Prova Final de Análise Estatística de Dados e Informações

(09/02/2025)

Nome:_____ Matrícula: _____

Questão 1 – (2,5 pontos)

Esta questão aborda a aplicação prática de um problema de Ciência de Dados utilizando **Regressão Linear**. O objetivo é prever preços de imóveis com base em dados reais da região de *King County*, nos Estados Unidos. A base de dados utilizada é a Previsão de Vendas de Imóveis em King County (EUA). Siga os passos abaixo para desenvolver sua solução:

Instruções:

1. **Análise Descritiva dos Dados** (20%)
   - Realize uma análise inicial da base de dados.
   - Inclua estatísticas descritivas (média, mediana, desvio padrão, etc.) e gráficos relevantes (distribuições, correlações, etc.).
2. **Construção do Modelo de Regressão Linear** (30%)
   - Construa um modelo de Regressão Linear para prever os preços dos imóveis.
   - Apresente os coeficientes do modelo, R^2 e outras métricas de avaliação.
3. **Interpretação dos Resultados** (10%)
   - Explique os resultados obtidos pelo modelo, destacando o impacto de cada variável nas previsões e explicações do fenômeno.
   - Verifique se os pressupostos da Regressão Linear (linearidade, homocedasticidade, normalidade dos resíduos, etc.) foram atendidos.
4. **Ajustes no Modelo** (30%)
   - Identifique possíveis problemas nos pressupostos do modelo.

- Apresente soluções para corrigir esses problemas, como transformações de variáveis ou ajustes no modelo.
- Reavalie o desempenho do modelo ajustado.

5. **Tomada de Decisão** (10%)
   - Com base no modelo final, explique como os resultados podem ser aplicados em um contexto de negócios.
   - Forneça exemplos de decisões estratégicas que poderiam ser tomadas com base nas previsões.

Questão 2 – (2,5 pontos)

Esta questão aborda a aplicação prática de um problema de Ciência de Dados utilizando **Machine Learning**. O objetivo é prever se os indivíduos irão cancelar suas reservas em uma rede de hotéis, utilizando o conjunto de dados Hotel Booking Demand. Siga os passos abaixo para desenvolver sua solução:

Instruções:

a) **Análise Descritiva dos Dados** (10%)

- Realize uma análise descritiva da base de dados.
- Inclua gráficos e tabelas para explorar as características dos dados.

b) **Modelo de Regressão Logística** (60%)

- Construa um modelo de Regressão Logística para prever o cancelamento das reservas.
- Apresente as métricas de desempenho do modelo, como acurácia, precisão, recall e F1-score.

c) **Análise das *Features*** (20%)

- Identifique as *features* mais importantes para o cancelamento das reservas.
- Interprete os resultados, destacando quais variáveis têm maior impacto na previsão.

d) **Justificativa do Método** (10%)

- Explique por que a Regressão Logística é mais apropriada para este problema em comparação à Regressão Linear.

Questão 3 – (2,0 pontos)

Esta questão aborda a aplicação prática de um problema de **ANOVA (Análise de Variância)** utilizando dados reais empregados em contextos empresariais. O objetivo é analisar as médias de quantidades e preços de produtos agrupados por países, utilizando o conjunto de dados Vendas de Varejo Online. Siga os passos abaixo para desenvolver sua solução:

Instruções:

a) **Análise Descritiva dos Dados** (10%)

- Realize uma análise inicial da base de dados.
- Inclua gráficos e tabelas que explorem as variáveis de interesse.

b) **Comparação entre Países (ANOVA)** (40%)

- Realize uma análise de variância (ANOVA) para comparar as médias de quantidade e preço dos produtos, agrupados por países.
- Apresente os resultados estatísticos, incluindo valores de F, p-valor e a interpretação dos mesmos.

c) **Ajustes no Modelo de ANOVA** (40%)

- Verifique os pressupostos da ANOVA (normalidade, homocedasticidade, etc.).
- Corrija possíveis problemas identificados e apresente um modelo ajustado.

d) **Interpretação e Tomada de Decisão** (10%)

- Interprete os resultados finais da análise.
- Destaque possíveis decisões estratégicas baseadas nos resultados encontrados.

Questão 4 – (3,0 pontos)

Esta questão aborda a aplicação prática de um problema de **Risco de Crédito**, utilizando dados reais aplicados em contextos de Análise de Dados. O objetivo é construir um modelo capaz de prever e explicar os fatores que levam bancos a classificarem clientes como bons ou maus pagadores. Para isso, utilizaremos o conjunto de dados Risco de Crédito. Siga as etapas abaixo para desenvolver sua solução:

Instruções:

a) **Discussão sobre o problema** (10%)

- Apresente uma breve discussão sobre o problema de risco de crédito, explicando sua importância no contexto bancário e econômico.

b) **Análise Descritiva dos Dados** (15%)

- Realize uma análise exploratória dos dados.
- Inclua estatísticas descritivas e gráficos que evidenciem padrões ou características relevantes.

c) **Definição e Seleção dos Modelos** (30%)

- Escolha modelos de previsão adequados para o problema.
- Justifique sua seleção com base nas características dos dados e no objetivo da análise.

d) **Explicabilidade das Variáveis – SHAP *value*** (35%)

- Analise as principais variáveis que influenciam a classificação de clientes.
- Inclua uma interpretação econômica e de negócios dessas variáveis no contexto do problema.

e) **Tomada de Decisão** (10%)

- Apresente recomendações estratégicas para a gestão do risco de crédito com base nos resultados do modelo.

Observação

1. A Prova deverá ser realizada no tempo proposto e deverá ser individual.
2. Qualquer identificação de cópia ou cola repercutirá em zero na nota final da prova.
3. Em cada questão é necessária a descrição do modelo geral requerido, com a consequente substituição dos dados especificados no problema, e resoluções.
4. A não-observância dos procedimentos ressaltados acima impedirá a justificativa dos resultados apresentados, bem como a atribuição do conceito correspondente.