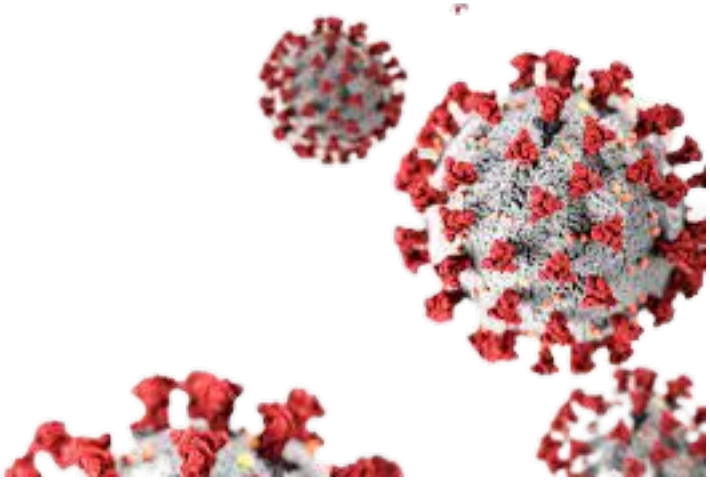
Data set 1 - Covid-19 Test based ML
Data set 2 - Online Payment Fraud Detection

8 project

Marie BAI, Alex CULPIN, Moussa SIDIBE

Covid-19 Test based ML

AI



LEARNING PLANET INSTITUTE

AGENDA

1. Presentation & EDA of the dataset
2. Representation Learning
3. Random Forest
4. SVM
5. Conclusion and comparaison

Presentation and EDA

- 11 features and 1 target
- 2499 observations

#	Column	Non-Null Count	Dtype
0	Country	2499 non-null	object
1	Age	2499 non-null	int64
2	Gender	2499 non-null	object
3	fever	2499 non-null	int64
4	Bodypain	2499 non-null	int64
5	Runny_nose	2499 non-null	int64
6	Difficulty_in_breathing	2499 non-null	int64
7	Nasal_congestion	2499 non-null	int64
8	Sore_throat	2499 non-null	int64
9	Severity	2499 non-null	object
10	Contact_with_covid_patient	2499 non-null	object
11	Infected	2499 non-null	int64

Presentation and EDA

```
The modalities taked by Gender are:['Male' 'Transgender' 'Female']
The modalities taked by Bodypain are:[1 0]
The modalities taked by Runny_nose are:[0 1]
The modalities taked by Difficulty_in_breathing are:[0 1]
The modalities taked by Nasal_congestion are:[0 1]
The modalities taked by Sore_throat are:[1 0]
The modalities taked by Severity are:['Mild' 'Moderate' 'Severe']
The modalities taked by Contact_with_covid_patient are:['no' 'not known' 'yes']
The modalities taked by Infected are:[0 1]
```

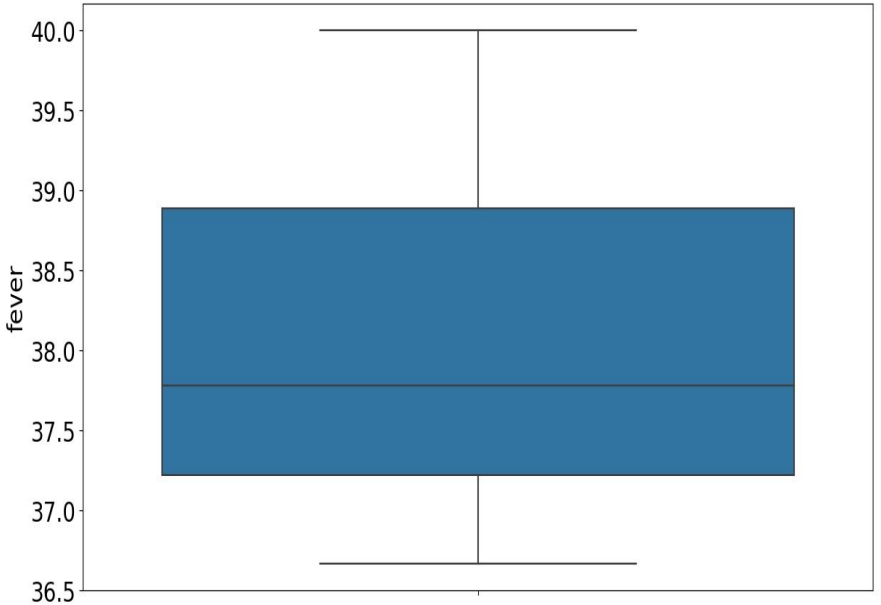
For binaries variables 1:Yes and 0:No

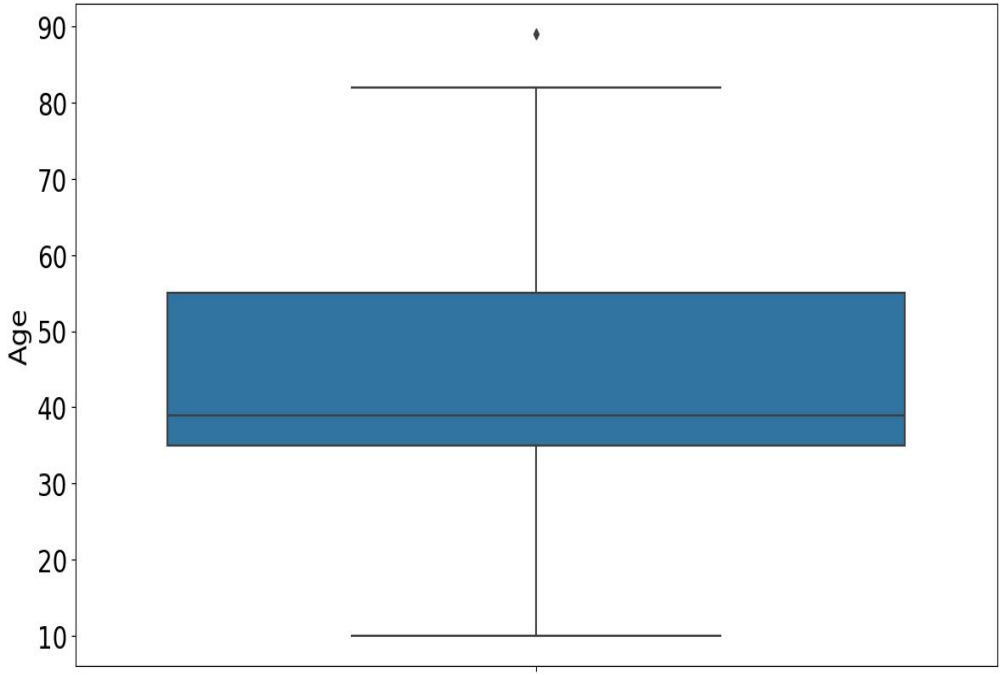
The dataset was cleaned, no NaN

LEARNING PLANET INSTITUTE

Presentation and EDA

- The distribution of temperature is quite normal. There are no outlier.

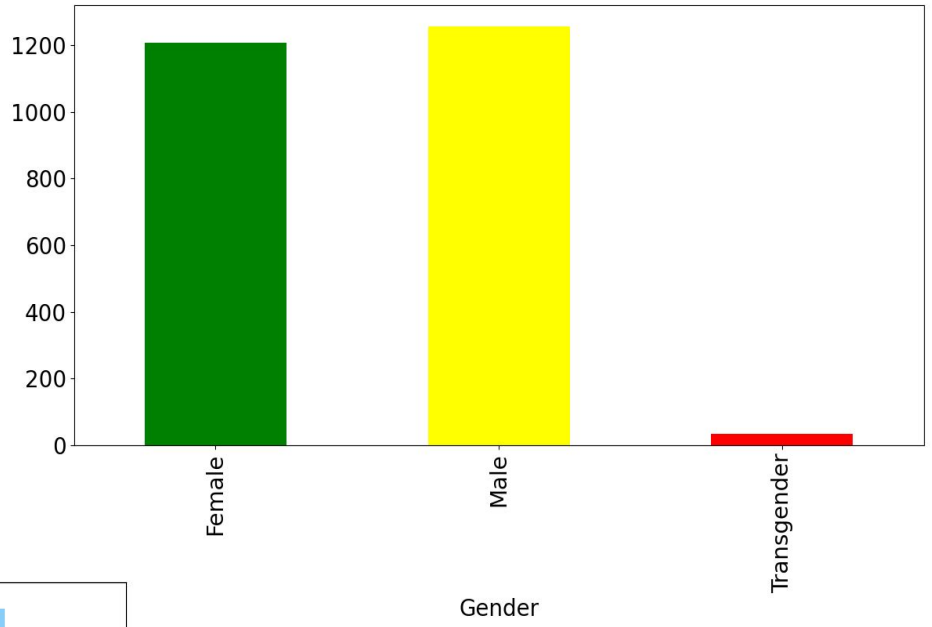


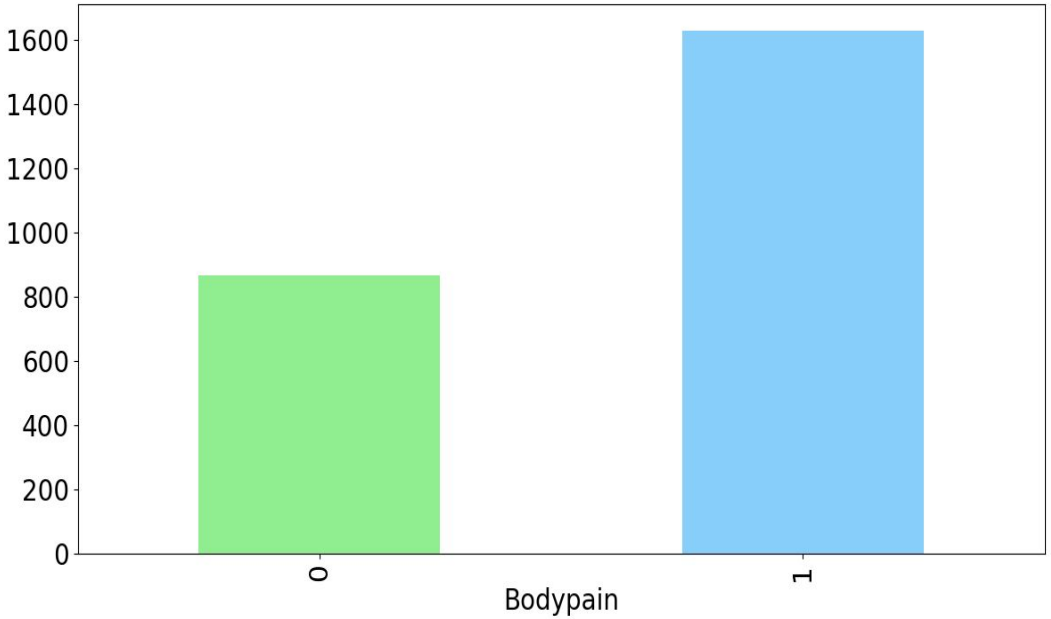


- Most of our indivius are aged 39 years old
- Age is between 10 and 89
- Just one outlier

