Mini Projeto Santander - DSA

In [1]:

```
# Biblioteca import para o projeto
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import preprocessing
```

In [2]:

```
df = pd.read_csv("data/train.csv")
```

In [3]:

```
df.head(2)
```

Out[3]:

	ID	var3	var15	imp_ent_var16_ult1	imp_op_var39_comer_ult1	imp_op_var39_comer_ult3
0	1	2	23	0.0	0.0	0.0
1	3	2	34	0.0	0.0	0.0

2 rows × 371 columns

Pré Processamento e Analise

In [4]:

```
df.shape
```

Out[4]:

```
(76020, 371)
```

In [5]:

```
df.describe()
```

Out[5]:

	ID	var3	var15	imp_ent_var16_ult1	imp_op_var39_come
count	76020.000000	76020.000000	76020.000000	76020.000000	76020.0(
mean	75964.050723	-1523.199277	33.212865	86.208265	72.3(
std	43781.947379	39033.462364	12.956486	1614.757313	339.3
min	1.000000	-999999.000000	5.000000	0.000000	0.0(
25%	38104.750000	2.000000	23.000000	0.000000	0.0(
50%	76043.000000	2.000000	28.000000	0.000000	0.0(
75%	113748.750000	2.000000	40.000000	0.000000	0.0(
max	151838.000000	238.000000	105.000000	210000.000000	12888.0

8 rows × 371 columns

In [6]:

```
#Colunas
colunas = list(df.columns)
```

In [7]:

```
df.duplicated()
# Portanto nao temos valores duplicados
```

Out[7]:

```
0        False
1        False
2        False
3        False
4        False
         ...
76015    False
76016    False
76017    False
76018    False
76019    False
Length: 76020, dtype: bool
```

In [8]:

```
def valor_missing(valor):
    if (valor.isnull == True):
        print("Temos valor nulo")
```

In [9]:

```
df.apply(valor_missing)
# Portanto nao temos valores missing
```

Out[9]:

```
ID                          None
var3                        None
var15                       None
imp_ent_var16_ult1          None
imp_op_var39_comer_ult1     None
                            ...
saldo_medio_var44_hace3     None
saldo_medio_var44_ult1      None
saldo_medio_var44_ult3      None
var38                       None
TARGET                      None
Length: 371, dtype: object
```

In [10]:

```
def limite_valor(coluna):
    print("Valor minimo: ", str(coluna.min()), "Valor maximo : ", str(coluna.max()), "
 da coluna : ", str(coluna.name))
```

