Curso: Ciência da Computação	Disciplina: M. A. Desempenho	Código/Turma: 36
Professor/a: Lyndainês		Data: 06/05/2025
Aluno/a: José Carlos Machado Magalhães Filho		Matrícula:2210416

Parte I – Análise Estatística Inicial

1. Tratamento de Dados Ausentes

Inicialmente, verificou-se a presença de dados ausentes no dataset. As seguintes colunas apresentavam valores faltantes:

- “latencia_ms”: 18 valores ausentes
- “armazenamento_tb”: 18 valores ausentes
- “tipo_hd”: 18 valores ausentes
- “tempo_resposta”: 1 valor ausente

Para garantir a integridade da análise, os dados ausentes foram tratados da seguinte forma, seguindo as melhores práticas:

- Variáveis Numéricas (“latencia_ms”, “armazenamento_tb”, “tempo_resposta”): substituídas pela mediana.
- Variáveis Categóricas (“tipo_hd”): substituídas pela moda.

Após este tratamento, confirmou-se que não restaram valores ausentes no dataset.

2. Análise Descritiva

Após a imputação dos dados ausentes, realizou-se uma análise descritiva das variáveis:

Variáveis Numéricas:

- “cpu_cores”: Média de 8.57 (2 a 15), desvio padrão 4.05.
- “ram_gb“: Média de 25.38 GB (4 a 64), desvio padrão 21.63.
- “latencia_ms“: Média de 157.14 ms (1.26 a 299.15), desvio padrão 75.65.
- “armazenamento_tb“: Média de 0.89 TB (0.256 a 2.0), desvio padrão 0.67.
- “tempo_resposta“: Média de 107.24 ms (28.68 a 412.04), desvio padrão 66.10.

Variáveis Categóricas:

- “sistema_operacional“: Linux (65), Windows (61), MacOS (59).
- “tipo_hd“: HDD (108), SSD (77).
- “tipo_processador“: AMD (63), Intel (63), Apple Silicon (59).

As variáveis categóricas presentes no dataset incluem “sistema_operacional”, “tipo_hd” e “tipo_processador”. Estas variáveis representam características qualitativas das máquinas analisadas e, para serem utilizadas em modelos de regressão, precisam ser convertidas para um formato numérico.

Este processo foi realizado por meio da técnica de codificação one-hot (também conhecida como dummy encoding), que cria colunas binárias indicando a presença ou ausência de cada categoria, exceto por uma categoria de referência.

Parte II – Modelo de Regressão e Diagnóstico

1. Modelo de Regressão Linear Múltipla (Modelo Inicial)

Foi ajustado um modelo de regressão linear múltipla para prever a variável “tempo_resposta” utilizando as demais variáveis como preditoras. As variáveis categóricas (”sistema_operacional”, ”tipo_hd”, ”tipo_processador”) foram transformadas em variáveis dummy (one-hot encoding), utilizando "Linux", "HDD" e "AMD" como categorias de referência, respetivamente.

Resultados do Modelo Inicial:

- R-squared: 0.669
- R-squared Ajustado: 0.654 (indicando que aproximadamente 65.4% da variabilidade do tempo de resposta é explicada pelo modelo).
- F-statistic: 4.52 (Prob (F-statistic): 1.93e-38, indicando que o modelo como um todo é estatisticamente significativo).

Coeficientes Significativos ($p < 0.05$):

- ”const”: 244.7960
- ”cpu_cores”: -11.4659 (cada núcleo adicional de CPU reduz o tempo de resposta em média 11.47 ms).
- ”ram_gb”: -1.4270 (cada GB adicional de RAM reduz o tempo de resposta em média 1.43 ms).

As variáveis ”latencia_ms”, ”armazenamento_tb” e as dummies das variáveis categóricas não se mostraram estatisticamente significativas neste modelo inicial.

2. Diagnóstico do Modelo Inicial

Multicolinearidade: O diagnóstico de multicolinearidade, através do Fator de Inflação da Variância (VIF), revelou VIF infinito para as variáveis

”sistema_operacional_MacOS” e ”tipo_processador_Apple Silicon”. Isto indica multicolinearidade perfeita, sugerindo a necessidade de remover pelo menos uma dessas variáveis para melhorar a estabilidade do modelo.

Heterocedasticidade: O teste de Breusch-Pagan indicou a presença de heterocedasticidade (p-valor F = 0.0171), significando que a variância dos erros do modelo não é constante.

