

# MASTER EN BUSINESS ANALYTICS

## MODULO III – FUNDAMENTOS TECNOLÓGICOS

### Práctica ETL + EDA + Modeling



**RAPIDMINER**

An  **ALTAIR Company**



Estudiante: Agustin Cervantes

Profesor: Christian Sucuzhanay

INTRODUCCIÓN A LA ACTIVIDAD .....	3
TAREAS A REALIZAR.....	3
TAREA 1 – IMPORTACIÓN DE DATOS .....	3
TAREA 2 – BALANCEO DE DATOS .....	5
TAREA 3 – RESULTADO DEL BALANCEO .....	6
TAREA 4 – PRUEBA DE MODELOS.....	7
TAREA 5 - PASO A PASO EN PYTHON .....	11
CONCLUSION.....	12

## INTRODUCCIÓN A LA ACTIVIDAD

La actividad tiene como finalidad, la preparación y análisis de un conjunto de datos de transacciones de tarjetas de crédito, a efectos de generar un modelo de machine learning capaz de interpretar casuísticas de fraude, que permitan establecer un modelo con una efectividad suficiente para predecirlo.

El dataset sobre el cual se trabajará contiene más de 6 millones de observaciones, por lo cual el desafío tecnológico se centra en la capacidad de analizar esa cantidad de registros de forma eficiente y efectiva. Por otra parte, el dataset presenta entre sus datos, una columna “IsFraud”, que toma valores 1 en caso de que la operación sea fraude y 0 en caso de que no lo sea.

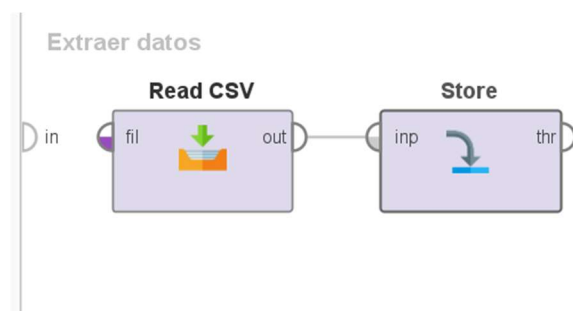
Será la mencionada columna, la que nos permitirá entrenar los modelos de Machine Learning que consideremos, en busca de alcanzar un accuracy mayor al 95% (objetivo de la presente actividad).

## TAREAS A REALIZAR

Se estarán realizando las tareas mencionadas por el profesor en la actividad señalada, en forma ordenada en esta sección.

### TAREA 1 – IMPORTACIÓN DE DATOS

Para la importación de los datos, se utilizarán los operadores Read CSV y Store:



En la importación y para simplificar el posterior procesamiento, se definió como “Label” y “binomial”, la columna “IsFraud”:

Import Data - Format your columns.

**Format your columns.**

Date format:  ☐ Replace errors with missing values ⓘ

	Orig nominal	oldbalance... real	newbalance... real	nameDest polynomial	oldbalance... real	newbalance... real	IsFraud binominal label	IsFlaggedFr... integer
1	31006815	170136.000	160296.360	M1979787155	0.000	0.000	0	0
2	36544295	21249.000	19384.720	M2044282225	0.000	0.000	0	0
3	05486145	181.000	0.000	C553264065	0.000	0.000	1	0
4	0083671	181.000	0.000	C38997010	21182.000	0.000	1	0
5	48537720	41554.000	29885.860	M1230701703	0.000	0.000	0	0
6	045638	53860.000	46042.290	M573487274	0.000	0.000	0	0
7	4988899	183195.000	176087.230	M408069119	0.000	0.000	0	0
8	12850431	176087.230	168225.590	M633326333	0.000	0.000	0	0
9	35012928	2671.000	0.000	M1176932104	0.000	0.000	0	0
10	2410124	41720.000	36382.230	C195600860	41898.000	40348.790	0	0
11	00366749	4465.000	0.000	C997608398	10845.000	157982.120	0	0
12	3177573	20771.000	17671.030	M2096539129	0.000	0.000	0	0
13	48232591	5070.000	2509.260	M972865270	0.000	0.000	0	0
14	16932897	10127.000	0.000	M801569151	0.000	0.000	0	0
15	26483832	503264.000	499165.220	M1635378213	0.000	0.000	0	0
16	3080434	15325.000	0.000	C476402209	5083.000	51513.440	0	0
17	1750706	450.000	0.000	M1731217984	0.000	0.000	0	0
18	37763630	24156.000	10008.140	M1877062007	0.000	0.000	0	0

no problems.