In [11]:

```
# Alguns saldos possuem valores negativos na conta
# Algumas variaveis tem apenas 1 valor
df.apply(limite_valor, axis=0)
```

```
Valor minimo:  1.0 Valor maximo :  151838.0  da coluna :  ID
Valor minimo:  -999999.0 Valor maximo :  238.0  da coluna :  var3
Valor minimo:  5.0 Valor maximo :  105.0  da coluna :  var15
Valor minimo:  0.0 Valor maximo :  210000.0  da coluna :  imp_ent_var16_ul
t1
Valor minimo:  0.0 Valor maximo :  12888.03  da coluna :  imp_op_var39_com
er_ult1
Valor minimo:  0.0 Valor maximo :  21024.81  da coluna :  imp_op_var39_com
er_ult3
Valor minimo:  0.0 Valor maximo :  8237.82  da coluna :  imp_op_var40_come
r_ult1
Valor minimo:  0.0 Valor maximo :  11073.57  da coluna :  imp_op_var40_com
er_ult3
Valor minimo:  0.0 Valor maximo :  6600.0  da coluna :  imp_op_var40_efect
_ult1
Valor minimo:  0.0 Valor maximo :  6600.0  da coluna :  imp_op_var40_efect
_ult3
Valor minimo:  0.0 Valor maximo :  8237.82  da coluna :  imp_op_var40_ult1
Valor minimo:  0.0 Valor maximo :  12888.03  da coluna :  imp_op_var41_com
er_ult1
Valor minimo:  0.0 Valor maximo :  16566.81  da coluna :  imp_op_var41_com
er_ult3
Valor minimo:  0.0 Valor maximo :  45990.0  da coluna :  imp_op_var41_efec
t_ult1
Valor minimo:  0.0 Valor maximo :  131100.0  da coluna :  imp_op_var41_efe
ct_ult3
Valor minimo:  0.0 Valor maximo :  47598.09  da coluna :  imp_op_var41_ult
1
Valor minimo:  0.0 Valor maximo :  45990.0  da coluna :  imp_op_var39_efec
t_ult1
Valor minimo:  0.0 Valor maximo :  131100.0  da coluna :  imp_op_var39_efe
ct_ult3
Valor minimo:  0.0 Valor maximo :  47598.09  da coluna :  imp_op_var39_ult
1
Valor minimo:  0.0 Valor maximo :  105000.0  da coluna :  imp_sal_var16_ul
t1
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var1_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var1
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  ind_var2_0
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  ind_var2
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var5_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var5
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var6_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var6
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var8_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var8
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var12_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var12
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var13_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var13_corto_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var13_corto
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var13_largo_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var13_largo
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var13_medio_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var13_medio
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var13
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var14_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var14
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var17_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var17
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var18_0
```

```
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var18
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var19
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var20_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var20
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var24_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var24
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var25_cte
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var26_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var26_cte
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var26
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var25_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var25
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  ind_var27_0
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  ind_var28_0
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  ind_var28
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  ind_var27
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var29_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var29
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var30_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var30
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var31_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var31
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var32_cte
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var32_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var32
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var33_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var33
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var34_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var34
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var37_cte
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var37_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var37
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var39_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var40_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var40
