

Trabalho Final de MEAD

Comentários Gerais

Prezados alunos, este é o trabalho final da disciplina Métodos Estatísticos de Apoio à Decisão do BI Master. Vocês devem resolvê-la e entregá-la no sistema, conforme especificado em cada questão.

Este trabalho pode ser resolvido em grupos de até 3 pessoas. Vocês devem entregá-la no sistema de forma individual, respeitando o prazo estipulado via postagem específica. Esse prazo é inadiável. Quaisquer dúvidas, entrem em contato com o professor.

Prof. Guilherme via prof.guilherme@ica.ele.puc-rio.br ou com o monitor via monitor.mead@ica.ele.puc-rio.br, acrescentando ao título do e-mail a tag [MEAD – código da turma] (Ex.: MEAD - 2019.3)

Questão 1 (1 pontos)

Maria irá se casar amanhã, em uma cerimônia ao ar livre no deserto. Nos últimos anos, choveu apenas 5 dias por ano no local da cerimônia. Infelizmente, a previsão do tempo é de chuva para o dia do casamento. Sabe-se que, quando de fato choveu, a previsão do tempo previu chuva em 90% dos casos e que quando não choveu, a previsão do tempo previu chuva em 10% dos casos. Qual é a probabilidade de que irá chover no dia do casamento?

Você pode entregar um arquivo .doc com a resolução do problema ou uma foto caso tenha resolvido de forma escrita.

Questão 2 (2 pontos)

Vamos analisar um dataset de resultados de partidas de futebol na Europa. Com o dataset `match_results.csv`, queremos saber se existe algum impacto significativo de jogar a partida em casa ou não (se o time ganha mais ao jogar em casa). Será necessário formular o teste de hipótese, especificando as hipóteses nula e alternativa, determinar um nível de significância para o problema, executar o teste e chegar em uma conclusão sobre o resultado.

Dica: nessa questão você irá precisar criar duas novas variáveis para começar o teste. O ideal é agrupar os resultados por time e calcular a média de vitórias em casa e a média de vitórias fora de casa e comparar essas médias.

A entrega será um arquivo jupyter notebook, com o teste de hipótese implementado e a conclusão descrita.

Questão 3 (7 pontos)

Nessa questão você terá duas opções para escolher. A primeira será um dataset próprio, onde a base de dados fica a critério do próprio aluno. Essa opção serve para o aluno escolher um dataset na sua área de trabalho ou interesse.

A segunda opção será utilizar uma base de dados de restaurantes de NY (ny_restaurant.csv). Essa base tem o resultado de uma pesquisa de avaliação de 168 restaurantes na cidade de Nova Iorque e consiste de alguns atributos:

- *Id*: identificador do restaurante
- *Restaurant*: nome do restaurante sendo avaliado
- *Price*: Preço médio de um jantar no restaurante
- *Food*: Nota média de 1 a 30 dos avaliadores para a qualidade da comida
- *Decor*: Nota média de 1 a 30 dos avaliadores para a decoração do restaurante
- *Service*: Nota média de 1 a 30 dos avaliadores para o serviço do restaurante
- *Location*: Variável categórica indicando a localização do restaurante

Nessa questão, subdividida em itens, vamos criar modelos de regressão múltipla ou regressão logística, dependendo do dataset escolhido.

a) Realizar uma análise descritiva dos dados plotando alguns gráficos e analisando algumas estatísticas básicas. (1 ponto)

b) Construir um modelo utilizando todas as variáveis do dataset. Após o ajuste do modelo será necessário interpretar cada coeficiente no impacto da variável resposta, dizendo a relação com a variável resposta e se são significantes ou não. O modelo no geral é significativo? Como podemos afirmar isso? Interprete o coeficiente de determinação ou o desvio (dependendo do modelo escolhido)? Para o dataset de restaurantes, queremos prever o preço do jantar de acordo com as outras variáveis. (2 pontos)

c) Investigar se as premissas do modelo foram violadas. (1 ponto)

d) Construir mais dois modelos diferentes do modelo saturado construído acima, justificando a escolha das variáveis. Não deixe de analisar se as premissas estão sendo violadas. (2 pontos)

e) Qual o melhor modelo? Comparar os modelos construídos com pelo menos três métricas, justificando a escolha do melhor modelo. (1 ponto)

A entrega dessa questão também será um arquivo jupyter notebook, com o modelo ajustado e as análises descritas.