



O CHATGPT NA CUSTOMIZAÇÃO DE DADOS PARA ENSINO DE DATA SCIENCE

DOI: 10.37702/2175-957X.COBENGE.2024.5158

Autores: ITALO PINTO RODRIGUES

Resumo: Este artigo investiga a aplicação de metodologias ativas de ensino no contexto da educação em estatística, destacando o uso de ferramentas tecnológicas avançadas, como o ChatGPT. A capacidade de gerar dados customizados por meio do ChatGPT é explorada, proporcionando um recurso valioso para simulações estatísticas e experimentos pedagógicos. O método proposto é dividido em três etapas: interação com o ChatGPT para obtenção de dados, preparação e organização dos dados, e modelagem estatística. A interação com o ChatGPT é detalhada, demonstrando como comandos específicos podem ser utilizados para gerar conjuntos de dados adaptados às necessidades dos alunos. A modelagem estatística inclui o uso de distribuições como a Weibull para analisar a variabilidade e a incerteza dos dados. Um estudo de caso ilustra a aplicação prática do método, confirmando a eficácia da ferramenta em apoiar o ensino de estatística. Os resultados mostram que a integração de IA e metodologias ativas pode transformar o ensino de disciplinas quantitativas, promovendo uma aprendizagem mais envolvente e contextualizada.

Palavras-chave: Ciência de Dados, Inteligência Artificial, ChatGPT, Aprendizagem Personalizada, Modelagem Estatística

O CHATGPT NA CUSTOMIZAÇÃO DE DADOS PARA ENSINO DE DATA SCIENCE

1 INTRODUÇÃO

A capacidade de coletar, analisar e interpretar dados é vital para a competitividade das empresas em um mundo cada vez mais orientado por dados (NUNES; SANTOS; ROCHA, 2023). Desse modo, a estatística é crucial para entender a variabilidade e a incerteza nos dados coletados, fornecendo as ferramentas necessárias para essa compreensão.

O ensino tradicional desta disciplina era centrado no docente. Contudo, essa abordagem está sendo substituída por métodos que promovem a participação ativa dos alunos no processo educativo (PRINCE; FELDER, 2006). As metodologias ativas de ensino, que promovem uma aprendizagem mais envolvente e dinâmica, estão reformulando o cenário educacional. Estratégias como aprendizagem baseada em projetos, gamificação e sala de aula invertida estão sendo implementadas para aumentar a motivação e a compreensão dos alunos (GOMEZ-DEL RIO; RODRIGUEZ, 2022).

Portanto, sendo a estatística uma habilidade essencial para os futuros profissionais de dados, tornar o ensino desta disciplina atrativo aos estudantes é imprescindível (SHAH, 2023).

No contexto das metodologias ativas de ensino, a convergência entre pedagogia e tecnologia avançada é exemplificada pela ascensão da Inteligência Artificial (IA) e dos modelos de linguagem. O ChatGPT, resultante dessa interseção, utiliza os avanços da IA para aprimorar a entrega educacional, estabelecendo um novo paradigma na integração de tecnologia com práticas pedagógicas inovadoras (BROWN et al., 2020; JAVAID et al., 2023).

Dessa forma, identificou-se que o ChatGPT possui a capacidade de gerar dados customizados, o que é particularmente valioso para simulações estatísticas e experimentos pedagógicos. A criação de conjuntos de dados específicos e adaptados às necessidades de estudantes, educadores ou pesquisadores permite uma abordagem mais prática e contextualizada do ensino de estatística.

O principal objetivo deste artigo é propor uma abordagem que permita ao usuário gerar dados customizados a partir do ChatGPT, de modo a apoiar o ensino de Estatística. O uso do método proposto será ilustrado a partir de um estudo de caso, proposto por Rodrigues e Campos, (2023).

2 METODOLOGIA

Para a obtenção do conjunto de dados, o método proposto se divide em três etapas principais: i) interação com o ChatGPT para obtenção de dados; ii) preparação dos dados, isto é, sua organização; e, finalmente, iii) a modelagem estatística. Uma das premissas adotadas neste estudo parte do pressuposto de que o usuário está registrado no ChatGPT 3.5 e tem conhecimento em como usar a ferramenta.

2.1 Interação com ChatGPT

A interação inicial se caracteriza por indicar ao ChatGPT como os dados devem ser enviados. De modo a obter resultados que permitam a reprodutibilidade, recomenda-se utilizar os comandos apresentados na Tabela 1. Os itens em negrito são os parâmetros que podem ser modificados a fim de que sejam gerados conjuntos de dados customizados ao contexto de cada estudante. Desta maneira, é possível criar uma generalização ao trabalho proposto por Rodrigues e Campos, (2023).

