Relatório de Pré processamento - inspeção de datasets

Cláudia Patrícia Silva Pimentel

E-mails: cacaupimentel@hotmail.com (mailto:cacaupimentel@hotmail.com), cacaupimentel@gmail.com (mailto:cacaupimentel@gmail.com), claudiapimentel@aluno.uema.br (mailto:claudiapimentel@aluno.uema.br)

1. Introdução

Trata-se de relatório de análise de datasets, cujo objetivo é analisar os tipos de dados, as tarefas, atributos e instâncias. Assim, este instrumento foi elaborado no jupyter notebook, sendo exportado para pdf, com o intuito de incluí-lo no SIGUEMA Acadêmico, conforme solicitação da Tarefa 1.

Neste sentido, este documento foi divido em mais 5 capítulos, além desta introdução, nos quais serão apresentados o resumo de cada um dos 5 datasets elecandos abaixo:

1. Default of credit card clients Data Set, disoponível em https://archive.ics.uci.edu/ml/index.html (https://archive.ics.uci.edu/ml/index.html);
2. Wholesale customers Data Set, disponível em https://archive.ics.uci.edu/ml/datasets/Wholesale+customers (https://archive.ics.uci.edu/ml/datasets/Wholesale+customers)
3. Post-Operative Patient Data Set em https://archive.ics.uci.edu/ml/datasets/Post-Operative+Patient (https://archive.ics.uci.edu/ml/datasets/Post-Operative+Patient);
4. Real-time Election Results: Portugal 2019 Data Set, disponível em https://archive.ics.uci.edu/ml/datasets/Real-time+Election+Results%3A+Portugal+2019 (https://archive.ics.uci.edu/ml/datasets/Real-time+Election+Results%3A+Portugal+2019);
5. Clickstream data for online shopping Data Set, em disponível em https://archive.ics.uci.edu/ml/datasets/clickstream+data+for+online+shopping (https://archive.ics.uci.edu/ml/datasets/clickstream+data+for+online+shopping)

2. Resumo e inspeções dos atributos de default of credit card clients Data Set (https://archive.ics.uci.edu/ml/index.html)

2.1 Informações sobre a Fonte

1. Autores: Yeh, IC, e Lien, CH;
2. Artigo: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients;
3. Publicação: Expert Systems with Applications, 36(2), 2473-2480;
4. Ano: 2009;
5. Objetivo da Pesquisa: foi um caso de comparação da precisão peditiva da probalilidade de inadimplência de clientes, da cidade de Taiwan, aplicando 6 médotos de mineração de dados (Yeh, Lien, 2009).;
6. Resultados da Pesquisa: os autores apresentaram um novo "Sorting Smoothing Method", traduzido por "Método de Suavização de Classificação", concluindo que "(...) entre as seis técnicas de mineração de dados, a rede neural artificial é a única que pode estimar com precisão a probabilidade real de inadimplência" (Yeh, Lien, 2009).

2.2 Importando o default of credit card clients Data Set (https://archive.ics.uci.edu/ml/index.html)

Foram instalados os pacotes do pandas, numpy e pandas_profiling, para exploraçao dos dados.

In [1]:

```
# Instalando o pacote pandas
%pip install pandas
```

```
Requirement already satisfied: pandas in c:\users\fcalp\anaconda3\lib\site-p
ackages (1.1.3)
Requirement already satisfied: pytz>=2017.2 in c:\users\fcalp\anaconda3\lib
\site-packages (from pandas) (2020.1)
Requirement already satisfied: numpy>=1.15.4 in c:\users\fcalp\anaconda3\lib
\site-packages (from pandas) (1.19.2)
Requirement already satisfied: python-dateutil>=2.7.3 in c:\users\fcalp\anac
onda3\lib\site-packages (from pandas) (2.8.1)
Requirement already satisfied: six>=1.5 in c:\users\fcalp\anaconda3\lib\site
-packages (from python-dateutil>=2.7.3->pandas) (1.15.0)Note: you may need t
o restart the kernel to use updated packages.
```

In [2]:

```
# Instalando o pacote numpy
%pip install numpy
```

```
Requirement already satisfied: numpy in c:\users\fcalp\anaconda3\lib\site-pa
ckages (1.19.2)
Note: you may need to restart the kernel to use updated packages.
```