Presentation and EDA

- We have 1207 females, 1257 Males and 34 Transgenres





- 1630 of observations suffer from body pain against 868

Representation learning

Correlation analysis

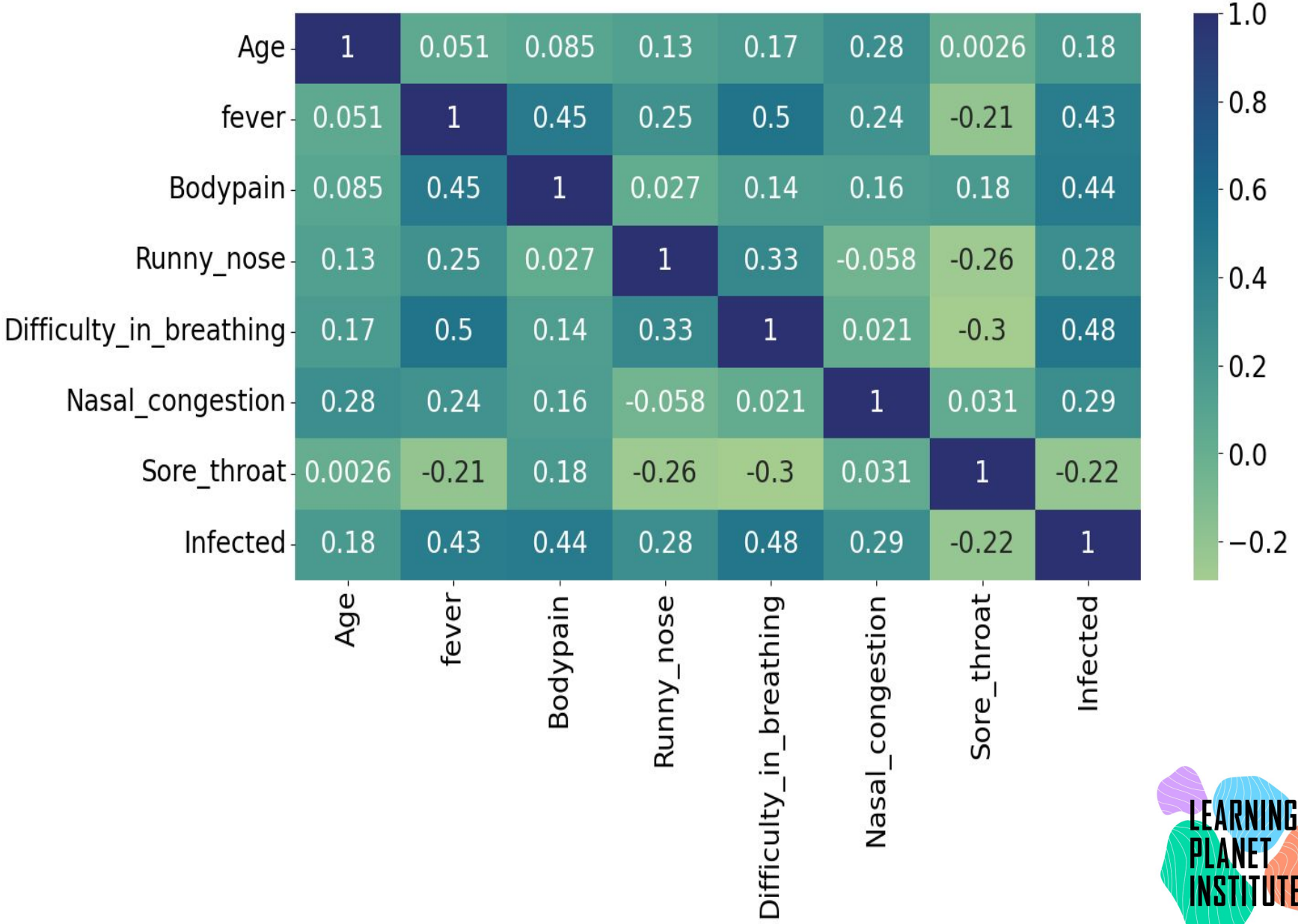
Data normalisation

With Standardscaler

- PCA for features that are high correlated to each other
- TSNE for none linear data reduction stracture

LEARNING PLANET INSTITUTE

Correlation analysis



Standardization & representation learning

Why we scale data before ?

- Firstly, PCA is based on covariance matrix
- Feature have different scale then different variance
- The feature with the high variance due to scale will be well represented than others
- Finally we reduced all variances to 1 and mean to 0

$$z = \frac{x - \mu}{\sigma}$$

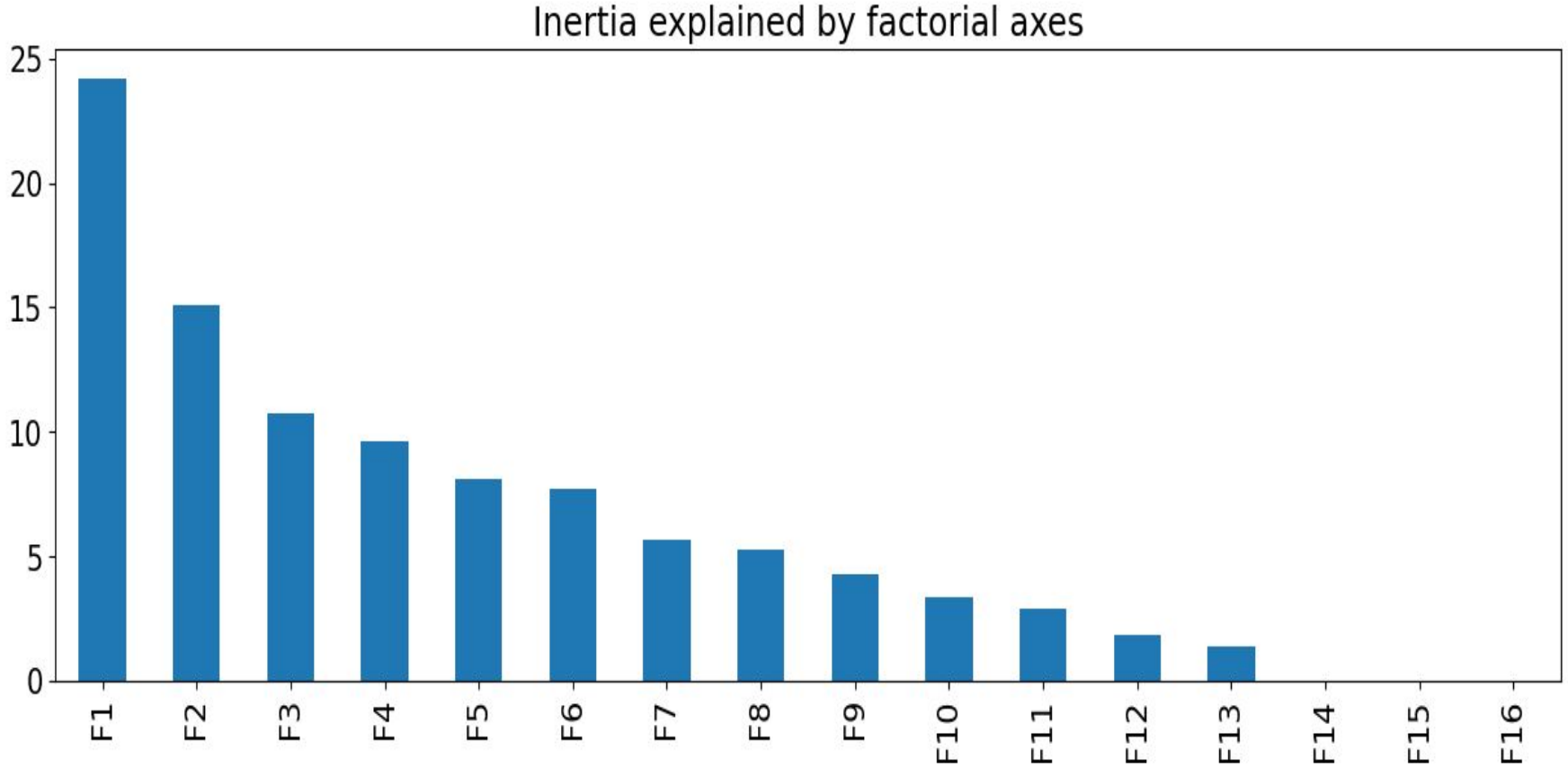
```
from sklearn.preprocessing import StandardScaler
```

```
sta = StandardScaler()
```


LEARNING PLANET INSTITUTE

representation learning

Explained variance by factorials axes

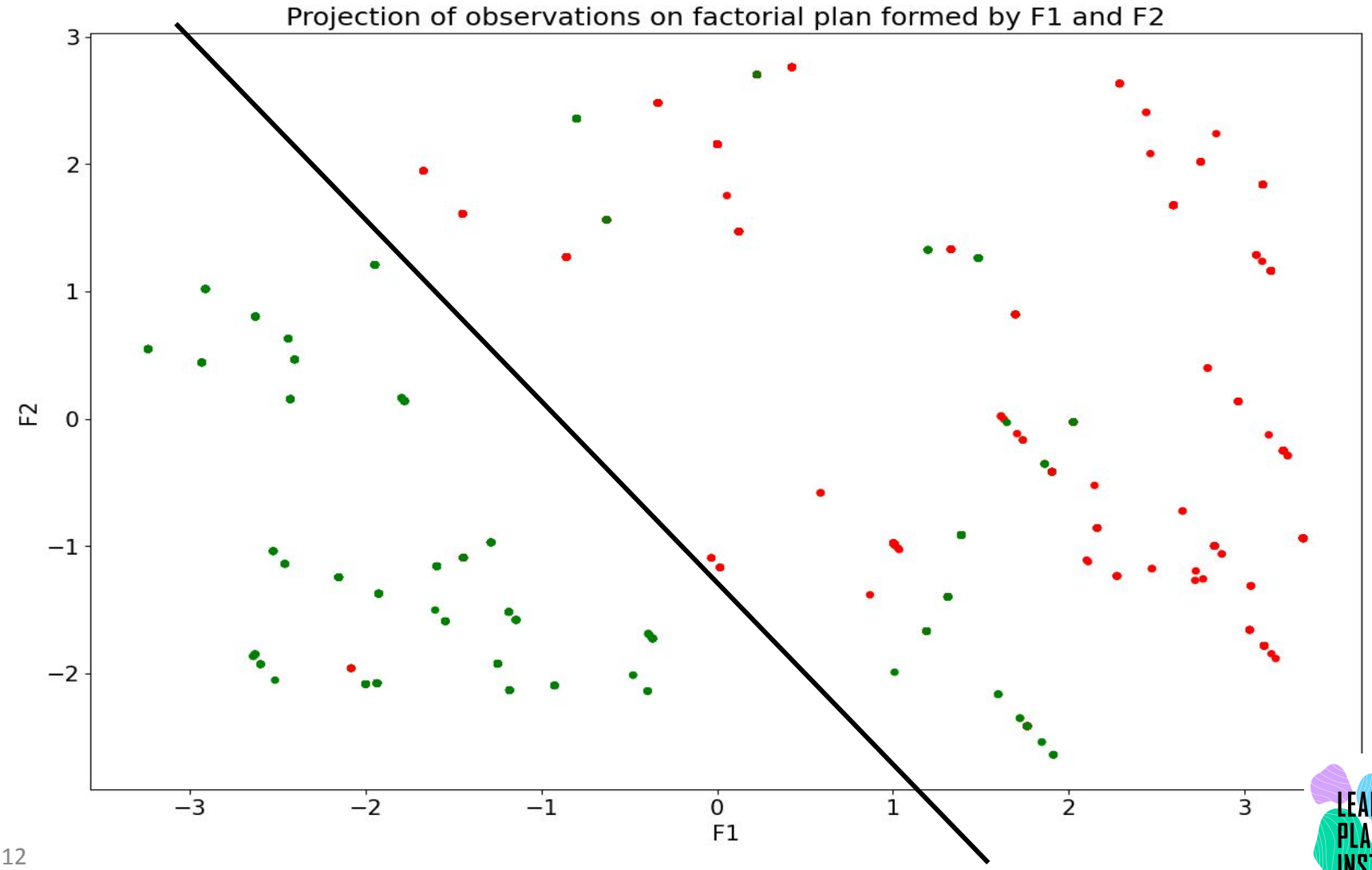


- The data is summarized into 13 components by 16.
- The two first components explain 40% of variability of the data.

