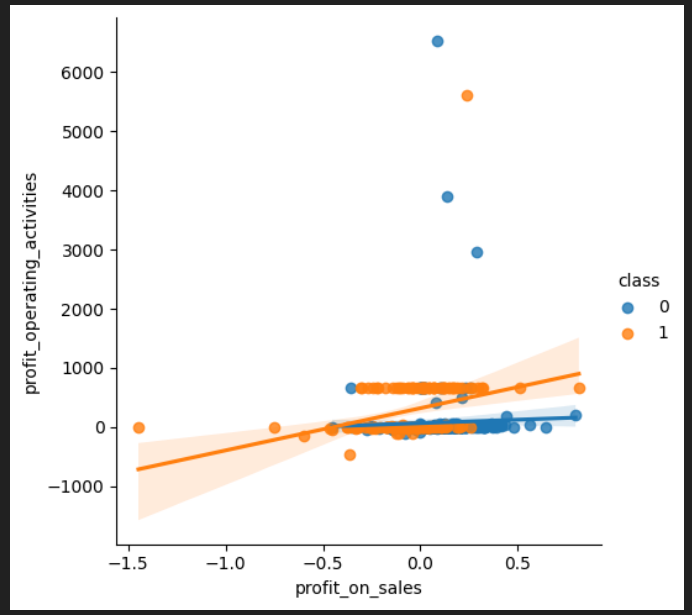


Pointplot and boxenplot of class and sales_minus_cost/profit_on_sales

- Using other seaborn functions, we can see the same observation as before : the average of the values is different according to the class. This indicates that these variables are useful and necessary in predicting whether a company is bankrupt or not.
- We can also conclude that the pointplot method is the most useful for this kind of observations.





