Lista de exercícios 3 de C24 - Inteligência Artificial

Análise exploratória de dados (EDA)

Exercícios teóricos

1. A frase "Garbage in, Garbage out" é frequentemente citada no contexto de ciência de dados. Qual é a relação direta entre esse conceito e a etapa de EDA?

A) A EDA serve apenas para criar visualizações bonitas para apresentações de negócios.
B) A EDA é uma curadoria que garante que o modelo de ML não aprenda com dados ruidosos ou irrelevantes, o que degradaria as previsões.
C) O princípio indica que, se a EDA for bem feita, o algoritmo de Machine Learning não precisa ser treinado.
D) Significa que o "lixo" (dados ruins) deve ser mantido para que o modelo aprenda a lidar com os erros do mundo real.

2. Um analista está trabalhando com uma coluna que contém "Níveis de Febre" (Baixa, Média, Alta). Como essa variável deve ser classificada e qual a restrição matemática aplicada a ela?

A) Qualitativa Nominal: operações aritméticas são permitidas.
B) Quantitativa Discreta: apenas a contagem é permitida.
C) Qualitativa Ordinal: existe uma hierarquia, mas operações aritméticas não são aplicáveis.
D) Quantitativa Contínua: permite o cálculo da média e da variância.

3. Ao utilizar o método df.describe() em um conjunto de dados, um cientista percebe que a média de uma coluna é muito superior à mediana. O que essa observação sugere sobre os dados?

A) Os dados estão perfeitamente distribuídos de forma normal.
B) A coluna contém apenas valores duplicados.
C) Há a provável presença de outliers com valores muito altos que estão distorcendo a média.
D) A biblioteca Pandas está com um erro de cálculo, pois média e mediana deveriam ser sempre iguais.

4. Durante a limpeza de dados, por que colunas que contêm IDs únicos ou timestamps constantes costumam ser removidas?

A) Porque ocupam muito espaço na memória e travam o algoritmo.
B) Porque são valores irrelevantes que não ajudam o objetivo do problema e podem atrapalhar o aprendizado do modelo.
C) Porque a biblioteca Pandas não consegue processar números e letras na mesma coluna.

D) Porque o documento afirma que os IDs são considerados outliers por definição.

5. Se um dataset possui uma coluna numérica com muitos valores ausentes e uma distribuição altamente assimétrica (com muitos outliers), qual seria a técnica de imputação mais robusta?

A) Preencher com a média.
B) Preencher com a mediana.
C) Remover todas as linhas com valores faltantes imediatamente.
D) Preencher com o valor zero para não alterar a soma total.

6. Qual técnica de preenchimento de valores faltantes é aplicável tanto a variáveis numéricas quanto a categóricas?

A) Mediana.
B) Média.
C) Moda.
D) KNNImputer.

7. Considere o exemplo do material de aula sobre o sensor de temperatura de um servidor que marca 500°C por um segundo. Qual deve ser a postura do analista?

A) Apagar o dado imediatamente, pois é fisicamente impossível.
B) Ignorar o valor e seguir com o treinamento do modelo.
C) Investigar antes de remover, pois pode representar uma falha crítica ou um evento raro relevante.
D) Substituir o valor pela média para suavizar o gráfico.

8. Em um Boxplot, o que representam os pontos localizados além dos "bigodes" (whiskers)?

A) A média aritmética dos dados.
B) O intervalo de confiança de 95%.
C) Possíveis outliers que estão fora do intervalo Q1 - 1.5 x IQR ou Q3 + 1.5 x IQR.
D) Os valores que representam a moda do conjunto de dados.

9. Para que a detecção de outliers via Z-Score seja considerada confiável, qual premissa sobre os dados deve ser assumida?

A) Que os dados seguem uma distribuição aproximadamente normal (Gaussiana).
B) Que os dados não possuem desvio padrão.
C) Que o dataset é composto apenas por variáveis qualitativas ordinais.
D) Que a média é sempre igual a zero.

10. Qual a principal vantagem do método IQR (Boxplot) sobre o método Z-Score na detecção de valores extremos?

A) O IQR é mais fácil de calcular manualmente.
B) O IQR é robusto a valores extremos e não pressupõe uma distribuição específica, enquanto o Z-Score é sensível a valores extremos.
C) O Z-Score só funciona com amostras pequenas.
D) O IQR identifica apenas erros de digitação, enquanto o Z-Score identifica fraudes.

