Licenciatura em Engenharia de Sistemas
UC: Análise de Dados
Teste da avaliação final
Data: 25/05/2023 **Duração: 60 minutos**

Nome: _____

Número: _____

- Este enunciado é composto por 9 páginas e está dividido em 2 partes: **Parte I** e **Parte II**.
- As questões têm igual cotação.
- Cada resposta errada a uma questão do tipo verdadeiro ou falso desconta 50% do valor da questão.
- Cada questão de escolha múltipla tem apenas uma resposta correta que deve ser assinalada com um círculo.
- Para alterar ou anular uma resposta, riscar, de forma clara, o que pretende que fique sem efeito. Respostas ilegíveis ou que não possam ser claramente identificadas são classificadas com zero valores.

Parte I

1. Numa amostra de grande dimensão onde a maioria das observações apresentam valores próximos e apenas uma pequena proporção de observações apresenta valores elevados, como se posiciona a média relativamente à mediana da amostra e como será a forma do respectivo histograma (admita uma distribuição unimodal)?

    A. A média será maior do que a mediana e o histograma será assimétrico com uma cauda longa para a esquerda.

    B. A média será maior do que a mediana e o histograma será assimétrico com uma cauda longa para a direita.

    C. A média será menor do que a mediana e o histograma será assimétrico com uma cauda longa para a esquerda.

    D. A média será menor do que a mediana e o histograma será assimétrico com uma cauda longa para a direita.

    E. A média será igual à mediana e o histograma será assimétrico com uma cauda longa para a esquerda.

    F. Nenhuma das opções anteriores.

2. Em geral, qual das seguintes afirmações NÃO é verdadeira?

    A. A média aritmética é sensível a valores extremos.

    B. A amplitude total e o desvio padrão são sensíveis a valores extremos.

    C. O desvio padrão é sempre não negativo e será tanto maior, quanta mais variabilidade houver entre os dados.

    D. Se a distribuição dos dados for bastante enviesada, não é conveniente utilizar a média como medida de localização, nem o desvio padrão como medida de variabilidade.

    E. Numa distribuição puramente simétrica, os valores da média e da mediana coincidem.

    F. O desvio padrão fornece informação sobre a variabilidade dos dados, uns relativamente aos outros.

3. Num problema de regressão múltipla, multicolinearidade representa _____.

   A. uma situação desejável onde há uma baixa correlação entre um par ou mais de variáveis independentes.

   B. uma situação indesejável onde há uma baixa correlação entre um par ou mais de variáveis independentes.

   C. uma situação desejável onde há uma alta correlação entre um par ou mais de variáveis independentes.

   D. uma situação indesejável onde há uma alta correlação entre um par ou mais de variáveis independentes.

   E. uma situação indesejável onde não há correlação entre a variável dependente e as variáveis independentes.

   F. Nenhuma das opções anteriores.

Questões 4 – 7. A estrutura económica de um sistema de ligas de Baseball permite que algumas equipas ganhem substancialmente mais do que outras, o que, por sua vez, permite que essas equipas gastem muito mais com os salários dos seus jogadores. Espera-se, portanto, que tais equipas tenham melhores jogadores e que ganhem mais jogos. Tendo em conta dados da época de 2020, estimou-se o seguinte modelo

$\hat{y} = 71,87 + 0,101 \times Valor - 0,060 \times Liga,$

onde Y é o número total de vitórias de cada equipa, *Valor* é o valor total pago à equipa (em milhões de unidades monetárias) e *Liga* é uma variável indicadora que é igual a 0 se a equipa jogar na Liga Nacional ou 1 se a equipa jogar na Liga Internacional. Admita que os pressupostos do modelo são verificados.

4. Se as equipas A e B jogarem na mesma liga e o valor pago à equipa A for 1 milhão de unidades monetérias maior do que o da equipa B, então esperaríamos que a equipa A vencesse, em média, _____.