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var41_0
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  ind_var41
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var39
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var44_0
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var44
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  ind_var46_0
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  ind_var46
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_var1_0
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_var1
Valor minimo:  0.0 Valor maximo :  7.0  da coluna :  num_var4
Valor minimo:  0.0 Valor maximo :  15.0  da coluna :  num_var5_0
Valor minimo:  0.0 Valor maximo :  15.0  da coluna :  num_var5
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var6_0
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var6
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_var8_0
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var8
Valor minimo:  0.0 Valor maximo :  111.0  da coluna :  num_var12_0
Valor minimo:  0.0 Valor maximo :  15.0  da coluna :  num_var12
Valor minimo:  0.0 Valor maximo :  18.0  da coluna :  num_var13_0
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_var13_corto_0
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_var13_corto
Valor minimo:  0.0 Valor maximo :  18.0  da coluna :  num_var13_largo_0
Valor minimo:  0.0 Valor maximo :  18.0  da coluna :  num_var13_largo
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var13_medio_0
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var13_medio
Valor minimo:  0.0 Valor maximo :  18.0  da coluna :  num_var13
```

```
Valor minimo:  0.0 Valor maximo :  111.0  da coluna :  num_var14_0
Valor minimo:  0.0 Valor maximo :  12.0  da coluna :  num_var14
Valor minimo:  0.0 Valor maximo :  36.0  da coluna :  num_var17_0
Valor minimo:  0.0 Valor maximo :  27.0  da coluna :  num_var17
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var18_0
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var18
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var20_0
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var20
Valor minimo:  0.0 Valor maximo :  9.0  da coluna :  num_var24_0
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_var24
Valor minimo:  0.0 Valor maximo :  33.0  da coluna :  num_var26_0
Valor minimo:  0.0 Valor maximo :  33.0  da coluna :  num_var26
Valor minimo:  0.0 Valor maximo :  33.0  da coluna :  num_var25_0
Valor minimo:  0.0 Valor maximo :  33.0  da coluna :  num_var25
Valor minimo:  0.0 Valor maximo :  117.0  da coluna :  num_op_var40_hace2
Valor minimo:  0.0 Valor maximo :  48.0  da coluna :  num_op_var40_hace3
Valor minimo:  0.0 Valor maximo :  234.0  da coluna :  num_op_var40_ult1
Valor minimo:  0.0 Valor maximo :  351.0  da coluna :  num_op_var40_ult3
Valor minimo:  0.0 Valor maximo :  249.0  da coluna :  num_op_var41_hace2
Valor minimo:  0.0 Valor maximo :  81.0  da coluna :  num_op_var41_hace3
Valor minimo:  0.0 Valor maximo :  468.0  da coluna :  num_op_var41_ult1
Valor minimo:  0.0 Valor maximo :  468.0  da coluna :  num_op_var41_ult3
Valor minimo:  0.0 Valor maximo :  249.0  da coluna :  num_op_var39_hace2
Valor minimo:  0.0 Valor maximo :  81.0  da coluna :  num_op_var39_hace3
Valor minimo:  0.0 Valor maximo :  468.0  da coluna :  num_op_var39_ult1
Valor minimo:  0.0 Valor maximo :  468.0  da coluna :  num_op_var39_ult3
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_var27_0
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_var28_0
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_var28
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_var27
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var29_0
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var29
Valor minimo:  0.0 Valor maximo :  114.0  da coluna :  num_var30_0
Valor minimo:  0.0 Valor maximo :  33.0  da coluna :  num_var30
Valor minimo:  0.0 Valor maximo :  36.0  da coluna :  num_var31_0
Valor minimo:  0.0 Valor maximo :  27.0  da coluna :  num_var31
Valor minimo:  0.0 Valor maximo :  12.0  da coluna :  num_var32_0
Valor minimo:  0.0 Valor maximo :  12.0  da coluna :  num_var32
Valor minimo:  0.0 Valor maximo :  12.0  da coluna :  num_var33_0
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_var33
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var34_0
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var34
Valor minimo:  0.0 Valor maximo :  36.0  da coluna :  num_var35
Valor minimo:  0.0 Valor maximo :  105.0  da coluna :  num_var37_med_ult2
Valor minimo:  0.0 Valor maximo :  114.0  da coluna :  num_var37_0
Valor minimo:  0.0 Valor maximo :  114.0  da coluna :  num_var37
Valor minimo:  0.0 Valor maximo :  33.0  da coluna :  num_var39_0
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_var40_0
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var40
Valor minimo:  0.0 Valor maximo :  33.0  da coluna :  num_var41_0
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_var41
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var39
Valor minimo:  0.0 Valor maximo :  114.0  da coluna :  num_var42_0
Valor minimo:  0.0 Valor maximo :  18.0  da coluna :  num_var42
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_var44_0
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var44
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_var46_0
