

I WANT ANSWERS



**CALL THE DATA SCIENTIST.
RIGHT NOW!**

Prática: Planejamento de sucessão

Planejamento de Sucessão

Descrição: Identificar funcionários com potencial para ocupar cargos de liderança no futuro e planejar a sucessão de cargos críticos

Dados sintéticos produzidos pelo ChatGPT, baseado no projeto Google Oxygen: Como a Google usou dados para ver se os gerentes fazem diferença?

- Idade do funcionário (valores inteiros de 18 a 75)
- Nível de educação (valores Superior, Médio, Doutorado, Especialização)
- Avaliação de desempenho (valores inteiros entre 0 e 5)
- Experiência em cargos de liderança (valores 0 ou 1)
- Habilidades e competências (valores inteiros entre 0 a 10)
- Participação em treinamentos de liderança (valores S ou N)
- Feedback de supervisores (valores reais entre 0.0 e 1.0)
- Satisfação no trabalho (valores inteiros entre 0 e 5)
- É um bom coach (valores S ou N)
- Empodera a equipe e não faz microgestão (valores S ou N)
- Exprime interesse e preocupação pelo sucesso e bem-estar pessoal dos membros da equipe (valores S ou N)
- É produtivo e orientado para os resultados (valores S ou N)
- É bom comunicador - escuta e compartilha informações (valores S ou N)
- Ajuda com desenvolvimento de carreira (valores S ou N)
- Tem uma visão clara e estratégia para a equipe (valores S ou N)
- Possui habilidades técnicas fundamentais que o ajudam a aconselhar a equipe (valores S ou N)
- **Lider (valores S ou N)**

Liderança

Projeto oxigênio do Google: aprenda transformar a cultura da sua equipe

Em 2008, uma equipe interna de pesquisadores lançou o Projeto Oxigênio do Google para entender o que leva um gerente a ser eficaz. Veja mais!

1. É um bom coach
2. Empodera a equipe e não faz microgestão
3. Exprime interesse e preocupação pelo sucesso e bem-estar pessoal dos membros da equipe
4. É produtivo e orientado para os resultados
5. É bom comunicador - escuta e compartilha informações
6. Ajuda com desenvolvimento de carreira
7. Tem uma visão clara e estratégia para a equipe
8. Possui habilidades técnicas fundamentais que o ajudam a aconselhar a equipe.

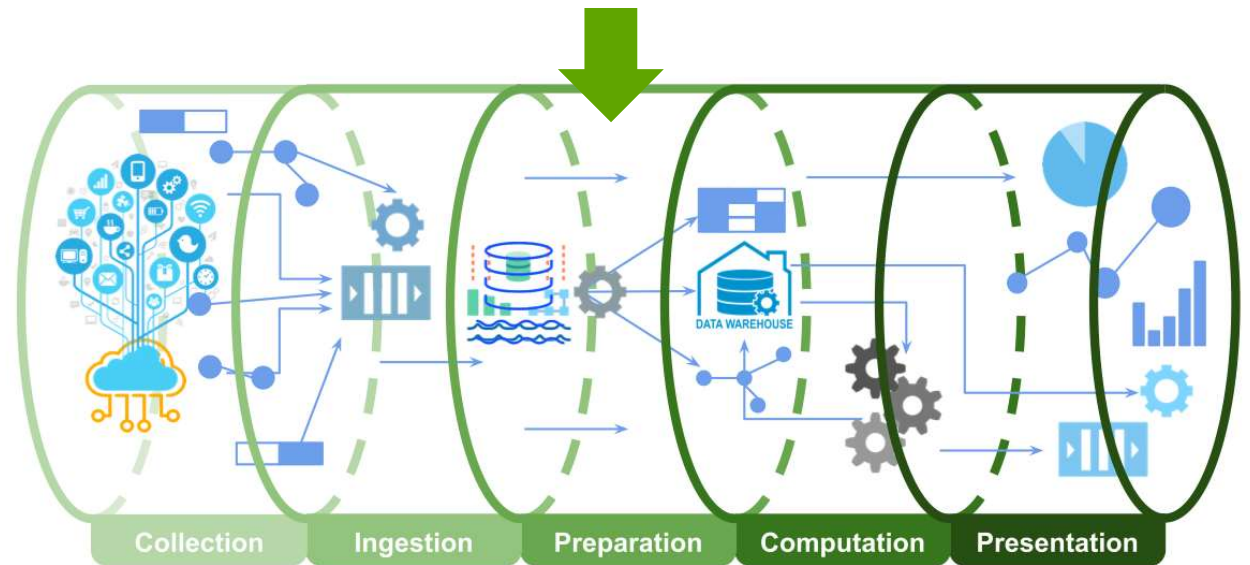
Dados tabulares (siméticos)

