

Análise de Dados Exploratória & Modelos de Machine Learning

Luis Martins (2242136@iscap.ipp.pt) – 2242136
João Pinto (2240514@iscap.ipp.pt) – 2240514

Pós Graduação em Business Analytics, ISCAP – Politécnico do Porto, Portugal

Resumo

O presente trabalho contém uma análise detalhada do dataset Bank_Data tendo como consideração conceitos de Análise de Dados Exploratória (ADE) e Machine Learning.

Durante a fase de ADE vamos analisar a estrutura dos dados, a sua composição e a interpretação das variáveis. Iremos tirar conclusões relativamente às correlações entre as respetivas variáveis e preparar todos os dados para a sua aplicação nos modelos de Machine Learning.

Relativamente aos modelos de machine learning, vamos analisar a aplicabilidade dos dados em todos os modelos para que seja possível ter uma comparação e perceber qual o modelo que melhor se adapta aos nossos dados.

Introdução

A análise de dados exploratória (ADE) e os modelos de machine learning são ferramentas essenciais para extrair insights valiosos de grandes volumes de dados, permitindo a tomada de decisões informadas em diversos setores. O presente relatório tem como objetivo analisar o dataset Bank_Data, que contém informações sobre clientes bancários, com o intuito de identificar padrões e características que influenciam a probabilidade de um cliente realizar depósitos a prazo.

Através da aplicação de técnicas de ADE, exploraremos a distribuição e as relações entre as variáveis do dataset, identificando possíveis outliers, valores ausentes e correlações significativas. Posteriormente, serão implementados diversos modelos de machine learning, como Logistic Regression, Decision Trees, Random Forest e Support Vector Machines (SVM), para prever o comportamento dos clientes e avaliar a eficácia de cada modelo.

Este estudo visa não só otimizar a compreensão sobre os fatores que impactam as decisões financeiras dos clientes, mas também fornecer uma análise quantitativa da performance dos modelos de machine learning aplicados.

Análise de Dados Exploratória

Importação de bibliotecas

Para garantir uma abordagem estruturada e eficiente na realização desta análise, foram importadas diversas bibliotecas que proporcionam funcionalidades adicionais, permitindo a manipulação de dados, a visualização de informações e a implementação de modelos de Machine Learning. Estas bibliotecas desempenham um papel fundamental na simplificação do código e na otimização dos processos analíticos.

Importação do Dataset

O primeiro passo desta análise consistiu na importação do conjunto de dados Bank_Data, que será objeto de estudo ao longo do relatório. O ficheiro encontra-se no formato CSV (Comma-Separated Values), o que permite uma fácil leitura e manipulação dos dados.

A importação deste dataset é essencial para iniciar o seu tratamento e explorar as suas principais características antes da aplicação de técnicas analíticas e modelos preditivos.

```
[ ] file_name = "bank_data.csv"

if os.path.exists(file_name):
    df = pd.read_csv(file_name, sep=";")
else:
    uploaded = files.upload()
    df = pd.read_csv(io.BytesIO(uploaded[file_name]), sep=";")
```

Figura 1 - Importação Dataset

Durante a leitura inicial da estrutura do dataset, verificou-se que os valores estavam separados por ponto e vírgula (;) em vez de vírgula (,) - o delimitador padrão em ficheiros CSV.

Para garantir uma importação correta dos dados, foi necessário especificar explicitamente o delimitador no código, assegurando assim a correta segmentação das colunas e a integridade da informação analisada.

Informações Gerais do Dataframe

Após a importação dos dados, iniciamos uma análise genérica ao nosso dataframe.

Inicialmente apuramos o número de linhas e colunas existentes no dataframe:

```
[ ] print(f"Number of rows: {df.shape[0]}")
    print(f"Number of columns: {df.shape[1]}")

    print("\n\n")
    print("Columns info: ")
    df.info()

Number of rows: 41188
Number of columns: 21
```

Figura 2 - Informação geral do DataFrame

Filas: 41188
Colunas: 21

```
Columns info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype
---  -
0   age                   41188 non-null  int64
1   job                   41188 non-null  object
2   marital               41188 non-null  object
3   education             41188 non-null  object
4   default               41188 non-null  object
5   housing               41188 non-null  object
6   loan                  41188 non-null  object
7   contact               41188 non-null  object
8   month                 41188 non-null  object
9   day_of_week           41188 non-null  object
10  duration              41188 non-null  int64
11  campaign              41188 non-null  int64
12  pdays                 41188 non-null  int64
13  previous              41188 non-null  int64
14  poutcome              41188 non-null  object
15  emp.var.rate          41188 non-null  float64
16  cons.price.idx        41188 non-null  float64
17  cons.conf.idx         41188 non-null  float64
18  euribor3m             41188 non-null  float64
19  nr.employed           41188 non-null  float64
20  y                     41188 non-null  object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

Figura 3 - Tipo de Variáveis

Classificação das Variáveis

Conforme mencionado anteriormente, o dataframe em análise é composto por 21 colunas, que representam as variáveis a considerar ao longo deste relatório. Estas variáveis podem ser agrupadas em duas categorias principais: numéricas e categóricas.

As variáveis numéricas são aquelas que representam valores quantitativos e permitem a realização de operações matemáticas. No presente dataset, as seguintes variáveis foram classificadas como numéricas:

- age, duration, campaign, pdays, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m e nr.employed.

Por sua vez, as variáveis categóricas representam atributos qualitativos e classificam os dados em diferentes grupos. As variáveis categóricas identificadas são:

- job, marital, education, default, housing, loan,
- contact, month, day_of_week, poutcome e y.

Análise Inicial dos Valores Únicos

Após a categorização das variáveis, procedeu-se à análise dos valores únicos presentes em cada uma delas. Constatou-se que a variável duration apresenta o maior número de valores distintos, totalizando 1.544 observações únicas.

Adicionalmente, verificou-se que nenhuma das variáveis contém exclusivamente valores nulos ou zero, pelo que, nesta fase, não se justifica a remoção de nenhuma coluna da análise.

Identificação de Valores Duplicados

Após a verificação da diversidade de valores em cada variável, foi realizada uma análise para identificar dados duplicados no dataset.