Previous Finish Cancel

El dataset con el total de datos tiene la siguiente información:

ExampleSet (/Local Repository/ue22303845/Data/whole\_dataset)

Result History

ExampleSet (/Local Repository/ue22303845/Data/whole\_dataset)

Filter (11 / 11 attributes): Search for Attributes

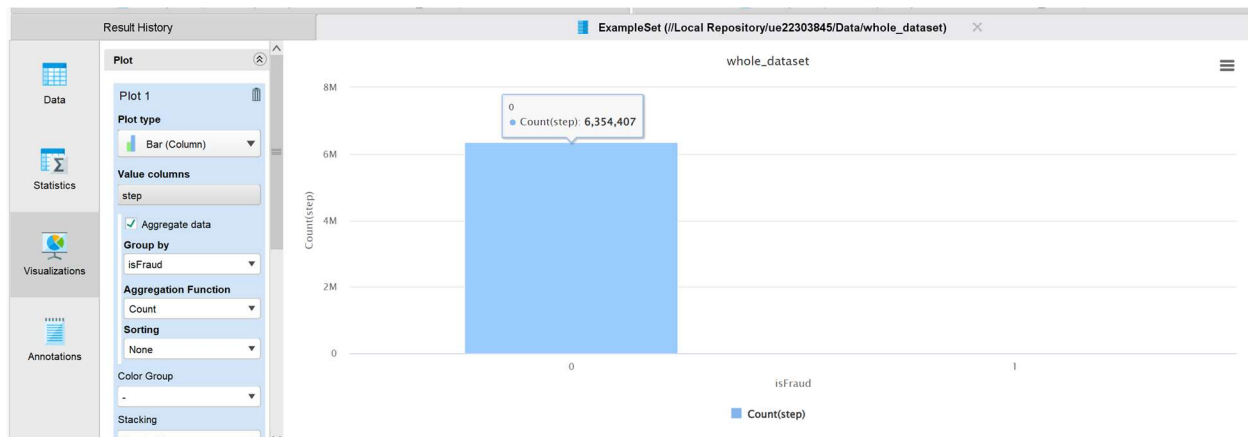
Name	Type	Missing	Statistics
Label IsFraud	Binominal	0	<p>Values: 0 (6354407), 1 (8213)</p>
step	Integer	0	<p>Min: 1, Max: 743, Average: 243.397</p>
type	Nominal	0	<p>Least: DEBIT (41432), Most: CASH_OUT (2237500)</p> <p>Values: CASH_OUT (2237500), PAYMENT (2151495), ...[3 more]</p>
amount	Real	0	<p>Min: 0, Max: 92445516.640, Average: 179861.904</p>
nameOrig	Nominal	0	<p>Least: C999999784 (1), Most: C1065307291 (3)</p> <p>Values: C1065307291 (3), C1462946854 (3), ...[6353305 more]</p>

Showing attributes 1 - 11

Examples: 6,362,620 Special Attributes: 1 Regular Attributes: 10

## TAREA 2 – BALANCEO DE DATOS

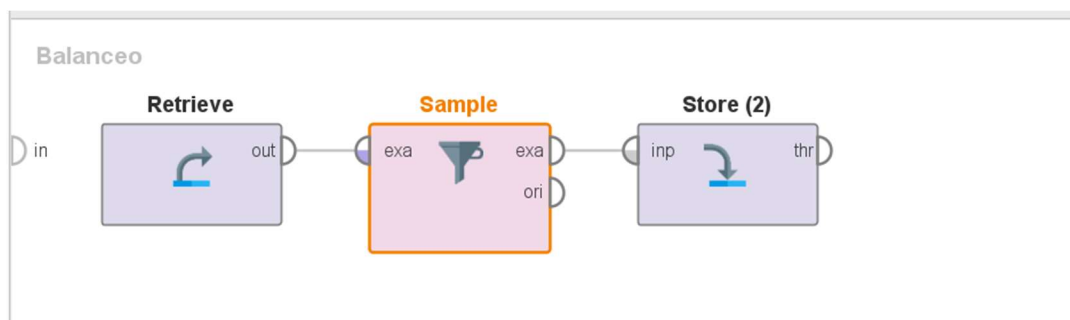
Como es posible apreciar en la figura que representa el valor “IsFraud”, éste se encuentra totalmente desbalanceado, habiendo 6.354.407 casos en los cuales no es fraude y 8.213 en los que si lo es:



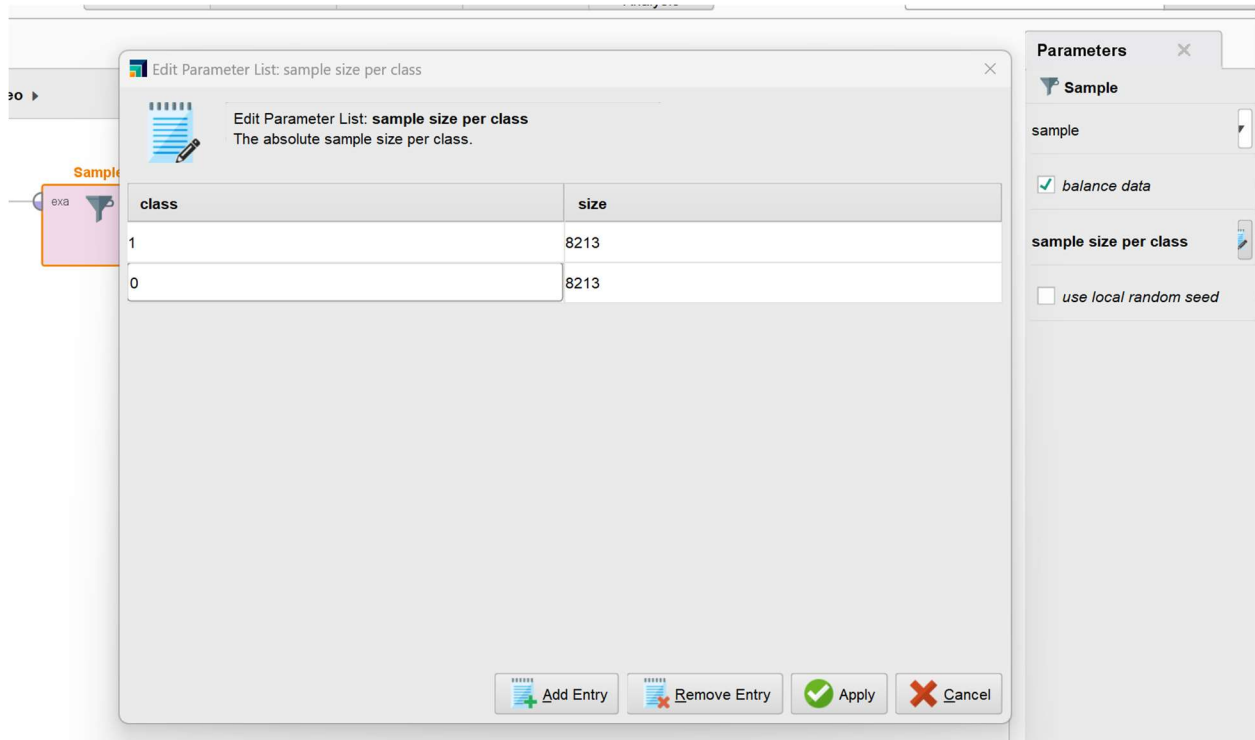
Entrenar nuestros modelos de Machine Learning de esta forma, generaría muy poca eficacia, ya que consideraría que casi el 100% de los casos no son fraude.

Para evitar este punto, se realiza un balanceo del dataset para entrenar nuestro modelo. Para esto, se genera una muestra sesgada del mismo, seleccionando el total de los casos de fraude y una cantidad idéntica de casos los “No fraude”. Las posibilidades técnicas limitan esta posibilidad en mi caso, por lo cual solo fue posible seleccionar casos “No fraude” de forma aleatoria, lo cual puede impactar directamente en el accuracy del modelo.

De todas formas, se presenta la forma en el cual se realizó:



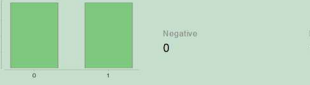
Tomando el dataset inicial, se realizó un muestreo de la siguiente manera:



Y se guardó el archivo como “balanced\_dataset”.