3. Modelo de Regressão Linear Múltipla (Modelo Ajustado)

Com base no diagnóstico de multicolinearidade, a variável ”tipo_processador_Apple Silicon” foi removida e um novo modelo (Modelo 2) foi ajustado.

Resultados do Modelo Ajustado:

- R-squared: 0.669
- R-squared ajustado: 0.656 (ligeiramente superior ao modelo inicial).
- F-statistic: 50.96 (Prob (F-statistic): 1.39e-39).

Coeficientes Significativos ($p < 0.05$):

- ”const”: 244.4136
- ”cpu_cores”: -11.4659
- ”ram_gb”: -1.4270

Os coeficientes significativos e as suas interpretações mantiveram-se consistentes com o modelo inicial. A remoção da variável ”tipo_processador_Apple Silicon” resolveu o problema de multicolinearidade perfeita associado a ela (VIF para ”sistema_operacional_MacOS” tornou-se 1.0), sem prejudicar o poder explicativo do modelo.

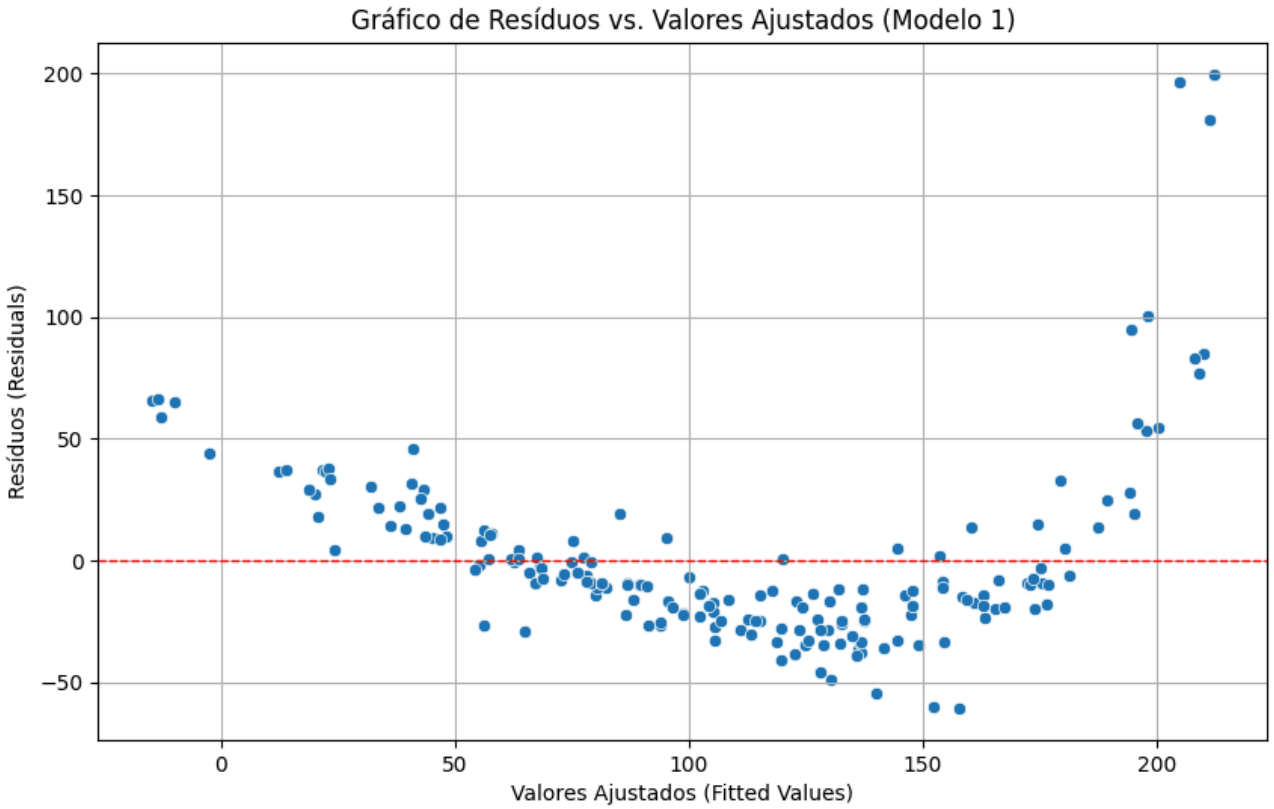

Figura 1 – Gráfico de Resíduos vs. Valores Ajustados.