```
[ ] df.duplicated().sum()  
↔ 12
```

Figura 4 - Identificação Duplicados

Eliminação de Valores Duplicados

A análise revelou a existência de 12 registos duplicados no dataframe. Embora o conjunto de dados não contenha informações pessoais detalhadas, a sua estrutura sugere que cada registo corresponde a um cliente único.

Ao examinar as variáveis em maior detalhe, verificou-se que algumas colunas, como duration (duração da chamada), age (idade), job (profissão) e marital (estado civil), apresentavam valores exatamente iguais nos registos duplicados.

Dado que a presença destes registos repetidos não acrescenta valor à análise e pode enviesar os resultados, optou-se por proceder à eliminação das 12 linhas duplicadas. Após esta operação, o dataframe passou a conter 41.176 registos.

Análise da Percentagem e Distribuição de Dados em Falta

A análise preliminar do dataset revelou que os valores em falta são, de forma geral, representados pelo termo “unknown”. Para facilitar a manipulação e o tratamento destes dados, optou-se por substituí-los por NaN (Not a Number), uma vez que as bibliotecas da Python, nomeadamente o pandas, oferecem funcionalidades mais eficientes para lidar com valores nulos.

Além disso, com base nas orientações fornecidas durante as aulas, foi possível identificar casos específicos de valores que podem ser considerados missing values:

- *pdays*: Esta variável apresenta determinados valores que podem ser interpretados como dados em falta. Especificamente, o valor 999 pode ser entendido como um missing value, uma vez que não representa um “não contacto” válido, dado o contexto da variável e a presença de outros valores numéricos legítimos. Assim, procedeu-se à substituição de 999 por NaN.
- *poutcome*: Esta variável categórica, que indica o resultado de campanhas anteriores, também contém valores em falta. O valor “nonexistent” pode ser interpretado como um missing value e, por conseguinte, foi igualmente substituído por NaN.

Após esta normalização, procedeu-se à análise da representatividade dos valores nulos em cada variável, permitindo compreender o impacto dos dados em falta na qualidade do dataset e orientar as estratégias de tratamento adequadas.

	Null Count	Null Percentage
age	0	0.000000
job	330	0.801438
marital	80	0.194288
education	1730	4.201477
default	8596	20.876239
housing	990	2.404313
loan	990	2.404313
contact	0	0.000000
month	0	0.000000
day_of_week	0	0.000000
duration	0	0.000000
campaign	0	0.000000
pdays	39661	96.320672
previous	0	0.000000
poutcome	35551	86.339130
emp.var.rate	0	0.000000
cons.price.idx	0	0.000000
cons.conf.idx	0	0.000000
euribor3m	0	0.000000
nr.employed	0	0.000000
y	0	0.000000

Figura 5 - Identificação Dados NaN

Análise de Valores em Falta

Com base nos resultados apresentados na figura acima, as seguintes conclusões podem ser tiradas relativamente às variáveis com dados em falta:

- **Job (0.80%):** A percentagem de valores em falta é muito reduzida, o que torna esta variável uma boa candidata à preservação. A estratégia sugerida é substituir os valores NaN pela moda (valor mais frequente) da variável.
- **Marital (0.19%):** Tal como a variável Job, a percentagem de valores NaN é extremamente baixa. Portanto, recomenda-se o mesmo tratamento, substituindo os valores em falta pela moda da variável.
- **Education (4.20%):** Embora apresente uma percentagem mais elevada de valores em falta, a variável ainda mantém uma quantidade significativa de dados válidos. Neste caso, poderá ser adequado substituir os valores NaN pela média ou pela moda, dependendo da análise mais detalhada.
- **Default (20.88%):** A variável Default apresenta uma percentagem considerável de valores NaN. Devido à magnitude desta falta de dados, é possível que esta variável precise ser removida do dataframe, uma vez que pode comprometer a análise.
- **Housing e Loan (2.40%):** Ambas as variáveis apresentam uma percentagem reduzida de valores em falta. Como tal, recomenda-se a substituição dos valores NaN, uma vez que a sua taxa de dados em falta não compromete substancialmente a integridade das variáveis.
- **Pdays (96.32%):** A variável Pdays contém uma percentagem altíssima de valores NaN, considerando-se que o valor 999 representa dados em falta. Devido a essa elevada proporção, esta variável é uma forte candidata a ser removida do dataframe, visto que não oferece informações úteis para a análise.
- **Poutcome (86.34%):** A variável Poutcome também apresenta uma grande quantidade de valores NaN. Assim como Pdays, é uma candidata a ser removida, dada a elevada percentagem de dados em falta.

Conclusões Finais

Com base na análise da distribuição de valores em falta, as variáveis Default, Pdays e Poutcome são as principais candidatas a serem removidas do dataframe. No caso da variável Default, será necessário realizar uma análise mais aprofundada para avaliar o impacto da sua remoção na análise global. Para Pdays e Poutcome, a análise bivariada será realizada para corroborar a decisão de remoção, que já se mostra clara com base na percentagem de dados em falta.

As restantes variáveis não apresentam valores NaN, pelo que não necessitam de tratamentos adicionais.

Explorar Características de Dados

Análise Univariada

Variáveis Numéricas

	age	duration	campaign	pdays	previous \
count	41176.000000	41176.000000	41176.000000	1515.000000	41176.000000
mean	40.023800	258.315815	2.567879	6.014521	0.173013
std	10.420680	259.305321	2.770318	3.824906	0.494964
min	17.000000	0.000000	1.000000	0.000000	0.000000
5%	26.000000	36.000000	1.000000	2.000000	0.000000
25%	32.000000	102.000000	1.000000	3.000000	0.000000
50%	38.000000	180.000000	2.000000	6.000000	0.000000
75%	47.000000	319.000000	3.000000	7.000000	0.000000
95%	58.000000	753.000000	7.000000	14.000000	1.000000
max	98.000000	4918.000000	56.000000	27.000000	7.000000
skew	0.784560	3.262808	4.762044	1.458564	3.831396
kurt	0.791113	20.243771	36.971857	2.564562	20.102164

	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000
mean	0.081922	93.575720	-40.502863	3.621293	5167.034870
std	1.570883	0.578839	4.627860	1.734437	72.251364
min	-3.400000	92.201000	-50.800000	0.634000	4963.600000
5%	-2.900000	92.713000	-47.100000	0.797000	5017.500000
25%	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	1.400000	93.994000	-36.400000	4.961000	5228.100000
95%	1.400000	94.465000	-33.600000	4.966000	5228.100000
max	1.400000	94.767000	-26.900000	5.045000	5228.100000
skew	-0.724061	-0.230853	0.302876	-0.709194	-1.044317
kurt	-1.062698	-0.829851	-0.359097	-1.406791	-0.003540

Figura 6 - Características Variáveis

Age

- Média: 40.02
- Mediana: 38.00
- Desvio-padrão (std): 10.42

A distribuição da variável Age apresenta uma média bastante próxima da mediana, indicando uma distribuição aproximadamente normal, com uma leve tendência à direita (média > mediana). O desvio-padrão é moderado, o que sugere uma variação moderada na idade dos clientes. A kurtosis relativamente baixa (inferior a 1) reforça a ideia de uma distribuição mais "normal".

Conclusão: A maior concentração de clientes situa-se entre as idades de 17 e 38 anos, sendo que a distribuição de idades para clientes com mais de 38 anos é mais dispersa. O número de outliers é reduzido e não exerce grande impacto na análise.

Duration

- Média: 258.32
- Mediana: 180.00
- Desvio-padrão (std): 259.28

A variável Duration apresenta uma média consideravelmente superior à mediana, indicando uma distribuição altamente assimétrica com forte tendência à direita, o que sugere a presença de outliers

extremos (confirmado pelo valor de skewness de 3.62). O desvio-padrão é elevado, semelhante ao valor da média, indicando uma grande variação nas durações das chamadas. A kurtosis elevada (muito superior a 3) mostra que há uma concentração significativa de valores próximos da média.

Conclusão: A grande maioria das chamadas tem uma duração inferior a 3 minutos, com alguns outliers de chamadas muito longas (valores extremos que distorcem a média).

Campaign

- Média: 2.568
- Mediana: 2.000
- Desvio-padrão (std): 2.770

A variável Campaign apresenta uma média superior à mediana, o que indica uma distribuição assimétrica à direita, com uma tendência a valores extremos (confirmado pelo skewness de 4.762). O desvio-padrão superior à média sugere uma grande variação nos dados, e a kurtosis de 36.97 revela que muitos valores se concentram perto da média, embora haja valores mais extremos.

