Relatório de inteligência artificial

Aluno: Luis Joaquim Vieira Borges
Matrícula:570637

1.Introdução

Explicação sobre os datasets:

Parte A

Contexto:

Os arquivos contêm dados que representam informações sobre clientes, para a análise de comportamento, segmentação ou modelagem preditiva.

Dados:

payment_data.csv:

id: ID do cliente
OVD_t1: número de vezes em atraso tipo 1
OVD_t2: número de vezes em atraso tipo 2
OVD_t3: número de vezes em atraso tipo 3
OVD_sum: total de dias em atraso
pay_normal: número de vezes em pagamento normal
prod_code: código do produto de crédito
prod_limit: limite de crédito do produto
update_date: data de atualização da conta
new_balance: saldo atual do produto
highest_balance: maior saldo do histórico
report_date: data do pagamento recente

customer_data.csv:

Dados demográficos do cliente e atributos de categoria que foram codificados.
fea_1: Categoria (ex: faixa etária, nível);
fea_3: Categoria (1, 2, 3);
fea_4: Valor monetário (ex: saldo, gasto);
fea_6: Contagem (ex: número de transações);
fea_7: Código especial ;
fea_8: Pontuação (ex: score de crédito);
fea_9: Categoria (1 a 5);
fea_10: Código (ex: ID de produto);

O rótulo é 1, o cliente apresenta alto risco de crédito. O rótulo é 0, o cliente apresenta baixo risco de crédito.

O que se deve analisar?

Tratar valores ausentes, analisar a distribuição da variável-alvo, fazer uma análise estatística exploratória e avaliar usando: Euclidiana, Manhattan,Chebyshev e Mahalanobis.

Parte B

Contexto:

Este conjunto de dados (housing.csv) contém informações sobre imóveis residenciais, de uma região específica, Boston, EUA. Ele será usado para análise com o uso modelos de regressão para prever preços de casas com base nas características socioeconômicas e estruturais.

Dados:

housing.csv:

RM (Average Number of Rooms): Número médio de cômodos por residência.

LSTAT (% Lower Status of the Population): Porcentagem de população de baixo status na área.

PTRATIO (Pupil-Teacher Ratio): Razão Pupilo-professor nas escolas do bairro.

MEDV (Median Value of Owner-Occupied Homes): Valor médio das residências ocupadas pelos proprietários em dólar

O que se deve analisar?

Tratar os valores ausentes, verificar a correlação entre as variáveis e a variável-alvo (MEDV), normalizar e padronizar os dados, calcular e comparar o RMSE e o R^2, visualizar e identificar os atributos mais relevantes os coeficientes do modelo Lasso e plotar um gráfico de importância dos atributos mais relevantes.

.

2.Observações

Parte A:

Para a análise dos dados será utilizado apenas o conjunto de dados “customer_data.csv”. O dataset “payment_data.csv “não foi utilizado para a análise por não haver variável-alvo definida e por apresentar dados completamente inconsistentes entre si, o que dificultaria a análise.

Primeiramente como “customer_data.csv” não possui nomenclaturas de suas variáveis, utilizando-se nome genéricos (ex:fea_1), foi preciso uma análise de comportamento dos dados com o intuito de visualizar possíveis significados ou padrões , assim temos:

fea_1: tipo da variável Inteiro, pode ser uma categoria (ex: faixa etária, nível), pois possui valores entre 1 e 7.
fea_3: tipo Inteiro, pode ser um grupo ou segmento.
fea_4: tipo Float, pode representar Valor monetário (ex: saldo, gasto), pois há um Faixa ampla (de 30.000 a 1.200.000)
fea_6: tipo Inteiro pode indicar uma Contagem (ex: número de transações), por que Varia de 4 a 15.
fea_7: tipo Inteiro, pode indicar um Código especial, em que -1 pode significar "não aplicável".
fea_8: tipo Inteiro, Pontuação (ex: score de crédito), em que Varia de 64 a 115
fea_9: tipo Inteiro, pode representar uma nota ou classificação (ex: 1-5 estrelas)
fea_10: tipo Inteiro, pode indicar Código(ex: ID de produto), pois possui Valores altos (ex: 151300, 72000)

não foi possível identificar:
fea_2: Pois há 149 valores ausentes;
fea_5: Pois os valores presentes são somente 1 e 2, em grande maioria 2;
fea_11: Por ter valores muito dispersos entre si.

