Universidade Estadual de Santa Cruz - UESC
Departamento de Ciências Exatas e Tecnológicas - DCET
CET083 - Probabilidade e Estatística
Curso de Ciência da Computação
Prof. José Cláudio Faria

Prova Prática
Pontuação total: 10
Prazo: 09/06/2022 - A ser marcado em aula

Nome: Igor Sousa dos Santos Santana	Matrícula: 201920070
Nome: Matheus Santos Silva	Matrícula: 201920235
Nome:	Matrícula:
Nome:	Matrícula:
Nome:	Matrícula:

Na resolução da prova use as funções geradoras de dados (`gerar_dados`, `gerar_dados_rl` e `gerar_tdf`), todas disponíveis no arquivo gerar_dados_v10.R na página da disciplina.

A função `gerar_dados` foi programada para gerar uma amostra aleatória estratificada de pessoas do município X. As variáveis aleatórias de interesse são: `Y1` (medido em *un*), `Y2` (medido em *un*) e `Sexo`. Adicionalmente, assuma que `Y1` e `Y2` não podem assumir valores reais negativos. Os dados são fictícios e tem finalidades exclusivamente didáticas para fins de avaliação prática em análise de dados.

Realizar a análise exploratória dos dados com respostas às seguintes questões:

1 AED: Apresentações tabulares e gráficas (2.0)

1.1 Diagrama de caixa (boxplot) para Y1 e Y2 (1.0)

1. (0.5) Antes e após a eliminação de possíveis outliers[1];
2. (0.5) Após a eliminação de possíveis outliers[2].

1.2 Para Y1 (1.0)

1. (0.5) Uma apresentação tabular contendo apenas as freqüências: absoluta (Fi), relativa (Fr, %) e acumulada (Fac, %), nessa ordem[2];
2. (0.5) Histograma e o polígono de freqüência acumulada dos dados[2].

2 AED: Medidas estatísticas básicas (3.0)

2.1 AED: Medidas determinadas a partir dos vetores (1.5)

Para as variáveis Y1 e Y2 elaborar apresentações tabulares[2] contendo as seguintes estimativas:

1. (0.5) Tendência central: média, mediana e moda;
2. (0.5) Posição: quartis e decis;
3. (0.5) Dispersão: amplitude total, variância, desvio padrão e coeficiente de variação.

2.2 AED: Medidas determinadas a partir de apresentações tabulares (1.5)

A função `gerar_tdf` foi programada para gerar uma tabela de distribuição de frequências do tipo comum, dessas que se encontra em publicações. Considere que esta tabela descreve um assunto de seu interesse - publicado - e que é necessário determinar as medidas estatísticas básicas com finalidades de entendimento e comparações.

Elabore uma apresentação tabular contendo:

1. (0.5) Tendência central: média, mediana e moda;
2. (0.5) Posição: quartis e decis;
3. (0.5) Dispersão: amplitude total, variância, desvio padrão e coeficiente de variação.

[1] Não distinguindo sexo
[2] Para cada sexo: M seguido de F

3 AED: Medidas estatísticas de associação e regressão linear (4.0)

Considere os dados gerados pela função `gerar_dados` para a questão subsequente:

3.1 Associação (1.5)

1. (0.5) Estimativas: covariância e correlação linear simples[2];
2. (0.5) Diagramas de dispersão dos dados[2,3];
3. (0.5) Um estudo semelhante foi realizado em um outro município, por outras pessoas. Contudo, as unidades de medida usadas foram: `Y1` $(100 * un)$ e `Y2` $(100 * un)$.

   Para comparar associações entre as variáveis de ambos os estudos, qual seria a medida estatística recomendada? Justifique.

3.2 Regressão linear (2.5)

Considere os dados gerados pela função `gerar_dados_rl` como uma amostra de um estudo da influência de uma variável fixa `X` (medido em un) sobre uma variável aleatória `Y` (medido em $un.dia^{-1}$).

Os dados são fictícios e tem finalidades exclusivamente didáticas para fins de avaliação prática em análise quantitativa de dados.

1. (1.0) Ajuste aos dados dois modelos de regressão linear: polinômios de grau I e II (ambos não forçado para a origem);
2. (0.5) Apresente um diagrama de dispersão dos dados[4] com o melhor modelo.
3. (0.5) Qual modelo melhor explica o fenômeno em estudo? Justifique com fundamentação estatística.
4. (0.5) Pelos critérios de ajustamento e escolha de modelos vistos em aula, os coeficientes de determinação (r^2) de modelos lineares ajustados (forçados e não forçados para a origem) são comparáveis? Justifique com fundamentação estatística.

