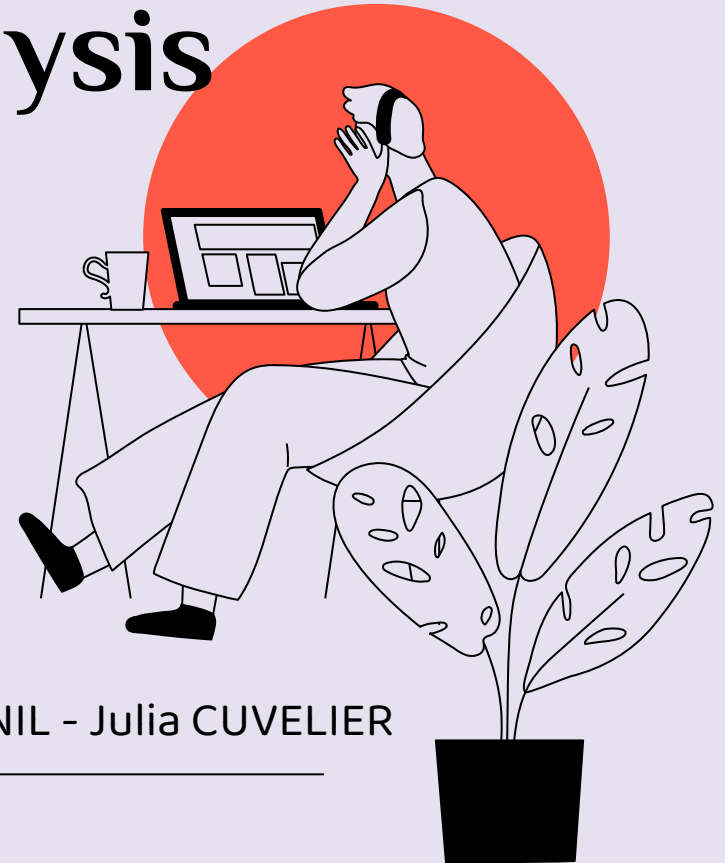


# Python for Data Analysis

## Final Project



DIA 2 - Kyllian ASSELIN DE BEAUVILLE - Arnaud COURNIL - Julia CUVELIER

---

# Table of contents

1

## **Introduction**

Our subject and dataset

2

## **General Informations**

About the dataset

3

## **Preliminary Analysis**

4

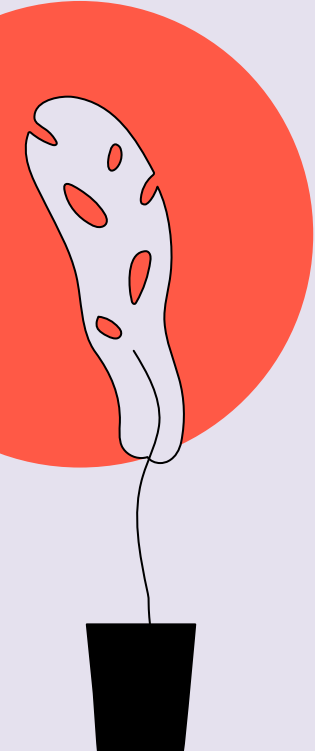
## **Preprocessing and PCA**

5

## **Modeling**

6

## **Conclusion**



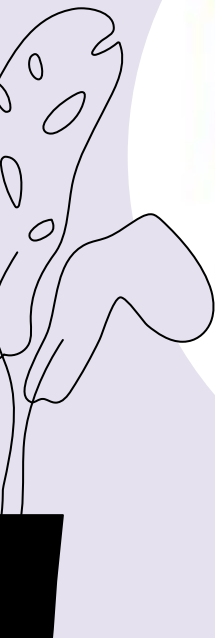


# 1 Introduction

Of our subject and dataset



# Taiwanese Company Bankruptcy Prediction



A company faces bankruptcy when they are unable to pay off their debts. The Taiwan Economic Journal for the years 1999 to 2009 has listed the details of company bankruptcy based on the business regulations of the Taiwan Stock Exchange. Our dataset to analyze will be this data.

Dataset :

<https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction>

We may wonder to what extent parameters such as those studied in the dataset (the profit, the sales, the revenue...) have an influence on the bankruptcy of a company, and therefore how they can help predict the bankruptcy?