In [3]:

```
# Instalando o pacote pandas_profiling
%pip install pandas_profiling
```

```
Requirement already satisfied: pandas_profiling in c:\users\fcalp\anaconda
3\lib\site-packages (2.11.0)
Requirement already satisfied: jinja2>=2.11.1 in c:\users\fcalp\anaconda3
\lib\site-packages (from pandas_profiling) (2.11.2)
Requirement already satisfied: phik>=0.10.0 in c:\users\fcalp\anaconda3\li
b\site-packages (from pandas_profiling) (0.11.2)
Requirement already satisfied: visions[type_image_path]==0.6.0 in c:\users
\fcalp\anaconda3\lib\site-packages (from pandas_profiling) (0.6.0)
Requirement already satisfied: matplotlib>=3.2.0 in c:\users\fcalp\anacond
a3\lib\site-packages (from pandas_profiling) (3.3.2)
Requirement already satisfied: confuse>=1.0.0 in c:\users\fcalp\anaconda3
\lib\site-packages (from pandas_profiling) (1.4.0)
Requirement already satisfied: numpy>=1.16.0 in c:\users\fcalp\anaconda3\l
ib\site-packages (from pandas_profiling) (1.19.2)
Requirement already satisfied: attrs>=19.3.0 in c:\users\fcalp\anaconda3\l
ib\site-packages (from pandas_profiling) (20.3.0)
Note: you may need to restart the kernel to use updated packages.Requireme
nt already satisfied: seaborn>=0.10.1 in c:\users\fcalp\anaconda3\lib\site
-packages (from pandas_profiling) (0.11.0)
```

In [7]:

```
# Importando as bibliotecas necessárias
import pandas as pd
import numpy as np
from pandas_profiling import ProfileReport
```

2.3 Procedimentos de limpeza do arquivo

Quando importado inicialmente, verificou-se que o arquivo identifica, na primeira linha, as variáveis binárias do estudo. Neste ponto foi necessário ignorar a primeira linha do arquivo, uma vez que esta linha possui a rotulagem das variáveis atribuidas no estudo dos autores Yeh e Lien (2009).

In [5]:

```
#importando o data set sem tratamento inicial apenas
cartaoini = pd.read_excel('DataSets/default of credit card clients.xls')
cartaoini
```

Out[5]:

	Unnamed: 0	X1	X2	X3	X4	X5	X6	X7	X8	X
0	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY
1	1	20000	2	2	1	24	2	2	-1	
2	2	120000	2	2	2	26	-1	2	0	
3	3	90000	2	2	2	34	0	0	0	
4	4	50000	2	2	1	37	0	0	0	
...	...	...	...	...	...	...	...	...	...	
29996	29996	220000	1	3	1	39	0	0	0	
29997	29997	150000	1	3	2	43	-1	-1	-1	
29998	29998	30000	1	2	2	37	4	3	2	
29999	29999	80000	1	3	1	41	1	-1	0	
30000	30000	50000	1	2	1	46	0	0	0	

30001 rows × 25 columns

In [6]:

```
#importando o data set
cartao = pd.read_excel('DataSets/default of credit card clients.xls', sheet_name="Data", sk
cartao
```

Out[6]:

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4
0	1	20000	2	2	1	24	2	2	-1	-1
1	2	120000	2	2	2	26	-1	2	0	0
2	3	90000	2	2	2	34	0	0	0	0
3	4	50000	2	2	1	37	0	0	0	0
4	5	50000	1	2	1	57	-1	0	-1	0
...	...	...	...	...	...	...	...	...	...	...
29995	29996	220000	1	3	1	39	0	0	0	0
29996	29997	150000	1	3	2	43	-1	-1	-1	-1
29997	29998	30000	1	2	2	37	4	3	2	-1
29998	29999	80000	1	3	1	41	1	-1	0	0
29999	30000	50000	1	2	1	46	0	0	0	0

30000 rows × 25 columns

In [10]:

```
#Incluindo os dados nos datasets
```

2.4 Inspeção e descrição dos Atributos

As aludidas variáveis são binárias e no estudo, segundo os autores, contou com uma revisão de literatura e utilizou 23 delas como explicação. sendo diponibiliza de forma simplificada como a seguinte descrição (Yeh, Lien, 2009):

X1: Coluna "LIMIT_BAL", é o valor do crédito concedido (dólar NT): inclui tanto o crédito ao consumidor individual quanto o crédito familiar (suplementar).

X2: Coluna "SEX", é o Gênero:

1 = masculino;

2 = feminino.

X3. Coluna "EDUCATION" é a escolaridade:

1 = pós-graduação;

2 = universidade;

3 = ensino médio;

4 = outros.

X4: Coluna "MARRIAGE" é o Estado civil:

1 = casado;

2 = solteiro;

3 = outros.