Conclusão: A maioria dos clientes foi contactada até 2 vezes, mas há um número considerável de clientes que foram contactados múltiplas vezes, com alguns casos extremos que influenciam significativamente a média.

Pdays

Devido ao reduzido número de valores válidos (apenas 1.515), a análise desta variável não proporciona insights significativos. Não foi possível tirar conclusões relevantes para a análise global.

Previous

- Média: 0.173
- Mediana: 0.000
- Desvio-padrão (std): 0.495

A variável Previous apresenta uma média ligeiramente superior à mediana, sugerindo assimetria positiva, corroborada pelo skewness de 3.831. O elevado desvio-padrão, comparado com a média e mediana, indica uma grande variação nos valores. A kurtosis de 20.102 confirma que a maioria dos clientes não foi contactada em campanhas anteriores, mas ainda assim há um pequeno número de clientes contactados múltiplas vezes, incluindo casos extremos de 7 contactos.

Conclusão: A maioria dos clientes não foi contactada após a última campanha, mas há uma pequena quantidade de clientes com múltiplos contactos.

Emp.var.rate

- Média: 0.0819%
- Mediana: 1.1000%

A variável apresenta uma mediana ligeiramente superior à média, indicando uma assimetria à esquerda (como podemos confirmar pelo valor negativo de skewness). Analisando o valor de Kurtosis concluímos que o mesmo é negativo, indicando que a curva é mais achatada que uma distribuição normal e, como tal, os valores estão bastante dispersos.

No contexto, os clientes forma contactados em diferentes momentos do ano e como tal é natural que existam diferentes taxas de empregabilidade. Como as taxas não são calculadas diariamente, mas sim por

um período mais longo, observa-se uma curva ondulada com alguns picos distantes nos diferentes valores que a taxa de empregabilidade tomou durante o ano.

Cons.price.idx

- Média: 93.5757
- Mediana: 93.7490

A variável Cons.price.idx apresenta uma média ligeiramente inferior à mediana, o que pode indicar uma distribuição assimétrica à esquerda. O gráfico não demonstra uma kurtosis estruturada, sugerindo que os dados não apresentam grandes concentrações nem grandes desvios.

Cons.conf.idx

- Média: -40.502
- Mediana: -41.800

A variável Cons.conf.idx reflete uma queda da confiança dos consumidores, com valores negativos indicando uma redução na confiança económica e no consumo. O gráfico também não mostra uma kurtosis estruturada.

Euribor3m

- Média: 3.6213%
- Mediana: 4.8570%

A Euribor3m apresenta uma mediana superior à média, indicando que houve períodos de maior crescimento na taxa de juros, o que impactou negativamente o consumo e outros indicadores económicos. O gráfico não apresenta uma kurtosis estruturada.

Nr.employed

- Média: 5167.03
- Mediana: 5191.00

O número de empregados no período analisado variou significativamente, com um desvio-padrão de 72.25, refletindo o impacto de outros indicadores económicos, como o índice de confiança. O gráfico não mostra uma kurtosis estruturada, sugerindo uma distribuição mais equilibrada, sem grandes concentrações ou desvios.

Análise das Variáveis Categóricas

Após a análise das variáveis numéricas, procedemos à análise das variáveis categóricas, utilizando countplots para melhor visualização e interpretação. A análise individualizada destas variáveis permite compreender a distribuição e o peso de cada categoria dentro do conjunto de dados.

Job

A variável Job apresenta uma distribuição bastante diversificada nas profissões dos clientes. As profissões mais predominantes incluem:

- Administrador

- Funções de atividades não associadas a trabalhos de escritório
- Técnicos

Esta distribuição sugere que a carteira de clientes é composta por uma variedade de ocupações, com maior destaque para funções administrativas e técnicas.

Education

A variável Education indica que a maioria dos clientes possui licenciatura ou ensino secundário como nível de escolaridade. Essa informação é útil para entender o perfil educacional dos clientes e pode ser um fator importante em análises subsequentes, como a propensão de aceitação de produtos financeiros.

Marital

Na variável Marital, a categoria casado é, de longe, a mais predominante. Isto indica que a maioria dos clientes do dataset está em estado civil casado, o que pode ser relevante ao se analisar o comportamento de consumo e a propensão para determinadas ofertas de produtos.

Housing

A variável Housing mostra que, embora os valores “yes” e “no” estejam relativamente próximos, a maioria dos clientes possui crédito habitação. Esta informação sugere que a oferta de crédito habitação é um produto bancário importante para a instituição, com muitos clientes já envolvidos nesse tipo de crédito.

Loan

Em contraste com a variável Housing, a variável Loan mostra que a maior parte dos clientes não possui crédito ao consumo. Este dado sugere uma possível oportunidade para ofertas de crédito ao consumo, visto que a maioria dos clientes não está envolvida com este produto financeiro.

Default

A variável Default indica que, apesar de um número significativo de registos em “unknown”, a maioria dos clientes não tem registo de incumprimento (default) nos seus créditos. Isso sugere que a maioria dos clientes é financeiramente responsável, o que é relevante para a avaliação de risco em novas ofertas de crédito.

Contact

Analisando a variável Contact, não se observa uma tendência clara em relação aos dias da semana. A distribuição é equilibrada, indicando que os clientes são contactados de forma quase uniforme durante os dias da semana.

Além disso, a maioria dos clientes tem contacto telefónico móvel, o que facilita a realização de campanhas de marketing e ofertas de produtos diretamente através do telefone.

Month

A variável Month mostra que a maioria dos contactos ocorreu no mês de Maio. Este é um dado relevante para as campanhas de marketing e pode indicar um pico de atividades promocionais ou campanhas especiais durante esse período.

Poutcome

A variável Poutcome, que indica o resultado da campanha anterior, não fornece conclusões claras devido ao número elevado de dados em falta. A falta de dados nesta variável impede uma análise mais aprofundada sobre a eficácia das campanhas anteriores.

Y (Target)

Finalmente, a variável Y, que representa a aceitação do produto bancário pelos clientes, mostra que a maioria dos clientes não aceitou a subscrição do produto. Esse é um dado relevante para entender a eficácia da campanha de marketing e a receptividade dos clientes aos produtos oferecidos pelo banco.

Conclusão Geral

A análise das variáveis categóricas oferece uma visão interessante sobre o perfil dos clientes, os seus comportamentos financeiros e a eficácia das campanhas passadas. É possível observar que uma parte significativa dos clientes tem crédito habitação, mas não crédito ao consumo. Além disso, a maioria não apresentou incumprimento (default) e foi contactada de forma equilibrada ao longo da semana. No entanto, a elevada percentagem de valores em falta em algumas variáveis, como Poutcome, limita a análise das campanhas anteriores.

Esses resultados são fundamentais para tomar decisões sobre estratégias de marketing futuras e para identificar oportunidades de melhoria na oferta de produtos financeiros aos clientes.

Análise Bivariada

Numérica vs Numérica

Na análise bivariada das variáveis numéricas, decidimos não utilizar o coeficiente de correlação de Pearson, dado que as características das variáveis numéricas, especialmente o elevado número de outliers em algumas delas, inviabilizam a aplicação deste método. Apenas a variável Age apresenta um número reduzido de outliers, mas mesmo assim, não consideramos apropriado usar Pearson para avaliar a correlação.

Em vez disso, optamos por realizar a análise bivariada utilizando as variáveis emp.var.rate (taxa de variação do emprego) e nr.employed (número de empregados). Estas duas variáveis são mais estáveis em relação à presença de outliers, o que facilita a análise.

Variáveis Seleccionadas:

- *emp.var.rate (Taxa de variação do emprego)*: Representa a variação percentual no nível de emprego durante o período de análise.
- *nr.employed (Número de empregados)*: Refere-se ao número total de empregados na economia durante o período de análise.