# 2 General informations

About the dataset

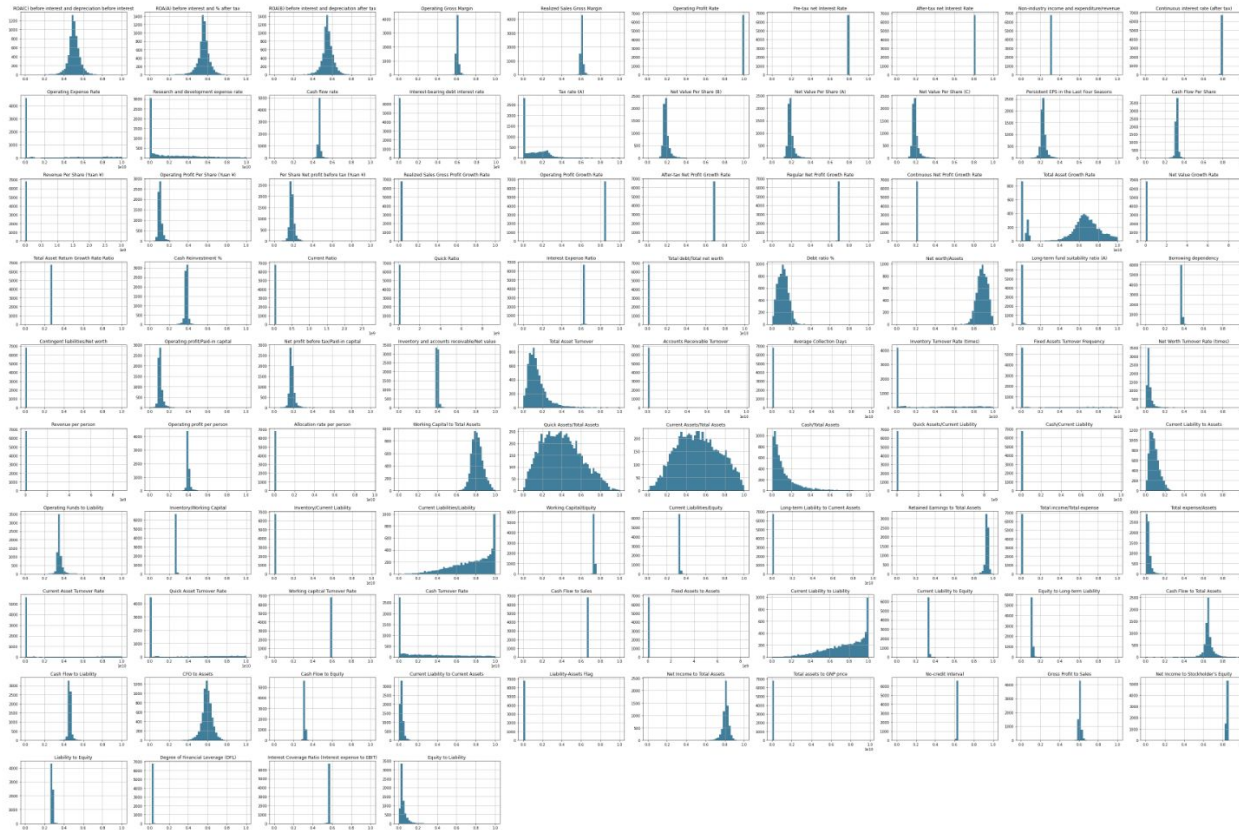


# Dataset :

Size	6819 x 96
Values Type	Int and Float
Missing values ?	No
Duplicated Values ?	No
Unique values ?	1
Standardized data ?	NO



# A quick look at the data



**Fig1:** general view of the distribution of the variables



# A quick look at the data

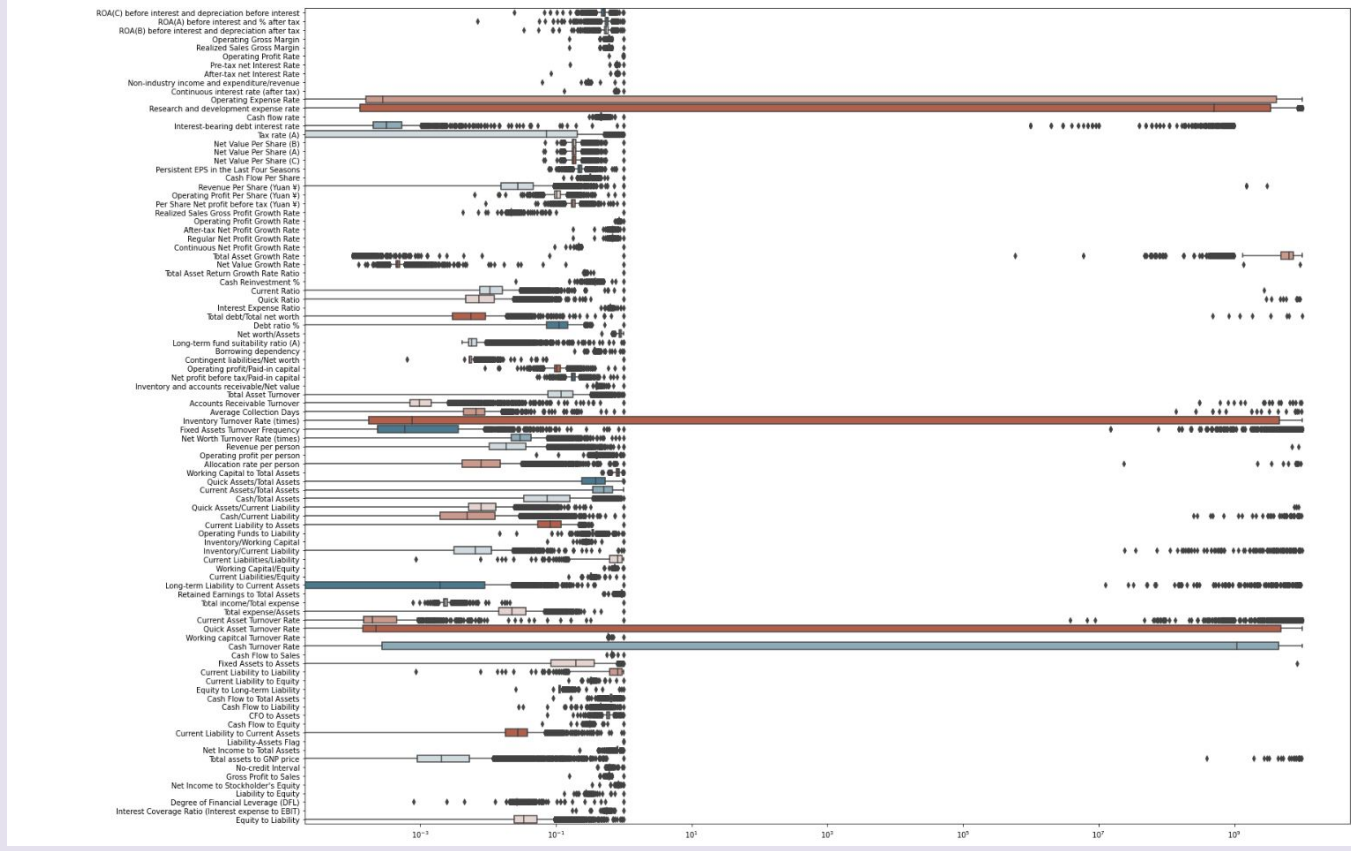


Fig2: boxplot graph of the distribution of the variables

# Comments :



- The data isn't standardized
- Values are very disparate and broad

# Our target

- We want to predict bankrupt
- Let's look at the repartition of the values for this column
- 2 values for this column :  
1 if the company is bankrupt, 0 if it isn't.