X5: Coluna "AGE" é a Idade, mensurada em ano.

X6 - X11: Colunas com o histórico de pagamentos anteriores. Rastreamos os registros de pagamentos mensais anteriores (de abril a setembro de 2005) da seguinte forma:

X6 = Coluna "PAY_0" é o status de reembolso em setembro de 2005;

X7 = Coluna "PAY_2" é a situação de amortização em agosto de 2005;

X8 = Coluna "PAY_3" é a situação de amortização em julho de 2005;

X9 = Coluna "PAY_4" é a situação de amortização em junho de 2005;

X10 = Coluna "PAY_5" é a situação de amortização em maio de 2005;

X11 = Coluna "PAY_6" é o estado de reembolso em abril de 2005.

A escala de medição para o estado de reembolso é:

-1 = pagamento em dia;

1 = atraso no pagamento por um mês;

2 = atraso no pagamento por dois meses; . . .;

8 = atraso no pagamento por oito meses;

9 = atraso no pagamento de nove meses ou mais.

X12-X17: Colunas com valor da fatura (dólar NT).

X12 = Coluna "BILL_AMT1" é o valor da fatura em setembro de 2005;

X13 = Coluna "BILL_AMT2" é o valor da fatura em agosto de 2005;

X14 = Coluna "BILL_AMT3" é o valor da fatura em julho de 2005;

X15 = Coluna "BILL_AMT4" é o valor da fatura em junho de 2005;

X16 = Coluna "BILL_AMT5" é o valor da fatura em maio de 2005;

X17 = Coluna "BILL_AMT6" é o valor da fatura em abril de 2005.

X18-X23: Colunas com valor do pagamento anterior (dólar NT).

X18 = Coluna "PAY_AMT1" é o valor pago em setembro de 2005;

X19 = Coluna "PAY_AMT2" é o valor pago em agosto de 2005;

X20 = Coluna "PAY_AMT3" é o valor pago em julho de 2005;

X21 = Coluna "PAY_AMT4" é o valor pago em junho de 2005;

X22 = Coluna "PAY_AMT5" é o valor pago em maio de 2005;

X23 = Coluna "PAY_AMT6" é o valor pago em abril de 2005.

Y: Coluna "default payment next month" é e a a probabilidade real de inadimplência como variável de resposta;

Segundo os autores, o resultado será dado (..) com a probabilidade real de inadimplência como variável de resposta (Y) e a probabilidade preditiva de inadimplência como variável independente (X), o resultado da regressão linear simples (Y = A + BX) mostra que o modelo de previsão produzido pela rede neural artificial tem o maior coeficiente de determinação; sua interceptação de regressão (A) é próxima a zero e o coeficiente de regressão (B) a um.

In [11]:

```
# Inspecionando os atributos do dataset
cartao.describe()
```

Out[11]:

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE
count	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000
mean	15000.500000	167484.322667	1.603733	1.853133	1.551867	35.485500
std	8660.398374	129747.661567	0.489129	0.790349	0.521970	9.217904
min	1.000000	10000.000000	1.000000	0.000000	0.000000	21.000000
25%	7500.750000	50000.000000	1.000000	1.000000	1.000000	28.000000
50%	15000.500000	140000.000000	2.000000	2.000000	2.000000	34.000000
75%	22500.250000	240000.000000	2.000000	2.000000	2.000000	41.000000
max	30000.000000	1000000.000000	2.000000	6.000000	3.000000	79.000000

8 rows × 25 columns

2.5 Analisando resultados

A seguir tens a análise completa pelo "ProfileReport".

In [13]:

```
profile = ProfileReport(cartao, title='Pandas Profiling Report', explorative=True)
profile.to_notebook_iframe()
```

A Jupyter widget could not be displayed because the widget state could not be found. This could happen if the kernel storing the widget is no longer available, or if the widget state was not saved in the notebook. You may be able to create the widget by running the appropriate cells.

A Jupyter widget could not be displayed because the widget state could not be found. This could happen if the kernel storing the widget is no longer available, or if the widget state was not saved in the notebook. You may be able to create the widget by running the appropriate cells.

A Jupyter widget could not be displayed because the widget state could not be found. This could happen if the kernel storing the widget is no longer available, or if the widget state was not saved in the notebook. You may be able to create the widget by running the appropriate cells.

Referências

Yeh, IC, & Lien, CH (2009). As comparações de técnicas de mineração de dados para a precisão preditiva da probabilidade de inadimplência de clientes de cartão de crédito. Expert Systems with Applications, 36 (2), 2473-2480.