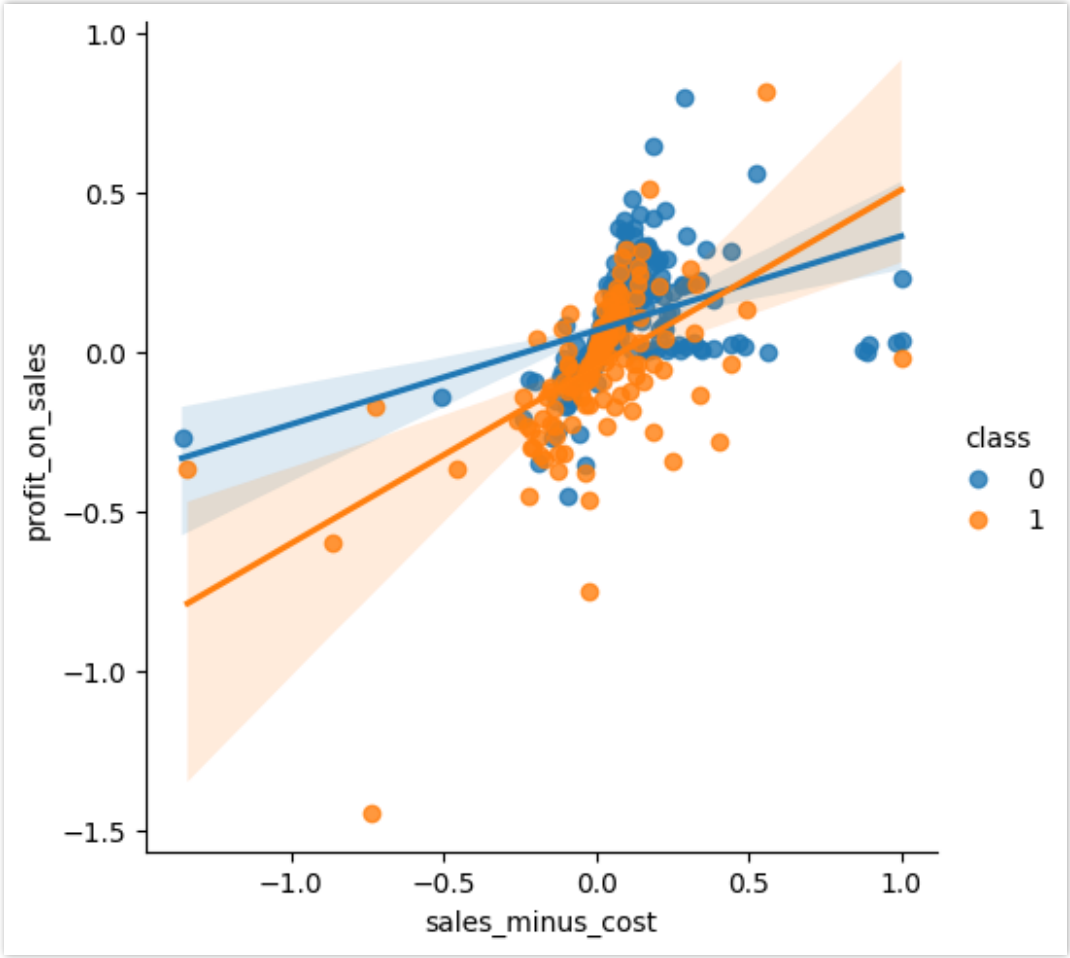


Lmplots of profit_operating_activities and profit_on_sales/total_liabilities/ net_profit

- With these graphs, we observe that regardless of the variable in X, the separation of classes is well observed: we obtain a bankruptcy when profit_operating_activities is worth 0, and the opposite when it is worth 1000

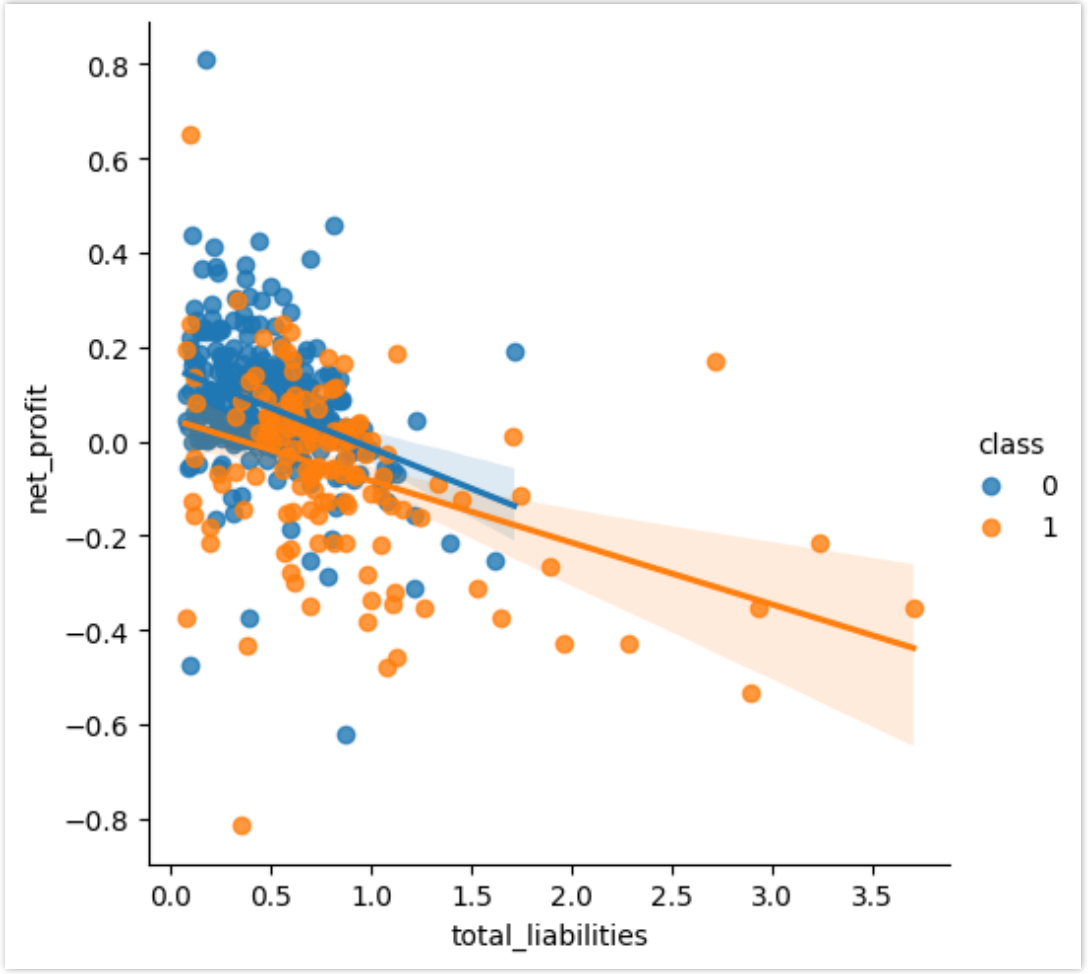
Lmplot of profit_on_sales and sales_minus_cost

