Instituto Superior de Engenharia
Politécnico de Coimbra

META 1

Pré-processamento de dados

DEPARTAMENTO DE ENGENHARIA INFORMÁTICA E DE SISTEMAS
MESTRADO EM ENGENHARIA INFORMÁTICA
UC: MACHINE LEARNING
2022/2023

ANDRÉ PROENÇA - 2016018783
ISABEL CATARINA CASTRO – 2018013160

Índice

1. Introdução 3
2. Dataset 3
    1. Análise de features 4
        Y – Target 4
        Age 4
        Job 5
        Marital 5
        Education 5
        Default 6
        Housing 6
        Loan 6
        Contact 6
        Month 7
        Day 7
        Duration 7
        Campaign 7
        Pdays 8
        Previous 8
        Poutcome 8
    2. Matrix de correlação 9
    3. Extra 9
3. Tratamento de Outliers 9
4. Transformação de Dados 10
    1. Abordagem aos dados numéricos 10
    2. Abordagem aos dados categóricos 10
    3. Definição da pipeline 10
5. Referências 10

1. Introdução

Este relatório, realizado no contexto da unidade curricular de *Machine Learning*, tem como objetivo descrever de forma clara e objetiva os métodos utilizados na análise e preparação do *dataset* escolhido.

Este *dataset* teve de ser sujeito a este pré-processamento para posteriormente ser submetido aos algoritmos *Supervised Learning* e *Unsupervised Learning*. Inicialmente, será feita uma apresentação do problema que pretendemos resolver com este *dataset*, seguido de uma análise geral dos dados e a devida exploração destes. Por fim, será descrito o processo da transformação dos dados.

2. Dataset

O *dataset* escolhido contém dados relacionados com campanhas de marketing direto de uma instituição bancária portuguesa. Estas campanhas foram realizadas através de contactos telefónicos e várias vezes foi necessário contatar o mesmo cliente, de forma, a saber se um "deposito a prazo bancário" seria subscrito ou não (sim/não). A partir da análise e processamento dos dados pretende-se prever se o cliente irá subscrever um deposito a prazo. (Bank Marketing Data Set, 2022)

Este dataset é composto por:

- Dados publicados em 2012
- Número de *features*: 17
- Número de instâncias: 45211
- Contem pelo menos 1 atributo de cada tipo (*numerical*, *categorical*, *ordinal*)

A seguinte listagem descreve os atributos deste dataset e qual o seu tipo:

- **age** (numeric)
- **job**: type of job (categorical)
- **marital**: marital status (categorical)
- **education** (categorical)
- **default**: has credit in default? (categorical)
- **housing**: has housing loan? (categorical)
- **loan**: has personal loan? (categorical)
- **contact**: contact communication type (categorical)
- **month**: last contact month of year (categorical)
- **day**: last contact day of the week (categorical)
- **duration**: last contact duration, in seconds (numeric)
- **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric)
- **previous**: number of contacts performed before this campaign and for this client (numeric)
- **poutcome**: outcome of the previous marketing campaign (categorical)
- **y**: has the client subscribed a term deposit? (binary)

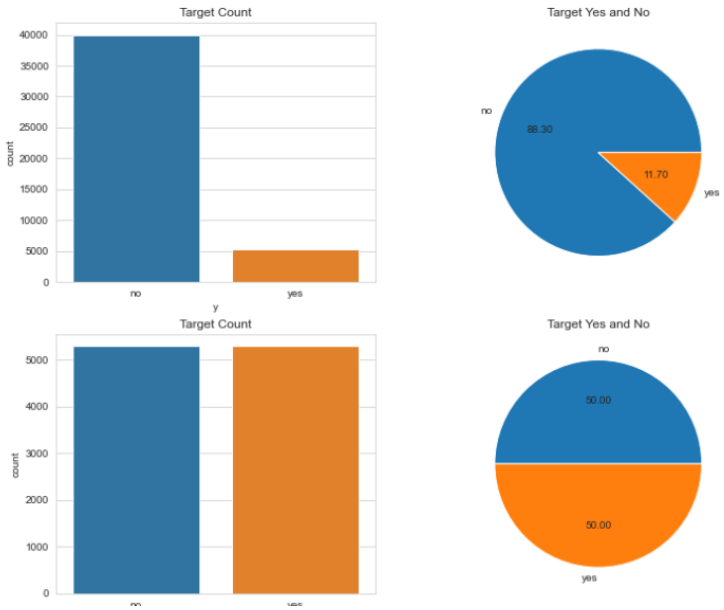
Na figura seguinte são apresentados alguns dos dados presentes, a partir do método *describe*() da biblioteca Pandas, vários sinalizadores – média, standard deviation (*std*) e quartis – para a avaliação da distribuição. Este dataset também não tem valores em falta ou NaN.