3.Resultados

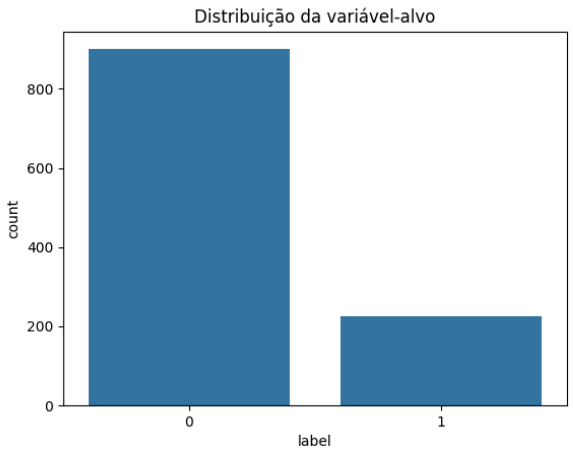
Parte A

Questão 1:

1. Carregue o dataset com pandas.
   Para isso foi usado df = pd.read_csv('customer_data.csv')

2. Trate valores ausentes.
   Utilizando missing = df.isnull().sum() foi possível identificar a quantidade exata de valores ausentes em cada coluna. Com isso foi evidenciado que a variável fea_2 possui 149 valores nulos, sendo a única com valores nulos em todo o dataset.

3. Analise a distribuição da variável-alvo.
   Com df['label'].value_counts(normalize=True) a variável-alvo label é analisada em termos de proporção de classes. O gráfico count plot mostra visualmente essa distribuição.

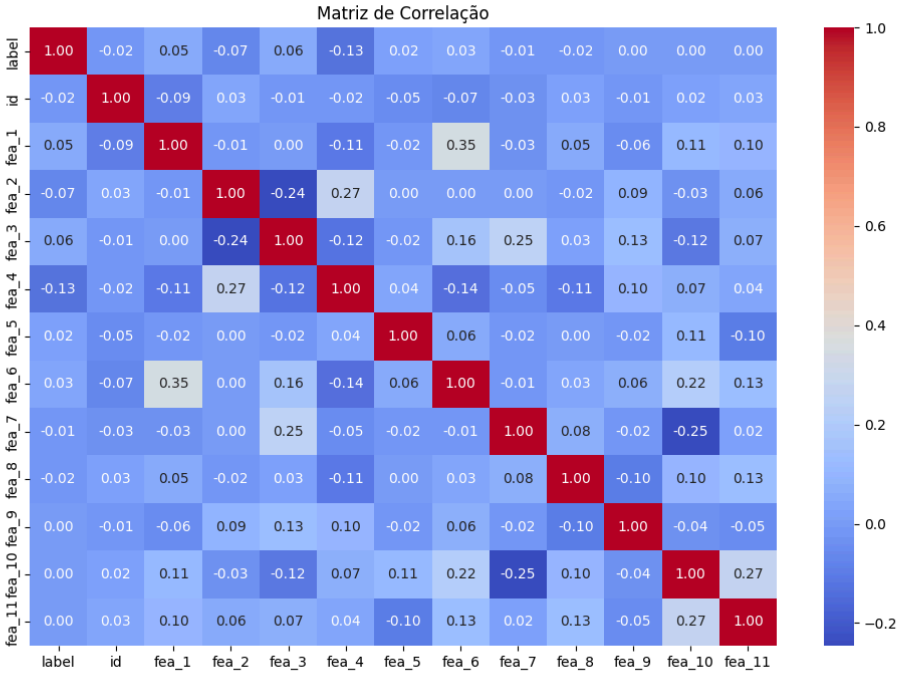


   Podendo perceber que 80% dos clientes possuem baixo risco de crédito e 20% apresentam alto risco de crédito.