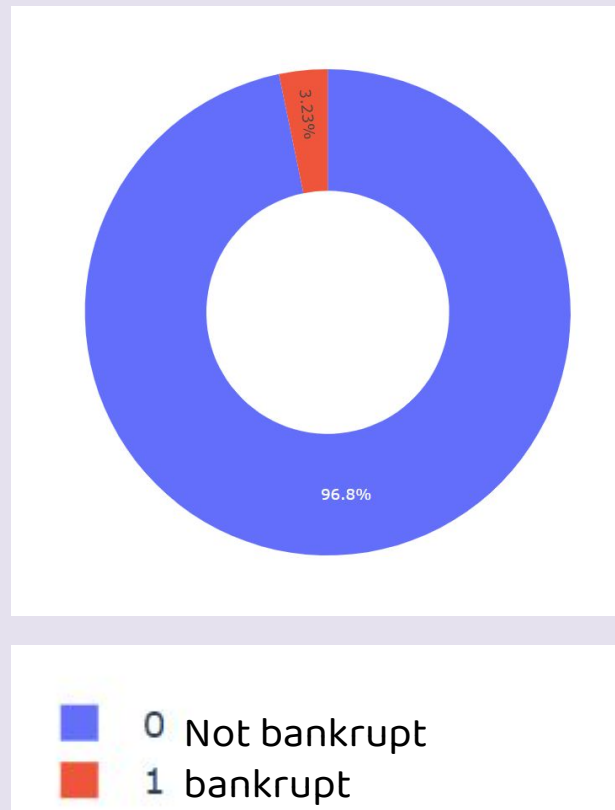


Fig3: Values repartition for bankrupt

# Comments :



- classification strongly imbalanced.
- nearly 97% of 0

# Correlation with the target

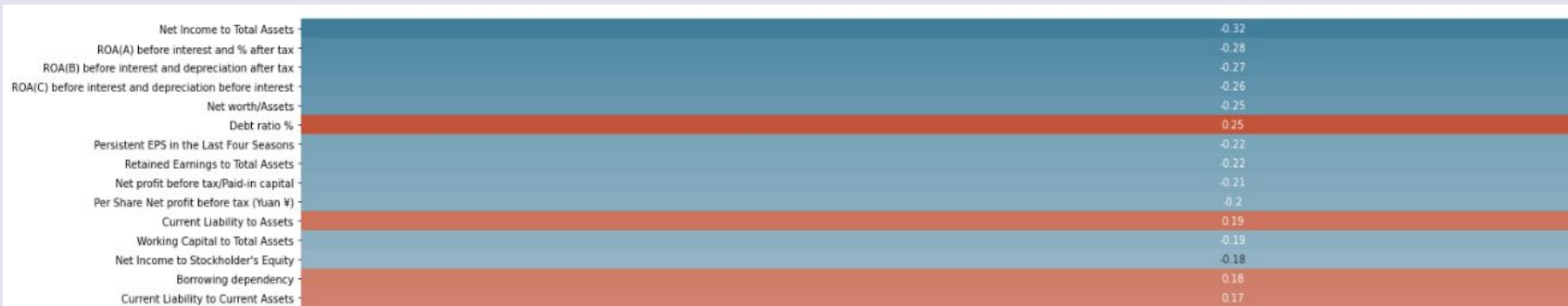


Fig3: Zoom on the heatmap showing the correlation of the variables with the target to better see the 15 more correlated parameters with the target

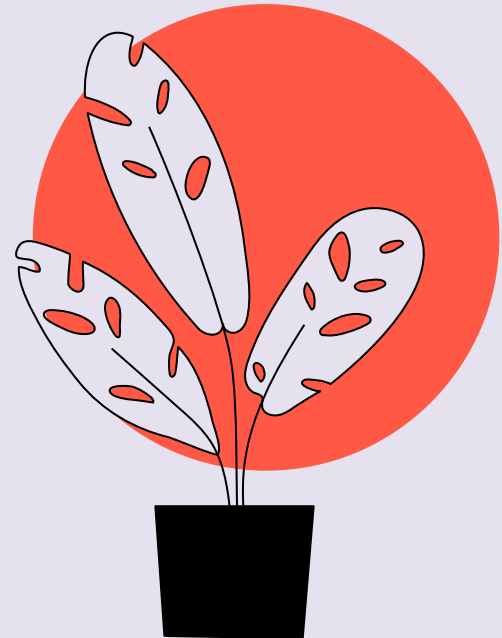
---

# 3 Preliminary Analysis



# Summary and analysis

- Too much features ( 96) so we will have to do a selection
- We have one unique value so we can delete it ( data cleaning)
- For our work, we will have to standardize the values
- Strongly imbalanced classification, we need to take this Into account because for example, predicting 0 every time Will give us a good result, but the model will be bad
- Among the features that seem to be the more correlated With bankrupt, we can find net income to total assets, debt ratio and features about ROA for example. This seems rather coherent.