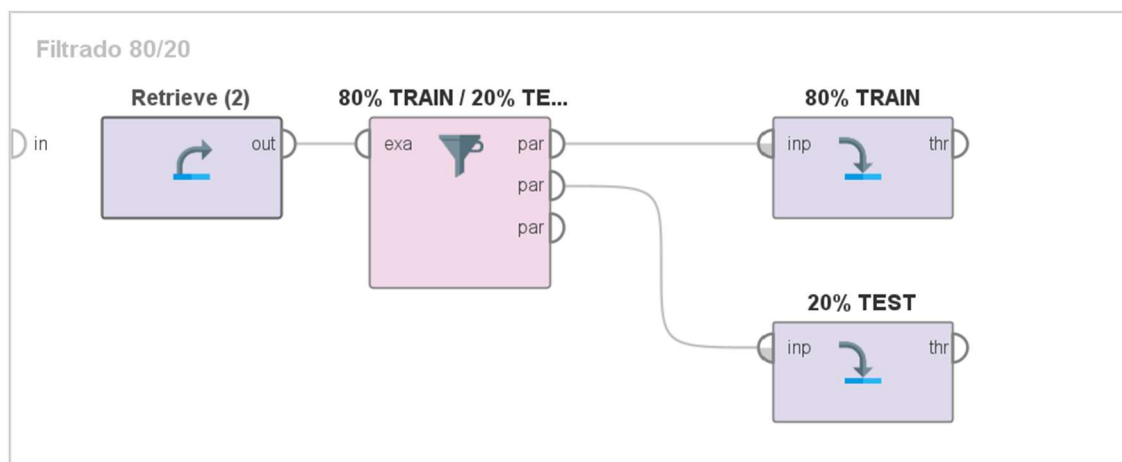
### TAREA 3 – RESULTADO DEL BALANCEO

El dataset “balanced\_dataset” se conforma con 16.426 casos y los datos de la columna “IsFraud” distribuidos de la siguiente manera:

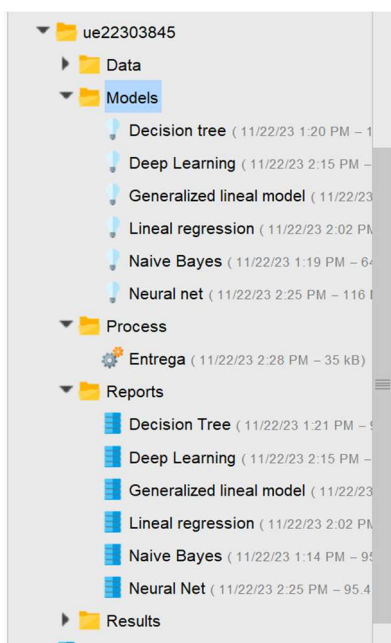
Name	Type	Missing	Statistics			Filter (11 / 11 attributes): <input type="text" value="Search for Attributes"/>
Label <b>IsFraud</b>	Binominal	0				Values 0 (8213), 1 (8213) <a href="#">Details...</a>
step	Integer	0	Min 1	Max 743	Average 307.752	
type	Nominal	0	Least DEBIT (52)	Most CASH_OUT (7052)	Values CASH_OUT (7052), TRANSFER (4775), ...[3 more]	
amount	Real	0	Min 0	Max 20187682.270	Average 820421.189	
nameOrig	Nominal	0	Least C999999784 (0)	Most C1000018372 (1)	Values C1000018372 (1), C1000036340 (1), ...[6353305 more]	
			Min	Max	Average	