- We observe that when the company is bankrupt, its profit_on_sales increases much slower than if it is not bankrupt.



Lmplot of profit_on_sales and sales_minus_cost

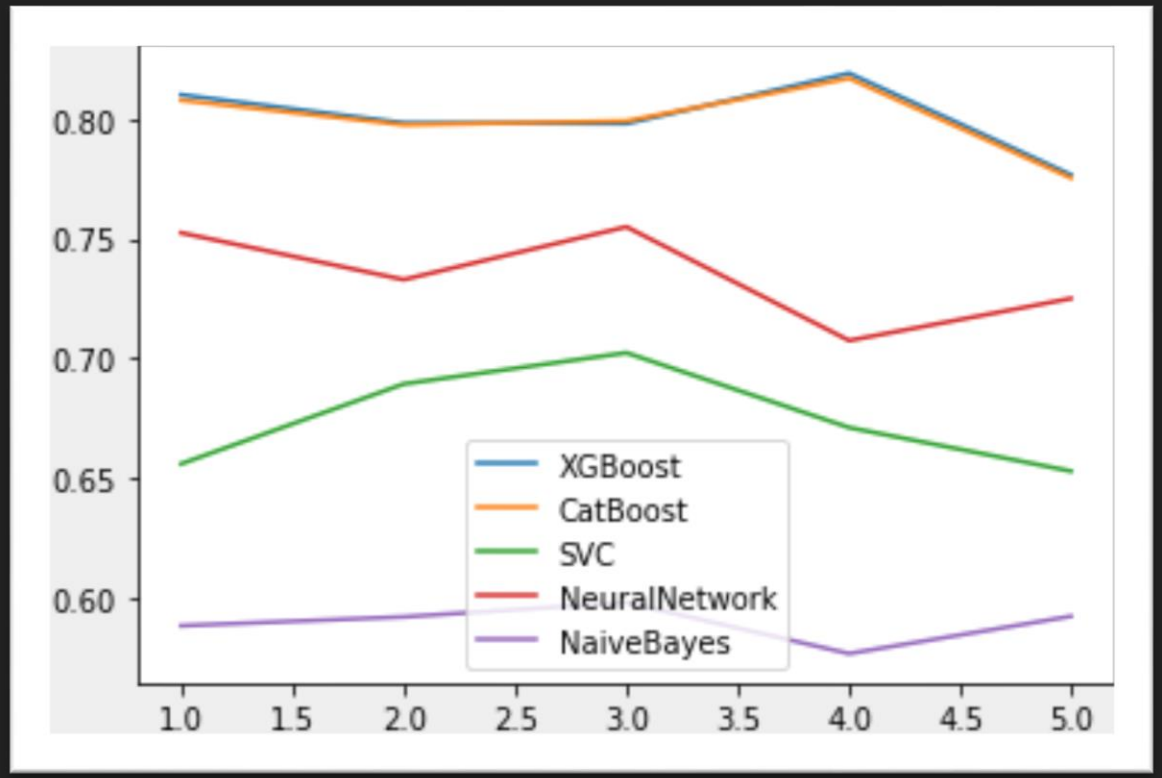
- We observe that when the company is bankrupt, its net_profit decreases much faster than if it is not bankrupt.