4

# Preprocessing and PCA

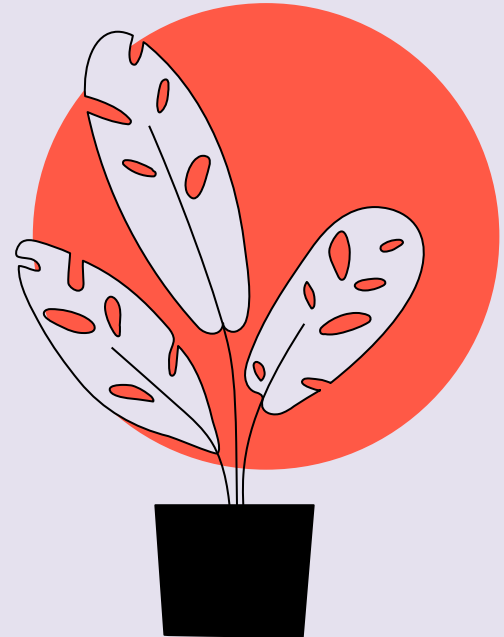




# Preprocessing

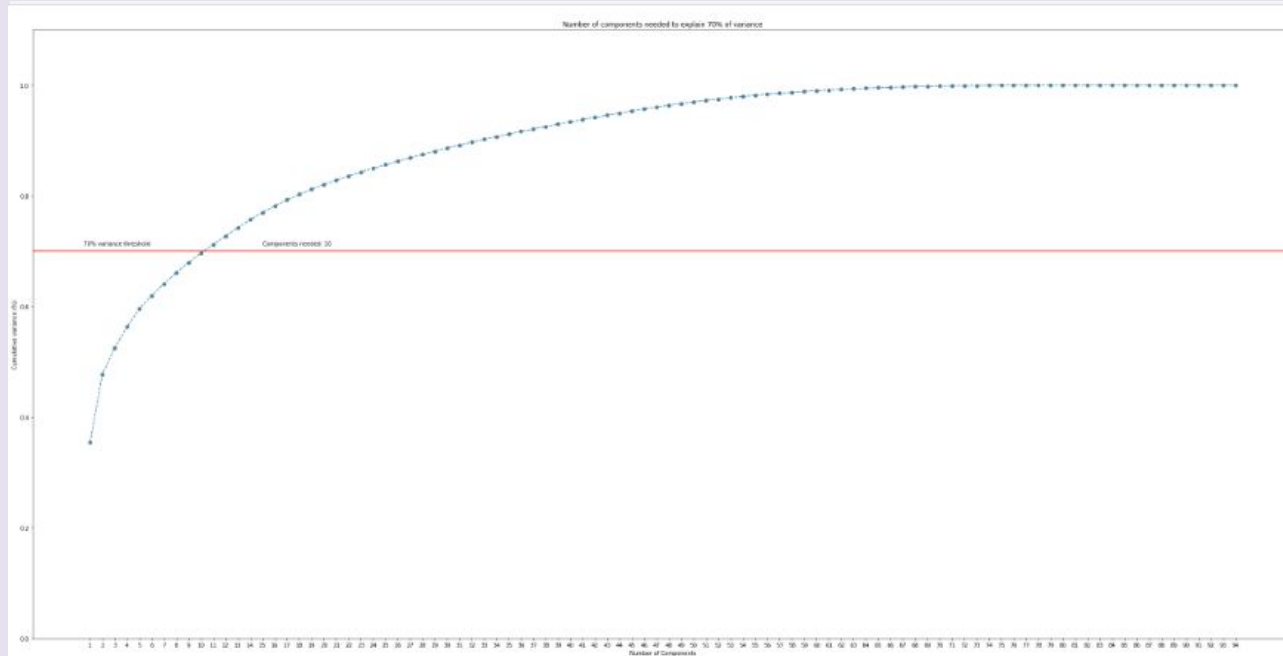
Before starting our work we :

- split features and target
- split our dataset into train and test set
- standardize the numeric features
- oversample the over represented classes to avoid making mistakes

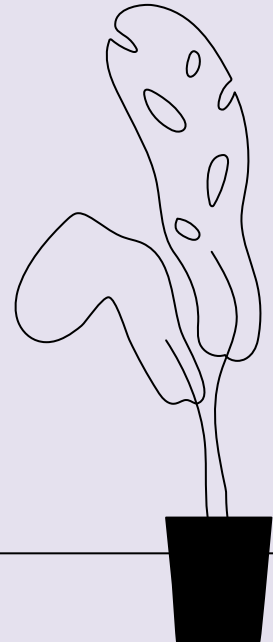


# PCA

We need to reduce the number of features so we will do a PCA.



Thanks to this graph, we can easily see how many components are needed to have 70% threshold variance. In this case, we will need 10 components.