## TAREA 4 – PRUEBA DE MODELOS

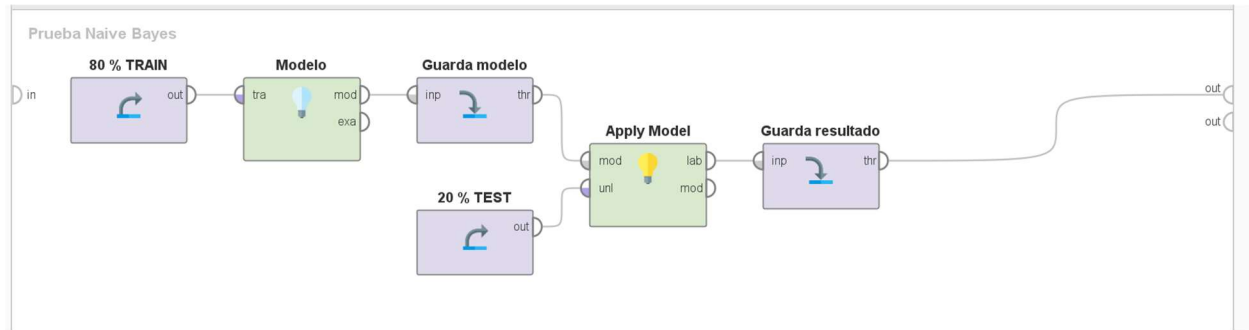
Para efectuar la prueba de modelos, inicialmente se realizó una subdivisión del muestreo, de modo que se separó un 80% para entrenar y 20% para testear el modelo:



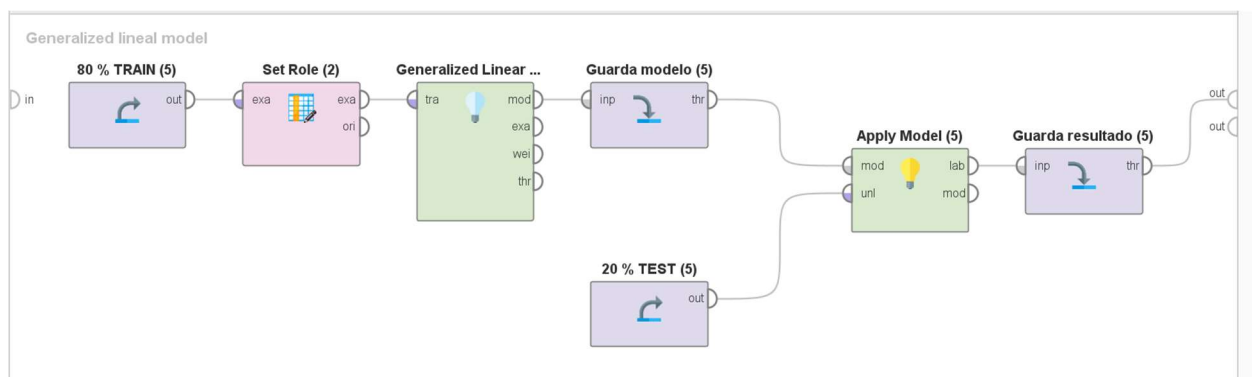
Con esos dos dataset, se fueron generando pruebas y testeos a 6 modelos diferentes, guardando los modelos en la carpeta “Modelos” y los resultados en “Reports”:



Todos los modelos se entrenaron con un proceso similar:



Salvo aquellos que no podían soportar polinomios, para los cuales hubo que hacer conversiones previas en los datos utilizando el operador “Set role”:



Los resultados resumidos de los diferentes modelos fueron:

### Decision Tree

Result History		ExampleSet (/Local Repository/ue22303845/Reports/Decision Tree)					Filter (14 / 14 attributes): Search for Attributes
	Name	Type	Missing	Statistics			
Data	Label IsFraud	Binominal	0	Negative 0	Positive 1	Values 0 (1643), 1 (1643)	
Statistics	Prediction prediction(IsFraud)	Binominal	0	Negative 0	Positive 1	Values 1 (1671), 0 (1615)	
Visualizations	Score confidence(0)	Real	0	Min 0	Max 1	Average 0.498	
	Score confidence(1)	Real	0	Min 0	Max 1	Average 0.502	
Annotations	step	Integer	0	Min 1	Max 743	Average 313.150	
	type	Nominal	0	Least DEBIT (15)	Most CASH_OUT (1410)	Values CASH_OUT (1410), TRANSFER (931), ...[3 more]	

Showing attributes 1 - 14

Examples: 3,286 Special Attributes: 4 Regular Attributes: 10



Result History

Data

Statistics

Visualizations

Annotations

ExampleSet (/Local Repository/ue22303845/Reports/Deep Learning)

ExampleSet (/Local Repository/ue22303845/Reports/Decision Tree)