4. Codifique variáveis categóricas, se necessário.
   df_encoded = pd.get_dummies(df, drop_first=True), A função get_dummies converte variáveis categóricas em variáveis binárias, removendo a primeira categoria para evitar multicolinearidade.

5. Faça uma análise estatística exploratória

print(df_encoded.describe()), retornou uma análise descritiva das variáveis:média, desvio-padrão, mínimo, máximo e quartis. Por fim, uma matriz de correlação foi plotada com o objetivo de identificar possíveis correlações entre as variáveis.



Com isso, é perceptível correlações nulas ou levemente expressivas tanto negativamente quanto fortemente, porém não muito relevantes para a análise.

6. Salve o arquivo como classificacao_ajustado.csv
   df_encoded.to_csv('classificacao_ajustado.csv', index=False), assim um novo dataset tratado será criado.

Questão 2:

1. Divida o dataset em treino e teste.
   X_train, X_test, y_train, y_test = train_test_split(
   X_scaled, y, test_size=0.3, random_state=42)
   Os dados foram divididos em 70% para treinamento e 30% para teste.

2. Implemente manualmente o KNN.

```
def knn_manual(X_train, y_train, X_test, k, tipo, VI=None):
    y_pred = []
    for x_test in X_test:
        distancias = [(calcular_distancia(x_test, x_train, tipo, VI), y)
                      for x_train, y in zip(X_train, y_train)]
        distancias.sort(key=lambda x: x[0])
        k_vizinhos = [label for _, label in distancias[:k]]
```

```
        pred = Counter(k_vizinhos).most_common(1)[0][0]
        y_pred.append(pred)
    return y_pred
```

3. Avalie usando:
    - Euclidiana : 0.7574
    - Manhattan : 0.7633
    - Chebyshev : 0.7781
    - Mahalanobis: 0.7633

4. Compare os resultados obtidos para os diferentes valores de distância considerando a métrica acurácia.

    Melhor Chebyshev (0.7781): ao considerar apenas a maior diferença entre os atributos, a Chebyshev indica que variáveis isoladas podem ser discriminativas.

    Manhattan e Mahalanobis (0.7633): essas métricas apresentaram desempenho iguais. Sabendo que Manhattan soma diferenças absolutas e é robusta a outliers e Mahalanobis leva em conta a variância e correlação. A igualdade implica que a covariância não tem grande impacto.

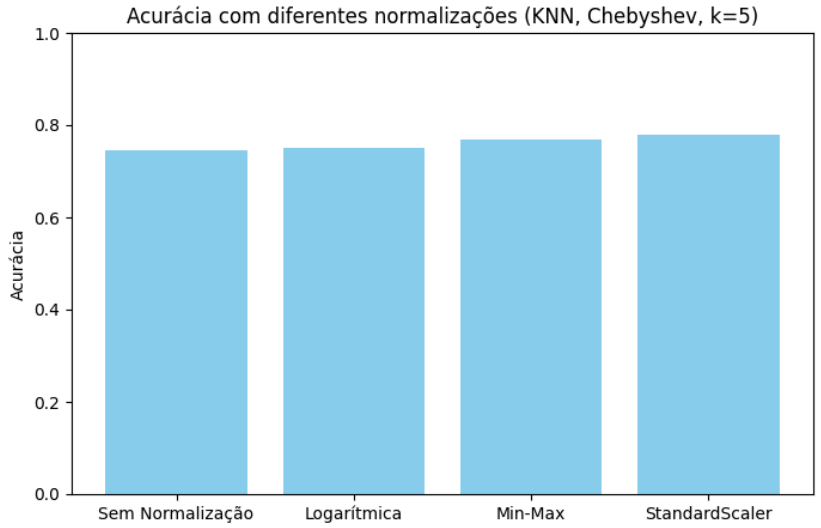
    Pior Euclidiana (0.7574): pois mesmo com normalização, a distância quadrática foi limitada.

