Statistical learning project

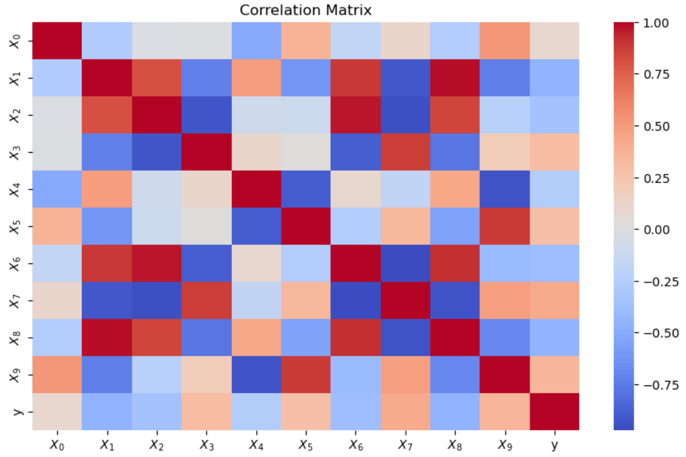
Thomas Labreur
Teacher : Franck Iutzeler

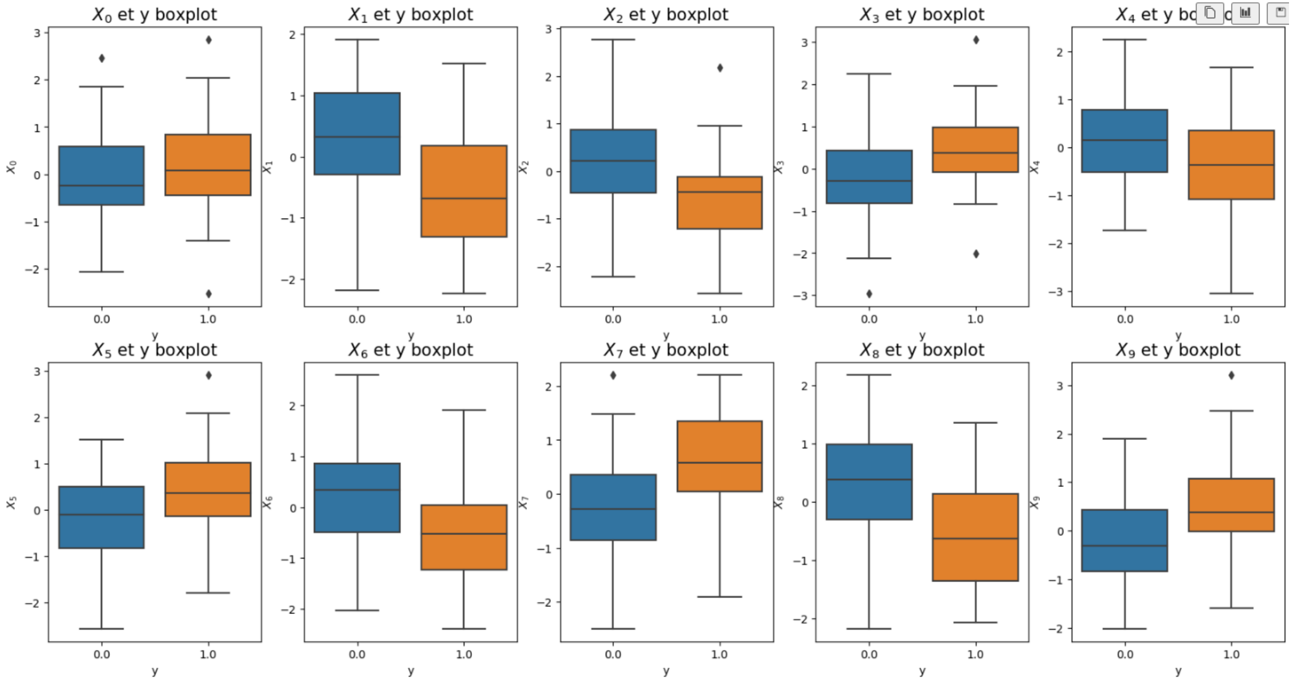
Goal : find the best classifier for a given dataset