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_var46
Valor minimo:  -0.9 Valor maximo :  3000000.0  da coluna :  saldo_var1
Valor minimo:  -2895.72 Valor maximo :  619329.15  da coluna :  saldo_var5
Valor minimo:  0.0 Valor maximo :  19531.8  da coluna :  saldo_var6
```

```
Valor minimo:  -4942.26 Valor maximo :  240045.0  da coluna :  saldo_var8
Valor minimo:  0.0 Valor maximo :  3008077.32  da coluna :  saldo_var12
Valor minimo:  0.0 Valor maximo :  450000.0  da coluna :  saldo_var13_cort
o
Valor minimo:  0.0 Valor maximo :  1500000.0  da coluna :  saldo_var13_lar
go
Valor minimo:  0.0 Valor maximo :  30000.0  da coluna :  saldo_var13_medio
Valor minimo:  0.0 Valor maximo :  1500000.0  da coluna :  saldo_var13
Valor minimo:  0.0 Valor maximo :  450000.0  da coluna :  saldo_var14
Valor minimo:  0.0 Valor maximo :  6119500.14  da coluna :  saldo_var17
Valor minimo:  0.0 Valor maximo :  3000000.0  da coluna :  saldo_var18
Valor minimo:  0.0 Valor maximo :  455858.16  da coluna :  saldo_var20
Valor minimo:  0.0 Valor maximo :  3008077.32  da coluna :  saldo_var24
Valor minimo:  0.0 Valor maximo :  69756.72  da coluna :  saldo_var26
Valor minimo:  0.0 Valor maximo :  69756.72  da coluna :  saldo_var25
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  saldo_var28
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  saldo_var27
Valor minimo:  0.0 Valor maximo :  19531.8  da coluna :  saldo_var29
Valor minimo:  -4942.26 Valor maximo :  3458077.32  da coluna :  saldo_var
30
Valor minimo:  0.0 Valor maximo :  6119500.14  da coluna :  saldo_var31
Valor minimo:  0.0 Valor maximo :  12210.78  da coluna :  saldo_var32
Valor minimo:  0.0 Valor maximo :  142078.8  da coluna :  saldo_var33
Valor minimo:  0.0 Valor maximo :  36000.0  da coluna :  saldo_var34
Valor minimo:  0.0 Valor maximo :  60000.0  da coluna :  saldo_var37
Valor minimo:  -0.9 Valor maximo :  8192.61  da coluna :  saldo_var40
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  saldo_var41
Valor minimo:  -4942.26 Valor maximo :  3008077.32  da coluna :  saldo_var
42
Valor minimo:  0.0 Valor maximo :  740006.61  da coluna :  saldo_var44
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  saldo_var46
Valor minimo:  0.0 Valor maximo :  99.0  da coluna :  var36
Valor minimo:  0.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_am
ort_var18_1y3
Valor minimo:  0.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_am
ort_var34_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_a
port_var13_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_a
port_var17_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_a
port_var33_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_c
ompra_var44_1y3
Valor minimo:  0.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_re
emb_var13_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_r
eemb_var17_1y3
Valor minimo:  0.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_re
emb_var33_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_t
rasp_var17_in_1y3
Valor minimo:  0.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_tr
asp_var17_out_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_t
rasp_var33_in_1y3
Valor minimo:  0.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_tr
asp_var33_out_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_imp_v
enta_var44_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_num_a
```

```
port_var13_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_num_a
port_var17_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_num_a
port_var33_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_num_c
ompra_var44_1y3
Valor minimo:  0.0 Valor maximo :  9999999999.0  da coluna :  delta_num_re
emb_var13_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_num_r
eemb_var17_1y3
Valor minimo:  0.0 Valor maximo :  9999999999.0  da coluna :  delta_num_re
emb_var33_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_num_t
rasp_var17_in_1y3
Valor minimo:  0.0 Valor maximo :  9999999999.0  da coluna :  delta_num_tr
asp_var17_out_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_num_t
rasp_var33_in_1y3
Valor minimo:  0.0 Valor maximo :  9999999999.0  da coluna :  delta_num_tr
asp_var33_out_1y3
Valor minimo:  -1.0 Valor maximo :  9999999999.0  da coluna :  delta_num_v
enta_var44_1y3
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  imp_amort_var18_hace3
Valor minimo:  0.0 Valor maximo :  15691.8  da coluna :  imp_amort_var18_u
lt1
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  imp_amort_var34_hace3
Valor minimo:  0.0 Valor maximo :  1096.02  da coluna :  imp_amort_var34_u
lt1
Valor minimo:  0.0 Valor maximo :  840000.0  da coluna :  imp_aport_var13_
hace3
Valor minimo:  0.0 Valor maximo :  450000.0  da coluna :  imp_aport_var13_
ult1
Valor minimo:  0.0 Valor maximo :  6083691.87  da coluna :  imp_aport_var1
7_hace3
Valor minimo:  0.0 Valor maximo :  432457.32  da coluna :  imp_aport_var17
_ult1
Valor minimo:  0.0 Valor maximo :  36000.0  da coluna :  imp_aport_var33_h
ace3
Valor minimo:  0.0 Valor maximo :  1260.0  da coluna :  imp_aport_var33_ul
t1
Valor minimo:  0.0 Valor maximo :  145384.92  da coluna :  imp_var7_emit_u
lt1
Valor minimo:  0.0 Valor maximo :  1039260.0  da coluna :  imp_var7_recib_
ult1
Valor minimo:  0.0 Valor maximo :  210001.35  da coluna :  imp_compra_var4
4_hace3