	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

1. Análise de features

Y – Target

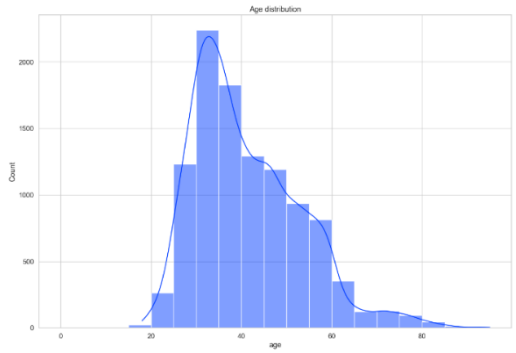
O y representa uma variável numérica binária, onde toma o valor de 0 caso o cliente não tenha subscrito um deposito bancário e toma o valor de 1 caso o cliente tenha subscrito o deposito bancário. Inicialmente ao analisarmos o nosso target, podemos concluir que o *dataset* era não balanceado, para o tornar balanceado fizemos um *undersampling* dos valores em maioria da variável y(target).



Age

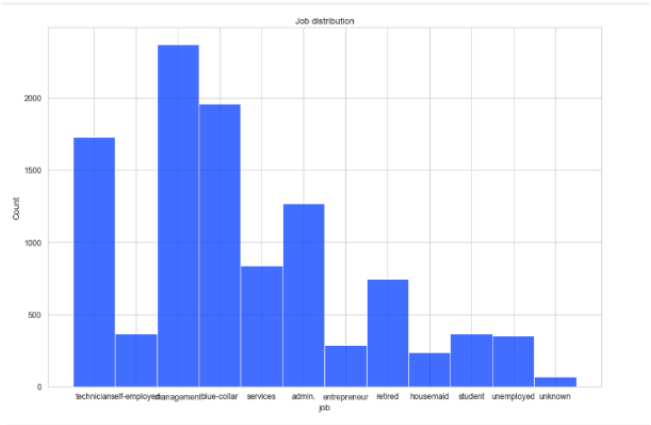
Variável numérica que representa a idade do cliente. A média de idades é 41 anos e tem um desvio padrão de 11. As seguintes figuras demonstram a distribuição de idades.

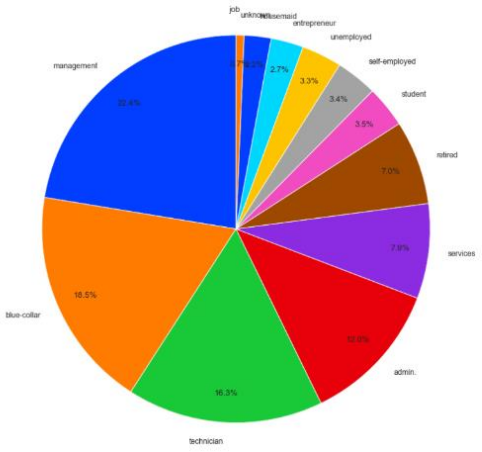
```
Mean:    41.16430327093968
Mode:    0    32
Name: age, dtype: int64
Median:  38.0
Variance:    143.335013795509
Std deviation:   11.97226017907684
Percentils (25, 50, 75):     0.00    18.0
0.25    32.0
0.50    38.0
0.75    49.0
1.00    95.0
```