4 Contextualização (1.0)

Localize um artigo científico (`periódico Qualis A ou B`) em área de seu interesse no qual a análise exploratória de dados (AED - possivelmente com medidas de associação e uso de regressão linear como modelo explicativo) teve papel preponderante. Discuta o artigo `com ênfase nos recursos da AED usados e também na adequação das normas básicas das apresentações gráficas e tabulares adotada pelo periódico`.

Observações:

- Para possibilitar a correção, anexe esta prova `devidamente preenchida` na primeira página das respostas.
- As normas para apresentações gráficas e tabulares são obrigatórias, serão observadas e corrigidas.
- Sugere-se (mas não é obrigatório) o uso do ambiente R na resolução das questões propostas.
- Cada hora de atrazo na entrega da avaliação implica na perda de 25%. Portanto, após 4 horas não entregue.

[3] Considere Y2 no eixo das ordenadas e Y1 no eixo das abscissas
[4] Considere Y no eixo das ordenadas e X no eixo das abscissas

Universidade Estadual de Santa Cruz - UESC
Departamento de Ciências Exatas e Tecnológicas - DCET
CET083 - Probabilidade e Estatística
Curso de Ciência da Computação
Prof. José Claudio Faria

Nomes: Igor Sousa dos Santos Santana, Matheus Santos Silva
Mátriculas: 201920070 201920235 201920235

1 AED: Apresentações tabulares e gráficas (2.0)

1.1 (1.0) Diagrama de caixa (boxplot) para Y1 e Y2

1.1.1 (0.5) Antes e após a eliminação de possíveis outliers - sem distinção de sexo

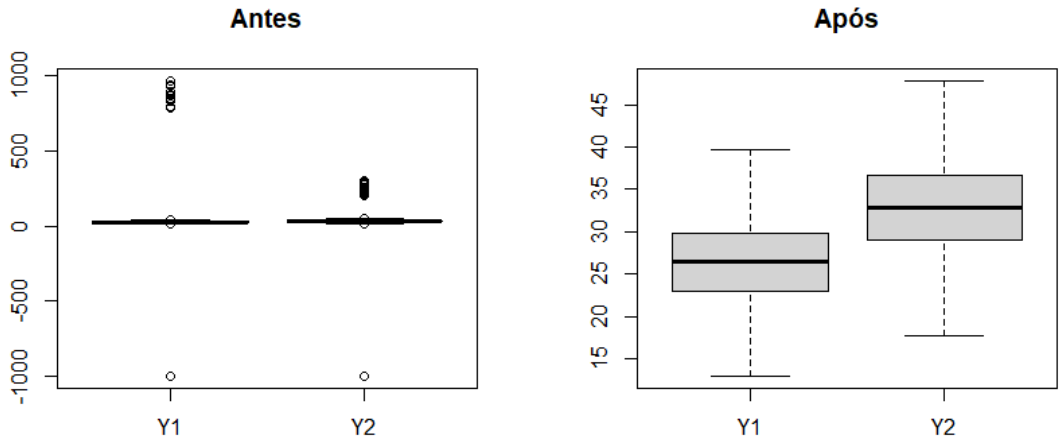


Figura 1 – Diagrama de caixa de Y1 (un) e Y2 (un) antes e após a eliminação de outliers, UESC/BA - 2022.