Questão 3:

Aplique as normalizações: logarítmica, minMax, StandardScaler.
testar_knn_normalizacao é responsável por aplicar a normalização e testar o desempenho do KNN.
As normalizações aplicadas foram:
Logarítmica: aplica log1p (log(x + 1)) nos atributos numéricos.
Min-Max: transforma os dados para o intervalo [0, 1], usando MinMaxScaler.
StandardScaler: centraliza os dados com média 0 e desvio padrão 1.

Reaplique o KNN manual com a melhor distância.
Após a implementação manual do KNN a distância utilizada foi Chebyshev, pois possui melhor métrica, que considera apenas a **maior diferença absoluta** entre os atributos de dois pontos.

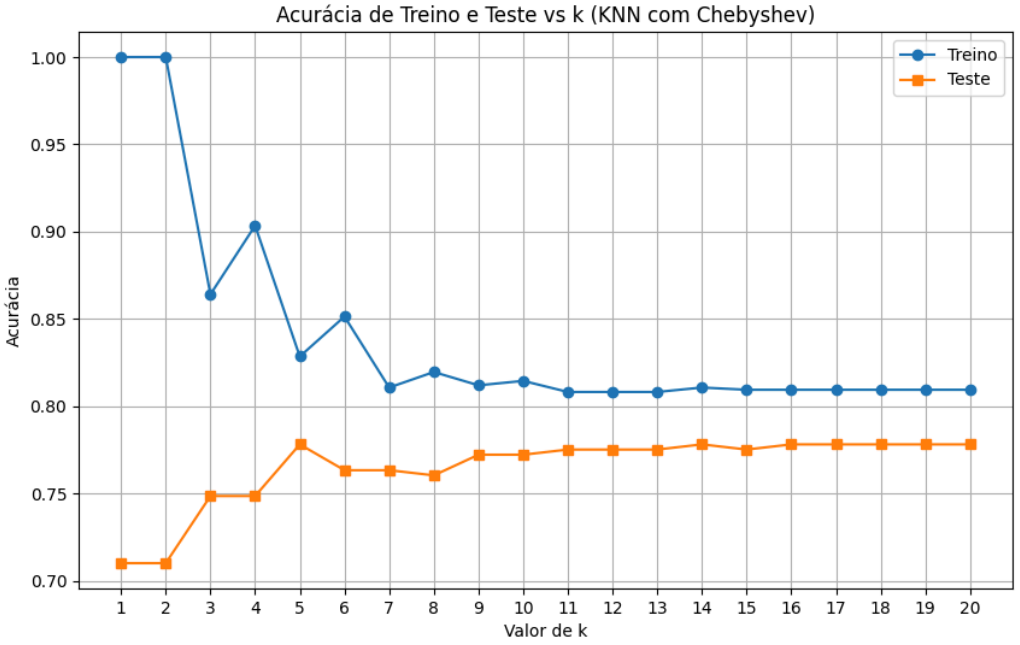
Compare os desempenhos com e sem normalização.
Para isso plotamos um gráfico de barras para comparar as normatizações
Acurácia com Sem Normalização: 0.7456
Acurácia com Logarítmica: 0.7515
Acurácia com Min-Max: 0.7692
Acurácia com Standard Scaler: 0.7781



Analise:

- A ausência de normalização apresentou o pior desempenho (74,56%)..
- A Min-Max e a Standard Scaler indicam que a escala dos atributos afeta diretamente o desempenho do KNN.
- A melhor performance foi com Standard Scaler (77,81%), pois o KNN é sensível à escala dos dados, pois utiliza a distância.

Faça também uma análise manual para descobrir o melhor valor de K para o KNN. Plote o gráfico de acurácia de treino e teste versus k.