Job

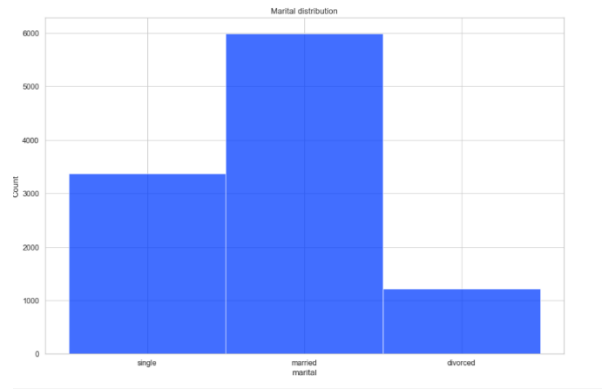
Variável categórica que representa o emprego do cliente. O emprego mais frequente é gestão seguido de blue-collar.

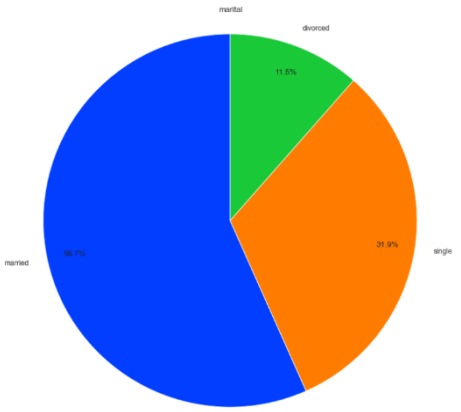




Marital

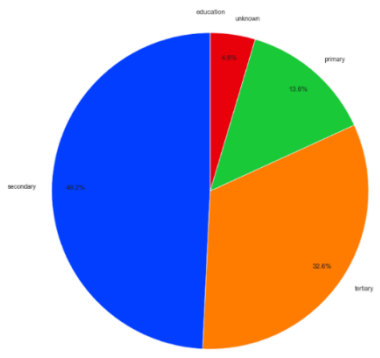
Variável categórica que representa o estado civil do cliente, o estado civil mais frequente neste *dataset* é o de casado. Cerca de 57% das pessoas são casadas, 32% solteira s e 12% são divorciadas.

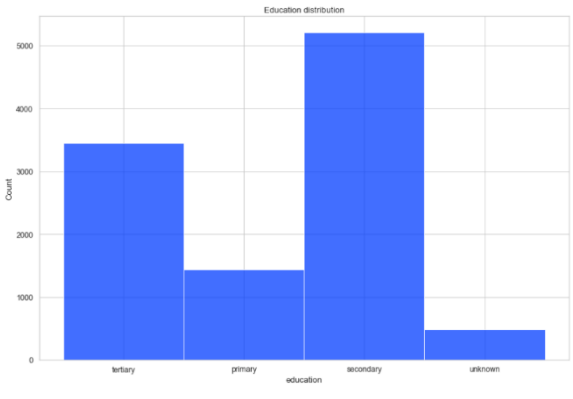




Education

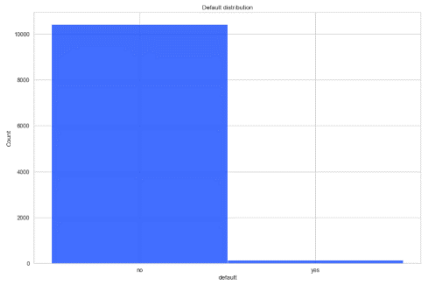
Variável categórica que representa qual as habilitações literárias do cliente, o nível de ensino com mais representação neste dataset é o de secundário. 50% tem o ensino secundário, 33% ensino básico e 14% o ensino primário.