1.1.2 (0.5) Após a eliminação de possíveis outliers - com distinção de sexo

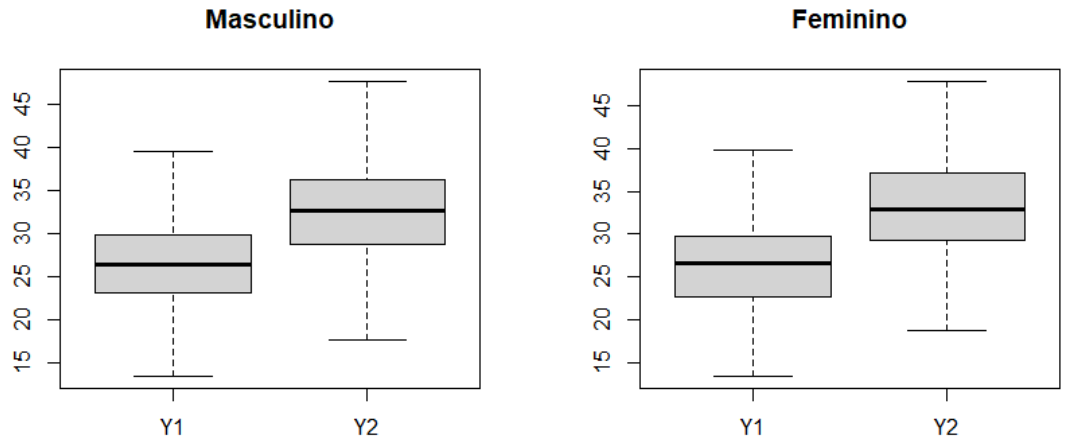


Figura 2 – Diagrama de caixa de Y1 (un) e Y2 (un) (sexo masculino e feminino, respectivamente), UESC/BA - 2022.

1.2 (1.0) Para Y1

1.2.1 (0.5) Apresentações tabulares

Tabela 1 – Tabela de distribuição de frequência de Y1 (un) (sexo masculino), UESC/BA - 2022

Class limits	f	rf	rf(%)	cf	cf(%)
[13.316,15.543)	12	0.01	0.92	12.00	0.92
[15.543,17.771)	25	0.02	1.91	37.00	2.83
[17.771,19.998)	71	0.05	5.43	108.00	8.26
[19.998,22.226)	140	0.11	10.70	248.00	18.96
[22.226,24.453)	201	0.15	15.37	449.00	34.33
[24.453,26.681)	238	0.18	18.20	687.00	52.52
[26.681,28.909)	214	0.16	16.36	901.00	68.88
[28.909,31.136)	174	0.13	13.30	1075.00	82.19
[31.136,33.364)	122	0.09	9.33	1197.00	91.51
[33.364,35.591)	73	0.06	5.58	1270.00	97.09
[35.591,37.819)	27	0.02	2.06	1297.00	99.16
[37.819,40.047)	11	0.01	0.84	1308.00	100.00

Tabela 2 – Tabela de distribuição de frequência de Y1 (un) (sexo feminino), UESC/BA - 2022

Class limits	f	rf	rf(%)	cf	cf(%)
[13.306, 15.745)	11	0.02	1.97	11.00	1.97
[15.745, 18.184)	27	0.05	4.85	38.00	6.82
[18.184, 20.623)	50	0.09	8.98	88.00	15.80
[20.623, 23.063)	62	0.11	11.13	150.00	26.93
[23.063, 25.502)	91	0.16	16.34	241.00	43.27
[25.502, 27.941)	87	0.16	15.62	328.00	58.89
[27.941, 30.380)	106	0.19	19.03	434.00	77.92
[30.380, 32.820)	70	0.13	12.57	504.00	90.48
[32.820, 35.259)	34	0.06	6.10	538.00	96.59
[35.259, 37.698)	16	0.03	2.87	554.00	99.46
[37.698, 40.137)	3	0.01	0.54	557.00	100.00

