

# Desafio Final

---

**Entrega** 11 set em 23:59      **Pontos** 100      **Perguntas** 15  
**Disponível** até 11 set em 23:59      **Limite de tempo** Nenhum

---

## Instruções

O Desafio 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.

Caso o teste seja iniciado e não enviado até o final do prazo de entrega, a plataforma enviará a tentativa não finalizada automaticamente, independente do progresso no teste. Fique atento ao seu teste e ao prazo final, pois novas tentativas só serão concedidas em casos de questões médicas.

O gabarito será disponibilizado partir de sexta, **11/09/2020**, às 23h59.

Bons estudos!

### 2. O arquivo abaixo contém o enunciado do desafio

[Enunciado do desafio final – Cientista de Dados.pdf](#) 

## Histórico de tentativas

	Tentativa	Tempo	Pontuação
MAIS RECENTE	<a href="#">Tentativa 1</a>	1.726 minutos	100 de 100

---

⚠ As respostas corretas estarão disponíveis em 11 set em 23:59.

Pontuação deste teste: **100** de 100

Enviado 9 set em 22:17

Esta tentativa levou 1.726 minutos.

### Pergunta 1

6,67 / 6,67 pts

Quantas instâncias e características existem, respectivamente, no dataset?

☐ (12, 5110)

☐ (7, 5000)

☒ (5110, 12)

☐ (5000, 7)

### Pergunta 2

6,67 / 6,67 pts

Quantas variáveis do tipo “string” estão presentes no dataset?

☒ 6

☐ 4

☐ 3

☐ 2

### Pergunta 3

6,67 / 6,67 pts

Qual é a idade (age) média dos entrevistados?

☐ 22,61 anos

☒ 43,22 anos

☐ 45,28 anos

☐ 55,12 anos

#### Pergunta 4

6,67 / 6,67 pts

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

☐

Existe no dataset uma maior quantidade de homens que sofreram AVC

☒

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

☐

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

☐

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

#### Pergunta 5

6,67 / 6,67 pts

Sobre o dataset é correto afirmar, EXCETO:



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



Existem dois tipos diferentes de classes de residências ("Residence\_type") presentes nesse dataset

### Pergunta 6

6,67 / 6,67 pts

Qual é o valor da mediana para a variável do nível médio de glicose do entrevistado ("avg\_glucose\_level")?

☐ 120☐ 271,74☒ 95☐ 78

### Pergunta 7

6,67 / 6,67 pts

Analisando o padrão de dispersão da variável do nível médio de glicose do entrevistado ("avg\_glucose\_level") é correto afirmar, EXCETO:



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



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



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

### Pergunta 8

6,67 / 6,67 pts

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



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



Pelo Boxplot não é possível identificar possíveis outliers



A mediana para essa variável corresponde ao valor de 68 anos



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

### Pergunta 9

6,67 / 6,67 pts

Quantas classes diferentes para a variável “work\_type” existem no dataset?

☐ 4

☐ 2

☐ 6

☒ 5

### Pergunta 10

6,67 / 6,67 pts

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

☐ Self-employed

☒ Private

☐ Govt\_job

☐ Never\_worked

### Pergunta 11

6,67 / 6,67 pts

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

☐ (30%, 70%)

☐ (20%, 80%)

☒ (80%, 20%)

☐ (70%, 30%)

### Pergunta 12

6,67 / 6,67 pts

Analisando as variáveis “bmi” e “smoking\_status”, é CORRETO afirmar:

☐ Existem oito classes distintas de “smoking\_status”

☐ A variável “bmi” possui apenas valores numéricos

☒ Ambas possuem instâncias com valores desconhecidos

☐ Ambas são variáveis numéricas

### Pergunta 13

6,67 / 6,67 pts

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

☒ Dentre os entrevistados que sofreram AVC, existem uma maior quantidade de indivíduos da classe que nunca fumaram (never smoked)

☐ Existem seis classes diferentes de “smoking\_status”

☐ Neste dataset existe uma maior quantidade de indivíduos que sofreram AVC



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

### Pergunta 14

6,67 / 6,67 pts

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



Os dados mostram que este dataset está balanceado



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



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



Existe uma maior quantidade de dados de indivíduos não hipertensos

### Pergunta 15

6,62 / 6,62 pts

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



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





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



A acurácia do modelo é superior a 90%



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

Pontuação do teste: **100** de 100