   A. 0,101 mais jogos que a B.

   B. 71,87 mais jogos que a B.

   C. 0,060 mais jogos que a B.

   D. 0,060 menos jogos que a B.

   E. 71,87 menos jogos que a B.

   F. Nenhuma das respostas anteriores está correta.

5. Suponha que se representaram os dados graficamente, juntamente com as retas de regressão para as equipas da Liga Nacional e da Liga Internacional. Qual seria o valor da inclinação da reta para as equipas da Liga Internacional?

   A. –0,060.

   B. 0,060.

   C. 0,941

   D. 0,101.

   E. 71,81.

   F. Nenhuma das respostas anteriores está correta.

6. O senso comum diz que equipas mais bem pagas devem ter uma forte tendência a vencer mais jogos, mas que o tipo de liga não deveria importar. Com base no senso comum descrito, em qual dos seguintes testes t, provavelmente, rejeita a hipótese nula?

   A. Teste t para *Valor*.

   B. Teste t para *Liga*.

   C. Em ambos os testes t.

   D. Em nenhum dos testes t.

   E. Teste t para o número total de vitórias.

   F. Nenhuma das respostas anteriores está correta.

7. Neste caso, as medidas de desempenho preditivo do modelo mais adequadas, de entre as seguinte, são _____.

   A. Acurácia; Sensibilidade; Precisão; Erro Quadrático Médio (MSE).

   B. Erro Quadrático Médio (MSE); Coeficiente de determinação (R^2); AUC.

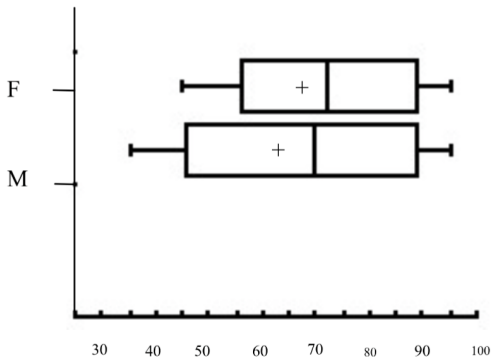
   C. Precisão; Sensibilidade; Especificidade; AUC.

   D. Erro Quadrático Médio (MSE); Raiz Quadrada do Erro Quadrático Médio (RMSE); Coeficiente de determinação (R^2).

   E. Precisão; Sensibilidade; Especificidade; AUC, F1 score, Erro Quadrático Médio (MSE); Raiz Quadrada do Erro Quadrático Médio (RMSE).

   F. Nenhuma das respostas anteriores está correta.

8. A distribuição das classificações dos estudantes num exame de Análise de Dados, por sexo, está apresentada na figura seguinte. Qual das seguintes afirmações é verdadeira?



   A. Cerca de 75% dos rapazes têm classificações abaixo de 80%.

   B. A distribuição das classificações dos rapazes é puramente simétrica.

   C. Cerca de 25% das raparigas têm classificações superiores a 55%.

   D. A classificação mediana dos rapazes é ligeiramente inferior à classificação mediana das raparigas.

   E. A variabilidade das classificações dos rapazes é superior, quando comparada com a das raparigas.

   F. Nenhuma das respostas anteriores está correta.

9. Da seguintes opções, qual (ou quais) será a melhor medida de tendência central para resumir a seguinte amostra: 5, 6, 6, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 10, 10, 11?

   A. Apenas a moda.

   B. Apenas a média.

   C. Apenas a mediana.

   D. A moda ou a média ou a mediana.

   E. A média ou o desvio padrão.

   F. Nenhuma das respostas anteriores está correta.

10. Uma explicação razoável para um conjunto de 75 classificações de alunos apresentar média igual a 81 e mediana igual a 68 é: _____