1.2.2 (0.5) Histograma e polígono de frequência acumulada

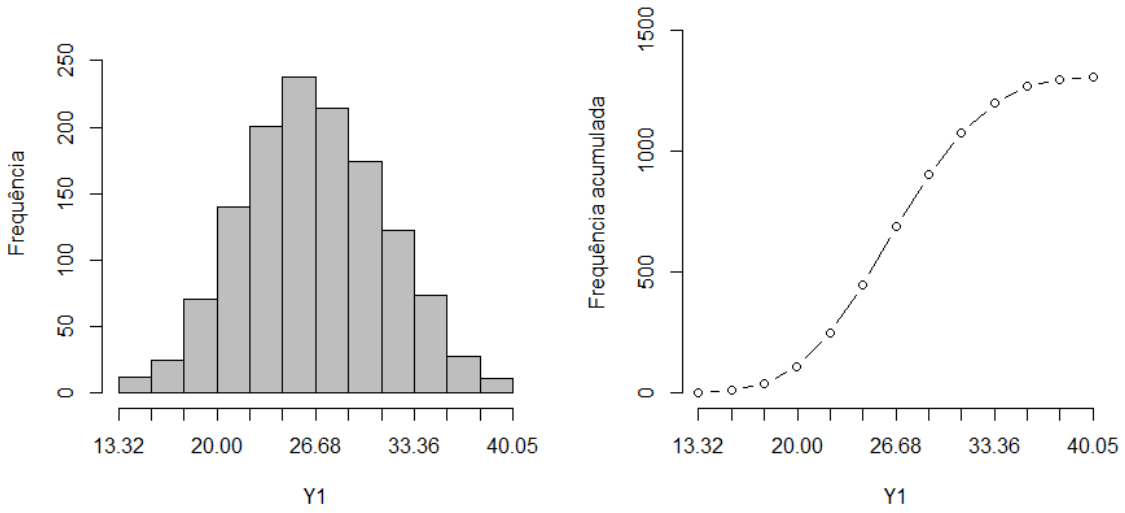


Figura 3 – Histograma e polígono de frequência acumulada de Y1 (un) (sexo masculino), UESC/BA - 2022.

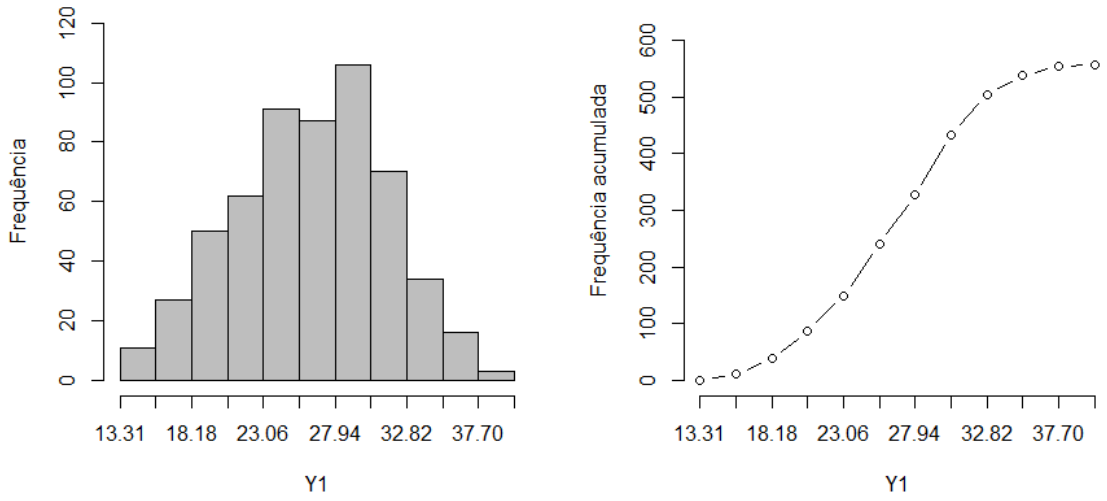


Figura 4 – Histograma e polígono de frequência acumulada de Y1 (un) (sexo feminino), UESC/BA - 2022.

2 AED: medidas estatísticas básicas (3.0)

2.1 (1.5) AED: Medidas determinadas a partir dos vetores

2.1.1 (0.5) Tendência central

Tabela 3 – Medidas de tendência central (sexo masculino), UESC/BA - 2022

	n	m	md	mo
Y1	1308	26.52	26.38	24.54
Y2	1308	32.63	32.63	35.37

Tabela 4 – Medidas de tendência central (sexo feminino), UESC/BA - 2022

	n	m	md	mo
Y1	557	26.28	26.68	19.45
Y2	557	33.27	32.90	30.95

2.1.2 (0.5) Posição

Tabela 5 – Quartis dos usuários (sexo masculino), UESC/BA – 2022

	25%	50%	75%
Y1	23.23	26.38	29.83
Y2	28.78	32.63	36.36

Tabela 6 – Quartis dos usuários (sexo feminino), UESC/BA – 2022

	25%	50%	75%
Y1	22.73	26.68	29.79
Y2	29.34	32.90	37.07

Tabela 7 – Decis dos usuários (sexo masculino), UESC/BA – 2022

	10%	20%	30%	40%	50%	60%	70%	80%	90%
Y1	20.45	22.42	23.88	25.15	26.38	27.67	29.04	30.39	32.89
Y2	25.68	27.90	29.71	31.13	32.63	34.17	35.45	37.26	39.72