LEARNING PLANET INSTITUTE

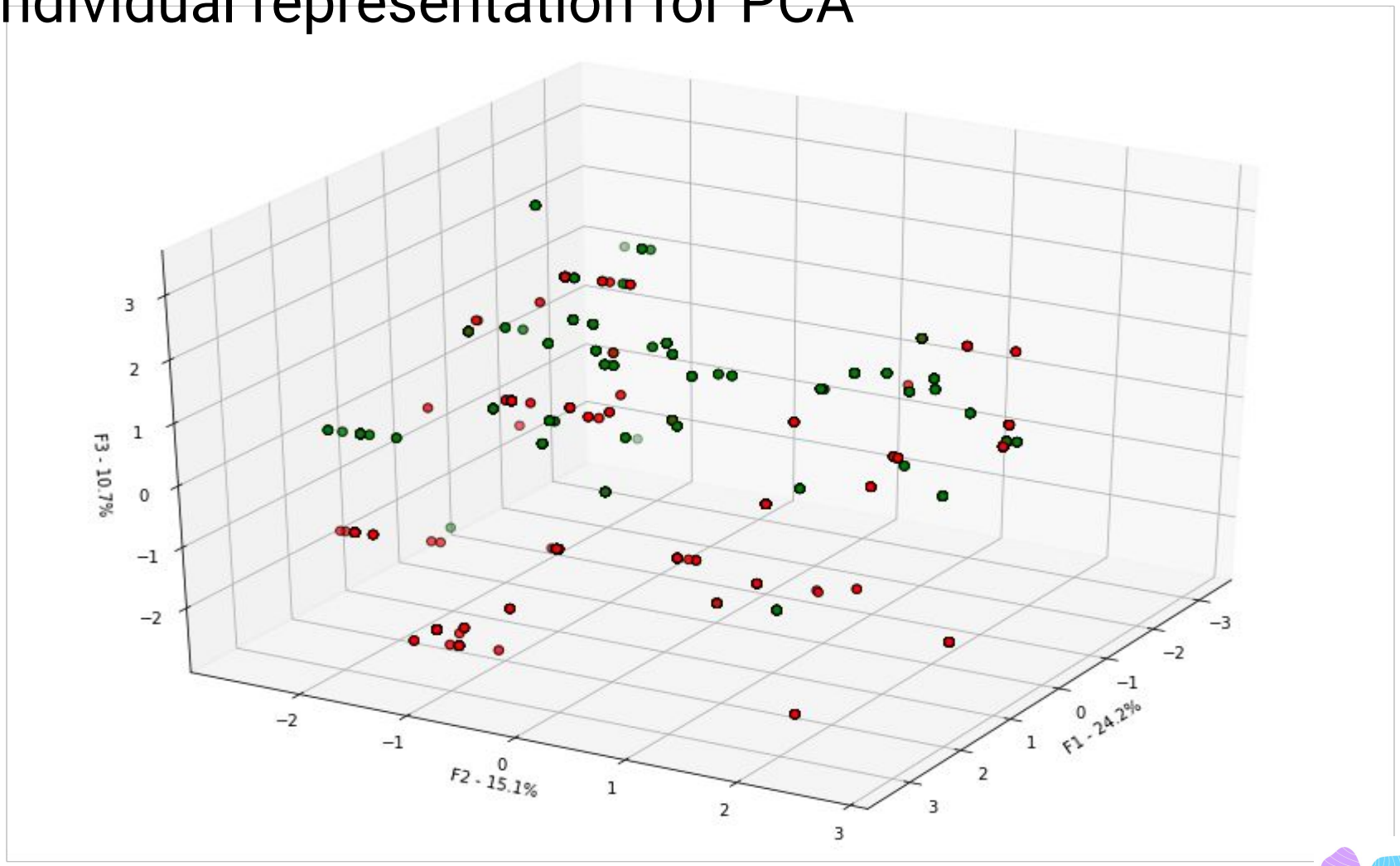
representation learning

Individual representation for PCA



LEARNING PLANET INSTITUTE

Individual representation for PCA

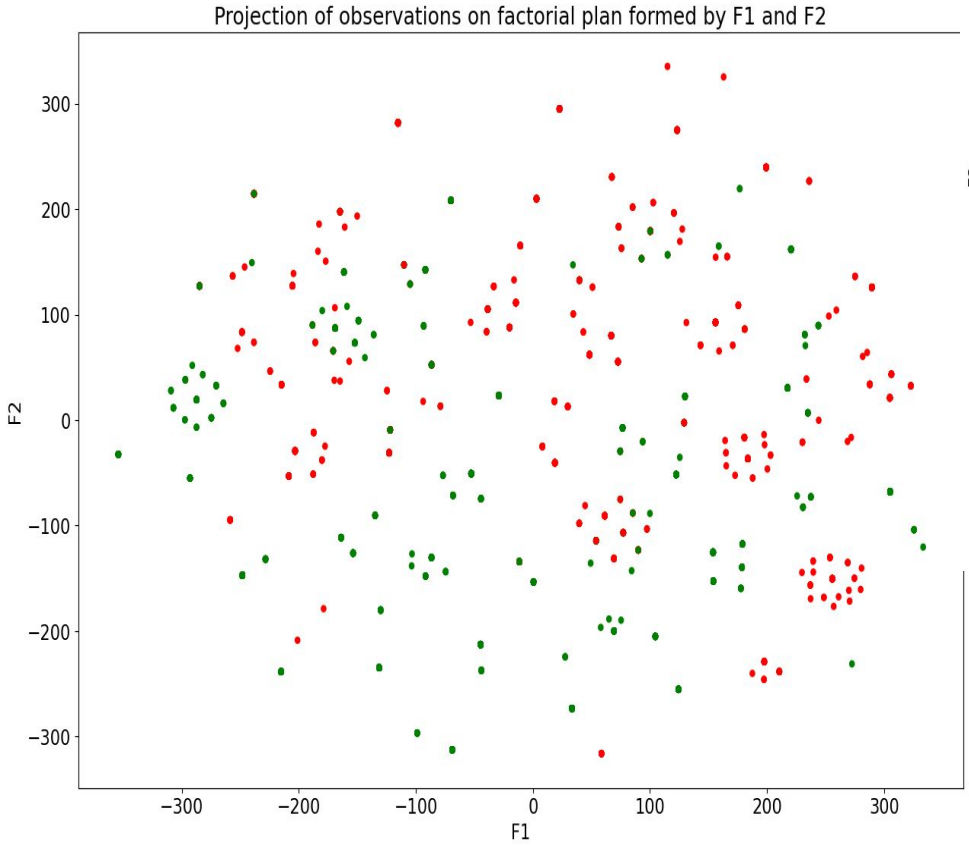


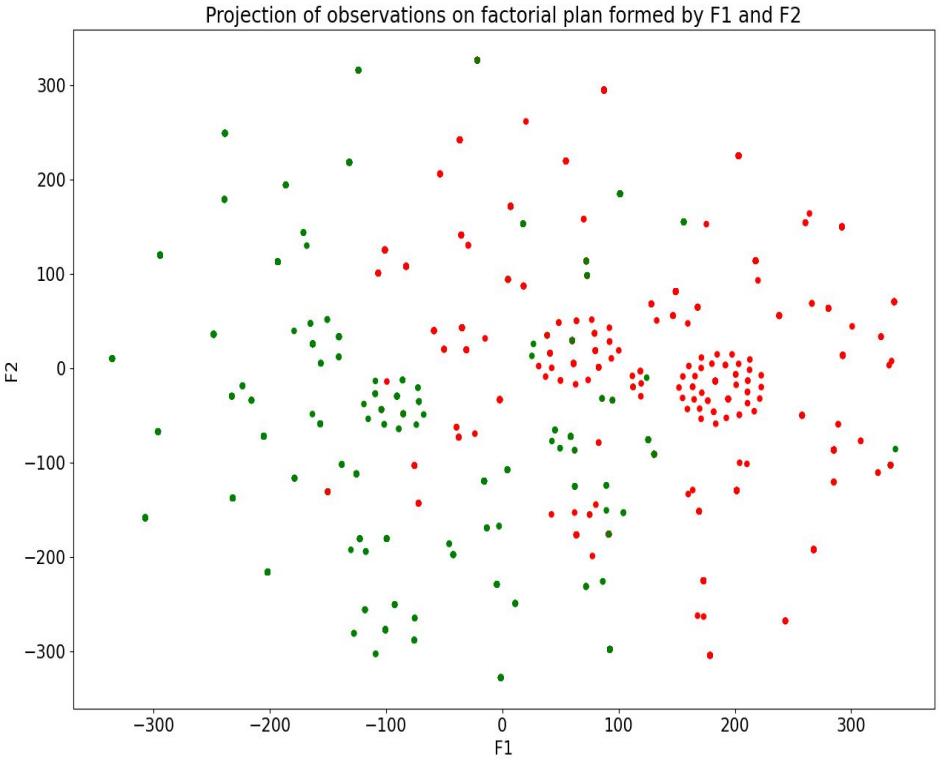
LEARNING PLANET INSTITUTE

representation learning

Individual representation for TSNE

Not standardized data

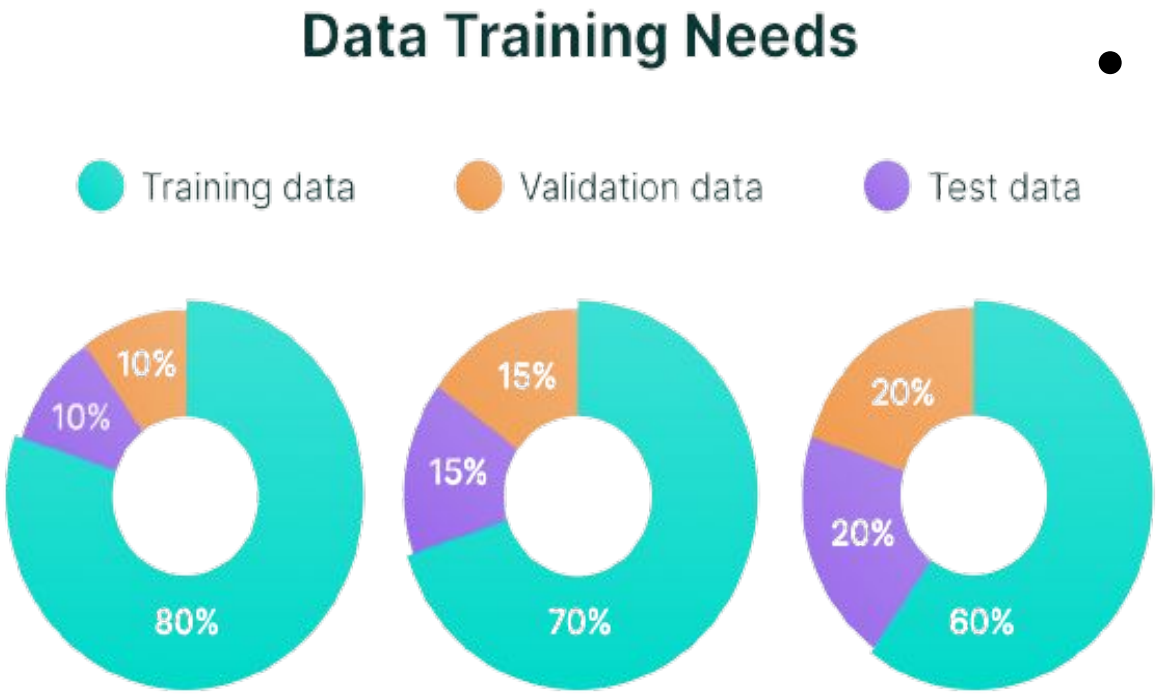




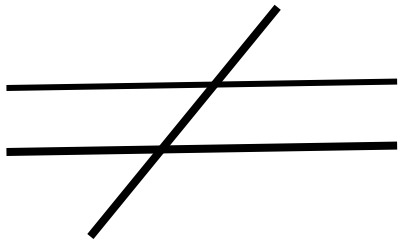
Standardized data

All features are summarized in two components

Data splitting process



- DL methods train and valid set are used to handle overfit or underfitting problem and we have fitting score and valid score for each epoch and test set is used for final evaluation