Valor minimo:  0.0 Valor maximo :  3410058.66  da coluna :  imp_compra_var
44_ult1
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  imp_reemb_var13_hace3
Valor minimo:  0.0 Valor maximo :  450000.0  da coluna :  imp_reemb_var13_
ult1
Valor minimo:  0.0 Valor maximo :  12027.15  da coluna :  imp_reemb_var17_
hace3
Valor minimo:  0.0 Valor maximo :  182132.97  da coluna :  imp_reemb_var17
_ult1
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  imp_reemb_var33_hace3
Valor minimo:  0.0 Valor maximo :  1200.0  da coluna :  imp_reemb_var33_ul
t1
Valor minimo:  0.0 Valor maximo :  1155003.0  da coluna :  imp_var43_emit_
ult1
```

```
Valor minimo:  0.0 Valor maximo :  2310003.0  da coluna :  imp_trans_var37
_ult1
Valor minimo:  0.0 Valor maximo :  96781.44  da coluna :  imp_trasp_var17_
in_hace3
Valor minimo:  0.0 Valor maximo :  133730.58  da coluna :  imp_trasp_var17
_in_ult1
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  imp_trasp_var17_out_h
ace3
Valor minimo:  0.0 Valor maximo :  69622.29  da coluna :  imp_trasp_var17_
out_ult1
Valor minimo:  0.0 Valor maximo :  49581.27  da coluna :  imp_trasp_var33_
in_hace3
Valor minimo:  0.0 Valor maximo :  13207.32  da coluna :  imp_trasp_var33_
in_ult1
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  imp_trasp_var33_out_h
ace3
Valor minimo:  0.0 Valor maximo :  3000.0  da coluna :  imp_trasp_var33_ou
t_ult1
Valor minimo:  0.0 Valor maximo :  209834.4  da coluna :  imp_venta_var44_
hace3
Valor minimo:  0.0 Valor maximo :  2754476.46  da coluna :  imp_venta_var4
4_ult1
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var7_emit_ult1
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var7_recib_ult1
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var10_ult1
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var10cte_ult1
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var9_cte_ult1
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var9_ult1
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var43_emit_ult1
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  ind_var43_recib_ult1
Valor minimo:  0.0 Valor maximo :  30000.0  da coluna :  var21
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_var2_0_ult1
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_var2_ult1
Valor minimo:  0.0 Valor maximo :  24.0  da coluna :  num_aport_var13_hace
3
Valor minimo:  0.0 Valor maximo :  30.0  da coluna :  num_aport_var13_ult1
Valor minimo:  0.0 Valor maximo :  12.0  da coluna :  num_aport_var17_hace
3
Valor minimo:  0.0 Valor maximo :  21.0  da coluna :  num_aport_var17_ult1
Valor minimo:  0.0 Valor maximo :  12.0  da coluna :  num_aport_var33_hace
3
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_aport_var33_ult1
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_var7_emit_ult1
Valor minimo:  0.0 Valor maximo :  24.0  da coluna :  num_var7_recib_ult1
Valor minimo:  0.0 Valor maximo :  9.0  da coluna :  num_compra_var44_hace
3
Valor minimo:  0.0 Valor maximo :  39.0  da coluna :  num_compra_var44_ult
1
Valor minimo:  0.0 Valor maximo :  60.0  da coluna :  num_ent_var16_ult1
Valor minimo:  0.0 Valor maximo :  123.0  da coluna :  num_var22_hace2
Valor minimo:  0.0 Valor maximo :  108.0  da coluna :  num_var22_hace3
Valor minimo:  0.0 Valor maximo :  96.0  da coluna :  num_var22_ult1
Valor minimo:  0.0 Valor maximo :  234.0  da coluna :  num_var22_ult3
Valor minimo:  0.0 Valor maximo :  78.0  da coluna :  num_med_var22_ult3
Valor minimo:  0.0 Valor maximo :  267.0  da coluna :  num_med_var45_ult3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_meses_var5_ult3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_meses_var8_ult3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_meses_var12_ult3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_meses_var13_corto
_ult3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_meses_var13_largo
```

```
_ult3
Valor minimo:  0.0 Valor maximo :  2.0  da coluna :  num_meses_var13_medio
_ult3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_meses_var17_ult3
Valor minimo:  0.0 Valor maximo :  2.0  da coluna :  num_meses_var29_ult3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_meses_var33_ult3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_meses_var39_vig_u
lt3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_meses_var44_ult3
Valor minimo:  0.0 Valor maximo :  438.0  da coluna :  num_op_var39_comer_
ult1
Valor minimo:  0.0 Valor maximo :  600.0  da coluna :  num_op_var39_comer_
ult3
Valor minimo:  0.0 Valor maximo :  210.0  da coluna :  num_op_var40_comer_
ult1
Valor minimo:  0.0 Valor maximo :  582.0  da coluna :  num_op_var40_comer_
ult3
Valor minimo:  0.0 Valor maximo :  24.0  da coluna :  num_op_var40_efect_u
lt1
Valor minimo:  0.0 Valor maximo :  24.0  da coluna :  num_op_var40_efect_u
lt3
Valor minimo:  0.0 Valor maximo :  438.0  da coluna :  num_op_var41_comer_
ult1
Valor minimo:  0.0 Valor maximo :  438.0  da coluna :  num_op_var41_comer_
ult3
Valor minimo:  0.0 Valor maximo :  90.0  da coluna :  num_op_var41_efect_u
lt1
Valor minimo:  0.0 Valor maximo :  156.0  da coluna :  num_op_var41_efect_
ult3
Valor minimo:  0.0 Valor maximo :  90.0  da coluna :  num_op_var39_efect_u
lt1
Valor minimo:  0.0 Valor maximo :  156.0  da coluna :  num_op_var39_efect_
ult3
