A Importância do Conhecimento Estatístico para Analistas e Cientistas de Dados

Olá! Seja bem-vindo(a) à nossa newsletter semanal sobre Ciência de Dados! Nesta edição, vamos destacar a importância crucial do conhecimento estatístico para analistas e cientistas de dados, enfatizando que esses conhecimentos são essenciais para uma atuação profissional eficaz.

A estatística é a espinha dorsal da análise de dados, permitindo que profissionais tomem decisões informadas baseadas em evidências. Para analistas e cientistas de dados, entender e aplicar princípios estatísticos é fundamental para a validação de modelos, a interpretação de resultados e a comunicação de insights. Na tabela abaixo, destacamos dez regras de ouro para aplicação profissional no campo da ciência de dados, baseadas em boas práticas estatísticas.

Nome da Regra	Descrição	Exemplo em Python	Principais Aplicações	Quando Não Utilizar

Verificação Contextual	Utilize os testes de suposições como uma das várias informações para decidir se uma suposição foi violada.	`import scipy.stats as stats`
		`ks_stat, p_value = stats.kstest(data, 'norm')`
		`print(p_value)`
			Avaliar suposições estatísticas antes de análises complexas, como ANOVA e regressão linear.	Quando os dados são claramente inadequados para a suposição em questão.

Análise de Outliers Consciente	Investigue outliers ao invés de removê-los cegamente, pois podem introduzir viés ou ocultar informações importantes.	`outliers = data[np.abs(data - data.mean()) > (3 * data.std())]` `print(outliers)`	Identificação e tratamento de outliers em análises exploratórias de dados.	Quando os outliers são evidentemente erros de medição ou entrada.

Verificação Pós-Preditores	Verifique a normalidade das variáveis dependentes após escolher os preditores.	`import statsmodels.api as sm`
		`model = sm.OLS(y, X).fit()`
		`residuals = model.resid`
		`sm.qqplot(residuals, line='45')`
			Modelagem estatística, incluindo regressão linear e ANOVA.	Quando a distribuição de Y, independente de X, é crítica para a análise.

Descritiva Antes de Testes	Execute sempre estatísticas descritivas e gráficos antes de realizar testes estatísticos.	`import matplotlib.pyplot as plt`
		`data.describe()`
		`plt.hist(data)`
		`plt.show()`
			Qualquer análise de dados inicial, incluindo preparação e limpeza de dados.	Em análises em que a descritiva não adiciona valor significativo ao entendimento dos dados.

Simplicidade Prudente	Use o teste estatístico mais simples que responda à questão da pesquisa e atenda às suposições.	`t_stat, p_value = stats.ttest_ind(group1, group2)` `print(p_value)`	Escolha de testes estatísticos para análises de comparação de grupos, como t-testes.	Quando a simplicidade compromete a validade ou a robustez dos resultados.
Transformações de Dados	Aplique transformações nos dados para atender às suposições dos testes, como log ou raiz quadrada.	`transformed_data = np.log(data)`	Análise de regressão, ANOVA, testes de normalidade.	Quando a transformação não melhora a distribuição ou interpretação dos dados.
Correlação Não é Causalidade	Entenda que correlação entre duas variáveis não implica em causalidade.	`correlation, p_value = stats.pearsonr(x, y)` `print(correlation)`	Análise exploratória de dados, modelos de predição.	Quando se necessita provar uma relação de causa e efeito.
Amostragem Representativa	Certifique-se de que a amostra seja representativa da população para garantir resultados válidos.	`sample = data.sample(n=100)`	Estudos de mercado, pesquisas de opinião, ensaios clínicos.	Quando a amostra não representa a população devido a viés de seleção.

Multicolinearidade	Verifique multicolinearidade em modelos de regressão múltipla e resolva se necessário.	`from statsmodels.stats.outliers_influence import variance_inflation_factor`
		`vif = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]`
		`print(vif)`
			Modelos de regressão múltipla, análise de variáveis preditoras.	Quando a multicolinearidade é baixa e não afeta significativamente os resultados.

Validação Cruzada	Utilize validação cruzada para avaliar o desempenho de modelos preditivos.	`from sklearn.model_selection import cross_val_score`
		`scores = cross_val_score(model, X, y, cv=5)`
		`print(scores.mean())`
			Modelos de machine learning, avaliação de desempenho de modelos.	Quando a validação cruzada não adiciona valor devido a um número muito pequeno de amostras.

Compreender e aplicar essas regras de boas práticas estatísticas ajudará você a evitar armadilhas comuns e a conduzir análises mais robustas e precisas. Essas práticas não são apenas teóricas; elas são ferramentas essenciais para lidar com os desafios do mundo real, onde os dados nem sempre são perfeitos.

Esperamos que você tenha achado estas informações úteis. Seja você um iniciante no mundo da Ciência de Dados ou um profissional experiente, entender essas regras pode abrir novas possibilidades em suas análises e implantações. Se você tiver alguma dúvida ou sugestão, não hesite em entrar em contato. Queremos garantir que nossos conteúdos sejam sempre relevantes e úteis para você.

Fique atento à nossa próxima edição, onde continuaremos a explorar ferramentas técnicas e dicas valiosas para aprimorar suas habilidades em Ciência de Dados. Até lá, continue explorando e aprendendo!

Saudações,

Prof. Dr. Dilermando Piva Jr
Coordenador de Ciência de Dados para Negócios / Fatec Votorantim
E-mail: f301.cdn@fatec.sp.gov.br