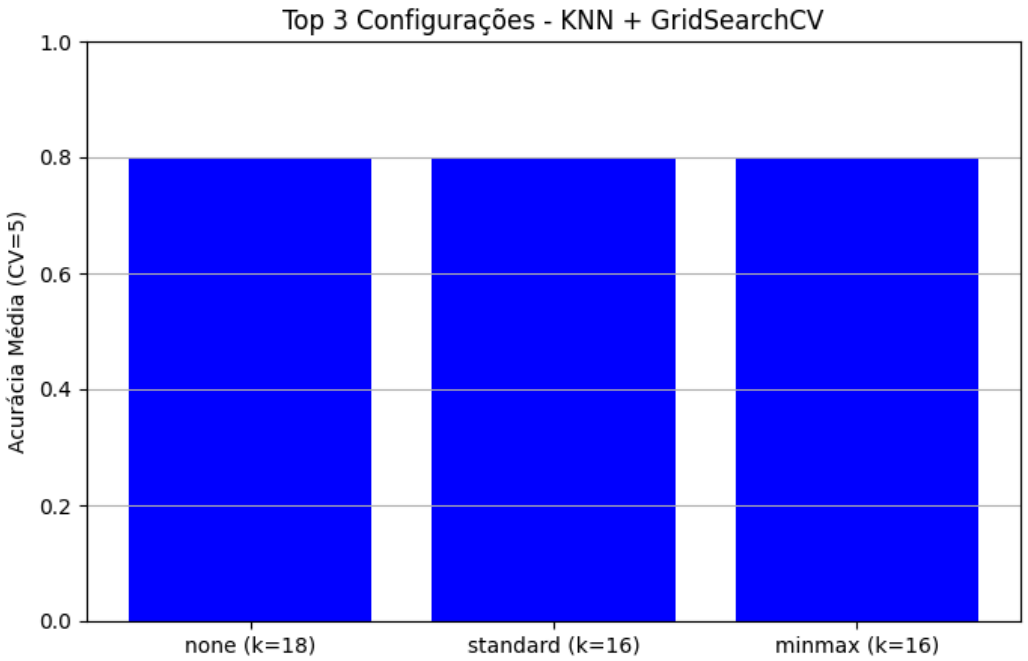
Com isso, foi possível notar que o knn=5(77,5%) é o melhor entre todos os demais valores.

Questão 4:

1. Use KNeighborsClassifier da Sklearn.
2. Teste valores de k entre 1 e 20 e as normalizações com GridSearchCV.
3. Plote gráfico dos resultados com as 3 melhores configurações.

Melhor configuração:

- logarítmica, Melhor k: 20, Acurácia: 0.7991
- standardscaler, Melhor k: 16, Acurácia: 0.8000
- minmax, Melhor k: 16, Acurácia: 0.8000
- Sem normalização, Melhor k: 18, Acurácia: 0.8009



O gráfico plotado demonstra que, dentre os 20 valores testados para o KNN, os melhores resultados foram obtidos com:

- k = 18 sem normalização (none),
- k = 16 com MinMaxScaler,
- k = 16 com StandardScaler.

Para a normalização **logarítmica**, o melhor valor de **k foi 20**, porém essa configuração não está entre as três melhores apresentadas no gráfico.

Questão 5:

Exiba: média, desvio padrão, matriz de confusão e classification_report.

Acurácia média: 0.7982;
Desvio padrão: 0.0036.
A acurácia média de 79,82% indica que o modelo acerta quase 80% dos exemplos em média nos 5 folds.
O desvio padrão muito baixo (0.0036) mostra que o desempenho é altamente estável entre as divisões dos dados.