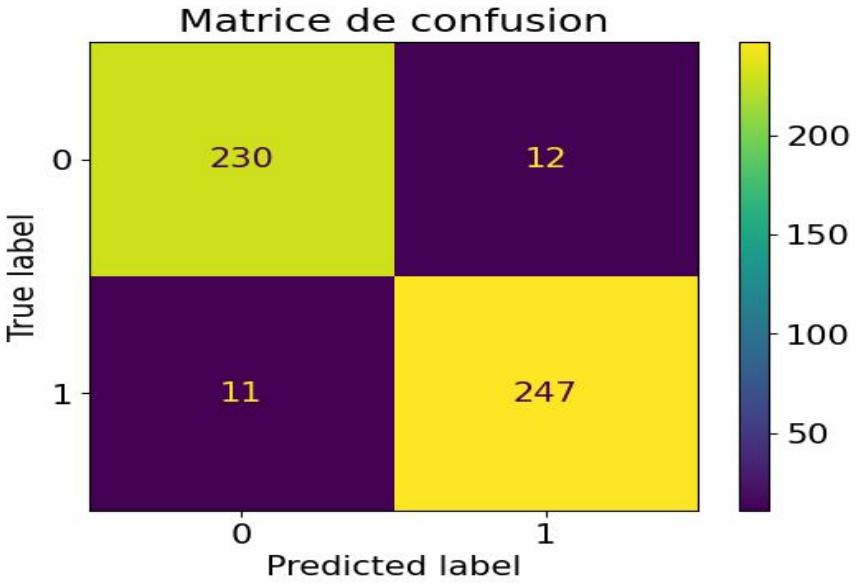


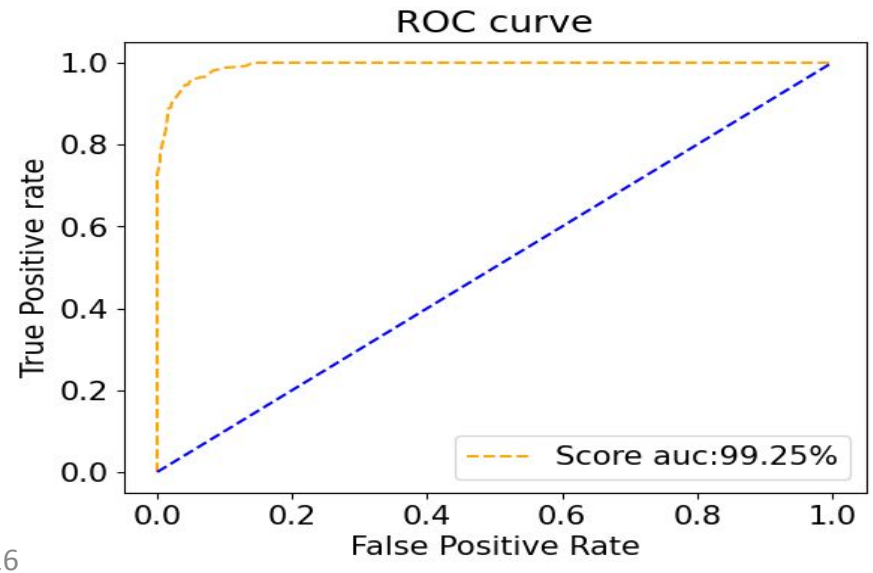
- Classic ML methods
  - Splitting depends the optimization strategy
    - Iterative and scratch optimisation need a validation set
    - GridSearch or another blackbox don't need validation set
      - Built-in cross validation is adopted to tune hyperparameters

LEARNING PLANET INSTITUTE

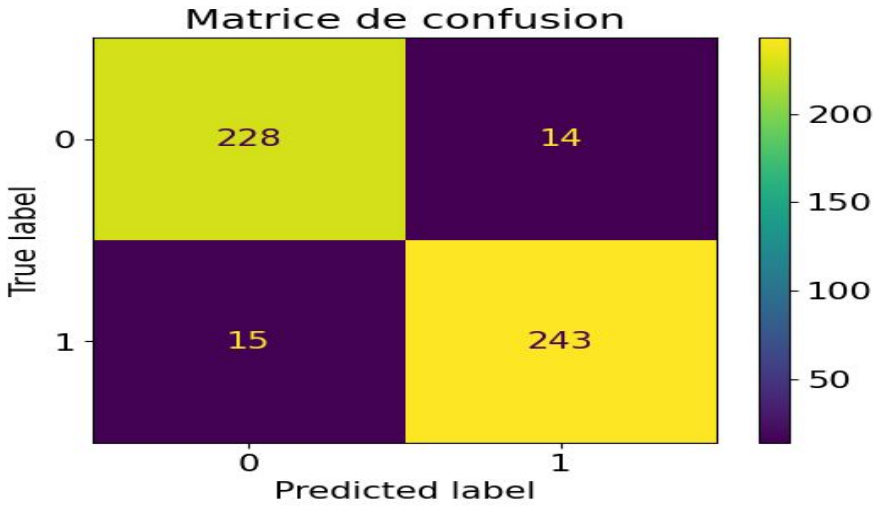
Random Forest

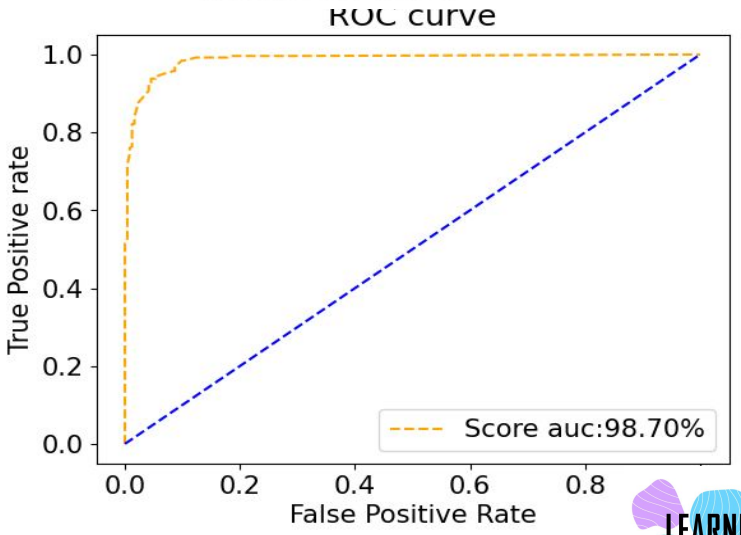
16 originals features:





Two TSNE components

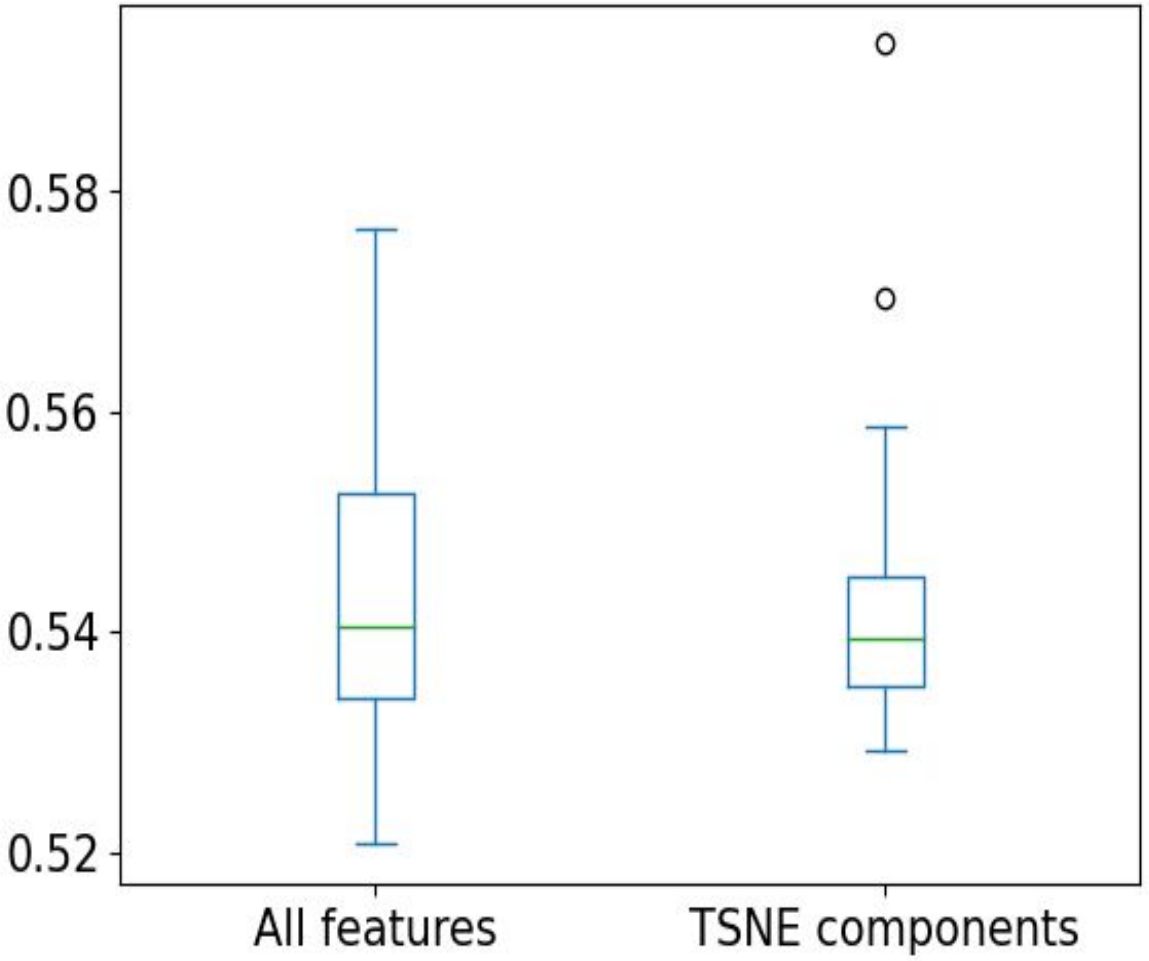




Time of fitting

In average fitting time randomforest take the same time to fit. That's due to built-in feature selection on each node the estimator