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_reemb_var13_hace3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_reemb_var13_ult1
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_reemb_var17_hace3
Valor minimo:  0.0 Valor maximo :  21.0  da coluna :  num_reemb_var17_ult1
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_reemb_var33_hace3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_reemb_var33_ult1
Valor minimo:  0.0 Valor maximo :  15.0  da coluna :  num_sal_var16_ult1
Valor minimo:  0.0 Valor maximo :  180.0  da coluna :  num_var43_emit_ult1
Valor minimo:  0.0 Valor maximo :  264.0  da coluna :  num_var43_recib_ult
1
Valor minimo:  0.0 Valor maximo :  93.0  da coluna :  num_trasp_var11_ult1
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_trasp_var17_in_ha
ce3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_trasp_var17_in_ul
t1
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_trasp_var17_out_h
ace3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_trasp_var17_out_u
lt1
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_trasp_var33_in_ha
ce3
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_trasp_var33_in_ul
t1
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  num_trasp_var33_out_h
ace3
Valor minimo:  0.0 Valor maximo :  3.0  da coluna :  num_trasp_var33_out_u
lt1
Valor minimo:  0.0 Valor maximo :  6.0  da coluna :  num_venta_var44_hace3
```

```
Valor minimo:  0.0 Valor maximo :  39.0  da coluna :  num_venta_var44_ult1
Valor minimo:  0.0 Valor maximo :  342.0  da coluna :  num_var45_hace2
Valor minimo:  0.0 Valor maximo :  339.0  da coluna :  num_var45_hace3
Valor minimo:  0.0 Valor maximo :  510.0  da coluna :  num_var45_ult1
Valor minimo:  0.0 Valor maximo :  801.0  da coluna :  num_var45_ult3
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  saldo_var2_ult1
Valor minimo:  -128.37 Valor maximo :  812137.26  da coluna :  saldo_medio
_var5_hace2
Valor minimo:  -8.04 Valor maximo :  1542339.36  da coluna :  saldo_medio_
var5_hace3
Valor minimo:  -922.38 Valor maximo :  601428.6  da coluna :  saldo_medio_
var5_ult1
Valor minimo:  -476.07 Valor maximo :  544365.57  da coluna :  saldo_medio
_var5_ult3
Valor minimo:  -287.67 Valor maximo :  231351.99  da coluna :  saldo_medio
_var8_hace2
Valor minimo:  0.0 Valor maximo :  77586.21  da coluna :  saldo_medio_var8
_hace3
Valor minimo:  -3401.34 Valor maximo :  228031.8  da coluna :  saldo_medio
_var8_ult1
Valor minimo:  -1844.52 Valor maximo :  177582.0  da coluna :  saldo_medio
_var8_ult3
Valor minimo:  0.0 Valor maximo :  3000538.14  da coluna :  saldo_medio_va
r12_hace2
Valor minimo:  0.0 Valor maximo :  668335.32  da coluna :  saldo_medio_var
12_hace3
Valor minimo:  0.0 Valor maximo :  3004185.6  da coluna :  saldo_medio_var
12_ult1
Valor minimo:  0.0 Valor maximo :  2272859.43  da coluna :  saldo_medio_va
r12_ult3
Valor minimo:  0.0 Valor maximo :  450000.0  da coluna :  saldo_medio_var1
3_corto_hace2
Valor minimo:  0.0 Valor maximo :  304838.7  da coluna :  saldo_medio_var1
3_corto_hace3
Valor minimo:  0.0 Valor maximo :  450000.0  da coluna :  saldo_medio_var1
3_corto_ult1
Valor minimo:  0.0 Valor maximo :  450000.0  da coluna :  saldo_medio_var1
3_corto_ult3
Valor minimo:  0.0 Valor maximo :  840000.0  da coluna :  saldo_medio_var1
3_largo_hace2
Valor minimo:  0.0 Valor maximo :  534000.0  da coluna :  saldo_medio_var1
3_largo_hace3
Valor minimo:  0.0 Valor maximo :  1500000.0  da coluna :  saldo_medio_var
13_largo_ult1
Valor minimo:  0.0 Valor maximo :  1034482.74  da coluna :  saldo_medio_va
r13_largo_ult3
Valor minimo:  0.0 Valor maximo :  7741.95  da coluna :  saldo_medio_var13
_medio_hace2
Valor minimo:  0.0 Valor maximo :  0.0  da coluna :  saldo_medio_var13_med
io_hace3
Valor minimo:  0.0 Valor maximo :  30000.0  da coluna :  saldo_medio_var13
_medio_ult1
Valor minimo:  0.0 Valor maximo :  18870.99  da coluna :  saldo_medio_var1
3_medio_ult3
Valor minimo:  -0.03 Valor maximo :  4210084.23  da coluna :  saldo_medio_
var17_hace2
Valor minimo:  0.0 Valor maximo :  2368558.95  da coluna :  saldo_medio_va
r17_hace3
Valor minimo:  0.0 Valor maximo :  3998687.46  da coluna :  saldo_medio_va
r17_ult1
Valor minimo:  0.0 Valor maximo :  3525776.88  da coluna :  saldo_medio_va
```

```
r17_ult3
Valor minimo:  0.0 Valor maximo :  10430.01  da coluna :  saldo_medio_var2
9_hace2
Valor minimo:  0.0 Valor maximo :  145.2  da coluna :  saldo_medio_var29_h
ace3
Valor minimo:  0.0 Valor maximo :  13793.67  da coluna :  saldo_medio_var2
9_ult1
Valor minimo:  0.0 Valor maximo :  7331.34  da coluna :  saldo_medio_var29
_ult3
Valor minimo:  0.0 Valor maximo :  50003.88  da coluna :  saldo_medio_var3
3_hace2
Valor minimo:  0.0 Valor maximo :  20385.72  da coluna :  saldo_medio_var3
3_hace3
Valor minimo:  0.0 Valor maximo :  138831.63  da coluna :  saldo_medio_var
33_ult1
Valor minimo:  0.0 Valor maximo :  91778.73  da coluna :  saldo_medio_var3
3_ult3
Valor minimo:  0.0 Valor maximo :  438329.22  da coluna :  saldo_medio_var
44_hace2
Valor minimo:  0.0 Valor maximo :  24650.01  da coluna :  saldo_medio_var4
4_hace3
Valor minimo:  0.0 Valor maximo :  681462.9  da coluna :  saldo_medio_var4
4_ult1
Valor minimo:  0.0 Valor maximo :  397884.3  da coluna :  saldo_medio_var4
4_ult3
Valor minimo:  5163.75 Valor maximo :  22034738.76  da coluna :  var38
Valor minimo:  0.0 Valor maximo :  1.0  da coluna :  TARGET
```