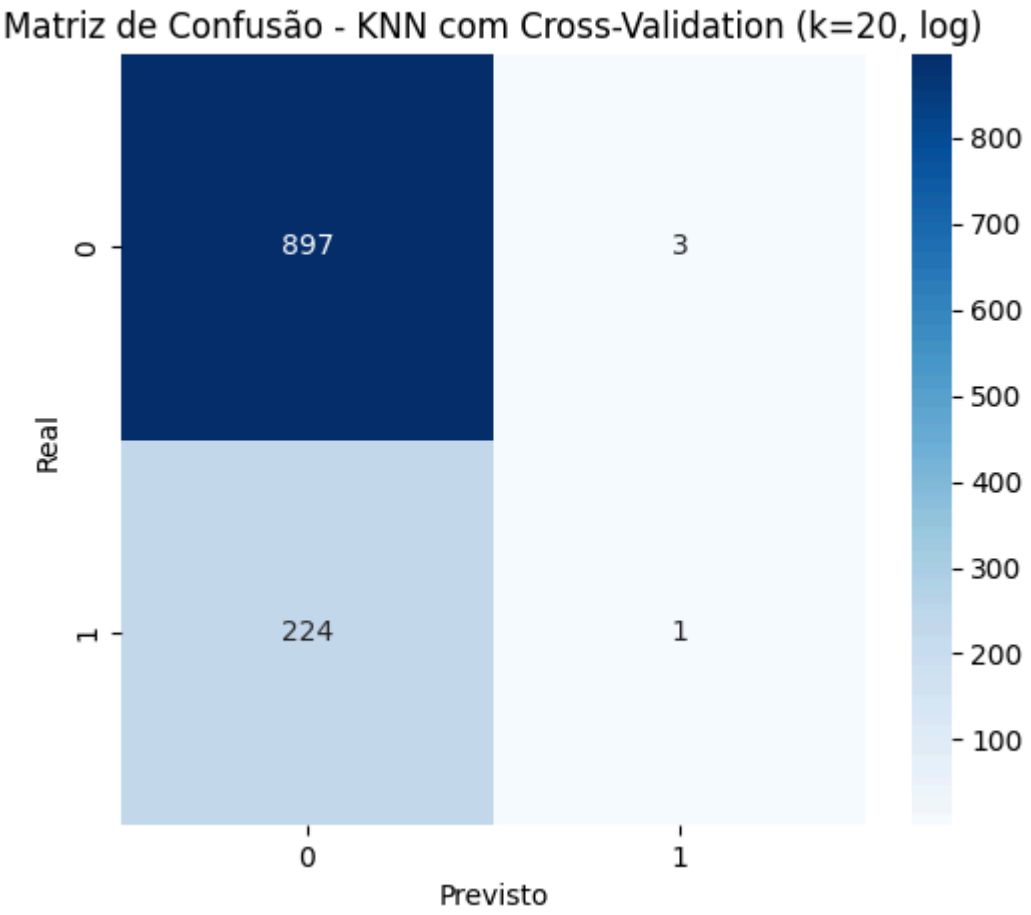
Classification Report:

Métrica	Classe 0	Classe 1	Média Macro	Média Ponderada
Precisão	**0.80**	**0.25**	**0.53**	**0.69**
Recall	**1.00**	**0.00**	**0.50**	**0.80**
F1-Score	**0.89**	**0.01**	**0.45**	**0.71**
Suporte	**900**	**225**	**—**	**1125**

Classe 0 (80% dos dados):
- Precisão 0.80: 80% das vezes que o modelo previu a classe 0, ele acertou.
- Revocação 1.00: o modelo acertou todos os exemplos reais da classe 0 .
- F1-score 0.89: equilíbrio entre precisão e revocação.

Classe 1 (20% dos dados):
- Precisão 0.25: apenas 25% das vezes que o modelo previu classe 1, ele acertou.
- Revocação 0.00: o modelo não conseguiu identificar corretamente nenhum exemplo da classe 1.
- F1-score 0.01: desempenho extremamente baixo para essa classe.

Conclusão:
- O modelo está muito enviesado para a (classe 0).
- A classe (classe 1) não está sendo aprendida, pois recall zero e F1-score praticamente nulo.
- Isso indica que o modelo está ignorando a classe 1,por desequilíbrio nas classes.



O modelo acertou 867 instâncias da classe 0 e errou 3, por outro lado na classe 1 errou 224 e acertou somente 1 instância. Diante disso, entende-se que o modelo não está reconhecendo a classe 1 perfeitamente.