Modeling

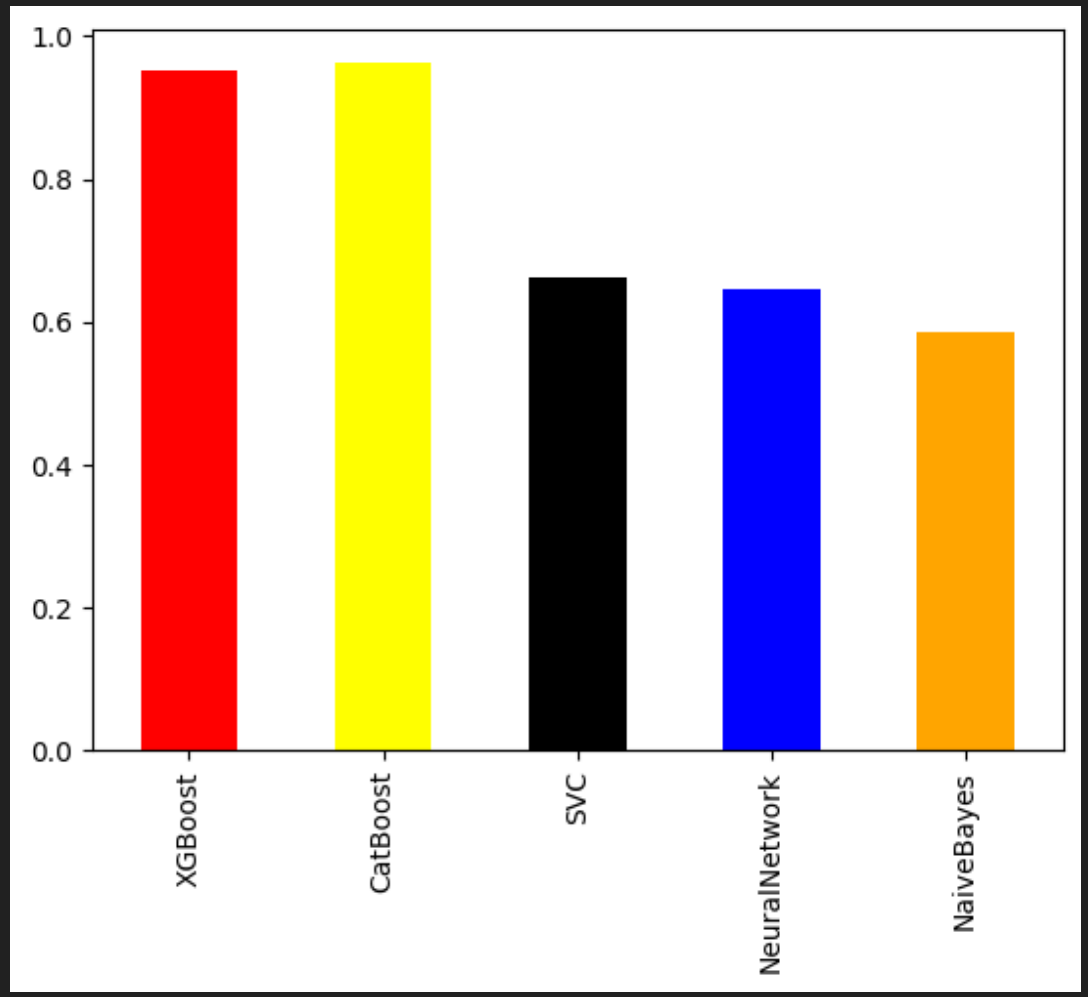
In order to improve the precision of our model, we performed a cross-validation with 5 different estimators:

- XGBoost
- CatBoost
- SVC
- NeuralNetwork
- NaiveBayes



Model comparison

	Score
XGBoost	0.951420
CatBoost	0.962122
SVC	0.663063
NeuralNetwork	0.645931
NaiveBayes	0.585095



The XGBoost classifier has the best accuracy / time ratio, we will use it in the API