Out[11]:

```
ID                          None
var3                        None
var15                       None
imp_ent_var16_ult1          None
imp_op_var39_comer_ult1     None
                            ...
saldo_medio_var44_hace3     None
saldo_medio_var44_ult1      None
saldo_medio_var44_ult3      None
var38                       None
TARGET                      None
Length: 371, dtype: object
```

In [12]:

```
#Temos um problema de desbalanceamento
df.TARGET.value_counts()
```

Out[12]:

```
0    73012
1     3008
Name: TARGET, dtype: int64
```

Tratando Valores unicos nas variaveis

In [13]:

```
valores_unicos = []
for coluna in colunas:
    if df[coluna].nunique() < 2:
        del df[coluna]
```

In [14]:

```
df.shape
```

Out[14]:

```
(76020, 337)
```

In [15]:

```
colunas = list(df.columns)
```

Tratando DELTA

In [16]:

```
# Avaliando a variavel "delta" por apresentar valores extremos
df.iloc[:,175:201].describe()
```

Out[16]:

	delta_imp_amort_var18_1y3	delta_imp_amort_var34_1y3	delta_imp_aport_var13_1y3	de
count	7.602000e+04	7.602000e+04	7.602000e+04	
mean	2.630887e+05	2.630887e+05	4.867140e+07	
std	5.129183e+07	5.129183e+07	6.959537e+08	
min	0.000000e+00	0.000000e+00	-1.000000e+00	
25%	0.000000e+00	0.000000e+00	0.000000e+00	
50%	0.000000e+00	0.000000e+00	0.000000e+00	
75%	0.000000e+00	0.000000e+00	0.000000e+00	
max	1.000000e+10	1.000000e+10	1.000000e+10	

8 rows × 26 columns

In [17]:

```
coluna_delta = list(df.iloc[:,175:201])
coluna_delta
```

Out[17]:

```
['delta_imp_amort_var18_1y3',
 'delta_imp_amort_var34_1y3',
 'delta_imp_aport_var13_1y3',
 'delta_imp_aport_var17_1y3',
 'delta_imp_aport_var33_1y3',
 'delta_imp_compra_var44_1y3',
 'delta_imp_reemb_var13_1y3',
 'delta_imp_reemb_var17_1y3',
 'delta_imp_reemb_var33_1y3',
 'delta_imp_trasp_var17_in_1y3',
 'delta_imp_trasp_var17_out_1y3',
 'delta_imp_trasp_var33_in_1y3',
 'delta_imp_trasp_var33_out_1y3',
 'delta_imp_venta_var44_1y3',
 'delta_num_aport_var13_1y3',
 'delta_num_aport_var17_1y3',
 'delta_num_aport_var33_1y3',
 'delta_num_compra_var44_1y3',
 'delta_num_reemb_var13_1y3',
 'delta_num_reemb_var17_1y3',
 'delta_num_reemb_var33_1y3',
 'delta_num_trasp_var17_in_1y3',
 'delta_num_trasp_var17_out_1y3',
 'delta_num_trasp_var33_in_1y3',
 'delta_num_trasp_var33_out_1y3',
 'delta_num_venta_var44_1y3']
```

In [18]:

```
for coluna in coluna_delta:
    df[coluna] = pd.Series([1 if x == 9999999999 else x for x in df[coluna]])
```