```
[ ] #Is there are relation between employment rate and number of employee people?
spearman_corr, _ = spearmanr(df["emp.var.rate"], df["nr.employed"]) # relação monotona quase perfeita
kendall_corr, _ = kendalltau(df["emp.var.rate"], df["nr.employed"]) # forte concordancia nos ranks
distance_corr = correlation(df["emp.var.rate"], df["nr.employed"]) # dependencia positiva (ainda que nao perfeita)
mi = mutual_info_regression(df[["emp.var.rate"]].values, df["nr.employed"]) # existe alguma partilha/overlap de informação entre as 2 colunas

print("Spearman: ", spearman_corr)
print("Kendall: ", kendall_corr)
print("Distance: ", distance_corr)
print("Mutual Info: ", mi)

Spearman: 0.9446874287648435
Kendall: 0.8451002483178448
Distance: 0.09305050148203986
Mutual Info: [1.62339249]
```

Figura 7 - Correlação variáveis numéricas

Na análise bivariada das variáveis emp.var.rate (taxa de variação do emprego) e nr.employed (número de empregados), foram utilizadas diversas métricas de correlação para avaliar a relação entre as duas variáveis. Os resultados obtidos com os diferentes métodos de correlação indicam uma forte dependência positiva entre elas.

Correlação de Spearman

A correlação de Spearman entre as duas variáveis resultou em valores muito próximos de 1, indicando uma correlação monótona quase perfeita. Isso sugere que à medida que uma variável aumenta, a outra também tende a aumentar de forma consistente. A correlação monótona implica que não há necessidade de uma relação linear para que as variáveis se movam juntas, apenas uma tendência constante de aumento ou diminuição.

Kendall's Tau

O valor obtido pela correlação de Kendall's Tau foi de 0.845, o que indica uma forte concordância nos rankings das variáveis. Com um valor tão elevado, podemos concluir que as variáveis têm uma relação estreita, e a mudança na posição de uma variável é frequentemente acompanhada pela mudança na posição da outra.

Correlação de Distância

Na correlação de distância, o valor próximo de 0 reforça os resultados anteriores, indicando que a dependência entre as variáveis não é influenciada por distorções ou variações não explicadas pelas outras métricas. Esse valor próximo de zero implica uma relação quase perfeita entre as variáveis, sem grande discrepância nas variações observadas.

Informação Mútua

O valor de informação mútua obtido foi 1.62, que sugere uma forte dependência entre as duas variáveis. Quanto mais afastado de 0 for o valor da informação mútua, maior a dependência entre as variáveis. Este valor confirma que as variáveis possuem uma forte relação, indicando que o aumento ou diminuição de uma delas está intimamente ligado ao comportamento da outra.

Em resumo, os resultados obtidos com diferentes métodos de correlação (Spearman, Kendall's Tau, correlação de distância e informação mútua) indicam que as variáveis emp.var.rate e nr.employed têm uma relação de dependência positiva quase perfeita. O aumento de uma das variáveis tende a ser acompanhado pelo aumento da outra, e essa relação é bastante consistente através dos diferentes métodos de análise.

Esses resultados são importantes para inferir que qualquer mudança nas condições de emprego, refletida na taxa de variação do emprego, provavelmente terá um impacto direto no número total de empregados na economia. Essa dependência entre as variáveis pode ser útil para modelagem preditiva, análise de risco ou tomada de decisões estratégicas em contextos económicos e de negócios.

Testes Estatísticos para Análise de Normalidade

Considerando o contexto do dataset, que possui um número considerável de registos (mais de 41.000 linhas), o teste Shapiro-Wilk não seria apropriado devido à sua limitação em grandes amostras. Portanto, optamos por aplicar o teste D'Agostino K2, que é mais adequado para amostras grandes e permite verificar a normalidade dos dados.

Variável "age"

Com base nas análises exploratórias realizadas previamente, a variável age foi identificada como uma candidata a seguir uma distribuição aproximadamente normal. A distribuição da variável foi observada como tendo uma ligeira tendência à direita, mas sem outliers extremos, o que sugere que os dados poderiam se aproximar de uma normalidade.

Teste D'Agostino K2

O teste D'Agostino K2 é utilizado para testar a normalidade dos dados com base em uma combinação de skewness (assimetria) e kurtosis (curtose). O teste avalia se os dados seguem uma distribuição normal, considerando essas duas métricas.

O resultado deste teste para a variável age indicaria se a distribuição dessa variável é significativamente diferente de uma normal. Se o p-valor for baixo (normalmente abaixo de 0.05), podemos rejeitar a hipótese nula de normalidade e concluir que a variável não segue uma distribuição normal. Caso contrário, com um p-valor alto, podemos não rejeitar a hipótese de que a variável segue uma distribuição normal.

```
p_value = normaltest(df['age'])
print(p_value)
```

NormaltestResult(statistic=3902.5978596235364, pvalue=0.0)

Figura 8 - Teste p_value

Contudo, podemos observar que, mesmo para esta variável, o valor de p_value é 0, devido ao facto de a distribuição normal nos valores ser muito ténue.

Numérica vs Categórica

```
[ ] #Is there are relation between pdays and y?
le = LabelEncoder()
numeric_binary_y = le.fit_transform(df['y'])

biseriatal_stat, biserial_p_value = pointbiseriatalr(numeric_binary_y, df["pdays"])
if (biseriatal_p_value < 0.05):
    print("As variáveis estão correlacionadas")
else:
    print("As variáveis não estão correlacionadas")

mi = mutual_info_classif(df[["pdays"]], df["y"])

print(f"Biseriatal Stat: {biseriatal_stat}, P-Value: {biseriatal_p_value}")
print("Mutual Info: ", mi)
```

As variáveis estão correlacionadas
Biseriatal Stat: -0.3249475863855851, P-Value: 0.0
Mutual Info: [0.03902582]

Figura 9 - Relação variáveis Numéricas vs Categóricas

Tendo em consideração a variável `pdays` e `Y`, verificamos que, embora o teste tenha sido significativo, mostrando alguma correlação negativa entre as variáveis, a informação mútua é muito baixa, o que nos permite concluir que a variável `pdays` não contém uma quantidade relevante de informação necessária para a previsão. Uma vez que o número de valores nulos desta variável é de quase 100%, é uma operação muito arriscada tentar preencher os valores desta mesma variável. Concluimos que, tal como analisado no início do relatório, faz sentido remover esta coluna.

```
df.drop(columns=['pdays'], inplace=True)
```

Figura 10 - Eliminação coluna `pdays`

Por sua vez, utilizando as variáveis `Y` e `euribor3m` verificamos o seguinte resultado:

```
As medianas são diferentes entre categorias
Kruskal Stat: 2930.314757187436, P-Value: 0.0
Biserial Stat: -0.30774039558468924, P-Value: 0.0
Mutual Info: [0.07433101]
```

Figura 11 - Relação Variáveis `Y` e `Euribor3m`

Este resultado demonstra que, no teste de Kruskal-Wallis, como o `p_value` é 0, rejeitamos a hipótese nula e concluímos que há diferenças significativas entre as medianas das categorias.

Na correlação biserial de rank, como o `p_value` = 0, a correlação é estatisticamente significativa. O valor negativo de (-0.3077) indica uma relação inversa, ou seja, à medida que uma categoria aumenta, a variável contínua tende a diminuir.

Usando a informação mútua, perante o valor de 0.0743, verificamos que existe alguma relação, mas como o valor é reduzido, a dependência não é muito forte.

Adicionalmente, fazendo uma análise das variáveis `duration` e `Y`, usando o coeficiente de correlação, verificamos que existe uma leve relação positiva entre o valor das duas variáveis, sugerindo que chamadas com maior duração tendem a contribuir para uma adesão ao depósito a prazo.

```
[ ] stat, p_value = pointbiserialr(numeric_binary_y, df["duration"])
    print(f"Coeficiente de correlção: {stat:.4f}, +-value: {p_value:.4f}")

Coeficiente de correlção: 0.4053, +-value: 0.0000
```

Figura 12 - Coeficiente Correlação

Categórica vs Categórica

Analisando a relação entre as diferentes variáveis categóricas e a variável a estimar `Y`, conseguimos tirar algumas conclusões interessantes sobre os dados:

- Pela análise do `p_value` do teste qui-quadrado (`chi2`), nenhuma das variáveis categóricas pode ser considerada independente da variável `Y`;

- O valor de informação mútua (mutual information) revela que nenhuma das variáveis possui informação particularmente relevante para o cálculo ou previsão do valor de Y, sobressaindo-se ainda assim o mês, a profissão e a forma de contacto como as variáveis categóricas mais correlacionadas com Y;
- A variável poutcome, apesar de apresentar uma ligeira correlação com Y (ainda que com uma informação mútua marginal de -0.03), não deve ser considerada válida nesta análise devido à percentagem muito elevada de valores desconhecidos (~86%). Esta alta taxa de valores em falta impede que possamos preenchê-los de forma fidedigna, levando a que esta variável possa ser excluída. Caso a percentagem de valores em falta fosse mais baixa, poderia valer a pena tentar perceber se esta variável teria uma correlação forte com alguma outra variável, permitindo assim uma tentativa mais realista de imputação dos valores em falta.
- Com um número de missing values perto dos 90%, qualquer tentativa de preenchimento desses valores seria muito pouco realista.