Parte III – Análise Crítica e Conclusões

1. Comparação de Modelos

O Modelo 2 (sem `tipo_processador_Apple Silicon`) apresentou um R^2 ajustado ligeiramente superior (0.656 vs. 0.654) e resolveu um problema crítico de multicolinearidade. Um teste F para comparar os dois modelos indicou não haver diferença estatisticamente significativa entre eles (p-valor = 0.9217). Portanto, o Modelo 2 é preferível devido à sua parcimônia e maior robustez estatística.

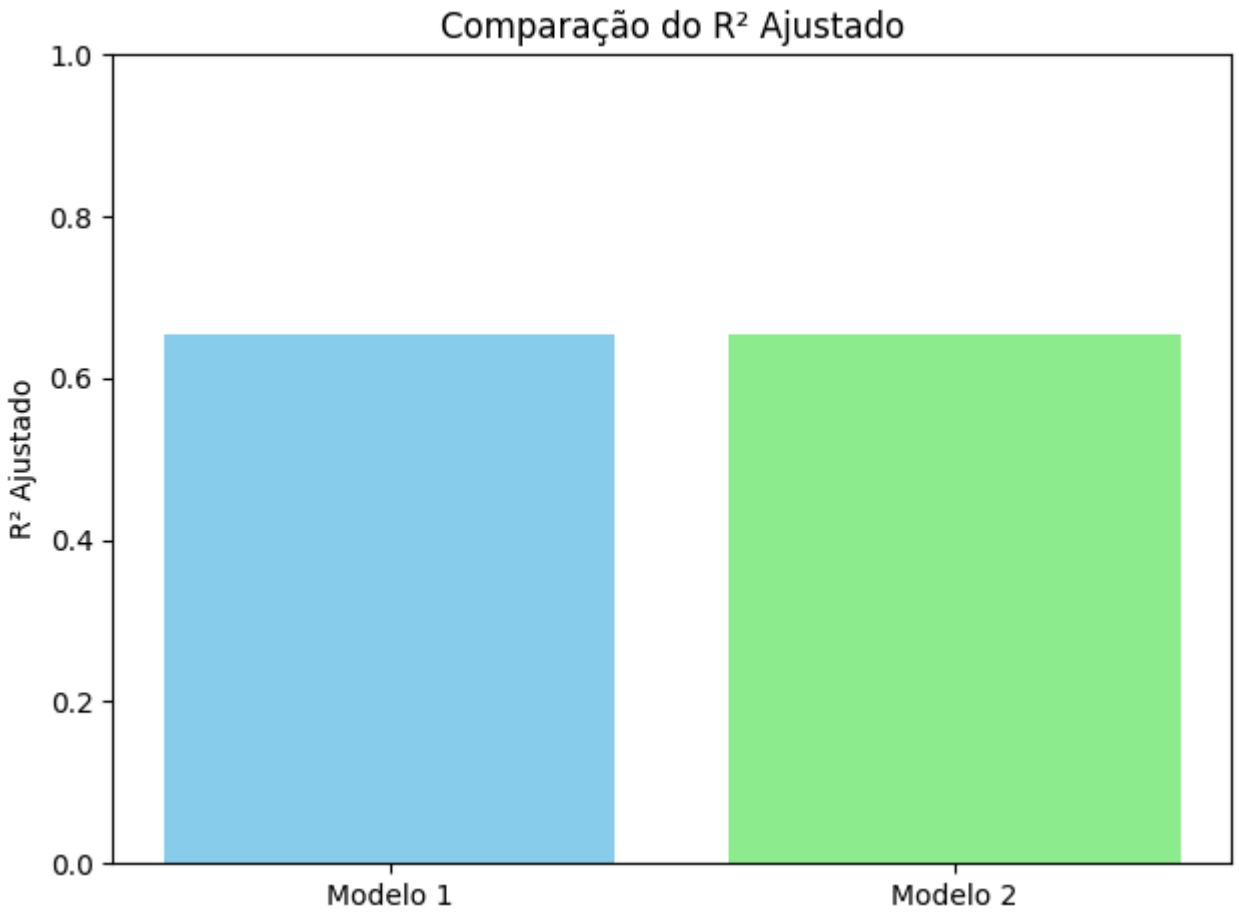

Figura 2 – Gráfico de Comparação do R^2 ajustado

2. Sugestões Práticas

Com base nos resultados do Modelo 2, as seguintes sugestões podem ser feitas para otimizar o tempo de resposta dos sistemas:

- Aumentar o número de núcleos da CPU (”cpu_cores”): Esta variável apresentou o impacto negativo mais substancial e estatisticamente significativo no tempo de resposta. Cada núcleo adicional está associado a uma redução considerável no tempo de resposta.
- Expandir a memória RAM (”ram_gb”): O aumento da RAM também demonstrou ser eficaz na redução do tempo de resposta, sendo uma melhoria recomendada, especialmente para sistemas com alta demanda.