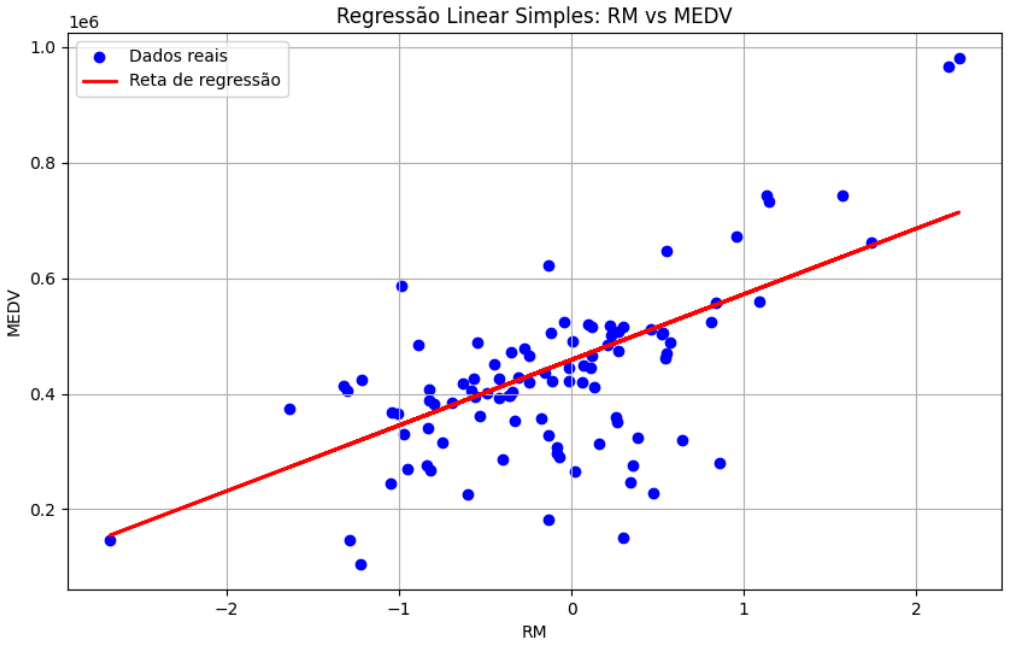
Parte B

Questão 1:

1. Trate valores ausentes.
   O conjunto não possui valores ausentes.
2. Verifique a correlação com a variável-alvo.
   Correlações com MEDV:
   RM 0.697209: forte correlação;
   PTRATIO -0.519034: correlação negativa moderada;
   LSTAT -0.760670:forte correlação negativa.

Questão 2:

1. Plote a reta de regressão (feature mais correlacionada).



Há uma tendência positiva: quanto mais cômodos existirm na residência mais alto será o valor médio. No entanto, os pontos estão um pouco dispersos em torno da reta, indicando que a variável RM explica apenas parte da variação em MEDV. Isso sugere que outros fatores podem influenciar no valor médio também.

2. Calcule RMSE e R^2.
   RMSE: 115595.95
   Representa o erro médio das previsões. Esse valor é alto, o que mostra que há uma diferença significativa entre os valores previstos e os reais.
   R^2: 0.3920
   indica que apenas 39,2% da variação no valor das casas é pelo número médio de cômodos. Isso sugere que RM tem uma certa influência sobre MEDV, mas muitos outros fatores também afetam os preços das casas como visto anteriormente.

Questão 3:

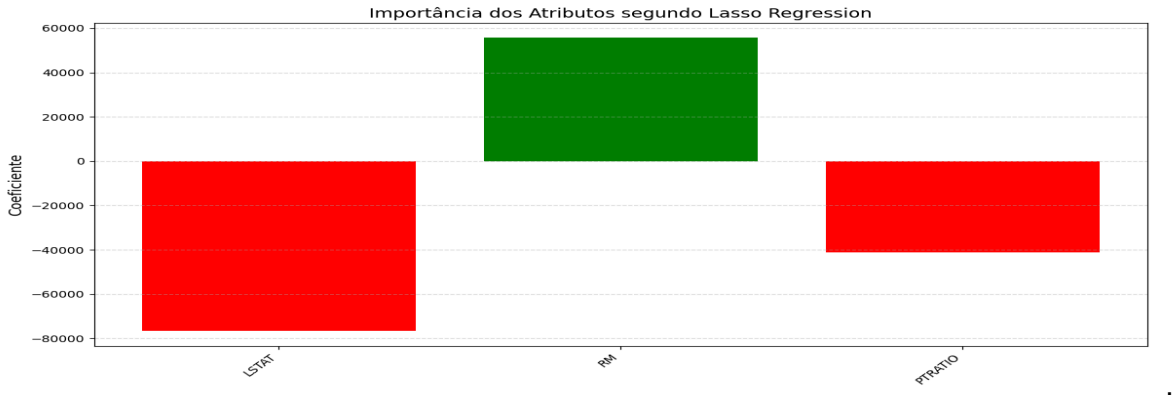
Compare RMSE e R² em uma tabela.