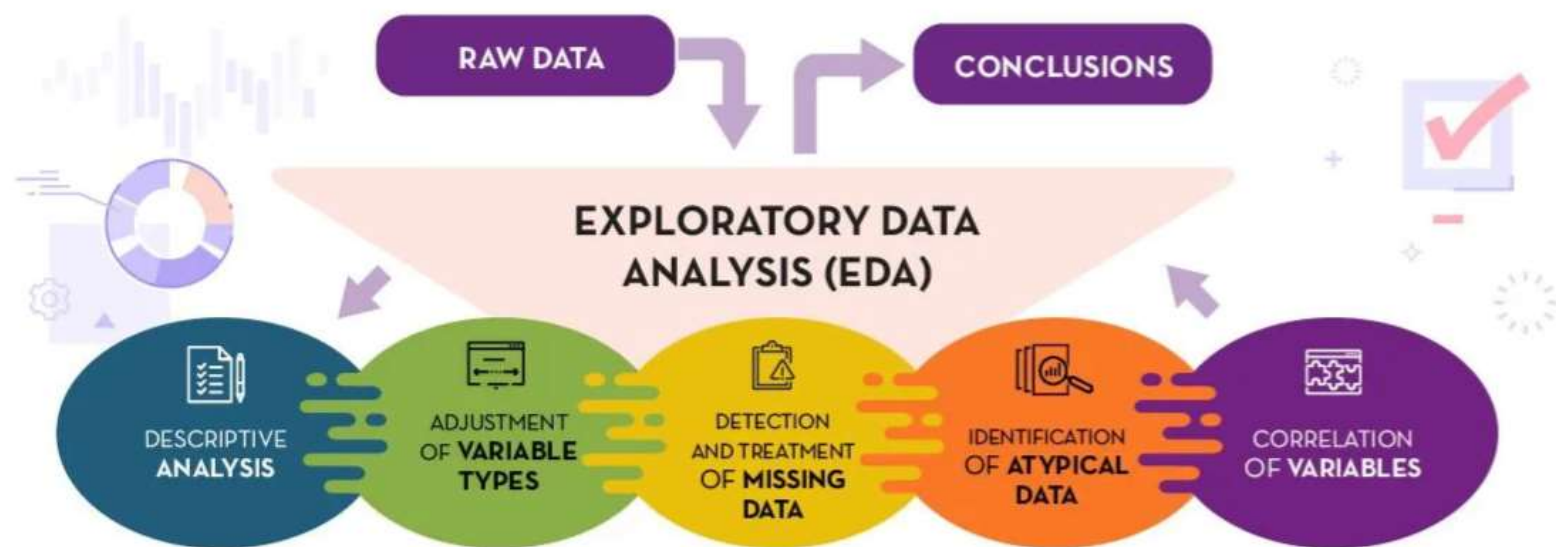
Modelo de classificação (sim/não)

KNN

Árvore de decisão

SVM





- Avaliação da Qualidade dos Dados:
 - A EDA ajuda a **localizar e resolver problemas com a qualidade dos dados**, incluindo inconsistências, outliers e valores ausentes.
- Engenharia e Seleção de Features:
 - Você pode **projetar novas features e decidir quais features incluir na sua análise** ao desenvolver uma compreensão mais profunda dos dados.
- Descoberta de Padrões:
 - Usando a EDA, é possível **encontrar relações, padrões e tendências** nos dados que podem não ser imediatamente aparentes.
- Construção e Avaliação de Modelos:
 - A EDA oferece **insights que ajudam na melhor seleção de modelos**, na melhoria da funcionalidade do modelo e na capacidade de avaliar a veracidade das suposições.