Tabela 1 – Comandos a serem enviados ao ChatGPT.

Comandos
Você precisa gerar, no formato de tabela, dados de [CONTEXTO] de uma empresa:
A empresa [CARACTERÍSTICA] .
Você deve indicar o [ITEM QUE SERÁ MODELADO] .
Além dessas informações você deve criar pelo menos duas colunas com mais dados sobre essa [CONTEXTO] .
Cada coluna deve ter 121 linhas com as informações de cada variável.
Você não pode indicar nenhum código para gerar a tabela em ambiente externo.
Os dados de [ITEM QUE SERÁ MODELADO] devem ter uma distribuição [TIPO DE DISTRIBUIÇÃO] , tente não ficar repetindo valores. Seja aleatório.
Você também deverá gerar as explicações de cada uma das colunas da tabela, considerando essa produção hipotética.
A tabela deve apresentar as 121 linhas sem resumo.
Você deve explicar os dados e a relação entre eles no final.

Fonte: Os Autores (2024)

O detalhamento é parte fundamental da interação com o ChatGPT. No entanto, a resposta do ChatGPT pode variar e nem todas as instruções serem atendidas, como por exemplo a instrução, "A tabela deve apresentar as 121 linhas sem resumo". A Figura 1 apresenta uma tabela resumida, sem considerar as 121 linhas solicitadas. Neste caso, pode ser que sejam necessárias outras interações para reafirmar as solicitações. A Figura 2 apresenta a instrução enviada para que fosse gerada a tabela completa.

Figura 1 – Tabela com 6 amostras em vez de 121.

Nº	Tempo até Falha (s)	Quantidade Produzida	Qualidade da Peça
1	152	76	Alta
2	98	49	Média
3	210	105	Alta
4	67	33	Baixa
5	134	67	Alta
...
121	176	88	Média

Fonte: Rodrigues e Campos, (2023).



Figura 2 – Prompt para melhorar a visualização dos dados.

Agora apresente a tabela com 121 linhas de amostragens sem apresentar "...” para resumir o conteúdo. Ao todo a tabela deverá possuir 122 linhas, sendo 1 para o cabeçalho e 121 para as amostras.

Fonte: Rodrigues e Campos, (2023).

Portanto, alguns ajustes finos, como o ajuste apresentado na Figura 2 podem ser necessários.

2.2 Interação com ChatGPT

Após utilizar o ChatGPT para gerar dados destinados à modelagem, recomenda-se o uso de softwares como Excel, Planilhas do Google ou programas mais avançados como Matlab, Python ou R. Dessa forma, os dados podem ser manipulados em um ambiente adequado para modelagem, conforme as especificações de cada software.

É importante destacar que, dependendo das configurações do navegador e do ChatGPT, pode ser necessário realizar alguns ajustes ao transferir os dados para esses softwares, como a substituição do separador decimal de ponto para vírgula e a concatenação de várias tabelas geradas pela ferramenta para obter o número desejado de amostras.

2.3 Modelagem Estatística

A modelagem estatística foi realizada considerando os dados em rol, isto é, ordenados do menor para o maior e foram realizadas as análises apresentadas na Tabela 2.

Tabela 2 – Modelagem estatística.

Item	Descrição	Equação
Média aritmética simples	A média frequentemente utilizada para resumir e representar de forma concisa um conjunto de dados.	1
Mediana	É uma medida de tendência central que descreve o valor que separa a metade superior da metade inferior de um conjunto de dados.	Se a quantidade de amostras for ímpar, Equação 2, se não, Equação e 3
Moda	A moda representa o valor mais recorrente em um conjunto de dados. Caso não haja repetições no conjunto, significa que ele não tem uma moda definida.	-

Fonte: Adaptada de Rodrigues e Campos, (2023) e Ramachandran e Tsokos, (2021).

$$\bar{x} = \frac{1}{K} \sum_{i=1}^K x_i \quad (1)$$

Onde, x_i são as amostras e K é a quantidade total de amostras

$$R_M = \frac{N + 1}{2} \quad (2)$$

Onde R_M representa a posição do valor da mediana.

$$M = \frac{d_{\frac{N}{2}} + d_{(\frac{N}{2})+1}}{2} \quad (3)$$

Onde M é a mediana, calculada pela média dos valores centrais, em que d representa o valor nas posições $\frac{N}{2}$ e $(\frac{N}{2}) + 1$.