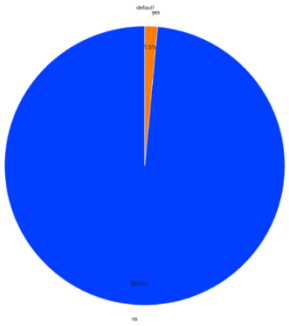




Default

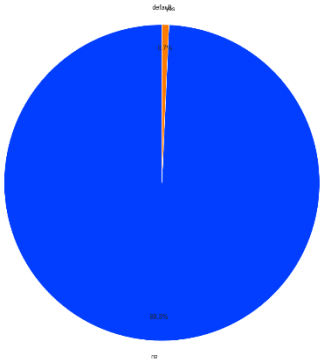
Variável categórica binária que representa se tem um crédito em atraso no banco, a maioria dos clientes analisados, não apresenta ter crédito em atraso.





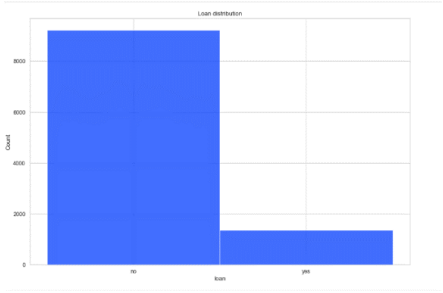
Housing

Variável categórica binária que representa se o cliente tem crédito a habitação ou não. Cerca de 99% das pessoas não tem crédito para habitação.



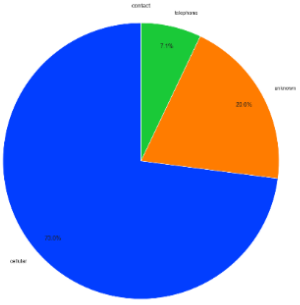
Loan

Variável categórica binária que presenta se o cliente tem crédito pessoal. 89% das pessoas não tem crédito pessoal.



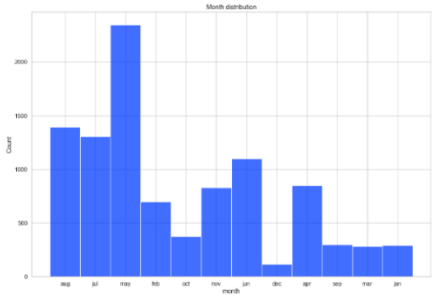
Contact

Variável categórica que representa o meio de contacto que foi feito com o cliente. 73% das chamadas foram realizadas por telemóvel, 7,1% por telefone e 20% não está reconhecido.



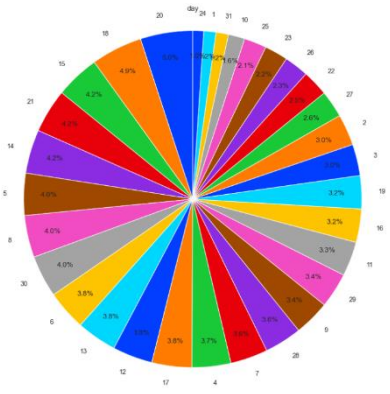
Month

Variável categórica ordinal que representa o último mês que o cliente foi contatado. O mês com mais contactos foi o mês de maio.



Day

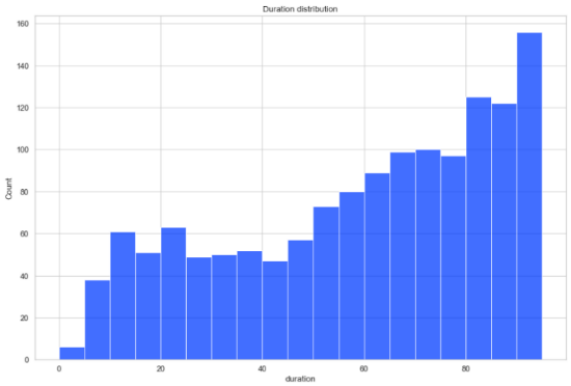
Variável categórica que representa o último dia do mês que o cliente foi contatado. O dia com mais contactos realizados foi o dia 20, com cerca de 5% de todas as chamadas realizadas.