```
df.drop(columns=['poutcome'], inplace=True)
```

Figura 13 - Eliminação coluna poutcome

Explorando a correlação entre as variáveis, concluímos que as variáveis euribor3m, nr.employed e emp.var.rate apresentam uma correlação bastante elevada entre si.

É também relevante destacar a relação entre o cons.price.idx e os indicadores macroeconómicos mencionados acima, ainda que essa correlação seja mais ténue.

Todas as outras correlações observadas são significativamente mais fracas, indicando uma menor dependência entre as restantes variáveis do dataset.

Tendo em conta a correlação das variáveis mencionadas anteriormente, realizamos uma análise mais detalhada para obter uma informação mais clara das relações entre essas variáveis e Y.

Observações principais:

- emp.var.rate apresenta uma distribuição com múltiplos picos, sugerindo a existência de períodos distintos na economia.
- euribor3m tem valores fortemente concentrados em pontos específicos, indicando possíveis momentos-chave na variação desta taxa.
- nr.employed mostra picos bem definidos, sugerindo que a recolha destes dados pode ter sido feita de forma programada ou em momentos específicos.
- cons.price.idx apresenta uma distribuição com vários picos, refletindo variações na inflação ao longo do tempo.

Relativamente à variável Y, verificámos que a resposta positiva à subscrição do produto bancário está mais concentrada em determinados valores de euribor3m e emp.var.rate, indicando que essas variáveis podem ter um impacto significativo na decisão final dos clientes.

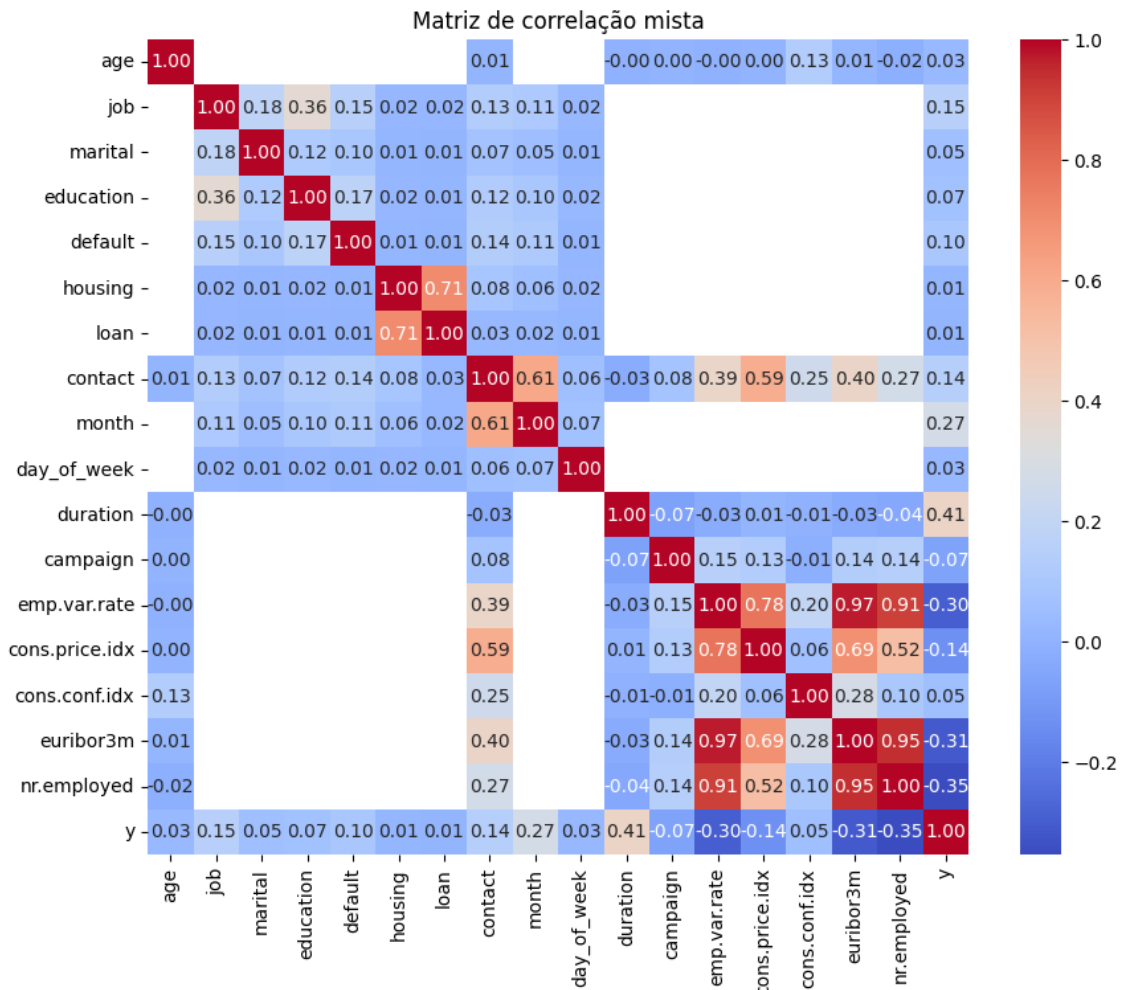


Figura 14 - Matriz de Correlação Mista

A análise da matriz de correlação mista revela que a variável duration é a que apresenta maior correlação com Y, indicando que quanto mais longa a duração do contacto, maior é a probabilidade de o cliente aceitar a oferta.

Outras variáveis com interações relevantes com Y são nr.employed e emp.var.rate:

- nr.employed tem uma correlação negativa com Y (-0.35), sugerindo que quando o número de empregados no mercado diminui, a taxa de aceitação aumenta. Este fenómeno pode estar relacionado com a maior consciencialização para a poupança em períodos de maior desemprego, levando à maior adesão a produtos como depósitos a prazo.

- emp.var.rate também apresenta uma correlação negativa com Y (-0.30), provavelmente pelo mesmo motivo, dada a forte correlação entre esta variável e nr.employed.

Além disso, observa-se uma forte relação entre emp.var.rate, nr.employed e euribor3m, reforçando a interligação entre esses indicadores macroeconómicos. Destaca-se ainda uma correlação positiva entre contact e month (0.59), sugerindo que o método de contacto e o mês da campanha podem ter impacto na aceitação da oferta.

Tratamento de Outliers e dados ausentes

Após a análise e a decisão de remover as colunas “previous”, “pdays” e “poutcome”, surge a necessidade de aprofundar a investigação sobre os dados ausentes nas colunas restantes. Para assegurar a qualidade e a integridade da nossa análise, é fundamental realizar uma abordagem mais detalhada.

Este próximo passo permitirá não só identificar quais colunas apresentam valores em falta, mas também definir a melhor estratégia para o seu preenchimento, de forma a manter a validade dos dados e evitar distorções nas conclusões. É essencial ponderar cuidadosamente como tratar esses valores ausentes, garantindo que qualquer intervenção respeite as características e padrões do conjunto de dados, promovendo uma análise mais precisa e confiável.

	Null Count	Null Percentage
age	0	0.000000
job	330	0.801438
marital	80	0.194288
education	1730	4.201477
default	8596	20.876239
housing	990	2.404313
loan	990	2.404313
contact	0	0.000000
month	0	0.000000
day_of_week	0	0.000000
duration	0	0.000000
campaign	0	0.000000
emp.var.rate	0	0.000000
cons.price.idx	0	0.000000
cons.conf.idx	0	0.000000
euribor3m	0	0.000000
nr.employed	0	0.000000
y	0	0.000000

Figura 15 - Valores em Falta

A análise da tabela revela que os dados em falta estão predominantemente em colunas categóricas, distribuídas por 3 colunas binárias (sim/não) e 3 colunas com múltiplos valores. Com base na matriz de correlação mista, que foi previamente explicada, e comparando as colunas com valores em falta com outras que podem fornecer informações relevantes, elaboramos a seguinte análise das relações entre as colunas:

Job: A coluna job apresenta a maior correlação com a coluna education, embora esta correlação seja modesta (0.36), o que indica uma relação fraca. Como a percentagem de valores em falta é reduzida (0.8%), optamos por utilizar a moda para preencher os valores ausentes.

Marital: Tal como a coluna job, a coluna marital mostra uma correlação muito fraca com outras variáveis. Considerando os 80 valores em falta, que representam uma pequena porção dos dados, também optamos pela imputação pela moda.

Education: A variável education apresenta uma quantidade substancial de valores em falta, o que dificulta a utilização de uma imputação simples. A correlação mais forte é com a variável job. Inicialmente, tentámos utilizar imputação múltipla através de um modelo preditivo (RandomForestRegressor), mas como esta abordagem não se mostrou viável, decidimos usar a moda, uma vez que o valor mais frequente (university degree) é significativamente mais comum que as restantes opções.

Default: A variável default apresenta o maior número de valores em falta. No entanto, por ser uma variável binária, é importante analisar a distribuição dos valores existentes antes de tomar uma decisão. Caso um dos valores tenha uma grande predominância, a imputação pela moda é uma opção. Se, por outro lado, ambos os valores binários (Yes/No) forem igualmente distribuídos, pode-se considerar a remoção da coluna, dado que, ao atribuir todos os valores “unknown” como “No”, a variância da coluna será reduzida a praticamente zero, tornando a variável redundante.

Housing e Loan: As variáveis housing e loan têm menos de 2.5% de valores em falta, o que é uma percentagem reduzida. Ambas são variáveis binárias, e, dado o baixo número de valores ausentes, podemos utilizar a moda para preencher os valores em falta, especialmente se um dos valores binários for muito mais frequente que o outro.

Conclusões:

- A coluna default deverá ser removida, dado que a imputação de todos os valores “unknown” como “No” resultaria numa coluna com variância quase nula, não acrescentando valor à análise.
- A variável loan pode ser preenchida com a moda, utilizando o valor mais frequente (“no”).
- Para a variável housing, apesar da distribuição mais equilibrada, como a percentagem de valores em falta é pequena, optámos por preencher com a moda (“yes”), uma vez que métodos computacionalmente mais complexos não trariam benefícios substanciais.

Outliers

Para o tratamento dos outliers iremos usar a distância interquartil para nos ser possível identificar os outliers.

