

ENTREGÁVEL – ESTATÍSTICA AVANÇADA
Data Expert | DNC

Você é um cientista de dados ‘free lancer’ que foi contatado por um grande milionário que está desejando leiloar alguns de seus diamantes. O dinheiro será todo doado para uma instituição de caridade.

A proposta de trabalho é desenvolver um modelo de previsão para o **preço mínimo** que poderá ser pedido para cada diamante, utilizando o dataset disponibilizado.

Para ser contratado, o milionário gostaria que você respondesse algumas perguntas antes, de modo que fosse possível acessar o seu conhecimento.

1. Qual é a média da variável ‘carat’?

- a) 0.5845
- b) 0.2253
- c) 0.7979 RESPOSTA CORRETA**
- d) 0.0186
- e) Não é possível calcular a média da variável ‘carat’

2. Qual é a correlação entre as variáveis ‘price’ e ‘depth’ ?

- a) 1
- b) 0
- c) 0.55974
- d) 0.921591
- e) -0.010647 RESPOSTA CORRETA**

3. Após o escalonamento das variáveis numéricas, qual é o valor máximo da variável ‘table’?

- a) 95
- b) 16.8 RESPOSTA CORRETA**
- c) 29
- d) 1
- e) 0.970772

4. Após a ‘dummyficação’ das variáveis categóricas, quantas colunas existem em um dataset com somente variáveis desse tipo?

a) 20

b) 17 RESPOSTA CORRETA

c) 3

d) 10

e) 1

5. A distribuição da resposta ‘price’ é normal e não precisa ser transformada.

a) verdadeiro

b) falso

6. Ao separar o dataset em dois pedaços, o de treinamento e o de teste, quantas linhas possuirá o dataset de teste? (Utilize uma proporção de 20% e argumento `random_state = 123456`) Dica: concatene os datasets e faça a matriz do modelo antes de fazer a separação entre treinamento e teste.

a) 24

b) 53940

c) 10788 RESPOSTA CORRETA

d) 100

e) nenhuma

7. Utilizando o método dos mínimos quadrados e ajustando o modelo com o logaritmo natural da resposta ‘price’ e os dados de ajuste, qual é a variável que tem o maior valor absoluto para a estatística ‘t’. (OBS: desconsidere o ‘intercept’).

a) x RESPOSTA CORRETA.

b) table

c) cut_Good

d) clarity_SI1

e) carat

8. Para o modelo ajustado na questão 07, qual é a variável que possui o maior VIF?

a) x RESPOSTA CORRETA.

b) table

- c) cut_Good
- d) clarity_SI1
- e) carat

9. Após remover todas as variáveis para manter os VIF's em no máximo 5, quantas variáveis sobraram no modelo? OBS: desconsidere o intercept. (Dica, comece removendo, uma por uma, as variáveis com maior VIF)

- a) 17
- b) 18 RESPOSTA CORRETA**
- c) 19
- d) 20
- e) 21

10. Qual é o R^2 para os dados de teste com o modelo sem as variáveis removidas na questão 9?

- a) 0.96 RESPOSTA CORRETA**
- b) 0.86
- c) 0.5
- d) 1
- e) 0.1

11. Utilizando os dados de teste, faça a análise de resíduos e responda: eles estão de acordo com a hipótese para a regressão linear?

- a) Sim, os resíduos são normais e homocedásticos.
- b) Não, os resíduos são normais porém heterocedásticos.
- c) Não, os resíduos são homocedásticos porém não normais.
- d) Não, os não são nem homocedásticos nem normais. RESPOSTA CORRETA.**
- e) Nenhuma das alternativas acima.

12. Utilizando a regressão regularizada tipo “ridge” e o dataset com TODOS os regressores, qual dentre os seguintes valores de alpha deixa o modelo com o maior R^2 para os dados de teste?

- a) 10 RESPOSTA CORRETA**

b) 100

c) 1000

d) 10000

e) 100000