11. Ao usar técnicas avançadas como KNNImputer ou IterativeImputer, qual cuidado fundamental deve ser tomado em um projeto de ML para evitar o "data leakage"?

A) Os valores faltantes devem ser preenchidos antes de carregar os dados.
B) Os valores faltantes devem ser calculados usando apenas o conjunto de treinamento.
C) Deve-se usar o conjunto de teste para estimar os valores do conjunto de treinamento.
D) Essas técnicas só devem ser usadas se o dataset tiver mais de 1 milhão de linhas.

12. Durante a análise exploratória de dados, é comum identificar relações estatísticas entre pares de variáveis. No entanto, a interpretação dessas relações deve ser feita com cautela. Qual afirmação expressa corretamente esse cuidado conceitual?

A) Uma correlação alta sempre indica que uma variável causa a outra.
B) Correlação só existe entre variáveis categóricas.
C) Correlação não implica causalidade.
D) Se a correlação for zero, as variáveis são obrigatoriamente idênticas.

13. Se um par de variáveis apresenta uma relação em formato de curva exponencial, mas sempre crescente, qual coeficiente apresentará um resultado mais fiel à força dessa relação?

A) Pearson, pois mede qualquer tipo de relação.
B) Spearman mede a relação monotônica e é menos sensível à forma da curva.
C) Z-Score, pois padroniza a curva.
D) Nenhum, pois correlações só funcionam para linhas retas perfeitas.

14. Em quais situações o coeficiente de Pearson resultará em um valor próximo de zero, mesmo que haja uma relação clara entre as variáveis?

A) Quando a relação é linear, positiva e forte.
B) Quando a relação é não linear (por exemplo, em formato de seno, de um círculo, logarítmico).
C) Quando os dados são qualitativos ordinais.
D) Sempre que houver mais de 100 linhas no dataframe.

15. Por que é importante detectar a multicolinearidade durante a EDA?

A) Para garantir que o gráfico de pizza fique colorido.

B) Porque variáveis altamente correlacionadas entre si são redundantes e podem prejudicar a predição, a interpretação dos coeficientes e a generalização dos modelos de ML.
C) Para aumentar o número de colunas no dataset final.
D) Porque a multicolinearidade impede a abertura de arquivos CSV.

16. Qual é a principal limitação do uso de diagramas de dispersão (scatter plots) em projetos de Big Data com centenas de variáveis?

A) Eles não aceitam cores diferentes para categorias.
B) Eles são limitados a variáveis qualitativas nominais.
C) É difícil escalar para alta dimensionalidade, o que torna a análise manual exaustiva e propensa a erros.
D) Os pontos em um scatter plot nunca se sobrepõem.

17. Por que os histogramas, embora úteis, não devem ser a única ferramenta para a decisão final sobre a remoção de outliers?

A) Porque eles não mostram a frequência dos dados.
B) Porque dependem do número de bins (caixas) e não definem limites objetivos, podendo esconder um outlier em um bin muito largo.
C) Porque histogramas só podem ser criados com a biblioteca Seaborn.
D) Porque eles invertem a ordem dos dados qualitativos.

18. No comando df.drop_duplicates(inplace=True), qual é a função do parâmetro inplace=True?

A) Criar uma cópia de segurança dos dados antes de deletar.
B) Fazer a alteração diretamente no próprio DataFrame original, sem criar um novo DataFrame.
C) Impedir que as duplicatas sejam removidas se houver valores nulos.
D) Mudar o tipo de dado de todas as colunas para "object".

19. Qual é a principal vantagem de utilizar o Pandas Profiling (ydata-profiling) em vez de realizar manualmente cada etapa da EDA?

A) Ele supre a necessidade de um cientista de dados no projeto.
B) Ele gera um relatório completo (estatísticas, correlações, alertas de nulos e de duplicatas) com uma única linha de código.
C) Ele corrige automaticamente todos os erros do banco de dados sem intervenção humana.
D) Ele transforma variáveis qualitativas em quantitativas automaticamente.

20. Qual é o posicionamento correto da EDA no ciclo de vida de um projeto de IA?

A) É a última etapa, realizada após a implantação do modelo em produção.
B) É uma etapa opcional, realizada apenas se o modelo de ML apresentar erros graves.