Para realizar a verificação da distribuição utiliza-se uma tabela de frequência para então ser gerado um histograma, conforme Tabela 3.

Tabela 3 – Tabela de Frequência.

Item	Descrição	Equação
Amplitude	Representa a variação entre o valor máximo e o mínimo observados.	4
Frequência absoluta	Refere-se à frequência com que um elemento se manifesta na amostra.	5
Frequência relativa		6
Frequência absoluta acumulada	Representa a acumulação da frequência absoluta do valor da variável i com todas as frequências que a precedem.	N_i
Frequência relativa acumulada		7
Classes	As classes servem como uma ferramenta para agrupar e simplificar a variedade de elementos presentes em uma amostra.	8
Intervalo de Classe		9

Fonte: Adaptada de Rodrigues e Campos, (2023) e Ramachandran e Tsokos, (2021).

$$H = Valor_{MÁXIMO} - Valor_{MÍNIMO} \quad (4)$$

$$n = \sum_1^K n_i \quad (5)$$

Onde n_i é a frequência absoluta, n número total de dados na amostra.

$$f_i = \frac{n_i}{n} \quad (6)$$

$$F_i = \frac{N_i}{n} \quad (7)$$

$$k \cong \sqrt{K} \quad (8)$$

$$h \cong \frac{H}{k} \quad (9)$$

Para a modelagem da distribuição foi escolhida a distribuição Weibull, dada pela Equação 10.

$$f(t) = \frac{\beta}{\eta} \left(\frac{t}{\eta}\right)^{\beta-1} * e^{-\left(\frac{t}{\eta}\right)^\beta} \quad (10)$$

Onde, β é o parâmetro de forma (ou inclinação) e η é o tempo de vida. O valor do erro RMS, indicado na Equação 11, é calculado comparando-se a frequência relativa obtida a partir dos dados reais e o valor obtido a partir da distribuição, neste caso, dado pela Equação 10.

$$E_{RMS} = \sqrt{\frac{1}{K} \sum_{i=1}^K (f_i - f(t))^2} \quad (10)$$

3 RESULTADOS E DISCUSSÃO

Para gerar o estudo de caso proposto por Rodrigues e Campos, (2023), foi utilizado o seguinte prompt, conforme Tabela 4. Os itens destacados, indicam aqueles que podem ser customizados conforme o contexto de cada estudante.

Tabela 4 – Comandos a serem enviados ao ChatGPT.

Comandos
<p>Você precisa gerar, no formato de tabela, dados de produção de uma empresa:</p> <p>A empresa fabrica peças para automóveis.</p> <p>Você deve indicar o tempo, em segundos, até o sistema de fabricação apresentar uma falha.</p> <p>Além dessas informações você deve criar pelo menos duas colunas com mais dados sobre essa produção.</p> <p>Cada coluna deve ter 121 linhas com as informações de cada variável.</p> <p>Você não pode indicar nenhum código para gerar a tabela em ambiente externo</p> <p>Os dados de tempo de produção devem ter uma distribuição Weibull, tente não ficar repetindo valores. Seja aleatório.</p> <p>Você também deverá gerar as explicações de cada uma das colunas da tabela, considerando essa produção hipotética.</p> <p>A tabela deve apresentar as 121 linhas sem resumo.</p> <p>Você deve explicar os dados e a relação entre eles no final.</p>

Fonte: Os Autores (2024)

Foi gerada uma tabela com as seguintes colunas e explicações fornecidas pelo ChatGPT:

- **Tempo de Produção (s):** Esta coluna apresenta o tempo necessário, em segundos, para a produção de cada parafuso. Os tempos foram gerados seguindo uma distribuição Weibull para simular a variabilidade real observada em processos de fabricação. Esta distribuição é comumente utilizada para modelar tempos de falha e durabilidade de produtos.
- **Defeitos por 1000 unidades:** Esta coluna indica a quantidade de defeitos encontrados por 1000 unidades produzidas. A variabilidade nesta coluna reflete a qualidade do processo de produção e a incidência de falhas.

- **Temperatura da Máquina (°C):** Esta coluna registra a temperatura da máquina de produção em graus Celsius durante a fabricação dos parafusos. A temperatura da máquina pode afetar a qualidade do produto final e a taxa de defeitos, sendo um parâmetro crítico para controle e otimização do processo.