# PCA

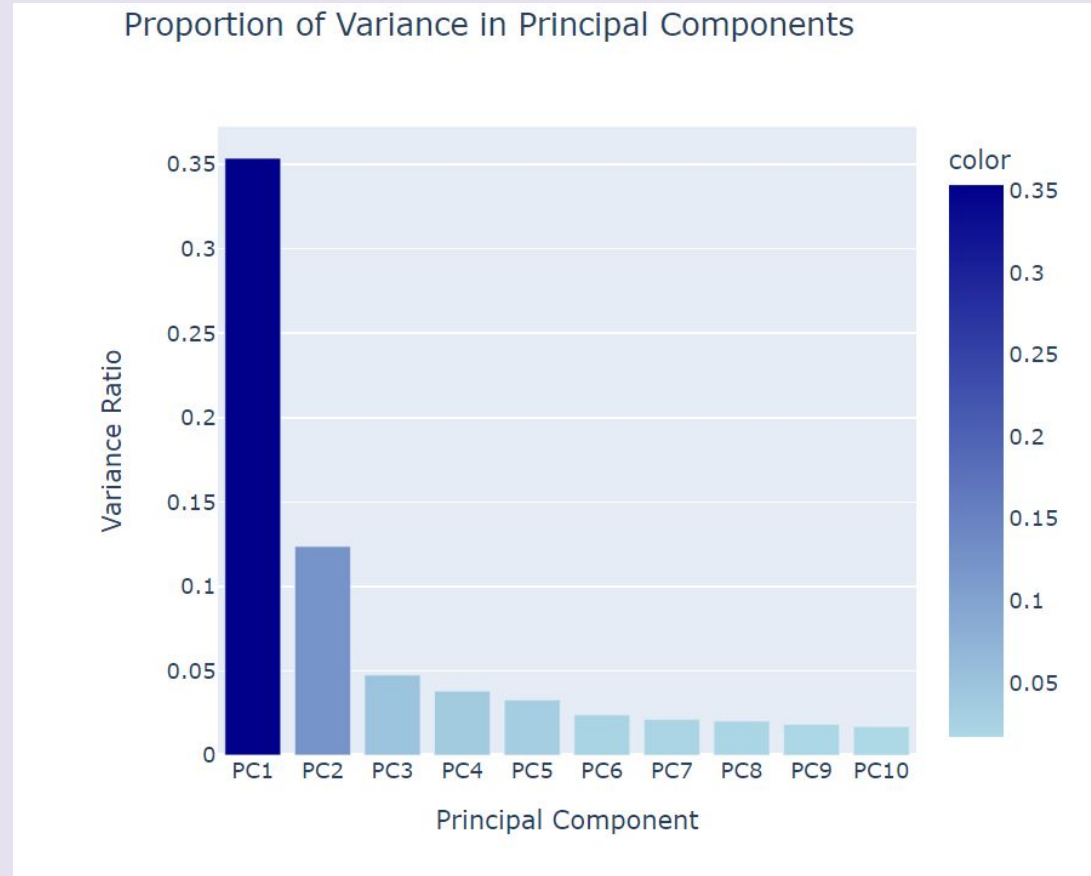
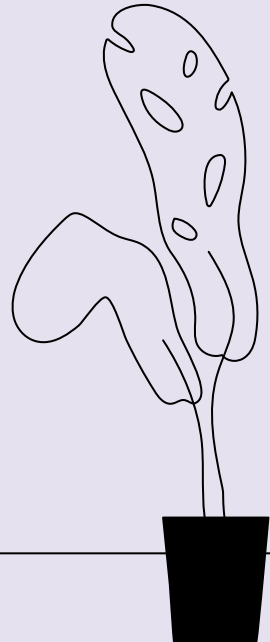


Fig5: The proportion of variance for the 10 principal components



# PCA

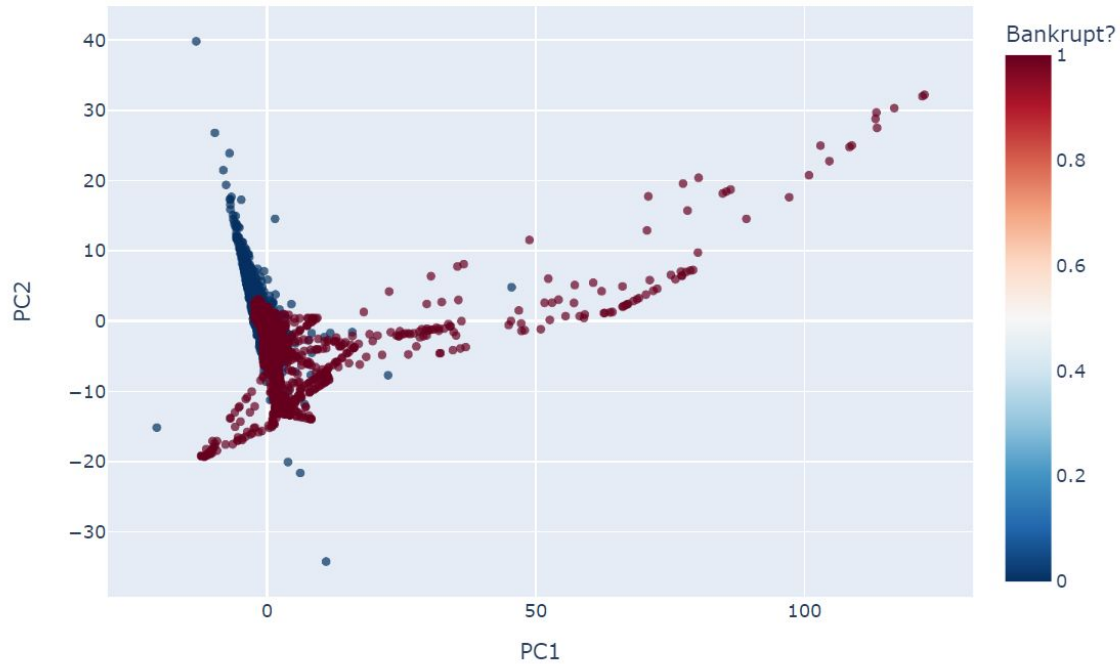
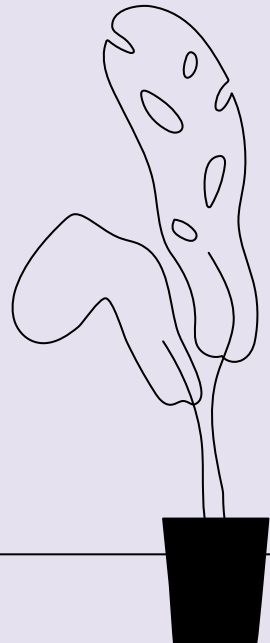


