Pergunta 1 — 6,67 pts

Quantas instâncias e características existem, respectivamente, no dataset?

- (5000, 7)
- (12, 5110)
- (5110, 12)
- (7, 5000)

Pergunta 2 — 6,67 pts

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

- 2
- 6
- 4
- 3

Pergunta 3 — 6,67 pts

Qual é a idade (age) média dos entrevistados?

- 45, 28 anos.
- 43, 22 anos.
- 22, 61 anos.

- 55, 12 anos.

Pergunta 4 — 6,67 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, uma maior quantidade de homens 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).

Pergunta 5 — 6,67 pts

É correto afirmar sobre o dataset, EXCETO:

- A variável bmi possui valores não numéricos.
- Existem dados categóricos e numéricos presentes neste dataset. Um exemplo de dados categóricos é o “Residence_type”.
- Existem dois tipos diferentes de classes de residências (“Residence_type”) presentes nesse dataset.
- O dataset está balanceado. Existem quantidades similares de instâncias de indivíduos que sofreram AVC e que não sofreram dessa enfermidade.

Pergunta 6 — 6,67 pts

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

- 271,74.
- 78.
- 95.
- 120.

Pergunta 7 — 6,67 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.
- O terceiro quartil corresponde ao valor de 120. Desse modo, podemos dizer que 75% dos dados estão abaixo desse valor.
- 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.

Pergunta 8 — 6,67 pts

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

- A mediana para essa variável corresponde ao valor de 68 anos.
- O primeiro quartil indica que 25% dos dados estão abaixo de 30 anos.
- Pelo Boxplot não é possível identificar possíveis outliers.

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

Pergunta 9 — 6,67 pts

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

- 2
- 4
- 5
- 6

Pergunta 10 — 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
- Never_worked
- Govt_job

Pergunta 11 — 6,67 pts

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

- (20%, 80%)
- (30%, 70%)
- (80%, 20%)
- (70%, 30%)

Pergunta 12 — 6,67 pts

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

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

Pergunta 13 — 6,67 pts

Após o agrupamento dos dados de “smoking_status” e ”stroke”, é CORRETO afirmar que:

- Existem seis classes diferentes de “smoking_status”
- Não é possível realizar o agrupamento, pois os dados possuem dimensões diferentes
- 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.

Pergunta 14 — 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 hipertensos.

Pergunta 15 — 6,62 pts

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

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