Duration

Variável numérica que representa a duração do último contacto com o cliente, em segundos. A média de duração da chamada é 376 segundos (~ 6 min) e tem um desvio padrão de 346. As seguintes figuras demonstram a distribuição da duração.

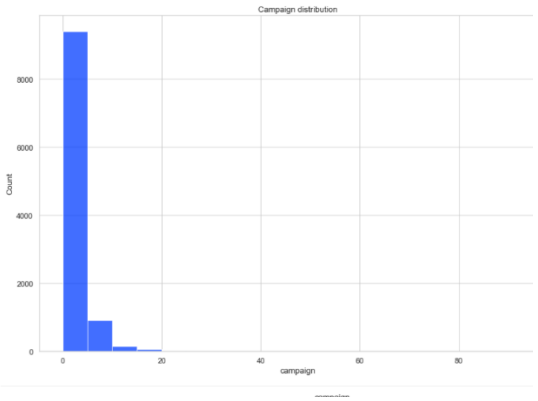
```
Mean:    376.2475988272167
Mode:    0    226
Name: duration, dtype: int64
Median:  259.0
Variance:    119743.21442168596
Std deviation:   346.03932496420975
Percentils (25, 50, 75):     0.00       2.0
0.25     144.0
0.50     259.0
0.75     503.5
1.00    3881.0
```



Campaign

Variável numérica que representa o número de contactos realizados durante a campanha para o cliente. A média é cerca de 2 contactos e tem um desvio padrão de 2,6. As seguintes figuras demonstram a distribuição.

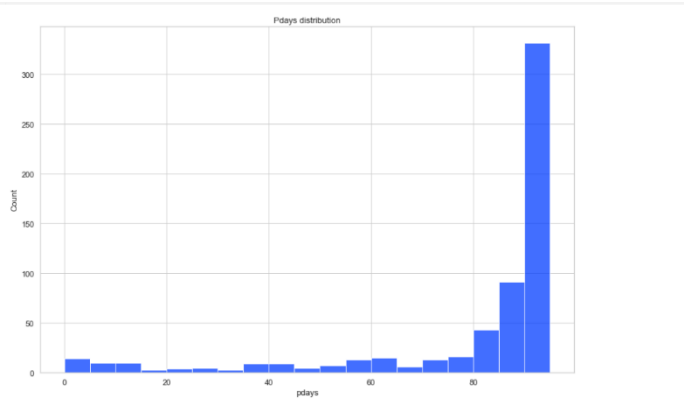
```
Mean:    2.497018457169901
Mode:    0    1
Name: campaign, dtype: int64
Median:  2.0
Variance:    6.846190465864992
Std deviation:   2.6165225903601503
Percentils (25, 50, 75):     0.00     1.0
0.25     1.0
0.50     2.0
0.75     3.0
1.00    43.0
```



Pdays

Variável numérica que representa o número de dias que passaram desde que o cliente foi contactado numa campanha prévia. A média de dias é 52 e tem um desvio padrão de 108. As seguintes figuras demonstram a distribuição.

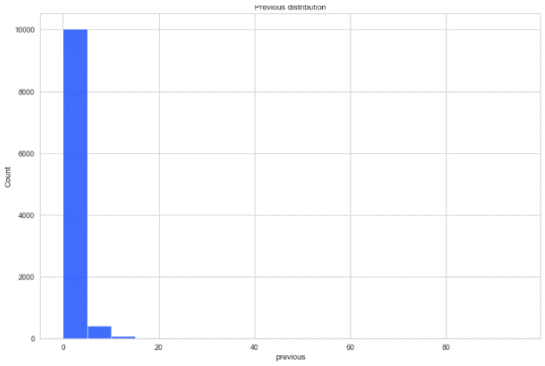
```
Mean:    51.662722453616055
Mode:    0   -1
Name: pdays, dtype: int64
Median:  -1.0
Variance:    11784.7525589422
Std deviation:   108.55760018967902
Percentils (25, 50, 75):     0.00     -1.00
0.25     -1.00
0.50     -1.00
0.75     42.25
1.00    854.00
Name: pdays, dtype: float64
```