In [19]:

```
df.iloc[:,175:201].describe()
```

Out[19]:

	delta_imp_amort_var18_1y3	delta_imp_amort_var34_1y3	delta_imp_aport_var13_1y3	de
count	76020.000000	76020.000000	76020.000000	
mean	0.000026	0.000026	-0.016956	
std	0.005129	0.005129	0.166389	
min	0.000000	0.000000	-1.000000	
25%	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	0.000000	
75%	0.000000	0.000000	0.000000	
max	1.000000	1.000000	5.500000	

8 rows × 26 columns

Normalização dos dados

In [20]:

```
min_max_scaler = preprocessing.MinMaxScaler()
```

In [21]:

```
target = df['TARGET']
```

In [22]:

```
x = df.drop(['TARGET'], axis=1).values #returns a numpy array
df_scaled = min_max_scaler.fit_transform(x)
df_scaled = pd.DataFrame(df_scaled, columns=colunas[0:336])
```

In [23]:

```
df_scaled.head(5)
```

Out[23]:

	ID	var3	var15	imp_ent_var16_ult1	imp_op_var39_comer_ult1	imp_op_var39_c
0	0.000000	0.999764	0.18	0.0	0.00000	
1	0.000013	0.999764	0.29	0.0	0.00000	
2	0.000020	0.999764	0.18	0.0	0.00000	
3	0.000046	0.999764	0.32	0.0	0.01513	
4	0.000059	0.999764	0.34	0.0	0.00000	

5 rows × 336 columns

PCA

In [24]:

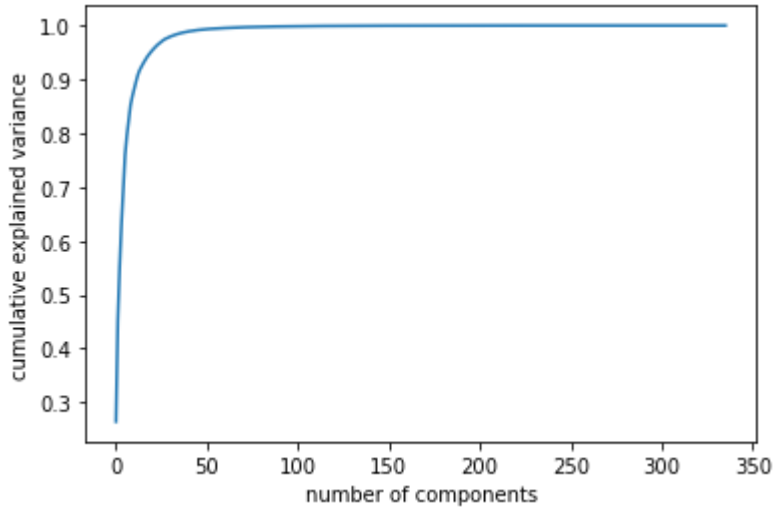
```
del df['ID']
```

In [25]:

```
from sklearn.decomposition import PCA
```

In [26]:

```
pca = PCA().fit(df_scaled)
plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.xlabel('number of components')
plt.ylabel('cumulative explained variance');
```



In [27]:

```
pca = PCA(n_components=50)
df_pca = pca.fit_transform(df_scaled)
```

In [28]:

```
pd.DataFrame(df_pca).shape
```

Out[28]:

```
(76020, 50)
```

Machine Learning

In [29]:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
```

```
C:\Users\bruno\Anaconda3\lib\site-packages\sklearn\ensemble\weight_boostin
g.py:29: DeprecationWarning: numpy.core.umath_tests is an internal NumPy m
odule and should not be imported. It will be removed in a future NumPy rel
ease.
  from numpy.core.umath_tests import inner1d
```

Divisão das variaveis de treino e teste

In [30]:

```
X_train, X_test, y_train, y_test = train_test_split(df_pca, target, test_size=0.30, ran
dom_state=42)
```

In [31]:

```
X_train.shape, y_train.shape
```

Out[31]:

```
((53214, 50), (53214,))
```

In [32]:

```
X_test.shape, y_test.shape
```

Out[32]:

```
((22806, 50), (22806,))
```

Algoritmo RandomForestClassifier

In [33]:

```
rfc = RandomForestClassifier(n_estimators = 500, random_state = 42)
rfc.fit(X_train, y_train)
```

Out[33]:

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gin
i',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=500, n_jobs=1,
            oob_score=False, random_state=42, verbose=0, warm_start=False)
```

In [34]:

```
# Prectitions com os valores de test
predictions = rfc.predict(X_test)
```

Error

In [35]:

```
# Calulo do erro
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print("Acuracia:",metrics.accuracy_score(y_test, predictions))
```

```
Acuracia: 0.9521178637200737
```