As demais variáveis analisadas, incluindo latência de rede, capacidade de armazenamento e as características categóricas (sistema operacional, tipo de HD, e as demais categorias de processador após o ajuste), não mostraram um impacto estatisticamente significativo no tempo de resposta dentro deste modelo específico e com este conjunto de dados.

3. Considerações Finais

A análise realizada permitiu identificar fatores chave que influenciam o tempo de resposta de sistemas computacionais. Embora o modelo ajustado explique uma porção considerável da variabilidade (aproximadamente 65.6%), a presença de heterocedasticidade sugere que futuras análises poderiam explorar transformações de variáveis ou o uso de regressão com erros robustos para aprimorar ainda mais as inferências.

Anexo com prints dos outputs

```
📁 Arquivo 'dataset_11.csv' carregado com sucesso!

🔍 Verificação de valores ausentes por coluna (antes do tratamento):
latencia_ms         18
armazenamento_tb    18
tipo_hd             18
tempo_resposta       1
dtype: int64

🔧 Preenchendo colunas numéricas com a mediana...
→ Coluna 'latencia_ms' preenchida com a mediana (164.59).
→ Coluna 'armazenamento_tb' preenchida com a mediana (0.51).
→ Coluna 'tempo_resposta' preenchida com a mediana (85.31).

🔧 Preenchendo colunas categóricas com a moda...
→ Coluna 'tipo_hd' preenchida com a moda ('HDD').

📊 Verificação final de valores ausentes após a imputação:
✅ Todos os dados ausentes foram preenchidos com sucesso.

📈 Estatísticas Descritivas - Variáveis Numéricas:
       cpu_cores      ram_gb  latencia_ms  armazenamento_tb  tempo_resposta
count  185.000000  185.000000   185.000000        185.000000      185.000000
mean     8.572973   25.383784   157.138919          0.888173      107.237595
std      4.048124   21.629005    75.645088          0.665306       66.101246
...
💾 Arquivo final salvo como: 'dataset_11_limpo_replicado_v2.csv'

✅ Colunas Numéricas Identificadas: ['cpu_cores', 'ram_gb', 'latencia_ms', 'armazenamento_tb', 'tempo_resposta']
✅ Colunas Categóricas Identificadas: ['sistema_operacional', 'tipo_hd', 'tipo_processador']
```

Figura 3 – Print do output da célula "Parte I - Análise Estatística"

```
=--Resumo Estatístico do Modelo 1 (com todas as variáveis)--=
                            OLS Regression Results
==============================================================================
Dep. Variable:         tempo_resposta   R-squared:                       0.669
Model:                            OLS   Adj. R-squared:                  0.654
Method:                 Least Squares   F-statistic:                     44.52
Date:                Tue, 06 May 2025   Prob (F-statistic):           1.93e-38
Time:                        22:14:32   Log-Likelihood:                -935.03
No. Observations:                 185   AIC:                             1888.
Df Residuals:                     176   BIC:                             1917.
Df Model:                           8
Covariance Type:            nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                            244.7960     12.919     18.949      0.000     219.300     270.292
cpu_cores                        -11.4659      0.711    -16.132      0.000     -12.869     -10.063
ram_gb                            -1.4270      0.141    -10.140      0.000      -1.705      -1.149
latencia_ms                       -0.0126      0.040     -0.312      0.756      -0.092       0.067
armazenamento_tb                  -3.7331      4.400     -0.848      0.397     -12.416       4.950
sistema_operacional_MacOS         -0.3824      3.890     -0.098      0.922      -8.059       7.294
sistema_operacional_Windows        5.4339      7.156      0.759      0.449      -8.690      19.557
tipo_hd_SSD                       -0.8341      5.826     -0.143      0.886     -12.331      10.663
tipo_processador_Apple Silicon    -0.3824      3.890     -0.098      0.922      -8.059       7.294
tipo_processador_Intel             3.0850      7.076      0.436      0.663     -10.880      17.050
...
=--Categorias de Referência Utilizadas (Dummy Encoding)--=
- sistema_operacional: A categoria de referência é 'Linux'
- tipo_hd: A categoria de referência é 'HDD'
- tipo_processador: A categoria de referência é 'AMD'
```