```
Column age with 4 outliers
Column duration with 1043 outliers
Column campaign with 1094 outliers
Column emp.var.rate with 0 outliers
Column cons.price.idx with 0 outliers
Column cons.conf.idx with 0 outliers
Column euribor3m with 0 outliers
Column nr.employed with 0 outliers
```

Figura 16 - identificação Outliers

Dado o reduzido número de outliers nas variáveis numéricas, optamos por uma estratégia de substituição direta para o tratamento destes valores discrepantes.

Idade (Age): A variável age apresenta uma distribuição relativamente próxima de uma distribuição normal. Apesar de alguns valores se situarem nos extremos, estes não estão excessivamente distantes do intervalo de valores mais comuns. Assim, para os outliers identificados, vamos substituí-los pelos percentis de 1% (limite inferior) e 99% (limite superior), uma abordagem que preserva a integridade da distribuição de dados, mantendo os valores dentro de um intervalo aceitável.

Duração (Duration) e Campanha (Campaign): Ambas as variáveis apresentam uma alta skewness e kurtosis acentuada, o que indica uma concentração significativa de valores em torno da média, com muitos outliers com valores extremamente altos. Para equilibrar a distribuição e reduzir o impacto destes valores extremos, optamos por substituir os outliers pela mediana. A mediana é uma medida robusta que não é

influenciada por valores extremos, oferecendo uma substituição mais adequada e representativa para os dados restantes.

Esta abordagem visa minimizar o impacto dos outliers nas variáveis, ao mesmo tempo que mantém a integridade dos dados para análises futuras.

Transformação de dados

Para a preparação dos dados e a aplicação dos modelos, será necessário realizar o encoding das variáveis categóricas e a normalização das variáveis numéricas, de forma a evitar que valores extremos ou desbalanceados afetem os resultados.

Discretização

A fim de reduzir a complexidade do modelo e agrupar valores que partilham padrões comuns, decidimos realizar a discretização baseada em domínio para duas variáveis: age e education. Esta abordagem será guiada por categorias amplamente reconhecidas e alinhadas com padrões sociais e educativos, o que ajudará a simplificar os dados, sem perder informações significativas.

Age:

A variável age apresenta uma grande variação, mas é comum que certos grupos etários apresentem comportamentos semelhantes em termos de decisões ou preferências. Para agrupar as idades de forma lógica, vamos basear-nos nas faixas etárias definidas pelo site do governo canadiano (<https://www.statcan.gc.ca/en/concepts/definitions/age2>), que, embora já não seja utilizado oficialmente, ainda oferece uma estrutura útil para a categorização. As faixas etárias serão:

- *Youth (Menos de 24 anos)*: Agrupa jovens que ainda estão em fase de aprendizagem ou início da vida adulta.
- *Adult (25 a 64 anos)*: Refere-se à fase adulta, onde as pessoas frequentemente estão em estágios de plena atividade profissional e familiar.
- *Senior (65 ou mais anos)*: Refere-se a pessoas na fase de reforma ou já em idade avançada.

Essa divisão simplifica a variável age, agrupando-a em três categorias amplas que mantêm a integridade dos comportamentos típicos dos diferentes grupos etários.

Education:

A variável education também pode ser agrupada de forma a reduzir o número de categorias, uma vez que, em termos de comportamento social e profissional, as pessoas com níveis de escolaridade semelhantes tendem a partilhar características em comum. A categorização será feita tendo em conta os níveis de escolaridade predominantes em Portugal:

- *Illiterate (Analfabeto)*: Refere-se às pessoas que não possuem qualquer grau de escolaridade formal.
- *Basic (Ensino básico)*: Inclui os graus de escolaridade mais baixos, como basic.4y, basic.6y, basic.9y, que correspondem à escolaridade obrigatória em Portugal.

- *High School (Ensino secundário)*: Agrupa os níveis de escolaridade intermediários, como high.school e professional.course, que correspondem ao ensino secundário e cursos profissionais.
- *University (Ensino superior)*: Engloba os indivíduos com formação universitária, ou seja, university.degree.

Este agrupamento permite uma visão mais geral do nível educacional, focando-se nas principais etapas da educação formal, sem a necessidade de discriminar todos os subníveis de cada categoria.

Variáveis Numéricas:

Age: Como esta variável segue uma distribuição aproximadamente normal, vamos aplicar o StandardScaler. Este método ajusta a variável para que tenha média 0 e desvio padrão 1, o que é ideal para variáveis com distribuição normal.

Duration e Campaign: Estas variáveis apresentam uma elevada concentração de valores em torno da média e alguns outliers. Para minimizar o impacto dos outliers, vamos utilizar o RobustScaler, que utiliza a mediana e o intervalo interquartil para escalonar os dados, tornando a transformação menos sensível aos valores extremos.

cons.price.idx, euribor3m e nr.employed: Por serem variáveis exclusivamente positivas e com presença de outliers, será aplicado o Min-Max Scaler, que ajusta os valores para o intervalo [0, 1]. Este método é adequado para dados não muito dispersos, mas que têm valores extremos.

emp.var.rate e cons.conf.idx: Estas variáveis apresentam valores negativos e, embora não contenham outliers extremos, não seguem uma distribuição normal. Dado que o Min-Max Scaler não é ideal para valores negativos e o StandardScaler não se aplica bem a distribuições não normais, vamos utilizar o RobustScaler, que é menos influenciado por valores extremos e também adequado para dados com distribuição assimétrica.

Variáveis Categóricas:

Job e Marital: Estas variáveis não têm uma ordem intrínseca entre as categorias. Portanto, vamos aplicar o One-Hot Encoding, que cria colunas binárias para cada categoria, permitindo representar as variáveis sem qualquer tipo de hierarquia ou ordenação.

Education: A variável education possui uma ordem natural entre as categorias (ex.: basic.4y < basic.6y < basic.9y < ...). Como existe uma relação de ordem, vamos utilizar o Ordinal Encoding, que atribui um valor numérico para cada categoria, respeitando a ordem implícita.

Housing, Loan, Y e Contact: Estas variáveis são binárias, ou seja, possuem apenas dois valores possíveis. Neste caso, o Label Encoding é apropriado, pois converte as categorias em valores numéricos (0 ou 1), sem implicar qualquer tipo de ordem.

Day_of_week: Embora day_of_week tenha uma aparente ordem (segunda-feira, terça-feira, ...), a distância entre os dias é a mesma em termos de sequência, o que não justifica o uso de Ordinal Encoding. Assim, será utilizado o One-Hot Encoding para criar uma coluna binária para cada dia da semana.

Month: Similar à variável day_of_week, a variável month não deve ser tratada como uma variável ordinal, uma vez que a distância entre os meses é constante. Logo, vamos usar o One-Hot Encoding para representar cada mês de forma independente.

Resumo das abordagens:

Normalização de Variáveis Numéricas:

- Age: StandardScaler
- Duration e Campaign: RobustScaler
- cons.price.idx, euribor3m, nr.employed: Min-Max Scaler
- emp.var.rate e cons.conf.idx: RobustScaler

Encoding de Variáveis Categóricas:

- Job e Marital: One-Hot Encoding
- Education: Ordinal Encoding
- Housing, Loan, Y e Contact: Label Encoding
- Day_of_week e Month: One-Hot Encoding

Este conjunto de transformações garantirá que as variáveis estejam adequadamente preparadas para a aplicação de modelos de machine learning, respeitando as suas características e a relação entre as categorias.

PCA

Para a preparação de utilização do PCA para a redução da dimensionalidade, com a aplicação do Scree plot, verificou-se que o ponto de cotovelo está nos 8 componentes.

Assim sendo, definimos este número como sendo o número de componentes necessários para utilizar no PCA.

Após a aplicação do PCA verificamos uma percentagem da variância explicada de 67.94%. Este indicador explica quanto da informação total dos dados originais é preservada pelas componentes principais escolhidas.

K-Means Model

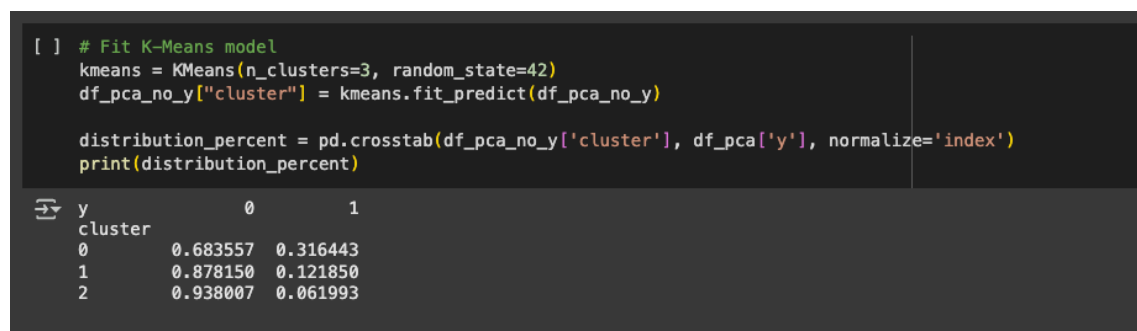


Figura 17 - K-Means Model

Após ter sido considerado para análise o Scree Plot e o PCA, conforme presente no *notebook*, aplicamos os resultados obtidos no K-Means.

Assim sendo, podemos concluir para os 3 clusters definidos, os seguintes dados:

- No cluster 0, 68,36% das amostras pertencem à classe 0 (No) da variável Y, enquanto 31,64% pertencem à classe 1 (Yes).
- No cluster 1, 87,82% das amostras pertencem à classe 0 (No) de Y, enquanto 12,19% pertencem à classe 1 (Yes).
- No cluster 2, 93,80% das amostras pertencem a No de Y, enquanto 6,20% pertencem a Yes.

Verifica-se que nos clusters construídos, todos eles demonstram dados da classe 0, o que pode indicar que as características dos dados estão mais relacionadas com a classe 0 (No). Pode também sugerir que a distribuição dos dados na variável alvo (Y) não é balanceada, com a classe 0 sendo mais comum ou com características mais distintas.