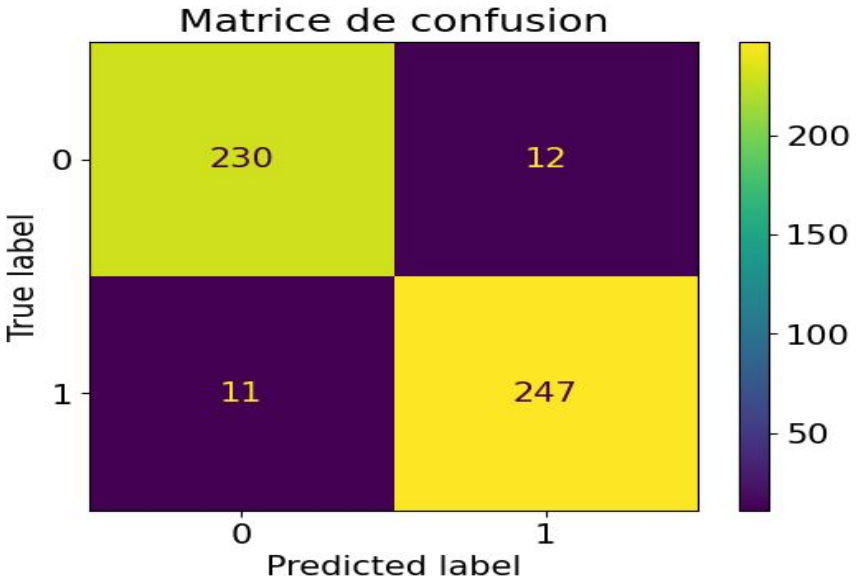
81 times of processing

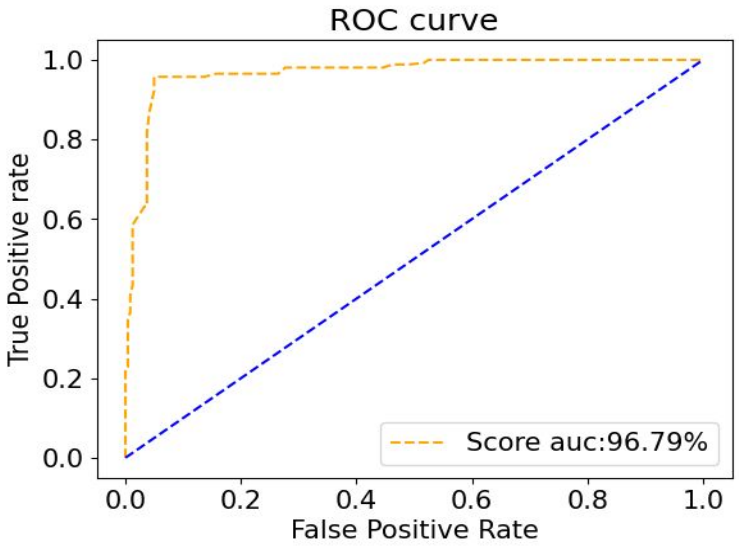


LEARNING PLANET INSTITUTE

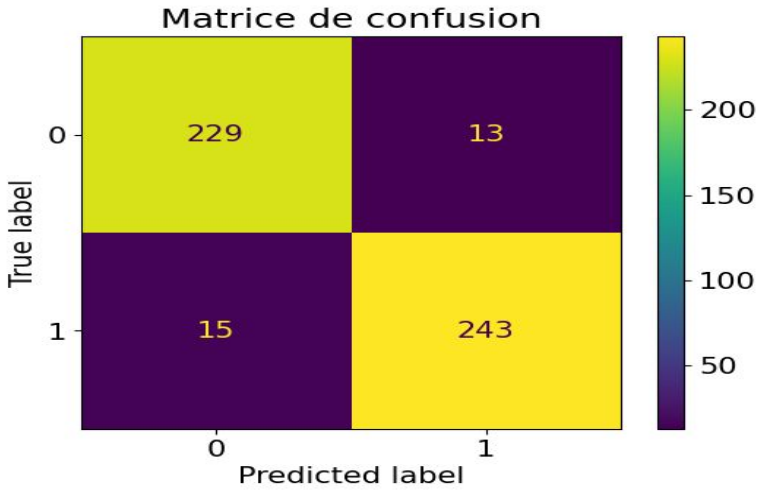
SVM

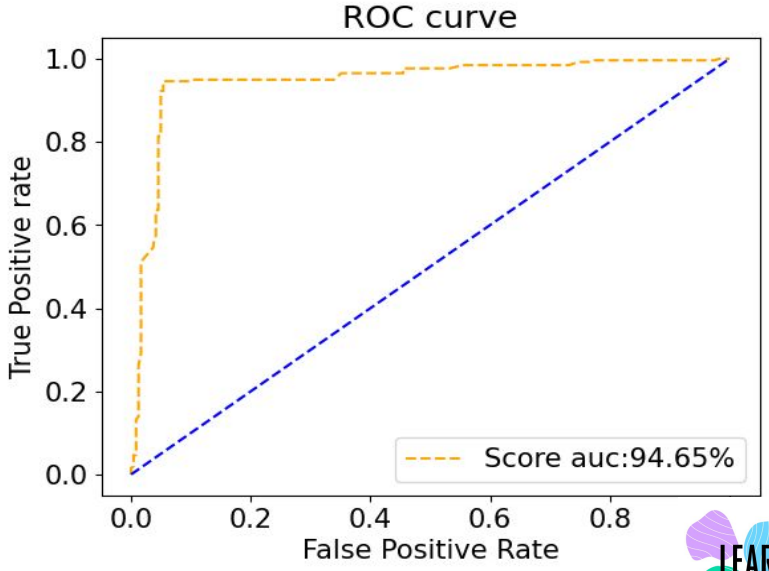
16 originals features:





Two TSNE components

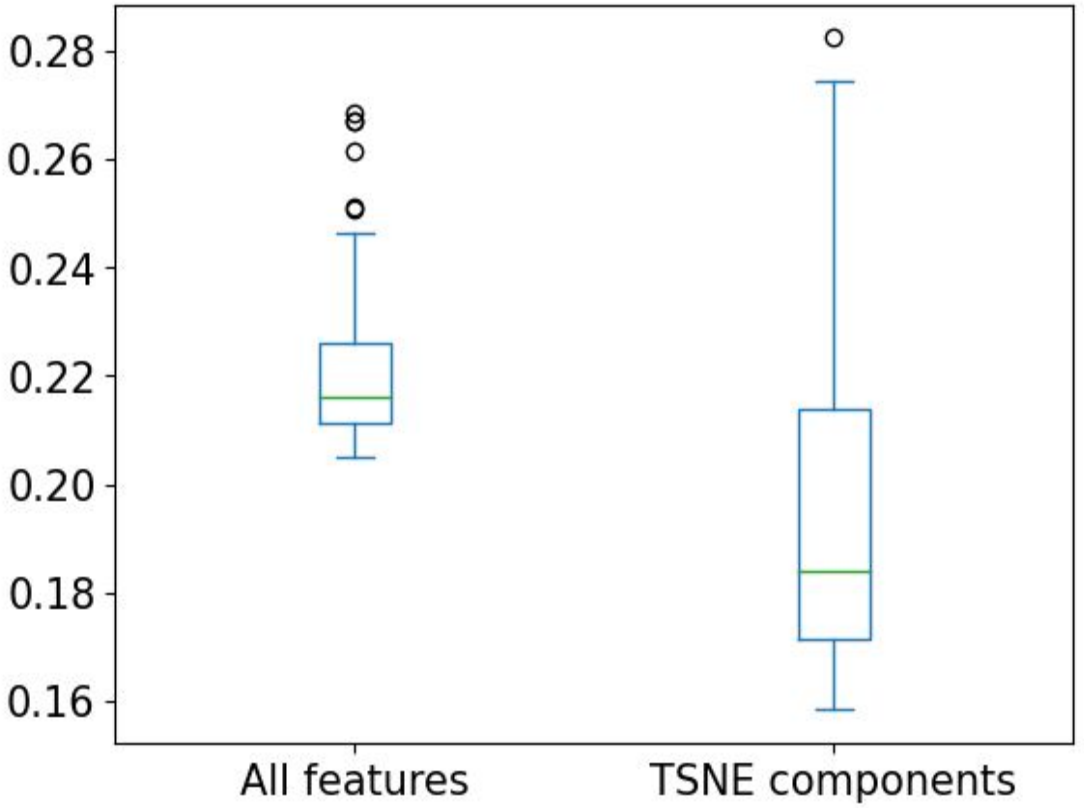




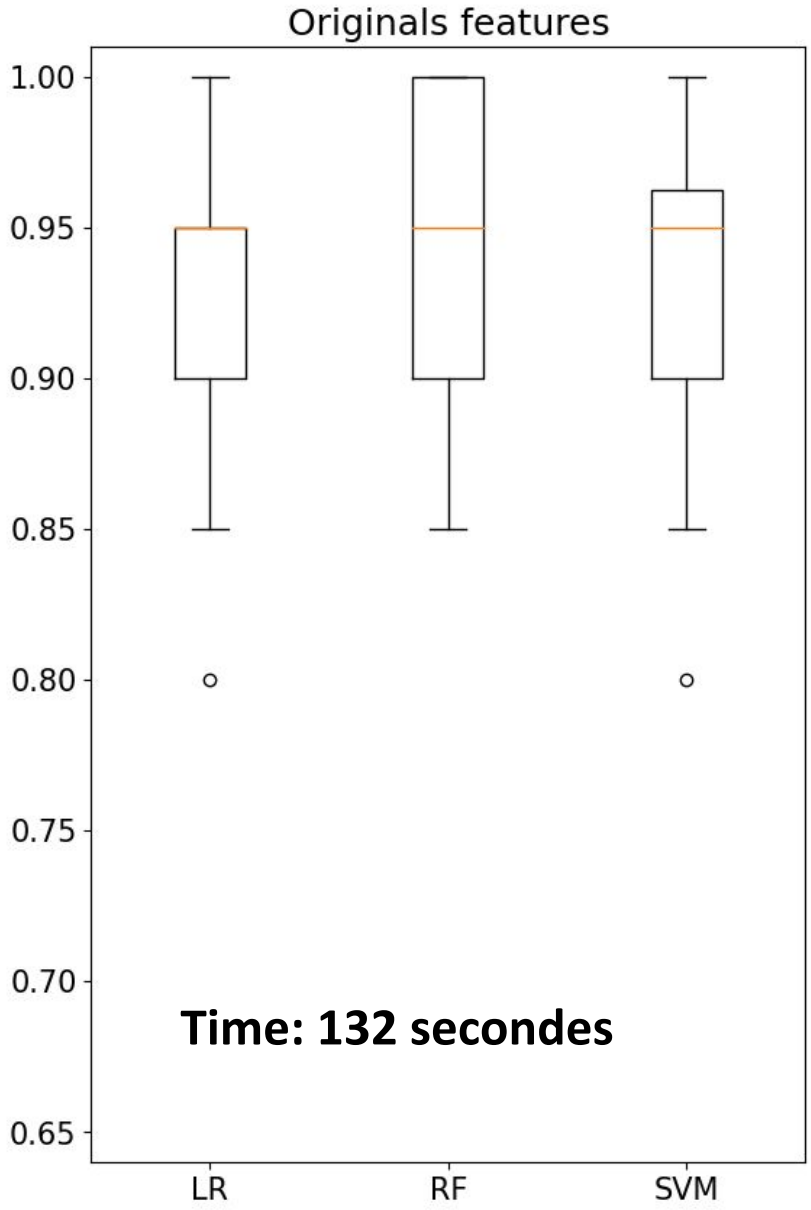
Time of fitting

In average SVM take a lot of time to fit with all features than on principal components