M2 SID

2024 - 2025

Linear correlations in train data



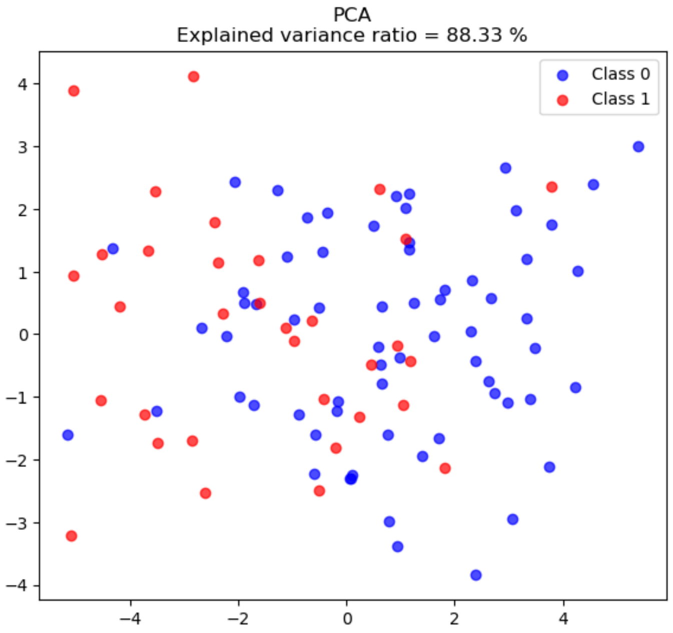


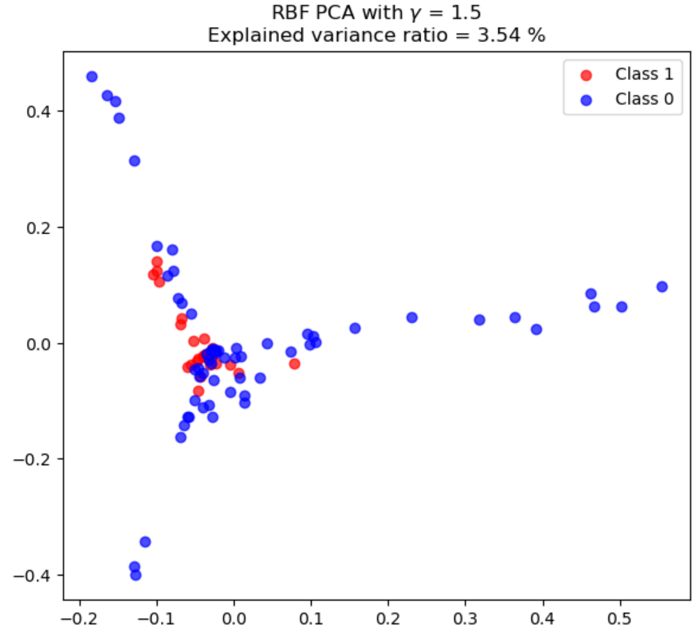
X1, X2, X3, X6, X7 and X8 seem to provide redundant information

Same for X4, X5 and X9

No individual linear correlation between Xi and y

Principal Components Analysis





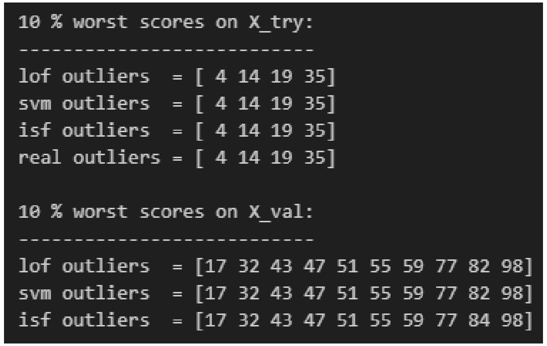
1) Classes are not linearly separable so linear models won’t be adapted

2) They still don’t seem to be separable in the Hilbert space associated with RBF kernel but it’s only a 2D projection. The two first principal components only explain 3,54 % of variance.