Conforme solicitado, o ChatGPT também apresentou a seguinte explicação entre os dados: “Os dados apresentados mostram a inter-relação entre o tempo de produção, a qualidade do produto e as condições operacionais da máquina. Um tempo de produção variado, conforme uma distribuição Weibull, é esperado devido às flutuações inerentes aos processos industriais. A taxa de defeitos reflete a eficiência e a estabilidade do processo, em que condições operacionais como a temperatura da máquina desempenham um papel significativo. Temperaturas inadequadas podem aumentar a incidência de defeitos, enquanto tempos de produção que desviam significativamente da média podem indicar problemas no processo. Analisar esses dados pode ajudar na identificação de padrões e na implementação de melhorias no processo de fabricação.”

A Tabela 5 apresenta o resultado dos itens elencados na Tabela 2. A Tabela 6 apresenta os resultados dos itens elencados na Tabela 3, e, a partir destes dados, foi construído o histograma, conforme Figura 3.

O erro médio quadrático (RMS) demonstra que a distribuição Weibull realmente se aproxima dos dados, conforme Figura 4. Os resultados apresentados estão relacionados ao comando “Você deve indicar o tempo, em segundos, até o sistema de fabricação apresentar uma falha”, isto é, foi modelado o tempo de produção de cada parafuso nesta fábrica hipotética.

Tabela 5 – Resultados da modelagem estatística.

Item	Resultado
Média aritmética simples	46,85 segundos
Mediana	46,77 segundos
Moda	48,60 segundos

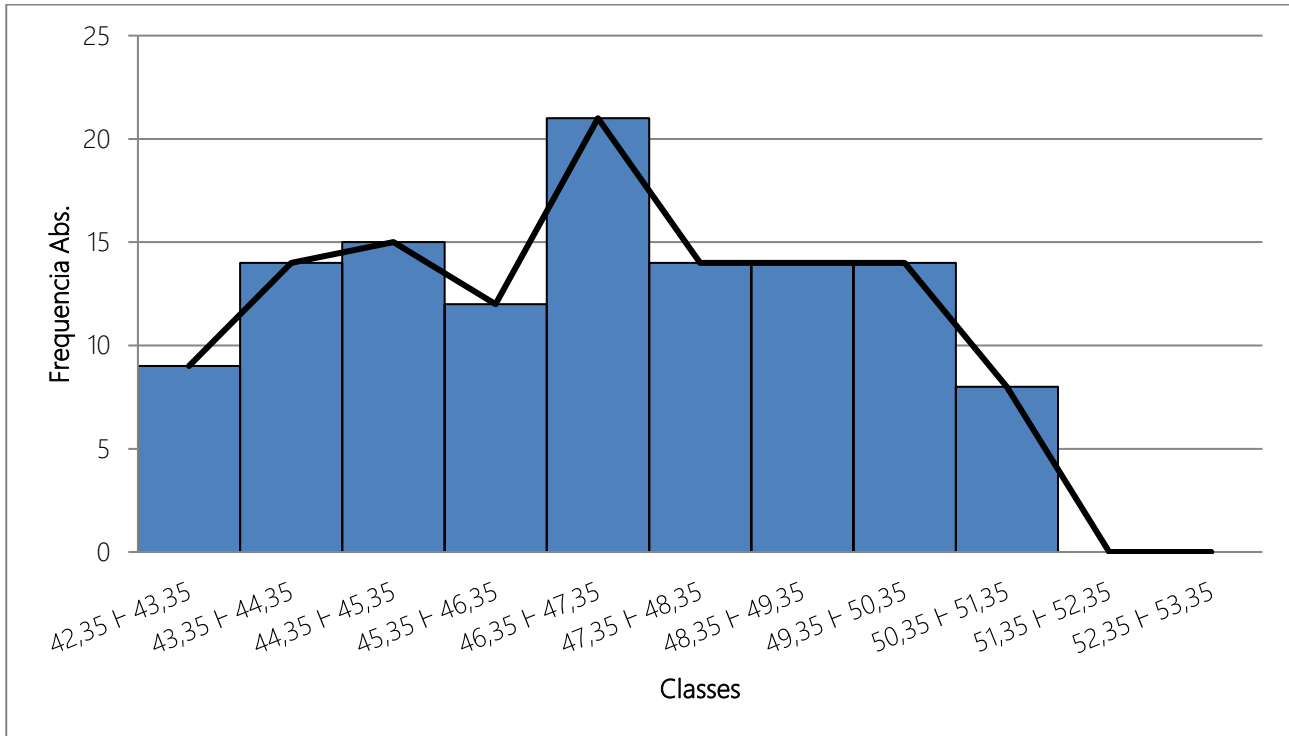
Fonte: Os Autores (2024).