C) É o primeiro passo em qualquer projeto de ciência de dados, executado antes da construção e do treinamento dos modelos.
D) É uma etapa de processamento de hardware que não envolve análise de software.

Exercícios práticos: EDA com o dataset do Titanic

O dataset inclui colunas como:

- PassengerId: identificador do passageiro
- Survived: 1 = sobreviveu, 0 = não sobreviveu
- Pclass: classe no navio (1, 2, 3)
- Name: nome do passageiro
- Sex: sexo do passageiro
- Age: idade em anos
- SibSp: irmãos/cônjuges no navio
- Parch: pais/filhos no navio
- Ticket: número do bilhete
- Fare: preço da passagem
- Cabin: número da cabine
- Embarked: porto de embarque (C, Q, S)

O dataset está disponível em:
https://github.com/zz4fap/c24_inteligencia_artificial/blob/main/data/titanic.csv

OBS.: Crie um notebook Jupyter para responder aos exercícios abaixo.

1. Inspeção Inicial

Objetivo: entender a estrutura dos dados.

Enunciado:

1. Carregue o dataset train.csv com Pandas.
2. Exiba o número de linhas e colunas.
3. Liste os tipos de dados de cada coluna.
4. Mostre as primeiras 10 linhas para inspeção visual.

2. Tipos de Variáveis

Objetivo: classificar corretamente as variáveis.

Enunciado:

1. Classifique cada coluna como numérica ou categórica.
2. Diga qual tipo de variável (qualitativa/quantitativa) e qual subtipo (discreta/contínua, nominal/ordinal).
3. Identifique colunas que dificilmente devem ser usadas em modelos de predição.

3. Valores Faltantes

Objetivo: identificar e tratar valores ausentes.

Enunciado:

1. Verifique quantos valores faltantes há em cada coluna.
2. Para as variáveis numéricas com valores faltantes (e.g., Age), preencha com a mediana.
3. Para variáveis categóricas com valores faltantes (e.g., Cabin, Embarked), preencha com a moda.
4. Relate como isso alterou o número de valores faltantes.

4. Análise de Outliers

Objetivo: detectar outliers em variáveis contínuas.

Enunciado:

1. Escolha Age e Fare para análise.
2. Faça um boxplot para cada uma.
3. Calcule IQR e filtre outliers (pontos além de 1.5 × IQR).
4. Quantos outliers foram identificados?
5. Discuta se esses valores podem ser erros ou casos reais de exceções válidas.

5. Distribuições e Histograma

Objetivo: visualizar distribuições de variáveis.

Enunciado:

1. Plote histogramas de Age e Fare.
2. Descreva a distribuição de idade (alguma assimetria?).
3. Comparar o padrão de distribuição das tarifas (por classe de passagem).

6. Correlação entre Variáveis

Objetivo: analisar relações entre atributos.

Enunciado:

1. Calcule as matrizes de correlação de Pearson e de Spearman apenas entre variáveis numéricas.
2. Plote heatmaps das matrizes de correlação.
3. Identifique quais pares de variáveis apresentam alta correlação entre si. Há risco de multicolinearidade?
4. Identifique quais variáveis têm correlação mais forte com Survived.

7. Análise de Grupos

Objetivo: comparar grupos nos dados.

Enunciado:

1. Calcule a taxa de sobrevivência por:

    - sexo (male, female)
    - classe (Pclass)
2. Plote gráficos de barras que mostrem essas comparações.
3. Comente quais grupos tiveram maior taxa de sobrevivência e quais foram os possíveis motivos.

8. Scatter Plot bivariado
Objetivo: explorar as relações entre duas variáveis.

Enunciado:
1. Crie um gráfico de dispersão de Age vs Fare.
2. Colora os pontos de acordo com a sobrevivência (e.g., cores diferentes para Survived).
3. O gráfico mostra algum padrão entre idade, tarifa e sobrevivência?

9. Desafio (Opcional)
Objetivo: criar novas features (como o tamanho da família e o título extraído do nome) para enriquecer a análise e verificar, por meio de visualizações, se essas variáveis ajudam a explicar as diferenças nas taxas de sobrevivência.

Enunciado:
1. Crie novas features como:
    - tamanho da família (SibSp + Parch + 1)
    - título extraído do nome (Mr, Mrs, Miss, etc.)
2. Avalie pelo menos uma nova visualização usando a nova variável.
3. Discuta como essa feature pode ajudar a explicar a diferença na sobrevivência.