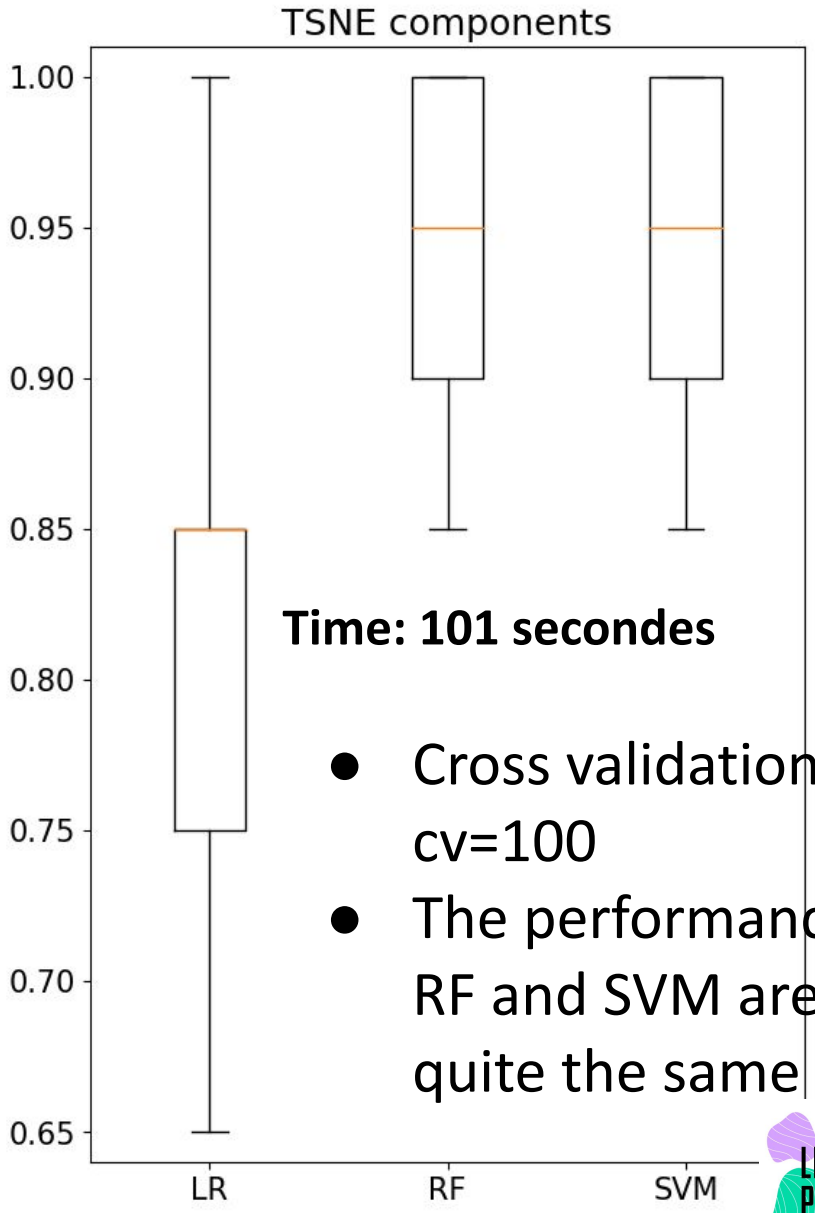
89 times processing



Conclusion



Time: 132 secondes



Time: 101 secondes

- Cross validation cv=100
- The performance of RF and SVM are quite the same

Online Payment Fraud Detection

LEARNING PLANET INSTITUTE

AGENDA

1 Dataset EDA

2 Correlation analysis

3 Compare models

4 PCA

5 t-SNE

EDA - Online Payment Fraud Detection

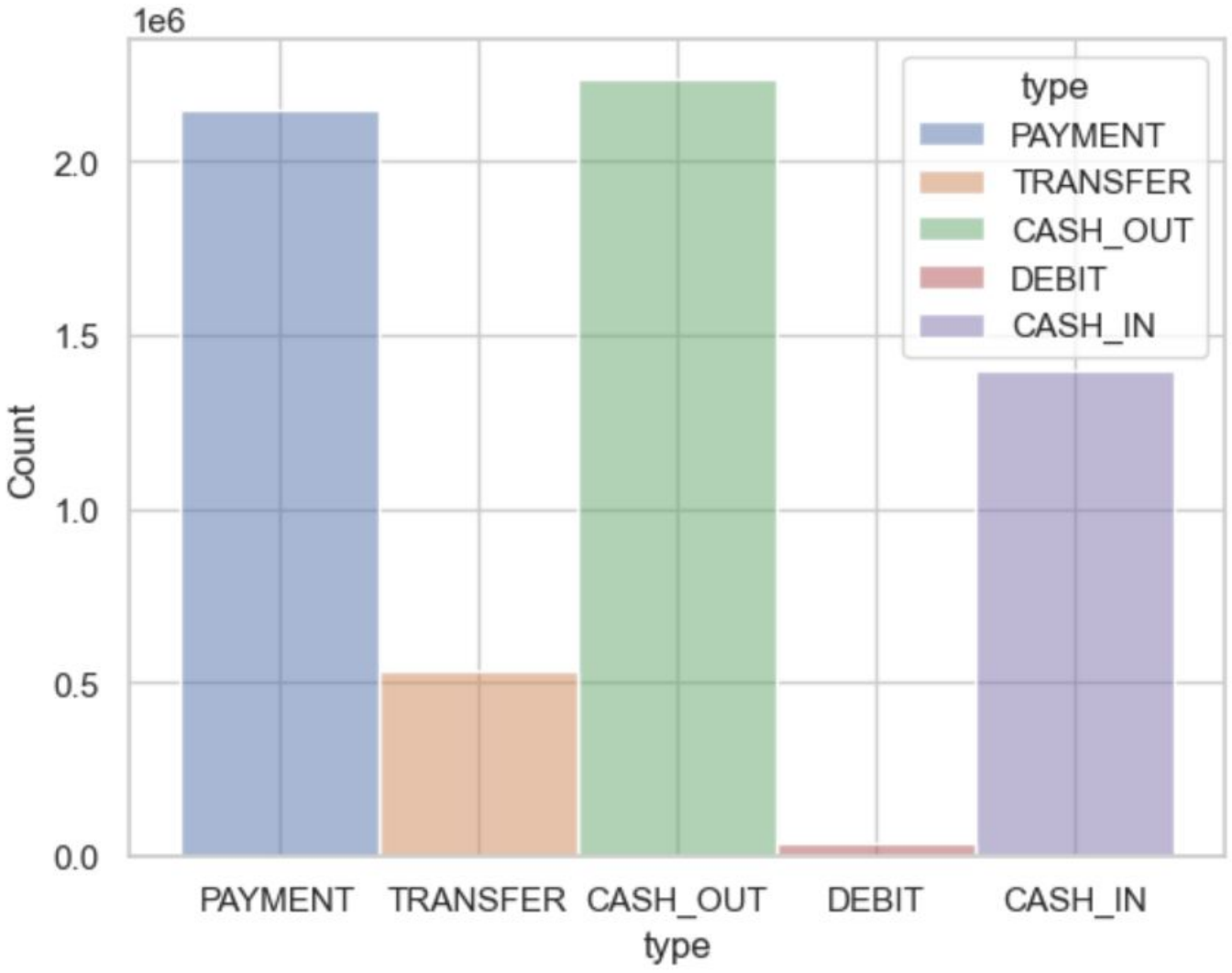
	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6362620 entries, 0 to 6362619
Data columns (total 10 columns):
 #   Column          Dtype
---  ------          -----
 0   step            int64
 1   type            object
 2   amount          float64
 3   nameOrig        object
 4   oldbalanceOrg   float64
 5   newbalanceOrig  float64
 6   nameDest        object
 7   oldbalanceDest  float64
 8   newbalanceDest  float64
 9   isFraud         int64
dtypes: float64(5), int64(2),
memory usage: 485.4+ MB
```

* the **amount** of the transactions

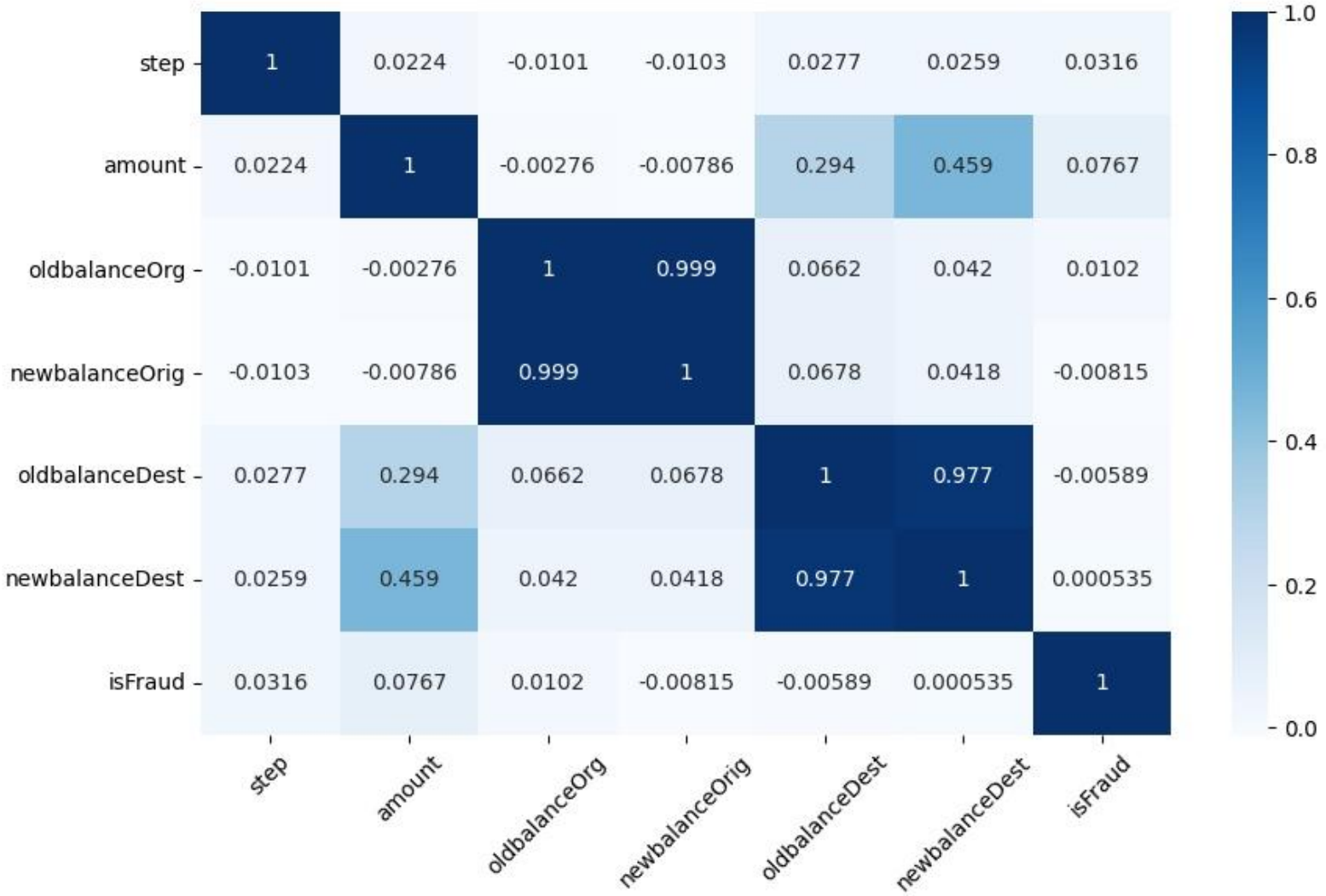
	step	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud
count	6362620.00	6362620.00	6362620.00	6362620.00	6.362620e+06	6.362620e+06	6362620.00
mean	243.40	179861.90	833883.10	855113.67	1.100702e+06	1.224996e+06	0.00
std	142.33	603858.23	2888242.67	2924048.50	3.399180e+06	3.674129e+06	0.04
min	1.00	0.00	0.00	0.00	0.000000e+00	0.000000e+00	0.00
25%	156.00	13389.57	0.00	0.00	0.000000e+00	0.000000e+00	0.00
50%	239.00	74871.94	14208.00	0.00	1.327057e+05	2.146614e+05	0.00
75%	335.00	208721.48	107315.18	144258.41	9.430367e+05	1.111909e+06	0.00
max	743.00	92445516.64	59585040.37	49585040.37	3.560159e+08	3.561793e+08	1.00

Type of payment



2237500 CASH_OUT
2151495 PAYMENT
1399284 CASH_IN
532909 TRANSFER
41432 DEBIT

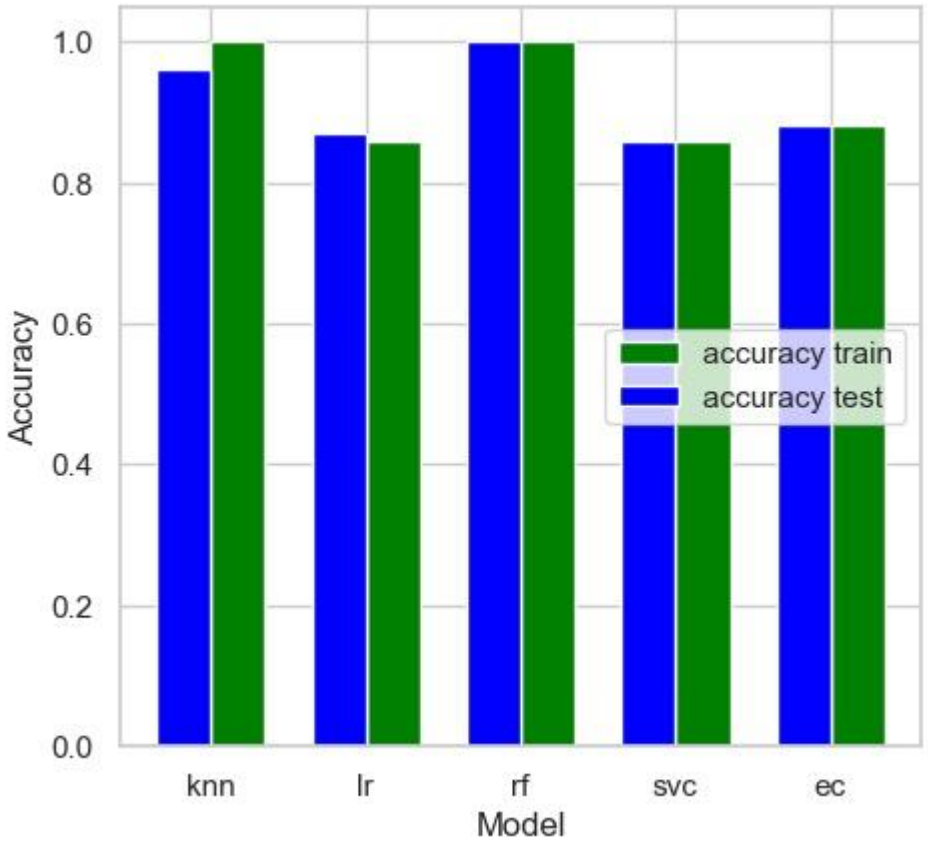
Correlation analysis



Compare models

```
0    8213
1    8213
```

Balancing data




```
acc_train_knn =  1.0
acc_test_knn =  0.96
[[2310  128]
 [  75 2415]]
              precision    recall  f1-score   support

           0       0.97      0.95      0.96      2438
           1       0.95      0.97      0.96      2490

    accuracy                           0.96      4928
   macro avg       0.96      0.96      0.96      4928
weighted avg       0.96      0.96      0.96      4928
```

```
acc_train_lr =  0.86
acc_test_lr =  0.87
[[2394   44]
 [ 586 1904]]
              precision    recall  f1-score   support

           0       0.80      0.98      0.88      2438
           1       0.98      0.76      0.86      2490

    accuracy                           0.87      4928
   macro avg       0.89      0.87      0.87      4928
weighted avg       0.89      0.87      0.87      4928
```

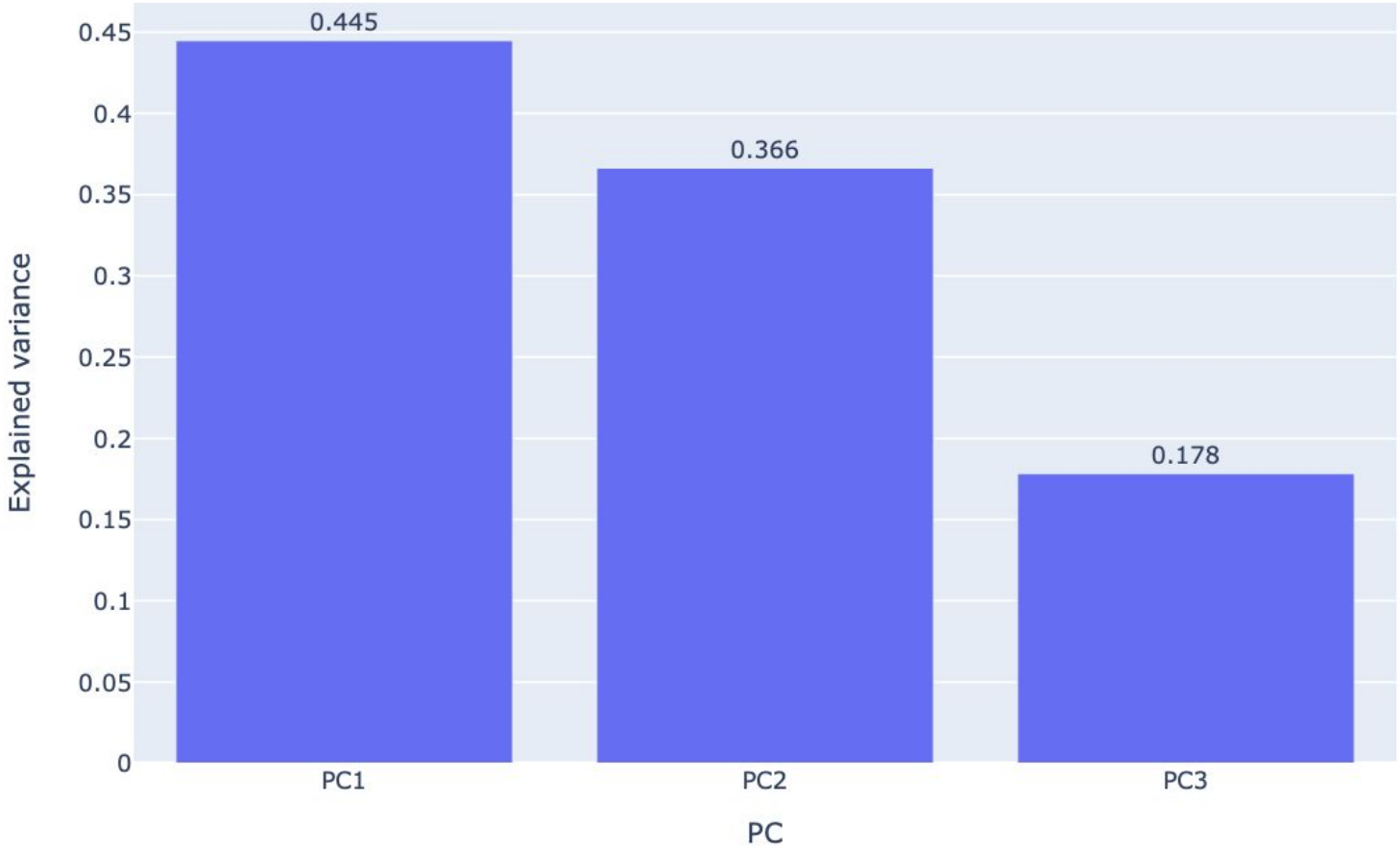
```
acc_train_rf =  1.0
acc_test_rf =  0.99
[[2412   26]
 [  12 2478]]
              precision    recall  f1-score   support

           0       1.00      0.99      0.99      2438
           1       0.99      1.00      0.99      2490

    accuracy                           0.99      4928
   macro avg       0.99      0.99      0.99      4928
weighted avg       0.99      0.99      0.99      4928
```

```
acc_train_svc =  0.86
acc_test_svc =  0.87
[[2389   49]
 [ 579 1911]]
              precision    recall  f1-score   support