À medida que os clusters aumentam em “pureza” (ou seja, a proporção da classe 0 aumenta), a classe 1 diminui drasticamente. Isso pode refletir que, embora as amostras de classe 1 sejam menos numerosas em geral, elas ainda possuem características comuns que podem ser agrupadas em certos clusters, mas de forma mais diluída ou dispersa.

Desafios e Oportunidades

A elaboração deste relatório levantou imensos desafios relativamente à interpretação do *Dataset* e à forma como construir um output de dados o mais fiáveis possível, respeitando os dados em estudo.

Estes desafios foram importantes para alavancar o nosso pensamento crítico, capacidade analítica e interpretabilidade dos dados.

A complexidade que surgiu para identificar a qualidade dos dados, a identificação de padrões ou até mesmo a escolha das técnicas certas a serem usadas, permitiu melhorar a tomada de decisão, a preparação dos dados e a formulação de hipóteses.

Conclusão

A Análise de Dados Exploratória realizada neste relatório permitiu obter uma compreensão detalhada do dataset *Bank_Data*. Foi possível identificar padrões e destacar fatores relevantes para a tomada de decisão.

A inspeção inicial dos dados permitiu identificar a qualidade, valores ausentes e duplicados do dataset. A forte correlação entre variáveis, como *euribor3m*, *nr.employed* e *emp.var.rate*, permitiu identificar variáveis que potenciem padrões para a ser utilizado em machine learning.

Tudo isto associado à visualização dos dados, leva a identificarem-se oportunidades para uma segmentação mais eficiente e estratégias personalizadas.

Em suma, toda esta análise realizada irá permitir iniciar o processo de machine learning com dados com mais qualidade para treinar modelos robustos e eficazes.

Modelos Machine Learning

Após o respetivo tratamento e conclusões iniciais obtidas com a Análise de Dados exploratória, aplicamos o dataset devidamente tratado aos modelos de machine learning.

No código apresentado no notebook, utilizamos seis modelos diferentes de Machine Learning: Logistic Regression, SVC(Support Vector Classifier), KNN(K-Nearest Neighbors), Decision Tree, Random Forest, Fradient Boosting e XGBoost. A análise baseia-se em duas métricas principais de avaliação de desempenho: Accuracy e AUC.

Resultados_dos_Modelos_de_Machine_Learning

Model	Train/Test	Class	Precision	Recall	F1-Score	Accuracy	AUC	Fit Elapsed Time
Logistic Regression	Train	n	0.9107080461230584	0.9794342716589232	0.943821704277978	0.896387370977535	0.8617575018560172	3.538001537322998
Logistic Regression	Test	n	0.9029926451940148	0.9801789401238816	0.9400039601346446	0.8896308887809616	0.8512291413127993	3.538001537322998
Logistic Regression	Train	y	0.5873886223440713	0.2336423118865866	0.3343085625121904	0.896387370977535	0.8617575018560172	3.538001537322998
Logistic Regression	Test	y	0.5885714285714285	0.2121524201853759	0.3118849356548069	0.8896308887809616	0.8512291413127993	3.538001537322998
SVC with kernel (rbf)	Train	n	0.901882345644556	0.993543317846406	0.9454965132723224	0.8982088646023072	0.8831262221325467	3955.074142932892
SVC with kernel (rbf)	Test	n	0.8945671049367403	0.9927047487955952	0.9410843609316892	0.8903593977659058	0.8084360423372079	3955.074142932892
SVC with kernel (rbf)	Train	y	0.7272727272727273	0.1374045801526717	0.2311396468699839	0.8982088646023072	0.8831262221325467	3955.074142932892
SVC with kernel (rbf)	Test	y	0.6954022988505747	0.1246138002059732	0.211353711790393	0.8903593977659058	0.8084360423372079	3955.074142932892
KNN	Train	n	0.9233873571382448	0.97707707023777	0.9494738239883146	0.9075895567698846	0.9237490157371778	9.665144205093384
KNN	Test	n	0.9124548736462094	0.9741225051617344	0.9422808068703816	0.8947304516755706	0.8734172630510545	9.665144205093384
KNN	Train	y	0.6586978636826043	0.3530534351145038	0.4597089101881434	0.9075895567698846	0.9237490157371778	9.665144205093384
KNN	Test	y	0.6083333333333333	0.3007209062821833	0.4024810475534114	0.8947304516755706	0.8734172630510545	9.665144205093384
Decision Tree	Train	n	0.9192791982526018	0.977691992347636	0.9475862525660552	0.903885853066181	0.8767910919669549	2.7403721809387207
Decision Tree	Test	n	0.9091259640102828	0.9735719201651756	0.9402459288800266	0.8908450704225352	0.8559096808123821	2.7403721809387207
Decision Tree	Train	y	0.6388274336283186	0.3148854961832061	0.4218407596785975	0.903885853066181	0.8767910919669549	2.7403721809387207
Decision Tree	Test	y	0.5789473684210527	0.2718846549948506	0.3700070077084793	0.8908450704225352	0.8559096808123821	2.7403721809387207
Random Forest	Train	n	0.9347166645152962	0.9895121617928396	0.9613342183869896	0.9292653309046752	0.9498359294972764	563.2047414779663
Random Forest	Test	n	0.9104878985785632	0.9786648313833448	0.943346158949184	0.8963088878096164	0.8925561872414259	563.2047414779663
Random Forest	Train	y	0.8427254098360656	0.4484732824427481	0.5854092526690391	0.9292653309046752	0.9498359294972764	563.2047414779663
Random Forest	Test	y	0.6370023419203747	0.2801235839340886	0.3891273247496423	0.8963088878096164	0.8925561872414259	563.2047414779663
Gradient Boosting	Train	n	0.9286720901928128	0.9905370319759496	0.9586074652031606	0.9239829993928356	0.9365237096321852	2284.6626300811768
Gradient Boosting	Test	n	0.9087189976987984	0.9783895388850654	0.9422681779015046	0.8942447790189413	0.8923648150103873	2284.6626300811768
Gradient Boosting	Train	y	0.8387660069848661	0.3928571428571428	0.5350909766060156	0.9239829993928356	0.9365237096321852	2284.6626300811768
Gradient Boosting	Test	y	0.6207729468599034	0.2646755921730175	0.3711191335740072	0.8942447790189413	0.8923648150103873	2284.6626300811768
XGBoost	Train	n	0.9237209302325582	0.9837728887674227	0.952801627872351	0.9133879781420764	0.9342089922653782	66.63068056106567
XGBoost	Test	n	0.9107097104791186	0.9785271851342052	0.9434012341583172	0.8964303059737737	0.9007600312716402	66.63068056106567
XGBoost	Train	y	0.7308781869688386	0.351690294438386	0.4748757592490337	0.9133879781420764	0.9342089922653782	66.63068056106567
XGBoost	Test	y	0.6372093023255814	0.282183316168898	0.3911491791577444	0.8964303059737737	0.9007600312716402	66.63068056106567
LightGBM	Train	n	0.9234914830141469	0.9834654277124898	0.9525353627264456	0.912902246508804	0.940836965767324	260.90895199775696
LightGBM	Test	n	0.9104649673370052	0.9783895388850654	0.9432059447983014	0.8960660514813016	0.9049912004213024	260.90895199775696
LightGBM	Train	y	0.7260894170911149	0.3497818974918211	0.4721251149954001	0.912902246508804	0.940836965767324	260.90895199775696
LightGBM	Test	y	0.634032634032634	0.2801235839340886	0.3885714285714285	0.8960660514813016	0.9049912004213024	260.90895199775696

Figura 18 - Resultado Modelos Machine Learning

Analisando individualmente a performance de cada um dos modelos chegamos à seguinte conclusão:

1. Regressão Logística

- Para a classe "n", apresenta um F1-Score de **0.9438** e uma Acurácia global de **89.64%**.
- No "y", o F1-Score é inferior, com **0.3343**, sugerindo dificuldades na predição desta classe.
- A área sob a curva ROC (AUC) é **0.8618**, indicando um desempenho moderado.
- Tempo de treino: **3.54 segundos**.

2. SVC com kernel RBF

- Para "n": F1-Score de **0.9455**, ligeiramente superior à Regressão Logística.
- Para "y": Desempenho mais fraco, com um F1-Score mais baixo.

- Acurácia global de **89.82%**.
- AUC de **0.8831**, sugerindo melhor capacidade discriminativa.
- Tempo de treino significativamente maior (**3955.07 segundos**).

3. K-Nearest Neighbors (KNN)

- Modelo com uma Acurácia global de **90.12%**.
- Acurácia global razoável, mas menor capacidade de generalização no conjunto de teste.
- AUC moderado, indicando sensibilidade à escolha de vizinhos.
- Tempo de treino razoável, mas pode ser afetado por aumento de dados.

4. Decision Tree

- Tempo de treino de **2.74s**
- Acurácia de **89.74%**
- Excelente capacidade de ajuste ao treino (possível overfitting).
- F1-Score elevado no treino, mas perda significativa no teste.
- Tempo de treino rápido.

5. Random Forest

- Acurácia de **91.28%**
- Melhor balanço entre acurácia e generalização.
- AUC elevado (**92.12%**), indicando boa capacidade discriminativa.
- Tempo de treino moderado.

6. Gradient Boosting

- Acurácia (**90.91%**) e F1-Score (**70.18%**) elevados.
- Melhor desempenho que Decision Tree e Random Forest, mas com tempo de treino mais longo.

7. XGBoost

- Um dos melhores desempenhos gerais.
- Elevada AUC (**91.75%**) e F1-Score (**69.06%**) para ambas as classes.
- Tempo de treino (**66.63s**) otimizado graças à sua eficiência computacional.

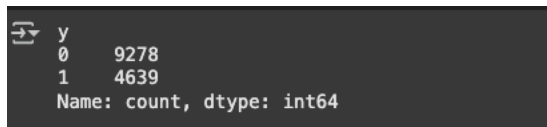
8. LightGBM

- Acurácia de **90.44%**
- Resultados semelhantes ao XGBoost, mas ligeiramente mais rápido.
- Boa capacidade de generalização e balanço entre predições.
- Tempo de treino de **260.91s**

Conclusões

Perante os resultados apresentados, o XGBoost demonstra ser um dos melhores modelos. Contudo, com o objetivo de termos uma melhoria generalizada na performance dos modelos, vamos utilizar undersampling para perceber se é possível prever com mais precisão quando um cliente aceita a proposta.

Assim sendo, vamos utilizar o método de undersampling Tomek Links e do Random undersampling, permitindo reduzir o tamanho da classe maioritária de forma seletiva.