Fig6: Observation on the 2 first components



# PCA

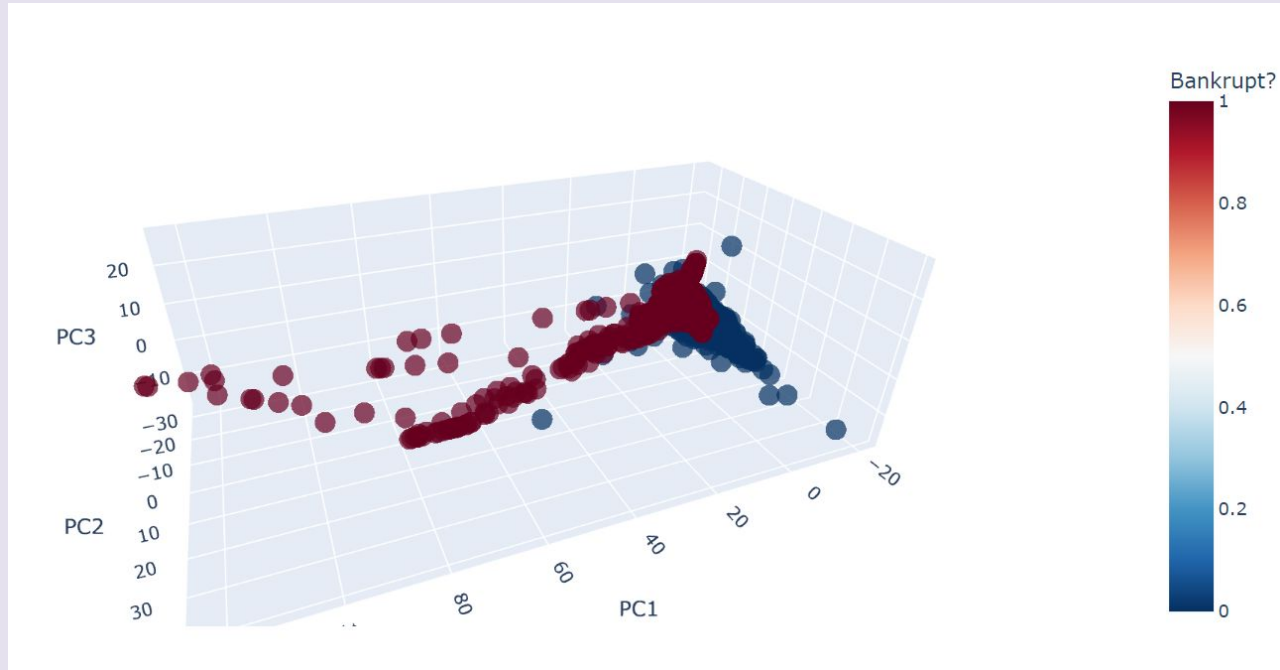
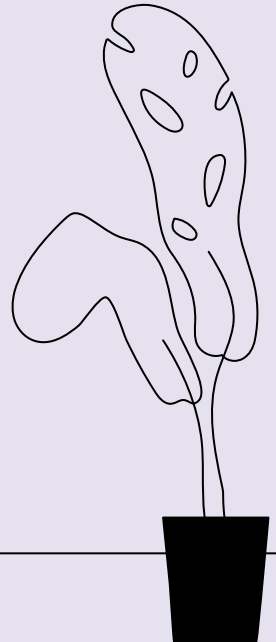


Fig7: Observation on the 3 first components





# 5

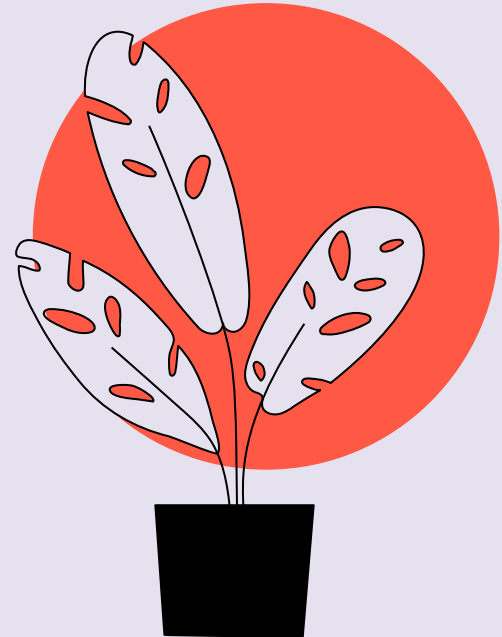
# Modeling

# For modeling :

- We choose f2-score because it seemed more appropriate

We tried :

- Logistic regression
- Support vector machine (Linear Kernel)
- K – nearest neighbors
- Support Vector Machine (RBF Kernel)
- Decision Tree
- Bagging
- Random Forest
- Gradient Boosting