Exploratory Data Analysis (EDA)

- Análise da **dimensão** da base de dados
 - Entender o tamanho e a estrutura da base de dados.

python

 Copiar código

```
df.shape
df.info()
df.memory_usage(deep=True)
```




Exploratory Data Analysis (EDA)

- Análise da **distribuição** dos dados
 - Compreender a distribuição de valores nas variáveis

	Id	MSSubClass	LotFrontage	LotArea	Street	Alley	OverallQual
count	1460.00	1460.00	1460.00	1460.00	1460.00	1460.00	1460.00
mean	730.50	56.90	70.20	10516.83	1.00	0.03	6.10
std	421.61	42.30	22.43	9981.26	0.06	0.17	1.38
min	1.00	20.00	21.00	1300.00	0.00	0.00	1.00
25%	365.75	20.00	60.00	7553.50	1.00	0.00	5.00
50%	730.50	50.00	70.00	9478.50	1.00	0.00	6.00
75%	1095.25	70.00	80.00	11601.50	1.00	0.00	7.00
max	1460.00	190.00	313.00	215245.00	1.00	1.00	10.00

python

 Copiar código

```
df.describe()
df['coluna'].hist()
sns.distplot(df['coluna'])
```

Exploratory Data Analysis (EDA)

- Análise de **valores únicos**
 - Identificar a variedade de valores únicos em cada coluna.



python

 Copiar código

```
df.nunique()  
df['coluna'].value_counts()
```


Exploratory Data Analysis (EDA)

- Análise de **valores ausentes**
 - Contar valores ausentes por coluna.

```
python Copiar código

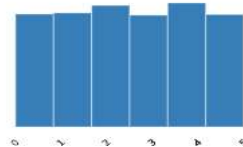
df.isnull().sum()
sns.heatmap(df.isnull(), cbar=False)
```

Satisfacao_trabalho

Real number (ℝ)

MISSING ZEROS

Distinct	6	Minimum	0
Distinct (%)	0.1%	Maximum	5
Missing	100	Zeros	794
Missing (%)	2.0%	Zeros (%)	15.9%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	2.5128571	Memory size	39.2 KiB



Remoção dos Dados Faltantes

v1	v2	v3	v4
2	55	44	casa
3		421	
2	23	12	apto
7		43	sítio
11			casa
65	12	21	casa

→

v1	v2	v3	v4
2	55	44	casa
2	23	12	apto
65	12	21	casa

Imputação pela Média ou Mediana

v1	v2	v3	v4
2	55	44	casa
3		421	
2	23	12	apto
7		43	sítio
11			casa
65	12	21	casa

→

v1	v2	v3	v4
2	55	44	casa
3	3.0	421	
2	23	12	apto
7	3.0	43	sítio
11	3.0	108.2	casa
65	12	21	casa

Imputação da Categoria mais Frequente

v1	v2	v3	v4
2	55	44	casa
3		421	
2	23	12	apto
7		43	sítio
11			casa
65	12	21	casa

→

v1	v2	v3	v4
2	55	44	casa
3		421	casa
2	23	12	apto
7		43	sítio
11			casa
65	12	21	casa

Exploratory Data Analysis (EDA)

- Análise do **balanceamento** nas classes
 - Contar frequência de cada classe.

Undersampling

Esse método consiste em **reduzir o número de observações da classe majoritária** para diminuir a diferença entre as categorias.

Oversampling

Ao contrário do *Undersampling*, o Oversampling **consiste em criar sinteticamente novas observações da classe minoritária**, com o objetivo de igualar a proporção das categorias.

SMOTE

A ideia por trás dela consiste em **criar observações intermediárias entre dados parecidos**, ou seja, se no *dataset* existem 2 pessoas, uma com altura 1,80 m e pesando 78 kg, a outra com 1,82 m e pesando 79 kg; o algoritmo do SMOTE adiciona uma “pessoa” intermediária medindo 1,81 m e pesando 78,3 kg.

python

Copiar código

```
df['classe'].value_counts()  
sns.barplot(x=df['classe'].value_counts().index, y=df['classe'].value_counts())
```

Lider

Categorical

HIGH CORRELATION

IMBALANCE

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	283.3 KiB




Modelos como **Gradient Boosting**, por exemplo, demonstram um desempenho melhor lidando com dados desbalanceados do que modelos como o **KNN** e **SVM**.