Tabela 6 – Tabela de frequências.

i	Classes	n_i	Ponto Médio	N_i	f_i	F_i	$f(t)$	Quadrado do Desvio (f_i e $f(t)$)
1	42,35 † 43,35	9	42,85	9	0,074	0,074	0,05015730023	0,00058674719
2	43,35 † 44,35	14	43,85	23	0,116	0,190	0,06944573499	0,00213968640
3	44,35 † 45,35	15	44,85	38	0,124	0,314	0,09213421625	0,00101332244
4	45,35 † 46,35	12	45,85	50	0,099	0,413	0,11545048927	0,00026493863
5	46,35 † 47,35	21	46,85	71	0,174	0,587	0,13387864564	0,00157411145
6	47,35 † 48,35	14	47,85	85	0,116	0,702	0,13962747897	0,00057240561
7	48,35 † 49,35	14	48,85	99	0,116	0,818	0,12588764574	0,00010373761
8	49,35 † 50,35	14	49,85	113	0,116	0,934	0,09291757458	0,00051915188
9	50,35 † 51,35	8	50,85	121	0,066	1,000	0,05212147580	0,00019583838
10	51,35 † 52,35	0	51,85	121	0,000	1,000	0,02008756789	0,00040351038
11	52,35 † 53,35	0	52,85	121	0,000	1,000	0,00464273796	0,00002155502
					1		E_{RMS}	0,025928231

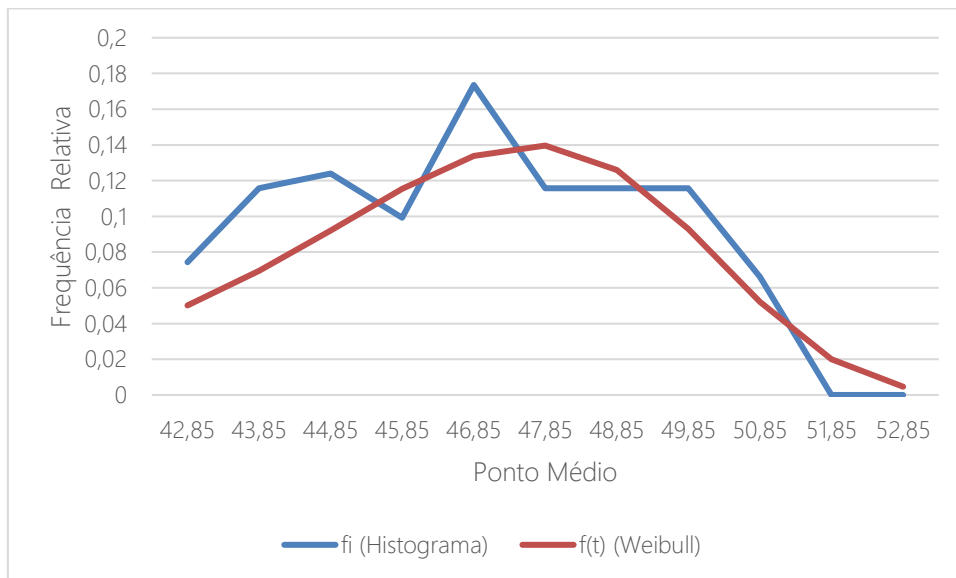
Fonte: Os Autores (2024).

Figura 3 – Histograma dos dados.



Fonte: Os Autores (2024).

Figura 4 – Comparação entre os dados gerados pelo ChatGPT e a Distribuição Weibull.



Fonte: Os Autores (2024).

4 CONSIDERAÇÕES FINAIS

O estudo apresentado demonstrou que a integração de metodologias ativas de ensino com tecnologias avançadas, como o ChatGPT, pode significativamente aprimorar o processo de ensino-aprendizagem em estatística. A capacidade do ChatGPT de gerar dados customizados oferece uma ferramenta robusta para simulações estatísticas e experimentos pedagógicos, atendendo às necessidades específicas de estudantes e educadores. A metodologia proposta, composta por três etapas principais — interação com o ChatGPT, preparação e organização dos dados, e modelagem estatística — mostrou-se eficaz na criação de conjuntos de dados adaptados ao contexto educacional.