# Evaluation

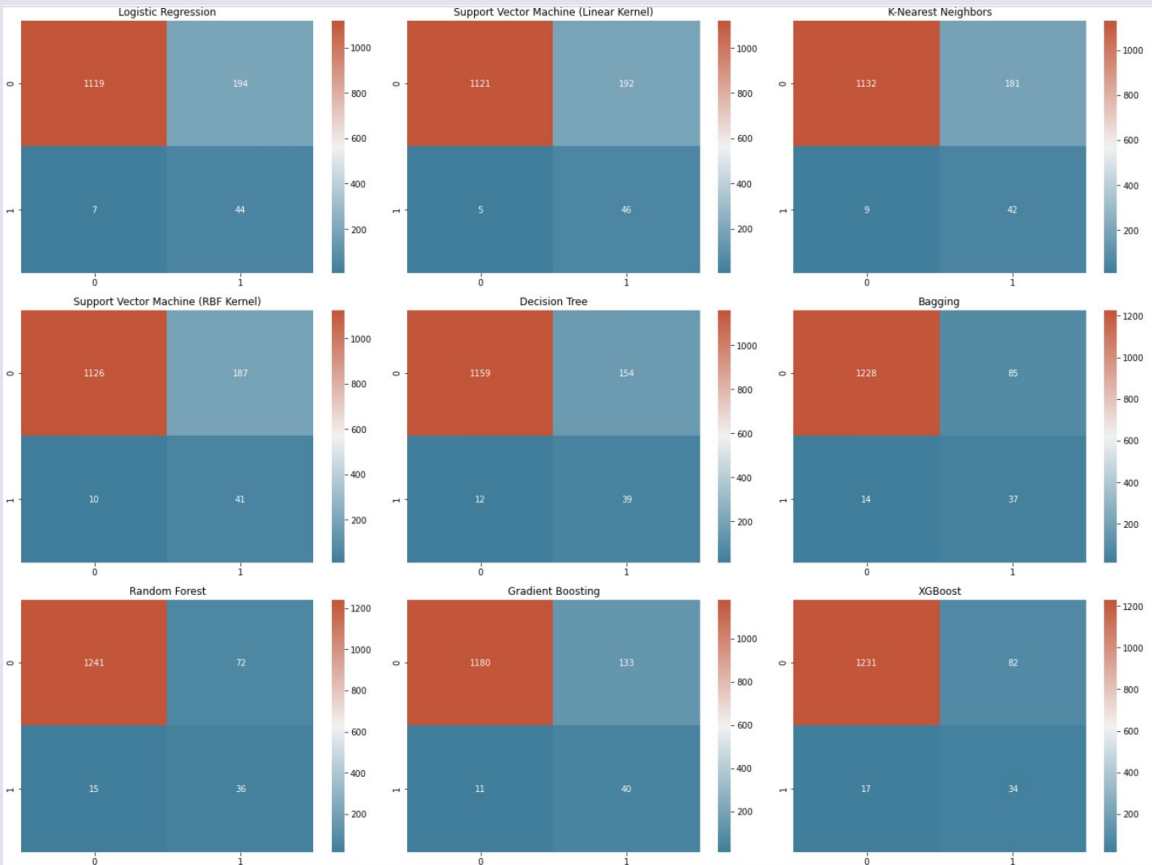
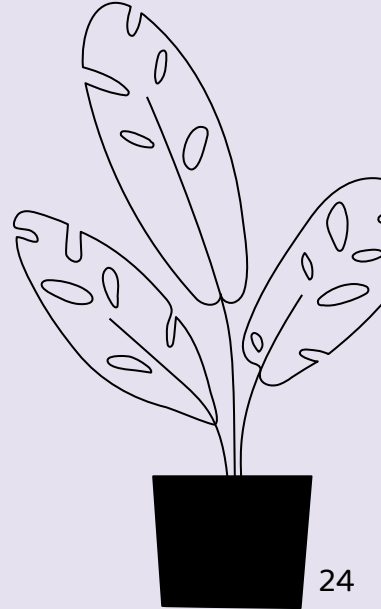


Fig8: Confusion matrix





# Evaluation

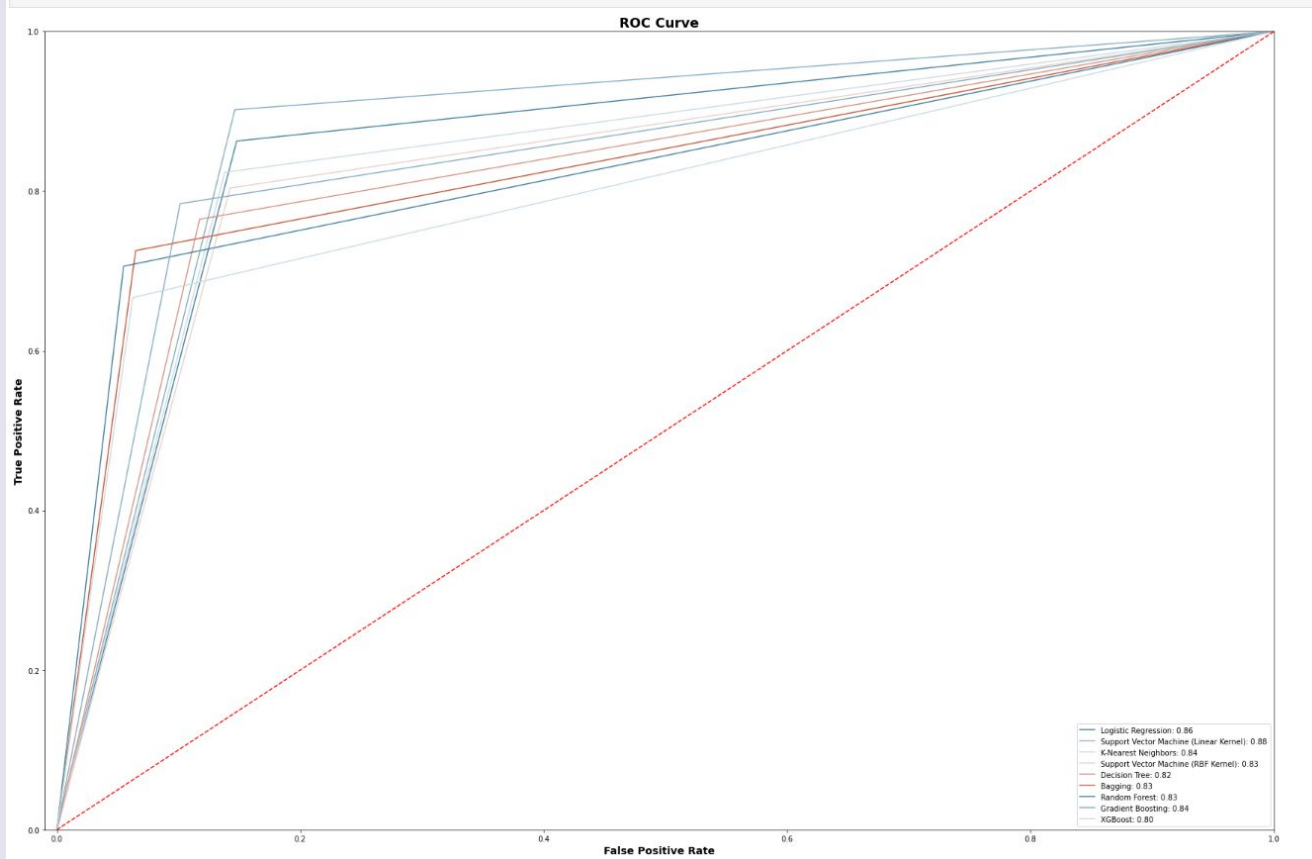
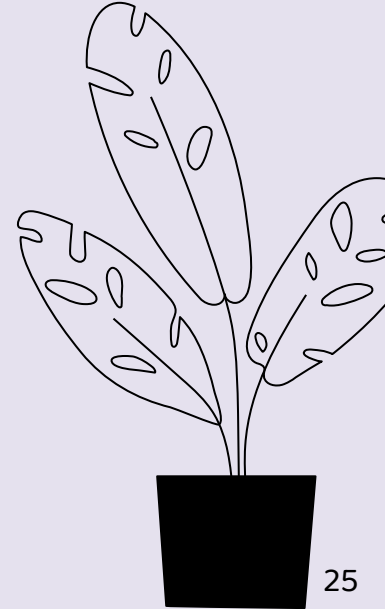