Previous

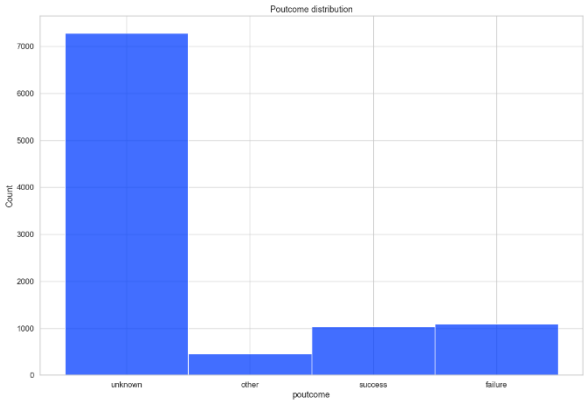
Variável numérica que representa o número de contactos realizados para o cliente antes desta campanha. A média é cerca de 1 contacto prévio e tem um desvio padrão de 2,15. As seguintes figuras demonstram a distribuição do número de contactos.

```
Mean:    0.8084059068534646
Mode:    0    0
Name: previous, dtype: int64
Median:  0.0
Variance:    4.651160991713653
Std deviation:   2.1566550469914407
Percentils (25, 50, 75):     0.00     0.0
0.25     0.0
0.50     0.0
0.75     1.0
1.00    58.0
Name: previous, dtype: float64
```



Poutcome

Variável categórica que representa o resultado da última campanha de marketing realizada. 74% é desconhecido, 11% é relativo ao falhanço e 11% a sucesso.

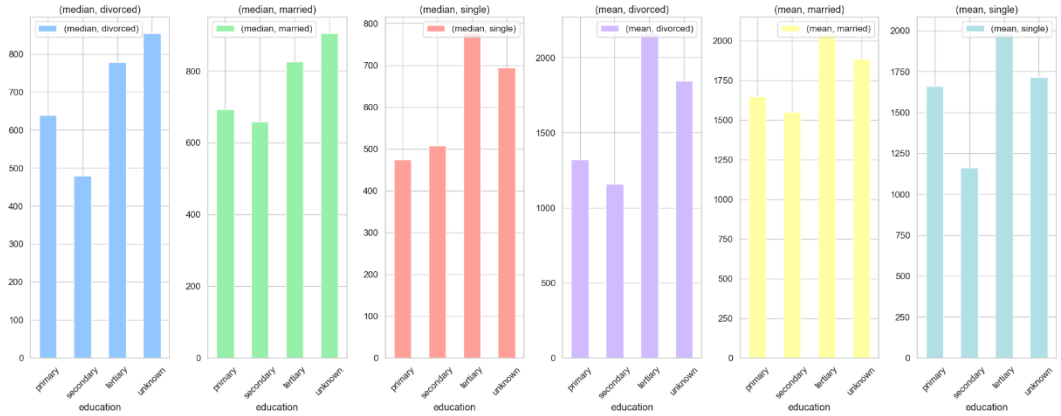


2. Matrix de correlação

Ao observar o *heatmap*, concluímos que as nossas variáveis têm pouca correlação, mas ainda assim as que mais se relacionam são os *days* e a *campaign*.

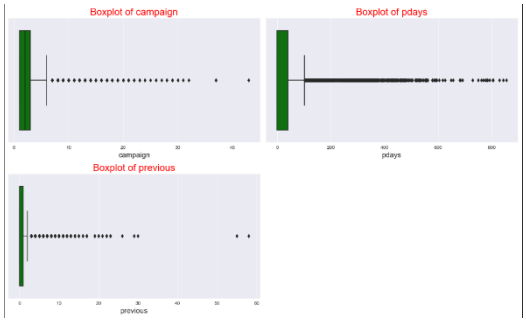
3. Extra

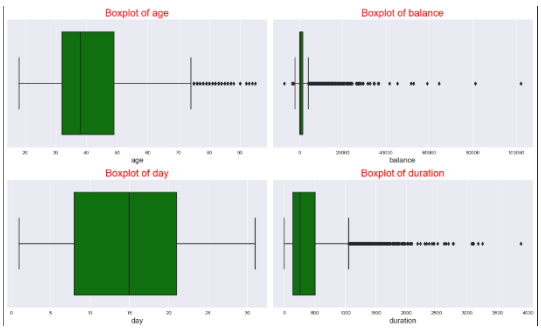
Comparar o saldo de cada cliente com as habilitações literárias e com o seu estado civil. Ao observar este gráfico, concluímos que quem tem um maior nível de habilitações literárias, em média tem um maior saldo de conta.