Name	Type	Missing	Statistics			Filter (14 / 14 attributes): <input type="text" value="Search for Attributes"/>
Label <b>isFraud</b>	Binominal	0	Negative 0	Positive 1	Values 0 (1643), 1 (1643)	
Prediction <b>prediction(isFraud)</b>	Binominal	0	Negative 0	Positive 1	Values 1 (1676), 0 (1610)	
Score <b>confidence(0)</b>	Real	0	Min 0	Max 1	Average 0.465	
Score <b>confidence(1)</b>	Real	0	Min 0	Max 1	Average 0.535	
<b>step</b>	Integer	0	Min 1	Max 743	Average 313.150	
<b>type</b>	Nominal	0	Least DEBIT (15)	Most CASH_OUT (1410)	Values CASH_OUT (1410), TRANSFER (931), ... [3 more]	

Showing attributes 1 - 14

Examples: 3,286    Special Attributes: 4    Regular Attributes: 20

Result History		ExampleSet (/Local Repository/ue22303845/Reports/Generalized linear model)				
	Name	Type	Missing	Statistics		
Data	isFraud	Binominal	0	Negative 0	Positive 1	Values 0 (1643), 1 (1643)
Statistics	prediction(isFraud)	Binominal	0	Negative 0	Positive 1	Values 0 (2101), 1 (1185)
Visualizations	score(confidence(0))	Real	0	Min 0	Max 1.000	Average 0.503
	score(confidence(1))	Real	0	Min 0.000	Max 1	Average 0.497
	step	Integer	0	Min 1	Max 743	Average 313.150

ExampleSet (/Local Repository/ue22303845/Reports/Lineal regression)

Filter (14 / 14 attributes):

	Name	Type	Missing	Statistics	
Data	Label				
	isFraud	Binominal	0	Negative 0	Positive 1
Statistics	Prediction				
	prediction(isFraud)	Binominal	0	Negative 0	Positive 1
Visualizations	Score				
	confidence(0)	Real	0	Min 0.256	Max 0.934
Annotations	Score				
	confidence(1)	Real	0	Min 0.066	Max 0.744
	step	Integer	0	Min 1	Max 743

Calculating...

Values  
0 (1643), 1 (1643)

Values  
0 (2441), 1 (845)

Average  
0.500

Average  
0.500

Average  
313.150

Showing attributes 1 - 14

Examples: 3,286 Special Attributes: 4 Regular Attributes: 10

Result History		ExampleSet (/Local Repository/ue22303845/Reports/Naive Bayes)			ExampleSet (/Local Repository/ue22303845/Reports/Linear regression)		
	Name	Type	Missing	Statistics		Filter (14 / 14 attributes):	
Data							
	Label						
	isFraud	Binominal	0	Negative 0	Positive 1	Values 0 (1643), 1 (1643)	
Statistics	Prediction						
	prediction(isFraud)	Binominal	0	Negative 0	Positive 1	Values 0 (2808), 1 (478)	
Visualizations	Score						
	confidence(0)	Real	0	Min 0	Max 1	Average 0.838	
	confidence(1)	Real	0	Min 0	Max 1	Average 0.162	
Annotations							
	step	Integer	0	Min 1	Max 743	Average 313.150	
				Least DEPT (45)	Most CASH_OUT (144)	Values CASH_OUT (144), TRANSFER (231), ... [2 more]	

## Neural Net

Result History		ExampleSet (/Local Repository/ue22303845/Reports/Neural Net)				
	Name	Type	Missing	Statistics		
Data	Label			Negative	Positive	Values
	<b>IsFraud</b>	Binominal	0	0	1	0 (1643), 1 (1643)
Statistics	Prediction			Negative	Positive	Values
	<b>prediction(IsFraud)</b>	Binominal	0	0	1	0 (1875), 1 (1411)
Visualizations	Score			Min	Max	Average
	<b>confidence(0)</b>	Real	0	0.000	1.000	0.501
Annotations	Score			Min	Max	Average
	<b>confidence(1)</b>	Real	0	0.000	1.000	0.499
	<b>step</b>	Integer	0	Min 1	Max 743	Average 313.150
	<b>type</b>	Nominal	0	Least DEBIT (15)	Most CASH_OUT (1410)	Values CASH_OUT (1410), TRANSFER (931), ...[3 more]