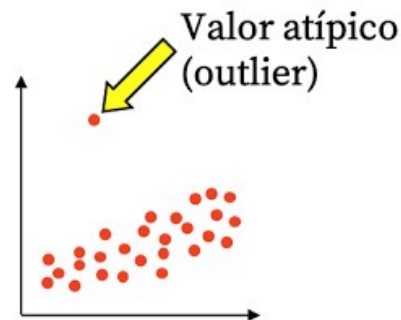
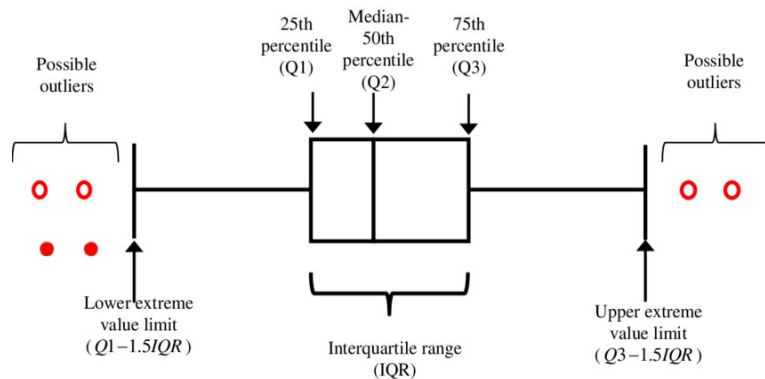
Exploratory Data Analysis (EDA)

- Identificação de outliers
 - Detectar valores atípicos que podem influenciar a análise.

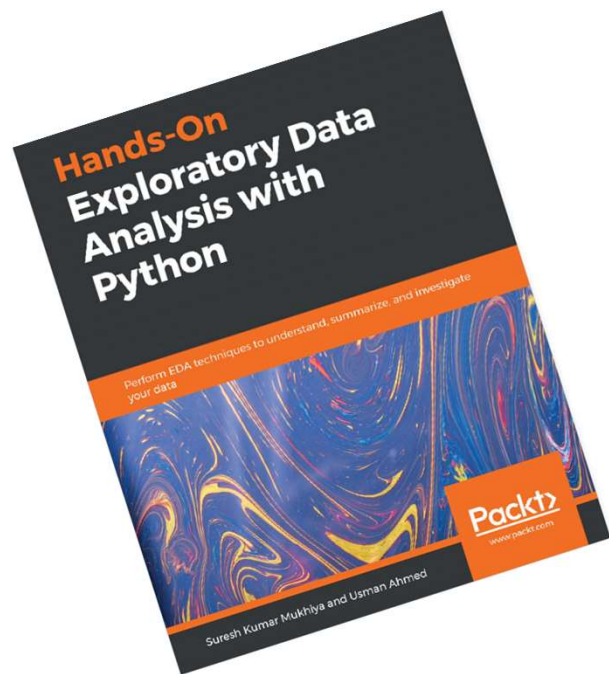
python

 Copiar código

```
sns.boxplot(x=df['coluna'])
```




- Opções de tratamento de outliers:
 - A **remoção de outliers** é útil para **erros de medição** claros, mas pode reduzir o tamanho da amostra
 - A **imputação** substitui valores discrepantes por estimativas razoáveis, **preservando o tamanho** da amostra
 - O **capping** limita os valores a um **threshold** máximo/mínimo, reduzindo a influência de outliers extremos
 - A **transformação** aplica funções que comprimem a **escala** dos dados, diminuindo o impacto de outliers



Exploratory Data Analysis (EDA)

- Bibliotecas de criação de relatórios exploratórios
 - Automatizar a criação de relatórios EDA.
- Bibliotecas:
 - Pandas Profiling
 - Sweetviz
 - D-Tale

python

 Copiar código

```
from pandas_profiling import ProfileReport
profile = ProfileReport(df, title="Pandas Profiling Report")
profile.to_notebook_iframe()
```