Figura 4 – Print do output da célula "Parte II – Modelo e Diagnóstico"

```
=-=-Cálculo do Fator de Inflação da Variância (VIF)-=-=
                        variavel       vif
0                      cpu_cores  4.423661
1                         ram_gb  2.106876
2                     latencia_ms  3.947534
3                armazenamento_tb  2.517450
4       sistema_operacional_MacOS       inf
5     sistema_operacional_Windows  1.746126
6                     tipo_hd_SSD  1.662376
7  tipo_processador_Apple Silicon       inf
8          tipo_processador_Intel  2.066306
=-=-Avaliação da Colinearidade com Base no VIF-=-=
Foram identificadas variáveis com VIF infinito — indicando colinearidade perfeita:
                        variavel  vif
4       sistema_operacional_MacOS  inf
7  tipo_processador_Apple Silicon  inf
Recomenda-se a exclusão de pelo menos uma dessas variáveis para estabilizar o modelo.

VIF - Variáveis ordenadas do maior para o menor valor:
                        variavel       vif
4       sistema_operacional_MacOS       inf
7  tipo_processador_Apple Silicon       inf
0                      cpu_cores  4.423661
2                     latencia_ms  3.947534
3                armazenamento_tb  2.517450
1                         ram_gb  2.106876
8          tipo_processador_Intel  2.066306
5     sistema_operacional_Windows  1.746126
6                     tipo_hd_SSD  1.662376
```

Figura 5 – Print do output da célula "Diagnóstico de Multicolinearidade"

```
Figura exportada como: residuos_vs_estimados_modelo1.png
=-=-Aplicação do Teste de Breusch-Pagan (Modelo 1)-=-=
Estatística LM (Lagrange Multiplier): 18.2647
p-valor associado (LM): 0.0322
Estatística F: 2.4099
p-valor associado (F): 0.0171
Conclusão: Como o p-valor F (0.0171) é inferior a 0.05, rejeita-se a hipótese nula.
Evidencia-se a presença de heterocedasticidade no modelo.
```

Figura 6 – Print do output da célula "Diagnóstico de Heterocedasticidade"

```
=-=-Resumo do Modelo 2 (com 'Apple Silicon' removido)-=-=
                            OLS Regression Results
==============================================================================
Dep. Variable:         tempo_resposta   R-squared:                       0.669
Model:                            OLS   Adj. R-squared:                  0.654
Method:                 Least Squares   F-statistic:                     44.52
Date:                Tue, 06 May 2025   Prob (F-statistic):           1.93e-38
Time:                        22:16:47   Log-Likelihood:                -935.03
No. Observations:                 185   AIC:                             1888.
Df Residuals:                     176   BIC:                             1917.
Df Model:                           8
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                        244.7960     12.919     18.949      0.000     219.300     270.292
cpu_cores                    -11.4659      0.711    -16.132      0.000     -12.869     -10.063
ram_gb                        -1.4270      0.141    -10.140      0.000      -1.705      -1.149
latencia_ms                   -0.0126      0.040     -0.312      0.756      -0.092       0.067
armazenamento_tb              -3.7331      4.400     -0.848      0.397     -12.416       4.950
sistema_operacional_MacOS     -0.7647      7.779     -0.098      0.922     -16.118      14.588
sistema_operacional_Windows    5.4339      7.156      0.759      0.449      -8.690      19.557
tipo_hd_SSD                   -0.8341      5.826     -0.143      0.886     -12.331      10.663
tipo_processador_Intel         3.0850      7.076      0.436      0.663     -10.880      17.050
==============================================================================
...
O Modelo 2 apresenta desempenho semelhante ao Modelo 1 em termos de R² ajustado.
Além disso, o Modelo 2 elimina o problema crítico de multicolinearidade, resultando em um número de condição bem mais estável.
A variável descartada ('tipo_processador_Apple Silicon') não era estatisticamente relevante no Modelo 1 (p=0.922).
Portanto, o Modelo 2 se destaca por ser mais enxuto e confiável.
```

Figura 7 – Print do output da célula "Parte III – Análise Crítica"