A interação detalhada com o ChatGPT revelou que, embora ajustes possam ser necessários, comandos bem definidos permitem a geração de dados consistentes e relevantes. A utilização de distribuições estatísticas, como a Weibull, na modelagem dos dados forneceu uma análise precisa da variabilidade e incerteza, essenciais para a compreensão de conceitos estatísticos avançados.

O estudo de caso confirmou a viabilidade prática do método, demonstrando que a ferramenta pode apoiar de maneira significativa o ensino de estatística. A análise dos resultados indicou que a combinação de IA com metodologias ativas pode transformar o ensino de disciplinas quantitativas, promovendo um aprendizado mais envolvente, contextualizado e eficaz.

Este trabalho reforça a importância da inovação pedagógica aliada à tecnologia, sugerindo que a adoção de ferramentas como o ChatGPT pode representar um avanço substancial na educação em ciência de dados e estatística. O protagonismo do estudante pode ser ressaltado ao considerarmos que o ensino de estatística poderá ser customizada à realizada específica de cada estudante.

Como trabalhos futuros, outras distribuições podem ser utilizadas, além de técnicas de modelagem orientada a dados, como Redes Neurais Artificiais.

AGRADECIMENTOS

Os autores agradecem: o Centro Universitário de Volta Redonda pelo apoio institucional para realizar esta pesquisa. I.P. Rodrigues também agradece à CAPES pelo apoio financeiro (Processo No. 88882.444522/2019-01) durante o doutorado.

REFERÊNCIAS

BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; AGARWAL, S.; HERBERT-VOSS, A.; KRUEGER, G.; HENIGHAN, T.; CHILD, R.; RAMESH, A.; ZIEGLER, D. M.; WU, J.; WINTER, C.; HESSE, C.; CHEN, M.; SIGLER, E.; LITWIN, M.; GRAY, S.; CHESS, B.; CLARK, J.; BERNER, C.; MCCANDLISH, S.; RADFORD, A.; SUTSKEVER, I.; AMODEI, D. Language Models are Few-Shot Learners. 28 maio 2020.

GOMEZ-DEL RIO, T.; RODRIGUEZ, J. Design and assessment of a project-based learning in a laboratory for integrating knowledge and improving engineering design skills. **Education for Chemical Engineers**, v. 40, p. 17–28, jul. 2022.

JAVOID, M.; HALEEM, A.; SINGH, R. P.; KHAN, S.; KHAN, I. H. Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. **BenchCouncil Transactions on Benchmarks, Standards and Evaluations**, v. 3, n. 2, p. 100115, jun. 2023.

PRINCE, M. J.; FELDER, R. M. Inductive Teaching and Learning Methods: Definitions, Comparisons, and Research Bases. **Journal of Engineering Education**, v. 95, n. 2, p. 123–138, abr. 2006.

RAMACHANDRAN, K. M.; TSOKOS, C. P. Descriptive statistics. Em: **Mathematical Statistics with Applications in R**. [s.l.] Elsevier, 2021. p. 1–40.

RODRIGUES, I. P.; CAMPOS, A. M. ChatGPT como Ferramenta de Geração de Dados para Modelagem Estatística: Uma Abordagem Prática na Disciplina de Data Science. In: Congresso Brasileiro De Ciências E Saberes Multidisciplinares, (2), 2023, Volta Redonda. **Anais...** Volta Redonda: Editora FOA, 2023

SHAH, C. The past, the present, and the future of information and data sciences: A pragmatic view. **Data and Information Management**, v. 7, n. 1, p. 100028, mar. 2023.

CHATGPT IN DATA CUSTOMIZATION FOR DATA SCIENCE EDUCATION

Abstract: *This article investigates the application of active teaching methodologies in the context of statistics education, highlighting the use of advanced technological tools such as ChatGPT. The ability to generate customized data through ChatGPT is explored, providing a valuable resource for statistical simulations and pedagogical experiments. The proposed method is divided into three stages: interaction with ChatGPT to obtain data, data preparation and organization, and statistical modeling. The interaction with ChatGPT is detailed, demonstrating how specific commands can be used to generate datasets tailored to students' needs. Statistical modeling includes the use of distributions such as Weibull to analyze the variability and uncertainty of the data. A case study illustrates the practical application of the method, confirming the tool's effectiveness in supporting the teaching of statistics. The results show that the integration of AI and active methodologies can transform the teaching of quantitative subjects, promoting a more engaging and contextualized learning experience.*

Keywords: *Data Science, Artificial Intelligence, ChatGPT, Customized Learning, Statistical modeling.*