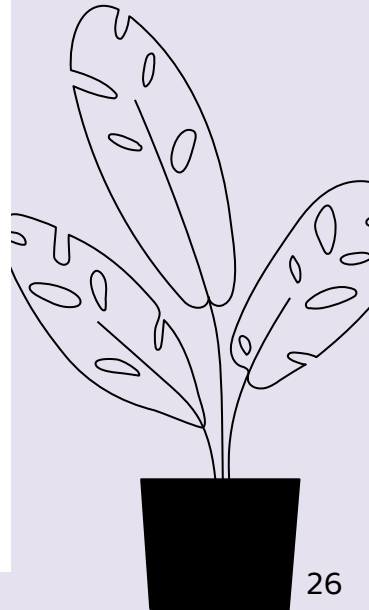
Fig9: ROC curves

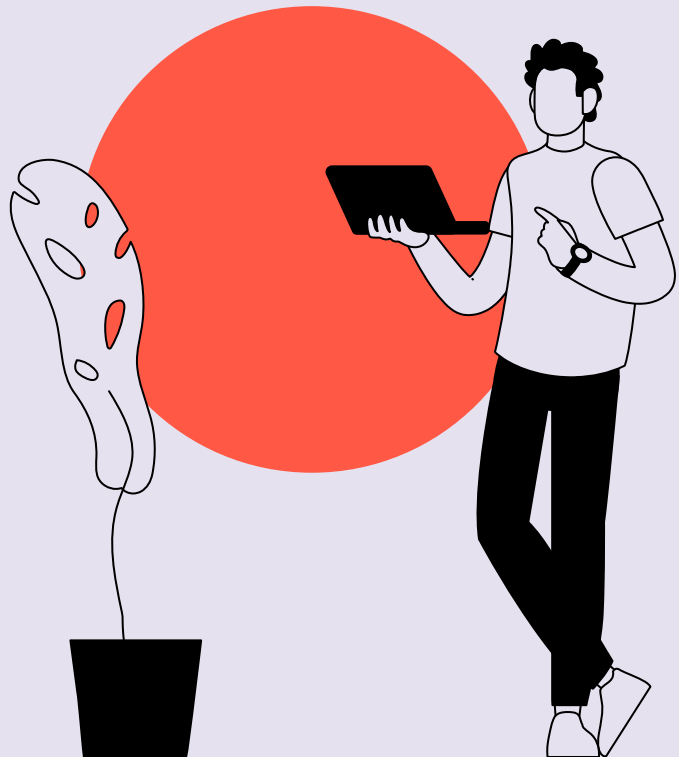


# Summary

Fig10: Summary and Accuracy

	Model	Parameters	Accuracy	F2-score
Ranking				
1	Random Forest	{'max_features': 'sqrt', 'n_estimators': 500}	0.94	0.58
2	Bagging	{'n_estimators': 200}	0.93	0.57
3	XGBoost	{'learning_rate': 0.1, 'n_estimators': 200, 'o...	0.93	0.53
4	Gradient Boosting	{'learning_rate': 0.1, 'n_estimators': 200}	0.89	0.53
5	Support Vector Machine (Linear Kernel)	{'C': 1, 'penalty': 'l2'}	0.86	0.52
6	Logistic Regression	{'C': 10, 'penalty': 'l2', 'solver': 'newton-cg'}	0.85	0.50
7	Decision Tree	{'max_depth': 10, 'min_samples_split': 2}	0.88	0.49
8	K-Nearest Neighbors	{'metric': 'manhattan', 'n_neighbors': 11, 'we...	0.86	0.49
9	Support Vector Machine (RBF Kernel)	{'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}	0.86	0.47





# 6 Conclusion



# Conclusion

To conclude we can say that the parameters studied during this study are good indicators of the bankruptcy or not of a company.

With the precautions we took to make our models as accurate as possible and usable on other similar datasets, we obtained rather good predictions with for the random forest model, an accuracy of 94%.