Modelo	RMSE (CV)	R² (CV)
Linear Regression	**99016.02**	**0.4238**
Ridge Regression	**98984.90**	**0.4241**
Lasso Regression	**99016.03**	**0.4238**

Identifique o melhor modelo.
Ridge Regression é o melhor modelo, pois apresenta o menor valor de RMSE e o maior valor de R² na tabela, indicando melhor desempenho preditivo.

Questão 4:

Plote um gráfico de importância



Discuta os resultados
O modelo Lasso:

Atributo	Coeficiente	Interpretação
RM	+56.000	Aumento no número médio de cômodos por casa está associado a um aumento no valor da residência.
LSTAT	-76.000	Maior porcentagem de população de baixa renda reduz o valor do imóvel.
PTRATIO	-39.000	Relações aluno/professor mais altas indicam queda no valor das casas.

4.Conclusões

Parte A

Os resultados esperados foram satisfeitos? Se não, qual o motivo?

Não pois o modelo preditivo não funcionou em customer_data,csve deixou os dados da classe enviesados

Qual a sua análise?

As variáveis apresentam baixa correlação entre si. Observa-se que 80% dos dados pertencem à classe 0, enquanto apenas 20% pertencem à classe 1, indicando um forte desbalanceamento. O melhor desempenho foi obtido utilizando a distância de Chebyshev com KNN = 5, atingindo uma acurácia de 77,81% sendo também o valor de *k* com melhor desempenho em comparação com os demais testados. A melhor performance foi observada com o uso do Standard Scaler.

Apesar da boa acurácia, o modelo está muito enviesado para a classe 0. A classe 1 não está sendo corretamente apreendida, apresentando *recall* zero e *F1-score* praticamente nulo. Isso evidencia que o modelo está ignorando a classe 1, resultado direto do forte desequilíbrio entre as classes

Parte B

Os resultados esperados foram satisfeitos? Se não, qual o motivo?

Sim, pois é um dataset feito para ser trabalhado com regressão.

Qual a sua análise?

A variável MEDV apresenta forte correlação positiva com RM (0,70), indicando que um maior número médio de cômodos está associado a valores mais altos das residências. Por outro lado, há correlação negativa moderada com PTRATIO (-0,52) e forte correlação negativa com LSTAT (-0,76), sugerindo que maiores taxas de população de baixa renda e relações aluno/professor mais altas estão associadas à queda no valor dos imóveis.

Embora RM tenha uma boa associação com MEDV, a dispersão em torno da reta indica que outros fatores também influenciam significativamente o valor das residências.

Entre os modelos avaliados, a Ridge Regression apresentou o melhor desempenho preditivo, com o menor RMSE e o maior R^2, destacando-se como o modelo mais eficaz para previsão do valor médio das casas.

5.Próximos passos

Depois desse relatório, quais os próximos passos você sugere para este projeto?

- Verificar se existem valores muito diferentes (outliers) que possam atrapalhar o modelo.
- Analisar os erros do modelo para ver se ele está deixando passar algum padrão.
- Fazer testes para saber se os erros seguem uma distribuição normal e se têm um comportamento estável.
- Criar novas variáveis ou combinar as que já existem pode melhorar os resultados.
- Testar outros modelos mais avançados, como Random Forest ou XGBoost, e comparar com o modelo atual.
- Para garantir que os resultados sejam confiáveis, é recomendado usar validação cruzada e outras medidas, como o erro médio.
- Ferramentas como SHAP ajudam a entender quais variáveis mais influenciam as previsões.
- Criar gráficos ou painéis interativos pode facilitar a apresentação.
- Se houver dados sobre localização, eles podem melhorar bastante o modelo.
- Por fim, transformar os valores da variável de saída pode ajudar se os erros forem muito diferentes entre si.