3) With parameter 1.5, one can obseve that class 1 data seems to behave differently than class 0 in the surrogate Hilbert Space. Maybe a kernel method with this parameter could be adapted.

Outlier detection

3 Algorithms tried :
LOF, OneClassSVM, IsolationForest

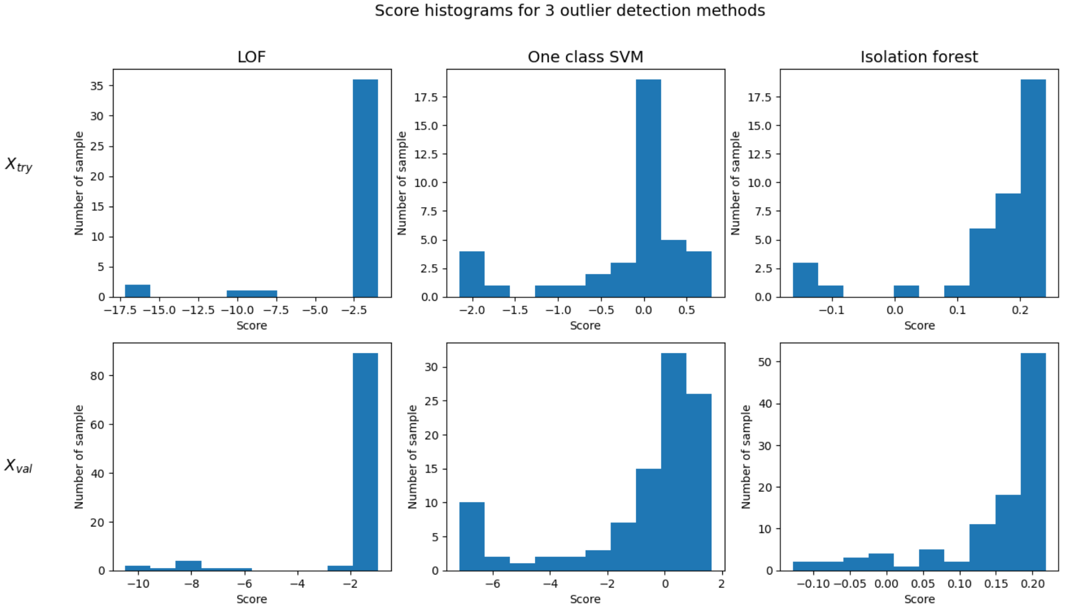


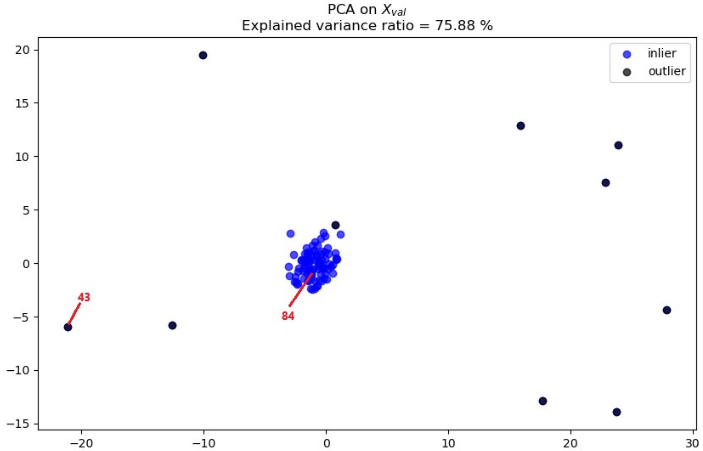
```
10 % worst scores on X_try:
---------------------------
lof outliers  = [ 4 14 19 35]
svm outliers  = [ 4 14 19 35]
isf outliers  = [ 4 14 19 35]
real outliers = [ 4 14 19 35]

10 % worst scores on X_val:
---------------------------
lof outliers  = [17 32 43 47 51 55 59 77 82 98]
svm outliers  = [17 32 43 47 51 55 59 77 82 98]
isf outliers  = [17 32 43 47 51 55 59 77 84 98]
```

