

Desafio Final

Entrega 26 jul em 23:59 **Pontos** 100 **Perguntas** 15
Disponível até 26 jul em 23:59 **Limite de tempo** Nenhum

Instruções

O Desafio do Módulo Final está disponível!

1. Instruções para realizar o desafio

Consulte a data de entrega no teste e em seu calendário.

Reserve um tempo para realizar a atividade, leia as orientações e enunciados com atenção. Em caso de dúvidas utilize o "Fórum de dúvidas do Desafio Final".

Para iniciá-lo clique em "Fazer teste". Você tem somente **uma** tentativa e não há limite de tempo definido para realizá-lo. Caso precise interromper a atividade, apenas deixe a página e, ao retornar, clique em "Retomar teste".

Clique em "Enviar teste" **somente** quando você concluí-lo. Antes de enviar confira todas as questões.

O gabarito será disponibilizado partir de domingo, **26/07/2020**, às 23:59.

Bons estudos!

2. O arquivo abaixo contém o enunciado do desafio

[Enunciado do Desafio Final - B. Analista de Dados.pdf](#) 

Histórico de tentativas

	Tentativa	Tempo	Pontuação
MAIS RECENTE	Tentativa 1	61 minutos	79,92 de 100

⚠ As respostas corretas estarão disponíveis em 26 jul em 23:59.

Pontuação deste teste: **79,92** de 100

Enviado 10 jul em 19:05

Esta tentativa levou 61 minutos.

Pergunta 1**6,66 / 6,66 pts**

Quantas instâncias e características existem, respectivamente, no dataset?

☐ (7, 5000).☐ (12, 5110).☐ (5000, 7).☒ (5110, 12).**Pergunta 2****6,66 / 6,66 pts**

Quantas variáveis do tipo "string" estão presentes no dataset?

☒ 6.☐ 4.☐ 2.☐ 3.**Pergunta 3****6,66 / 6,66 pts**

Qual é a idade (age) média dos entrevistados?

☐ 55,12 anos.

☒ 43,22 anos.

☐ 22,61 anos.

☐ 45,28 anos.

Pergunta 4

6,66 / 6,66 pts

Sobre a distribuição de AVC em relação ao sexo (gender) dos entrevistados, é CORRETO afirmar:



Apesar da pouca diferença, existe uma maior quantidade de mulheres que sofreram AVC.



Existe no dataset apenas dois tipos de gêneros, homens e mulheres.



Não podem ser identificadas diferenças entre os gêneros, pois o dataset está equilibrado (mulheres=homens).



Existe no dataset uma maior quantidade de homens que sofreram AVC.

Pergunta 5

6,66 / 6,66 pts

Sobre o dataset é correto afirmar, EXCETO:



Existem dois tipos diferentes de classes de residências (“Residence_type”) presentes nesse dataset.



Existem dados categóricos e numéricos presentes neste dataset. Um exemplo de dados categóricos é o “Residence_type”.



A variável bmi possui valores não numéricos.



O dataset está balanceado. Existem quantidades similares de instâncias de indivíduos que sofreram AVC e que não sofreram dessa enfermidade.

Pergunta 6

6,66 / 6,66 pts

Qual é o valor da mediana para a variável do nível médio de glicose do entrevistado (“avg_glucose_level”)?



271,74.



120.



95.



78.

Pergunta 7

6,66 / 6,66 pts

Analisando o padrão de dispersão da variável do nível médio de glicose do entrevistado (“avg_glucose_level”) é correto afirmar, EXCETO:



Existem valores que correspondem a possíveis outliers. Esses valores certamente devem ser eliminados do dataset, pois sempre causam problemas.



A dispersão dos dados no terceiro quartil é maior que no segundo, pois existe uma maior quantidade de valores diferentes no terceiro quartil.



Os possíveis outliers estão localizados após o limite superior do boxplot.



O terceiro quartil corresponde ao valor de 120. Desse modo, podemos dizer que 75% dos dados estão abaixo desse valor.

Pergunta 8

6,66 / 6,66 pts

Analisando a dispersão dos dados para a variável idade ("age"), é correto afirmar, EXCETO:



O maior existente para a idade dos entrevistados corresponde a 82 anos.



Pelo Boxplot não é possível identificar possíveis outliers.



O primeiro quartil indica que 25% dos dados estão abaixo de 30 anos.



A mediana para essa variável corresponde ao valor de 68 anos.

Pergunta 9

6,66 / 6,66 pts

Quantas classes diferentes para a variável “work_type” existem no dataset?

☐ 6.

☐ 4.

☐ 2.

☒ 5.

Pergunta 10

6,66 / 6,66 pts

Dentre as classes de tipos de trabalhos existentes (work_type), qual é aquela que possui uma maior quantidade de instâncias?

☒ Private.

☐ Never_worked.

☐ Govt_job.

☐ Self-employed.

Pergunta 11

6,66 / 6,66 pts

Qual foi, respectivamente, o percentual de dados utilizados para o treinamento e teste do modelo?

☐ (30%, 70%).

☐ (70%, 30%).

☒ (80%, 20%).

☐ (20%, 80%).

Pergunta 12

6,66 / 6,66 pts

Analisando as variáveis “bmi” e “smoking_status”, é CORRETO afirmar:

- ☐ Existem oito classes distintas de “smoking_status”.
- ☐ Ambas são variáveis numéricas.
- ☒ Ambas possuem instâncias com valores desconhecidos.
- ☐ A variável “bmi” possui apenas valores numéricos.

Incorreta

Pergunta 13

0 / 6,66 pts

Após o agrupamento dos dados de ‘smoking_status’ e ‘stroke’, é CORRETO afirmar que:

- ☐ Existem seis classes diferentes de “smoking_status”.
- ☐ Dentre os entrevistados que sofreram AVC, existem uma maior quantidade de indivíduos da classe que nunca fumaram (never smoked).
- ☐ Neste dataset existe uma maior quantidade de indivíduos que sofreram AVC.



Não é possível realizar o agrupamento, pois os dados possuem dimensões diferentes.

Incorreta**Pergunta 14****0 / 6,66 pts**

Sobre a relação entre a hipertensão (hypertension) e o AVC (stroke) presente neste dataset, é CORRETO afirmar:



A proporção entre indivíduos hipertensos e não hipertensos no dataset é a mesma.



Existe uma maior quantidade de dados de indivíduos não hipertensos.



Os dados mostram que este dataset está balanceado.



A proporção de incidência de AVC é maior nos indivíduos que sofrem de hipertensão.

Incorreta**Pergunta 15****0 / 6,76 pts**

Sobre o algoritmo de regressão logística aplicado para a previsão da ocorrência de AVC, é correto afirmar, EXCETO:



A acurácia do modelo é superior a 90%.



A árvore de decisão também poderia ser aplicada para esse modelo de classificação.



A regressão logística não deveria ser aplicada ao problema, pois ela trabalha apenas com dados categóricos.



Como o dataset está desbalanceado, a acurácia (accuracy) resultante pode estar enviesada.

Pontuação do teste: **79,92** de 100