           0       0.80      0.98      0.88      2438
           1       0.97      0.77      0.86      2490

    accuracy                           0.87      4928
   macro avg       0.89      0.87      0.87      4928
weighted avg       0.89      0.87      0.87      4928
```

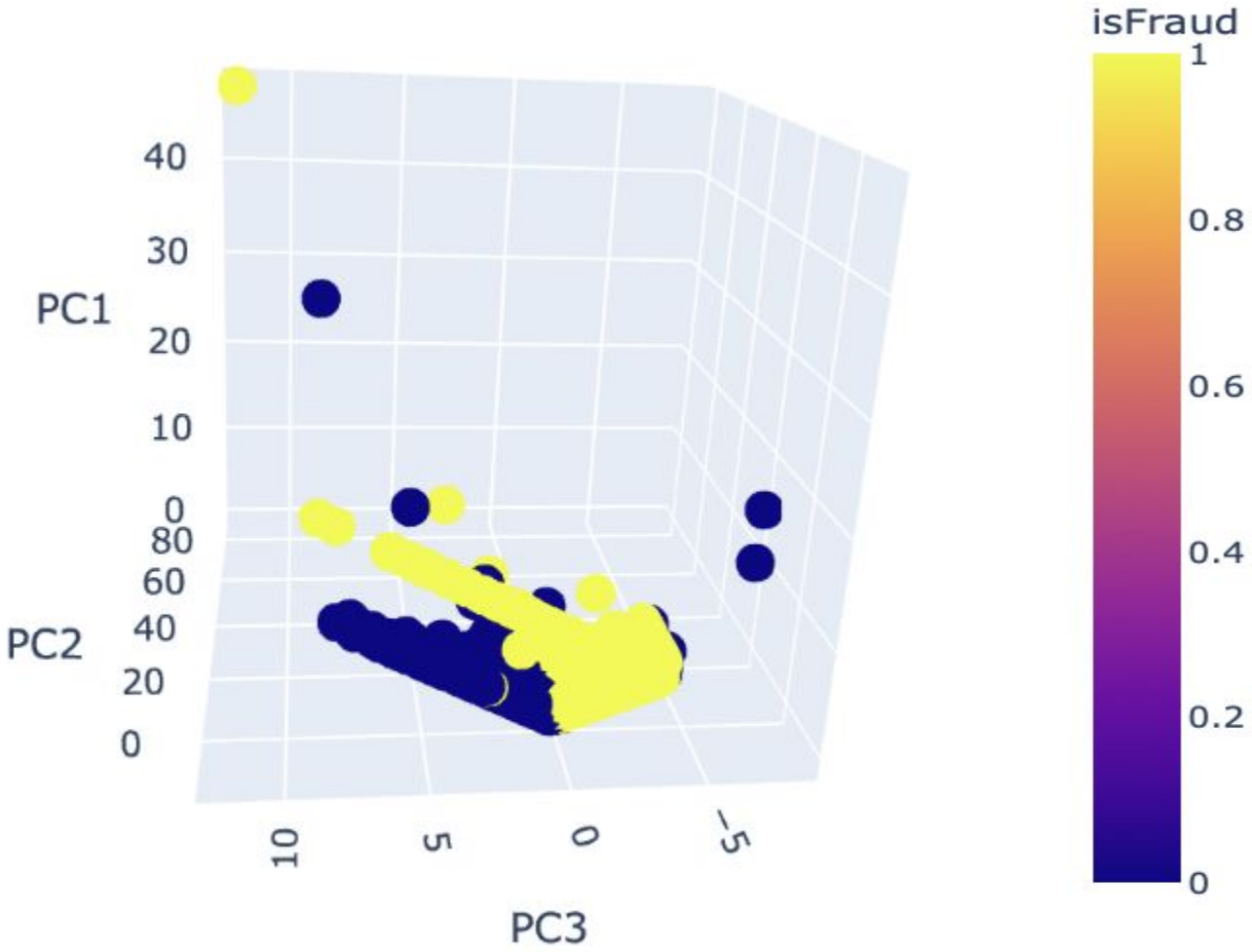
```
acc_train_ec =  0.88
acc_test_ec =  0.88
[[2415   23]
 [ 574 1916]]
              precision    recall  f1-score   support

           0       0.81      0.99      0.89      2438
           1       0.99      0.77      0.87      2490

    accuracy                           0.88      4928
   macro avg       0.90      0.88      0.88      4928
weighted avg       0.90      0.88      0.88      4928
```


LEARNING PLANET INSTITUTE

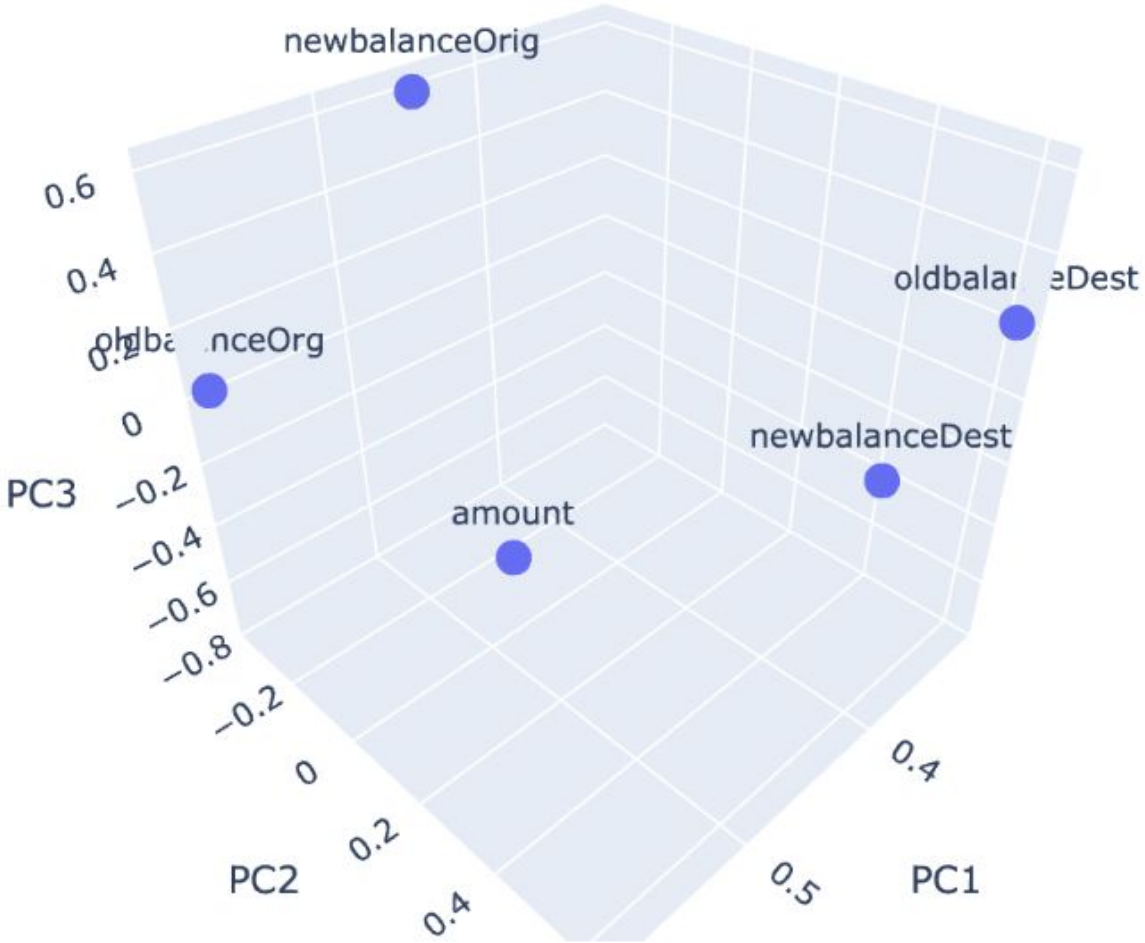
Principal component analysis - impact



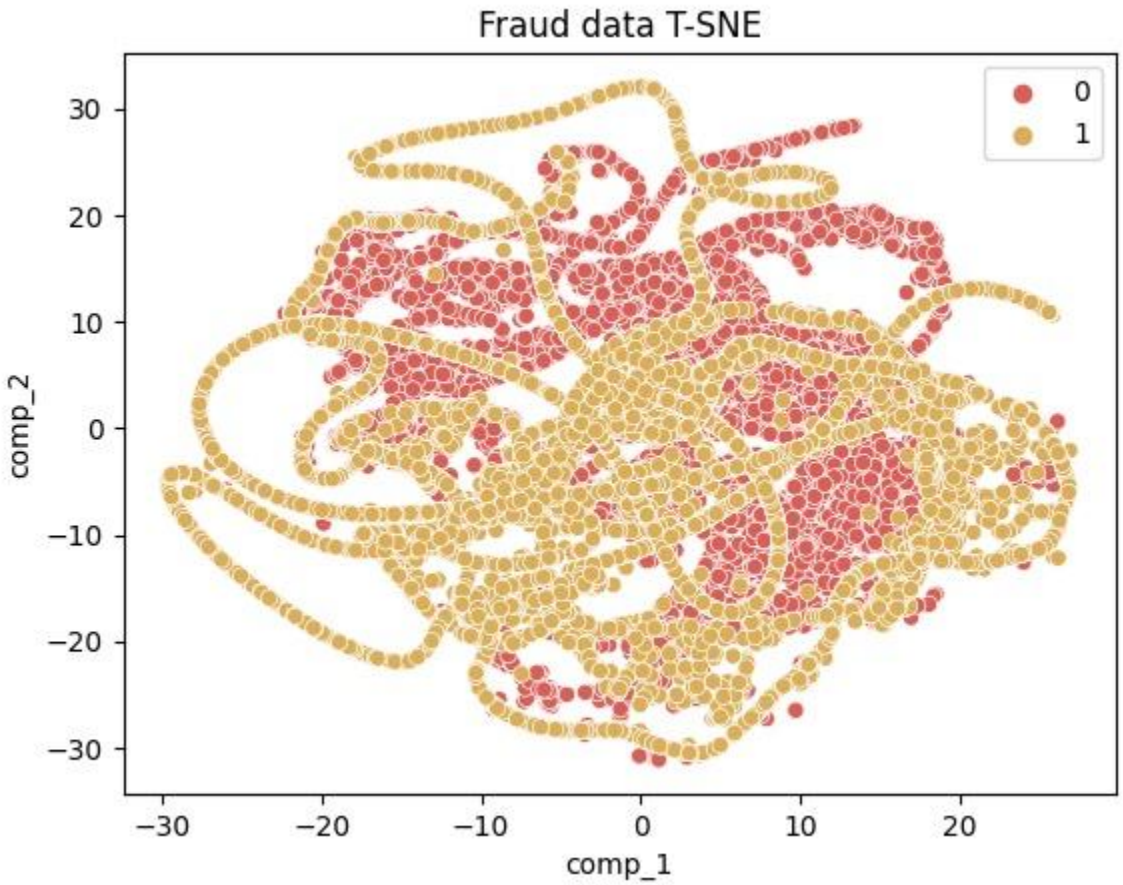
Principal component analysis



Principal component analysis



T-distributed Stochastic Neighbor Embedding



Thank you for your attention