Agree and right on X_try

Almost agree on X_val





According to LOF

To pick X_val outliers, I choose to trust the result of LOF and OneClassSVM because :

- 2 methods over 3 provided this result ;

- Those 2 method’s scores better split between outlier and inlier (see histograms) ;

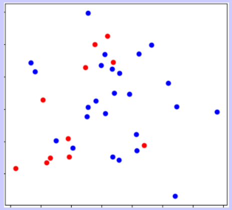
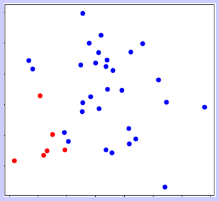
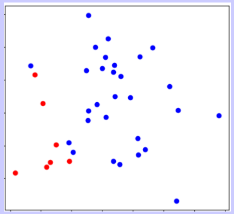
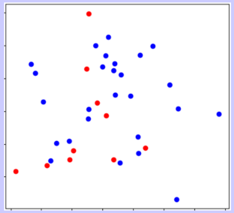
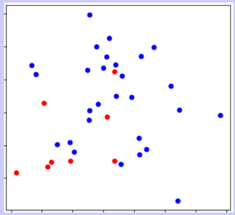
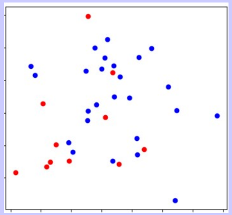
- Visually, the 43th point seems more likely to be an outlier than the 84th (see PCA).

Classifiers comparison

PERFORMANCES

	5-folds cross validation on X		Test on X_try			
	Best params (By grid search)	Mean f-score	precision	recall	f-score	accuracy
Linear SVC	C : 1	0,48	0,83	0,46	0,59	0,81
RBF SVC	C: 10 | gamma: 0,01	0,46	0,71	0,46	0,56	0,78
KNN	n_neighbors: 1 | weigths: ‘uniform’	0,48	0,5	0,46	0,48	0,69
Random Forest	max_depth : 5 | max_features: ‘sqrt' | min_samples_leaf: 2 | min_samples_split: 10 | n_estimators: 100	0,54	0,62	0,46	0,53	0,75
XGBoost	colsample_bytree: 0.6 | gamma: 0.5 | learning_rate: 0.1 | max_depth: 5 | n_estimators: 200 | subsample: 0.6	0,58	0,55	0,55	0,55	0,72

PREDICTIONS

Truth	Linear svc	RBF SVC	KNN	Random forest	XGBoost
					

Model selection and error

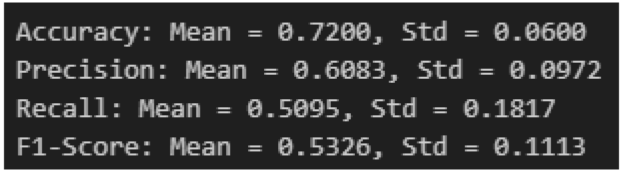
Criterion for model selection :

- As class 1 only represents 33 % of the dataset, accuracy isn’t a good metric. Linear SVC has 80 % accuracy by predicting 0 for the majority of the dataset.
- Thus, we have to pay attention to recall, which penalizes predicting 0 instead of 1.
- F1-score is also a good metric that takes recall into account, and precision too.
- I used it for cross-validation, the mean F1-core over the 5 folds is a more robust metric than f1-score on X-try (it’s more independent from the training data).

➡ With this criterion, XGBoost is the best classifier.

Error estimation :

- **1) Prediction error :** obtained by 5-folds cross-validation on (X,y).



```
Accuracy: Mean = 0.7200, Std = 0.0600
Precision: Mean = 0.6083, Std = 0.0972
Recall: Mean = 0.5095, Std = 0.1817
F1-Score: Mean = 0.5326, Std = 0.1113
```

- **2) Outlier detection error :** Based on visual representation on slide 4.

After scaling, outliers are easily identified, except maybe for 1 over 10, so I predict 10 % error in the worst case.