```
y
0    9278
1    4639
Name: count, dtype: int64
```

Figura 19 - Dataframe com dados balanceados

Dados balanceados vs Dados desbalanceados

Regressão Logística

O balanceamento melhorou significativamente a predição da classe minoritária (Y), aumentando o F1-Score e o Recall, reduzindo ligeiramente a acurácia global.

SVC com Kernel RBF

Verifica-se uma melhoria da capacidade discriminativa (AUC), tornando o modelo mais justo entre classes, com significativa redução no tempo de treino.

Random Forest

O Random Forest beneficiou do balanceamento, melhorando as classes minoritárias.

XGBoost e LightGBM

Ambos os modelos melhoraram a distribuição de predição entre classes e melhoraram o AUC e Recall para o Y.

Conclusão dados balanceados

O balanceamento dos dados melhorou a predição da classe minoritária, tendo sido visível pelo aumento do Recall e F1-Score para Y.

A acurácia geral diminuiu ligeiramente, tendo sido expectável, visto o balanceamento dos dados deixou de favorecer a classe maioritária.

Em relação ao tempo de treino não houve melhorias muito significativas.

Os modelos assentes em ensembles (Random Forest, XGBoost, LightGBM) demonstraram melhorias consistentes nos dados balanceados.

Desafios e Oportunidades

A construção de modelos de machine learning apresenta diversos desafios e oportunidades. Os principais desafios que encontramos estão relacionados com o desbalanceamento dos dados, escolha do algoritmo adequado, tempo de treino, optimização de hiperparâmetros, entre outros.

Estas dificuldades tornam-se fundamentais para conhecer a forma como os modelos se comportam e as potencialidades que cada um oferece.

Tal como referimos na análise dos modelos, o desbalanceamento dos dados levou à oportunidade de melhorar significativamente a nossa capacidade de encontrar soluções que permitissem estabilizar o comportamento dos modelos para prevenir o mais possível o overfitting.

Conclusão

Durante a análise dos modelos de machine learning confirmamos aquilo que já era visível durante a análise de dados exploratória, que o dataset demonstrava um desbalanceamento entre as classes.

Assim sendo, tomamos a decisão de balancear esses dados, o que teve um impacto bastante positivo na classe minoritária.

Tanto com os dados balanceados como desbalanceados, os modelos mais estáveis foram o XGBoost e o LightGBM, mantendo um bom desempenho em ambas as abordagens, sendo indicados com as melhores opções para o conjunto de dados analisado.

Verificamos que os modelos mais simples são os que têm menor impacto no tempo de treino, tanto em dados balanceados como desbalanceados.

Em suma, concluímos que o balanceamento melhorou significativamente a equidade nas previsões sem comprometer a generalização dos modelos.