    A. as classificações são, globalmente, muito baixas.

    B. as classificações são, globalmente, muito altas.

    C. existem alunos (poucos) com classificações muito altas.

    D. existem alunos (poucos) com classificações muito baixas.

    E. maior parte dos alunos tem classificação superior a 81.

    F. Nenhuma das respostas anteriores está correta.

Questões 11 – 15. Um investigador está interessado em saber se as variáveis *GRE* (classificação no exame de acesso ao curso de graduação), *GPA* (classificação média do curso de graduação) e prestígio da instituição de graduação, *rank* afetam a admissão num curso de pós-graduação. A variável de resposta, *admit*, é uma variável binária (1- Admitido; 0 - Não admitido). Foi estimado um modelo logístico em *R* para o propósito, entando o *output* obtido abaixo:

```
## Call:
## glm(formula = admit ~ gre + gpa + rank, family = "binomial",
##     data = mydata)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.627  -0.866  -0.639   1.149   2.079
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.98998    1.13995   -3.50  0.00047 ***
## gre          0.00226    0.00109    2.07  0.03847 *
## gpa          0.80404    0.33182    2.42  0.01539 *
## rank2       -0.67544    0.31649   -2.13  0.03283 *
## rank3       -1.34020    0.34531   -3.88  0.00010 ***
## rank4       -1.55146    0.41783   -3.71  0.00020 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 458.52  on 394  degrees of freedom
## AIC: 470.5
```

11. Das seguintes afirmações, qual é a verdadeira?

    A. Ao nível de significância de 5%, apenas uma variável apresenta um efeito estatisticamente significativo na probabilidade de admissão, mas ao nível de 3% já são duas a apresentar esse efeito.

    B. Ao nível de significância de 5%, três variáveis apresentam um efeito estatisticamente significativo na probabilidade de admissão, mas ao nível de 3% apenas uma apresenta esse efeito.

    C. Quatro variáveis apresentam um efeito estatisticamente significativo na probabilidade de admissão, ao nível de 5%.

    D. Cinco variáveis apresentam um efeito estatisticamente significativo na probabilidade de admissão, ao nível de 5%.

    E. Ao nível de significância de 5%, cinco variáveis apresentam um efeito estatisticamente significativo na probabilidade de admissão, mas ao nível de 3% apenas três apresentam esse efeito.

    F. Nenhuma das respostas anteriores está correta.

12. Ao nível de significância de 5%, qual das seguintes afirmações é falsa?

    A. Para cada aumento de unidade em *gre*, o logaritmo de chances de admissão (versus não admissão) aumenta, aproximadamente, em 0,002.

    B. Para um aumento de uma unidade no *gpa*, as probabilidades logarítmicas de ser admitido na pós-graduação aumentam, aproximadamente, em 0,804.

    C. Ter frequentado uma instituição de graduação com classificação 2, em comparação com uma instituição com classificação 1, altera as probabilidades logarítmicas de admissão em, aproximadamente, -0,675.

    D. Ter frequentado uma instituição de graduação com classificação 3, em comparação com uma instituição com classificação 1, altera as probabilidades logarítmicas de admissão em, aproximadamente, -1.340.

    E. Ter frequentado uma instituição de graduação com classificação 4, em comparação com uma instituição com classificação 1, altera as probabilidades logarítmicas de admissão em, aproximadamente, -1,551.

    F. Pelo menos, uma das respostas anteriores é falsa.

13. Das seguintes afirmações, qual é a verdadeira?

    A. As chances de ser admitido na pós-graduação (versus não ser admitido) aumentam em, aproximadamente, 0,81 por cada aumento de uma unidade em *gpa*.

    B. As chances de ser admitido na pós-graduação (versus não ser admitido) diminuem em, aproximadamente, 0,81 por cada aumento de uma unidade em *gpa*.

    C. As chances de ser admitido na pós-graduação (versus não ser admitido) aumentam em, aproximadamente, 0,04 por cada aumento de uma unidade em *gpa*.

    D. As chances de ser admitido na pós-graduação (versus não ser admitido) aumentam em, aproximadamente, 2,23 por cada aumento de uma unidade em *gpa*.

    E. As chances de ser admitido na pós-graduação (versus não ser admitido) diminuem em, aproximadamente, 2,23 por cada aumento de uma unidade em *gpa*.