## TAREA 5 - PASO A PASO EN PYTHON

## Práctica ETL + EDA + Modeling

## ▼ Installing libraries

```
[ ]: pip install pandas
```

## ▼ Load data

```
[1]: import pandas as pd
```

```
[6]: pwc = pd.read_csv('credit_card_bal.csv')
```

```
[8]: pwc.describe()
```

```
[8]:
```

	step	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
count	16426.000000	1.642600e+04	1.642600e+04	1.642600e+04	1.642600e+04	1.642600e+04	16426.000000	16426.000000
mean	306.185194	8.236570e+05	1.252718e+06	5.348437e+05	8.341153e+05	1.264052e+06	0.500000	0.000974
std	192.704918	1.852158e+06	3.277629e+06	2.539971e+06	3.226697e+06	3.592841e+06	0.500015	0.031196
min	1.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	0.000000
25%	162.000000	3.662393e+04	1.067450e+04	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	0.000000
50%	283.000000	1.706524e+05	1.202234e+05	0.000000e+00	0.000000e+00	1.244370e+05	0.500000	0.000000
75%	408.000000	5.482430e+05	7.965316e+05	0.000000e+00	5.214951e+05	1.115317e+06	1.000000	0.000000
max	743.000000	1.511569e+07	5.958504e+07	4.958504e+07	2.362305e+08	2.367265e+08	1.000000	1.000000

```
[9]: pwc.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16426 entries, 0 to 16425
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  -
0   step                  16426 non-null  int64
1   type                  16426 non-null  object
2   amount                16426 non-null  float64
3   nameOrig              16426 non-null  object
4   oldbalanceOrg         16426 non-null  float64
5   newbalanceOrig        16426 non-null  float64
6   nameDest              16426 non-null  object
7   oldbalanceDest        16426 non-null  float64
8   newbalanceDest        16426 non-null  float64
9   isFraud               16426 non-null  int64
10  isFlaggedFraud        16426 non-null  int64
dtypes: float64(5), int64(3), object(3)
memory usage: 1.4+ MB
```

```
[10]: pwc.columns
```

```
[10]: Index(['step', 'type', 'amount', 'nameOrig', 'oldbalanceOrg', 'newbalanceOrig',
        'nameDest', 'oldbalanceDest', 'newbalanceDest', 'isFraud',
        'isFlaggedFraud'],
        dtype='object')
```

```
[12]: pwc.corr

[12]: <bound method DataFrame.corr of
0      18  CASH_IN  180078.01  C1791832105  2796833.20  2976911.22
1      258  PAYMENT  2138.35  C634754263  0.00  0.00
2      183  TRANSFER  342675.47  C1778106585  342675.47  0.00
3      577  TRANSFER  141730.20  C1677850525  141730.20  0.00
4      742  TRANSFER  4009058.39  C1044665079  4009058.39  0.00
...     ...     ...     ...     ...     ...
16421  141  CASH_OUT  14898.80  C1882828175  104665.00  89766.20
16422  188  CASH_IN  437986.25  C1072308575  50474.00  488460.25
16423  586  CASH_OUT  0.00  C1303719003  0.00  0.00
16424  355  CASH_OUT  42483.97  C1595793252  42483.97  0.00
16425  374  PAYMENT  6069.42  C1139940003  0.00  0.00

      nameDest  oldbalanceDest  newbalanceDest  isFraud  isFlaggedFraud
0  C1756248403  1169682.86  989604.85  0  0
1  M205383539  0.00  0.00  0  0
2  C144558690  0.00  0.00  1  0
3  C607738036  0.00  0.00  1  0
4  C750074708  0.00  0.00  1  0
...     ...     ...     ...     ...
16421  C1045368778  2756474.78  2771373.59  0  0
16422  C1532905677  49843.00  0.00  0  0
16423  C900608348  1328472.86  1328472.86  1  0
16424  C636454309  51495.75  93979.72  1  0
16425  M17941836  0.00  0.00  0  0

[16426 rows x 11 columns]>

#Usefull commands

[ ]: pwd = current directory
ls = list files
cd = change directory
mkdir = make a new directory
```

## CONCLUSION

El resultado de los 6 modelos no fue satisfactorio para el objetivo de la actividad, siendo el mayor accuracy el del modelo de Naive Bayes con una confianza del 83% en la detección de casos no fraude.