Tabela 8 – Decis dos usuários (sexo feminino), UESC/BA – 2022

	10%	20%	30%	40%	50%	60%	70%	80%	90%
Y1	19.40	21.53	23.59	25.07	26.68	28.03	29.25	30.64	32.68
Y2	26.39	28.34	30.13	31.24	32.90	34.40	36.29	38.31	40.71

2.1.3 (0.5) Dispersão

Tabela 9 – Dispersão dos usuários (sexo masculino), UESC/BA – 2022

	a.t	variância	d.padrão	c.v
Y1	26.20	22.66	4.76	17.95
Y2	29.93	28.79	5.37	16.45

Tabela 10 – Dispersão dos usuários (sexo feminino), UESC/BA – 2022

	a.t	variância	d.padrão	c.v
Y1	26.30	26.02	5.10	19.41
Y2	29.01	31.72	5.63	16.93

2.2 AED: Medidas determinadas a partir de apresentações tabulares (1.5)

Tabela 11 – Tabela de distribuição de frequência reconstruída de publicação, UESC/BA – 2022

Class limits	f	rf	rf(%)	cf	cf(%)
[10, 20)	7	0.03	3.04	7	3.04
[20, 30)	17	0.07	7.39	24	10.43
[30, 40)	27	0.12	11.74	51	22.17
[40, 50)	37	0.16	16.09	88	38.26
[50, 60)	47	0.20	20.43	135	58.70
[60, 70)	39	0.17	16.96	174	75.65
[70, 80)	28	0.12	12.17	202	87.83
[80, 90)	19	0.08	8.26	221	96.09
[90, 100)	9	0.04	3.91	230	100.00

2.2.1 (0.5) Tendência central

Tabela 12 – Medidas de tendência central

	m	md	mo
medida	55.78	55.74	55.56

2.2.2 (0.5) Posição

Tabela 13 – Medidas de posição: quartis

	25%	50%	75%
quartil	41.76	55.74	69.62

Tabela 14 – Medidas de posição: decis

	10%	20%	30%	40%	50%	60%	70%	80%	90%
decil	29.41	38.15	44.86	50.85	55.74	60.77	66.67	73.57	82.63

2.2.3 (0.5) Dispersão

Tabela 15 – Medidas de dispersão

	a.t	variância	d.padrão	c.v
medida	75.50	381.92	19.54	35.03

3 AED: Medidas estatísticas de associação e regressão linear (4.0)

3.1 (1.5) Associação

3.1.1 (0.5) Estimativas: covariância e correlação linear simples

Tabela 16 – Matriz de variâncias e covariâncias (sexo masculino), UESC/BA – 2022

	Y1	Y2
Y1	22.66	20.38
Y2	20.38	28.79

Tabela 17 – Matriz de variâncias e covariâncias (sexo feminino), UESC/BA – 2022

	Y1	Y2
Y1	26.02	-26.61
Y2	-26.61	31.72

Tabela 18 – Matriz de correlações lineares simples (sexo masculino), UESC/BA – 2022

	Y1	Y2
Y1	1.00	0.80
Y2	0.80	1.00

Tabela 19 – Matriz de correlações lineares simples (sexo feminino), UESC/BA – 2022