3. Tratamento de Outliers

Ao analisar features numéricas através de um boxplot, conseguimos perceber que algumas dessas features tem outliers, estes outliers podem prejudicar o desempenho na fase de treino dos nossos algoritmos de previsão, por esse motivo vamos remover esses outliers.





Ao observar as plot boxes das nossas variáveis numéricas chegamos a conclusão de remover os seguintes valores:

- Remover todas as linhas onde o balance onde assumia valores negativos e acima de 400000.
- Remover todas as linhas onde a duração do último contacto foi acima de 2500 dias.
- Remover todas as linhas em que o número de contactos numa campanha foi superior a 40.

- Remover a coluna de pdays, devido ao facto de não ser uma feature interessante, pois segundo uma análise pormenorizada percebemos que a grande parte dos valores são -1(cliente novo, nunca antes contatado) e que o motivo de ser –1 influencia também as colunas do *previous*, pois obviamente o contacto feito anteriormente também irá ser 0 e o *poutcomes*, que relaciona o resultado de uma campanha anterior também irá ser sempre *unknow* porque o cliente é novo e nunca tinha participado numa campanha anterior.

4. Transformação de Dados

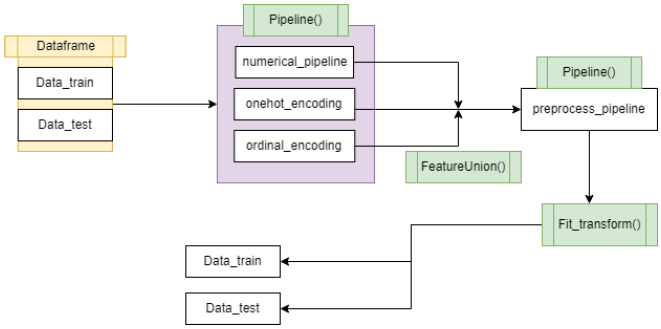
1. Abordagem aos dados numéricos

A todos os dados numéricos, foi aplicada uma normalização, através da função StandardScaler do sklearn. Não foram adicionadas novas features ao dataset pelo facto de a corelação entre as variáveis numéricas não originar features úteis para a previsão do target.

2. Abordagem aos dados categóricos

Aos dados categóricos foram aplicadas diferentes abordagens, para transformar os dados. As colunas (*default,housing,loan*) foi aplicado um *one hot encoding* que atribui 0 a um valor e 1 a outro valor. As colunas(*month,contact,poutcome,job,marital,education*) foi aplicado um *ordinal encoding* que codifica cada tipo de valor num número por ordem.

3. Definição da pipeline



5. Referências

Bank Marketing Campaign || Opening a Term Deposit. (09 de 11 de 2022). Obtido de Kaggle: https://www.kaggle.com/code/janiobachmann/bank-marketing-campaign-opening-a-term-deposit#notebook-container

Bank Marketing Data Set. (09 de 11 de 2022). Obtido de UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/bank+marketing

Random Oversampling and Undersampling for Imbalanced Classification. (09 de 11 de 2022). Obtido de Machine Learning Mastery: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/

StandardScaler, MinMaxScaler and RobustScaler techniques – ML. (09 de 11 de 2022). Obtido de geeksforgeeks: https://www.geeksforgeeks.org/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/

Under sampling. (09 de 11 de 2022). Obtido de Imbalanced-learn documentation: https://imbalanced-learn.org/stable/under_sampling.html