	Y1	Y2
Y1	1.00	-0.93
Y2	-0.93	1.00

3.1.2 (0.5) Diagrama de dispersão dos dados

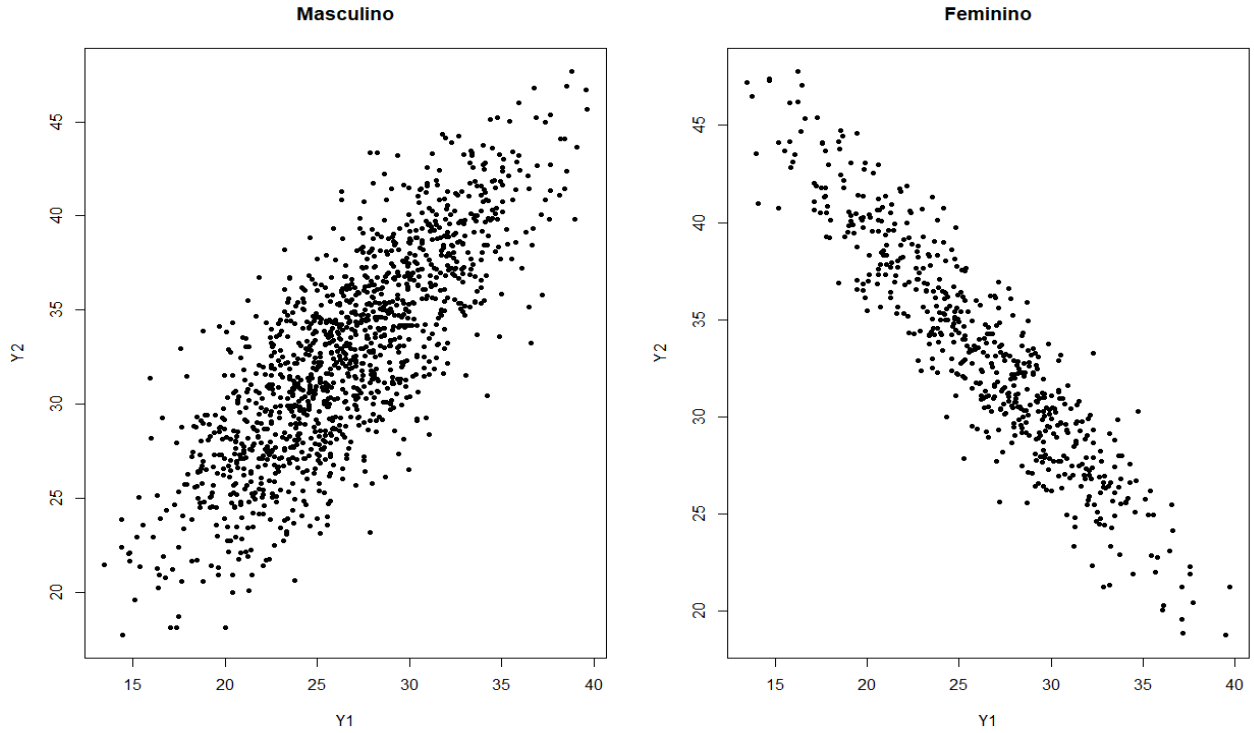


Figura 5 – Diagrama de dispersão de Y1 (un) e Y2 (un) (sexo masculino e feminino, respectivamente), UESC/BA - 2022.

3.1.3 (0.5) Comparação de estudos semelhantes

A medida de estatística recomendada seria o coeficiente de correlação, pois os estudos estariam em unidades de medidas diferentes e o coeficiente de correlação é uma medida que não é influenciada pelas unidades de medidas das variáveis, nem pelo tamanho da amostra caso esses estudos apresentem tamanhos diferentes.

3.2 Regressão linear (2.5)

3.2.1 (1.0) Ajustamento

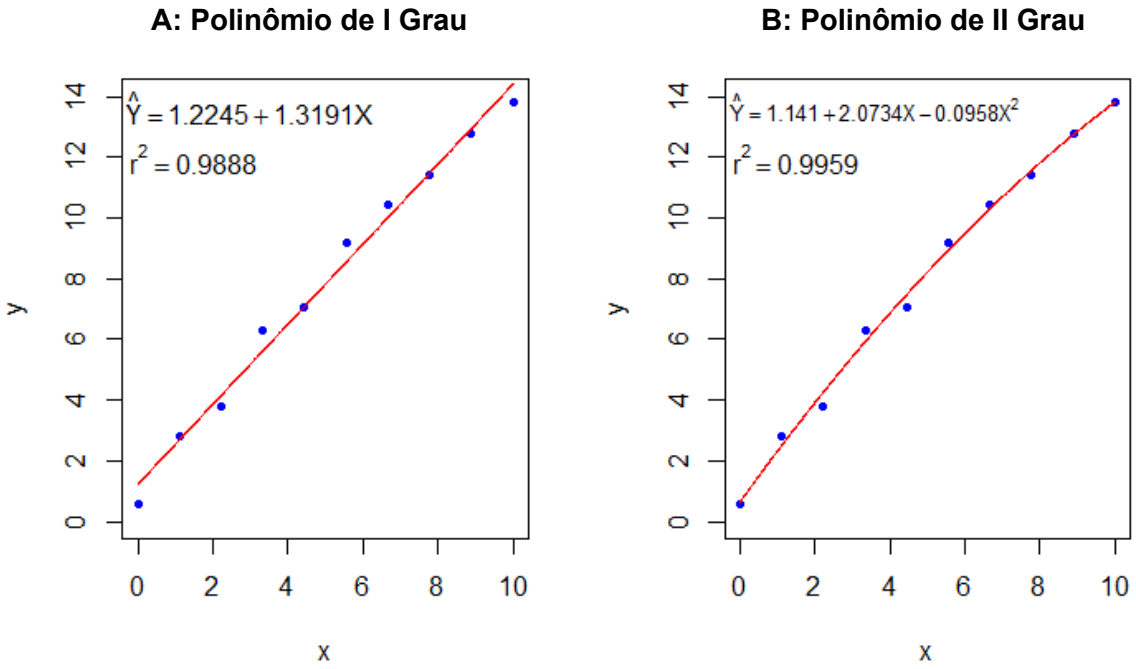
Tabela 20 – Polinômio grau I, UESC/BA - 2022

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.22	0.30	4.15	0.00
X	1.32	0.05	26.52	0.00

Tabela 21 – Polinômio grau II, UESC/BA - 2022

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.63	0.26	2.48	0.04
X	1.72	0.12	14.41	0.00
I(X^2)	-0.04	0.01	-3.47	0.01

3.2.2 (0.5) Diagrama de dispersão com modelos ajustados



3.2.3 (0.5) Qual modelo melhor explica o fenômeno em estudo?

O modelo que melhor explica o fenômeno é o Polinômio de II Grau, pois os pontos ficam mais agrupados, mais próximos, ou seja, o coeficiente de correlação (r^2) é maior, sendo mais ajustada para o caso estudado.

3.2.4 (0.5) Critérios de ajustamento e escolha de modelos

Os coeficientes de determinação de modelos lineares ajustados não são comparáveis, pois o modelo linear ajustado não forçado possui um somatório de desvios "ε" em sua composição, enquanto o modelo linear forçado assume esse erro como sendo 0 e força o gráfico a passar pela origem.

4 Contextualização (1.0)

Artigo: REVISÃO SISTEMÁTICA: APLICAÇÃO DE REDES NEURAIS PARA PREVISÃO DO CONSUMO DE ENERGIA ELÉTRICA

O artigo utilizado nesta questão tem como principal objetivo realizar uma revisão sistemática de 15 artigos que utilizaram redes neurais para a previsão de consumo de energia elétrica, nele são feitas análises a respeito das entradas, configurações das redes e performances das redes utilizadas nos 15 artigos. Para apresentar os resultados, o autor do artigo utiliza 4 tabelas e 5

gráficos. A apresentação dos resultados é feita em duas partes, avaliação da qualidade da metodologia e avaliação das características dos modelos. Na fase de avaliação da qualidade da metodologia é inicialmente feita com as tabelas 1 e 2, onde são mostrados, respectivamente, informações sobre ano e periódico de publicação, e avaliação do Qualis, SJR e JCR obtidas pelos artigos, e em seguida é apresentado na figura 2 um gráfico de pizza com a distribuição de frequência das avaliações dos artigos mostrados na tabela 2, e através do gráfico pode ser observado que quase metade dos artigos obtiveram boas avaliações, A1. A segunda etapa, avaliação das características dos modelos, inicia com a figura 3 mostrando a distribuição de frequência das entradas utilizadas nos artigos em um gráfico em colunas, e nele são visualizados as variáveis de entrada agrupadas nas categorias data, consumo, climáticos, população, financeiro, imagens e ambientais. A apresentação seguinte é feita na tabela 3 que mostra como foi feita a divisão do conjunto de dados entre treinamento e validação em cada artigo. As duas apresentações seguintes são sobre a quantidade de camadas ocultas e neurônios utilizados pelas redes dos artigos e estão, respectivamente, na figura e tabela 4. As duas últimas figuras apresentadas são um gráfico em colunas e um gráfico de pizza, que nessa ordem apresentam uma distribuição de frequência das métricas utilizadas para avaliar os modelos e uma avaliação das adequações dos modelos ao problema proposto. Em relação as normas adotadas pelo periódico, todos os gráficos estão de acordo com as normas, com legendas abaixo da figura espaçado por um espaço simples e apresenta todos os elementos mínimos. Contudo para as tabelas apesar de apresentarem todos os elementos mínimos ocorre um erro de separação do título e demais elementos em páginas distintas nas tabelas 3 e 4.

Link do Artigo: https://revistas2.uepg.br/index.php/ret/article/view/17956