FIOCRUZ

Concurso Público Fiocruz 2023

Pesquisador em Saúde Pública

Prova Discursiva

PE 64

Métodos Quantitativos

Espelho de Resposta

Pontuação de cada Questão Discursiva conforme Anexo II do Edital nº 3, de acordo com a Unidade detentora da vaga.

Espera-se que o candidato, no desenvolvimento do tema, tenha feito considerações técnicas adequadas sobre os seguintes pontos:

Questão 01

a) O algoritmo de árvore de decisão é uma técnica de aprendizado de máquina aplicada em tarefas de classificação e regressão. Ele opera dividindo iterativamente o conjunto de dados em subconjuntos menores com base nas características das variáveis explicativas, a fim de construir uma estrutura em forma de árvore que represente as decisões a serem tomadas para alcançar uma conclusão ou prever um resultado. Os componentes principais de uma árvore de decisão incluem nós e ramos. Os nós são pontos de decisão na árvore, onde ocorrem as divisões nos dados com base em condições específicas relacionadas às variáveis preditoras. Existem dois tipos principais de nós: nós de decisão, onde ocorrem as divisões nos dados, e nós de folha, onde são feitas as predições. Os ramos conectam os nós na árvore, representando o fluxo dos dados com base nas condições estabelecidas nos nós de decisão. Cada ramo representa uma escolha ou resultado possível da condição definida no nó de decisão.

   Principais etapas de construção do modelo:

   - Divisão da amostra em treinamento e teste: a amostra de dados é dividida em dois conjuntos distintos: o conjunto de treinamento e o conjunto de teste. O primeiro é utilizado para construir a árvore de decisão e ajustar seus hiperparâmetros, enquanto o segundo é empregado para avaliar o desempenho do modelo.
   - Critério de divisão: durante a construção da árvore, o algoritmo busca encontrar as melhores variáveis e pontos de divisão para separar os dados em subgrupos mais homogêneos em relação à variável resposta (no caso, a ocorrência de óbito em mulheres com câncer de mama). Diferentes critérios, como o índice de Gini e a entropia, medem a impureza dos subgrupos resultantes.

   - Critério de parada: para evitar o *overfitting* do modelo aos dados de treinamento, critérios de parada são definidos para determinar quando a construção da árvore deve cessar. Isso pode incluir limites para a profundidade máxima da árvore ou o número mínimo de observações em um nó para continuar a divisão, entre outros.
   - Ajuste de hiperparâmetros: hiperparâmetros, como profundidade máxima da árvore e número mínimo de amostras necessárias para dividir um nó, são ajustados para encontrar

a combinação ideal de valores. Técnicas como *grid search* ou otimização bayesiana são comumente utilizadas para esse fim.

- Avaliação do modelo: após a construção da árvore de decisão, o desempenho do modelo é avaliado utilizando o conjunto de teste. Métricas de avaliação comuns, como acurácia, precisão, recall (sensibilidade), especificidade, F1-score e área sob a curva ROC, fornecem uma medida objetiva de quão bem o modelo generaliza para novos dados e podem ajudar a identificar possíveis problemas, como *overfitting* ou *underfitting*.

b) Para mitigar o problema de alta variância em modelos de árvores de decisão, uma estratégia eficaz é o método de *Bagging (Bootstrap Aggregating)*. *Bagging* é uma técnica de ensemble que envolve a geração de múltiplas amostras de treinamento, cada uma sendo uma seleção aleatória com reposição dos dados originais. Cada amostra é utilizada para construir uma árvore de decisão. Em seguida, as predições dessas árvores são combinadas por meio de votação para problemas de classificação. Uma vantagem do *Bagging* é a melhoria na generalização do modelo, uma vez que a combinação das predições de várias árvores tende a resultar em um desempenho preditivo melhor em dados não observados. Além disso, o *Bagging* demonstra robustez em relação a dados de treinamento ruidosos e outliers, pois cada árvore é treinada em uma amostra diferente dos dados. No entanto, o Bagging apresenta desvantagens em termos de complexidade computacional, visto que a construção de múltiplas árvores de decisão e a combinação de suas previsões podem ser intensivas em recursos computacionais, especialmente para conjuntos de dados grandes. Além disso, a interpretabilidade do modelo pode ser comprometida, uma vez que a combinação de várias árvores pode tornar o modelo final mais difícil de ser interpretado do que uma única árvore de decisão.

   Outra abordagem que pode ser utilizada é a técnica de Florestas Aleatórias (*Random Forests*). Consiste na construção de várias árvores de decisão durante o treinamento, onde cada árvore é treinada em uma amostra aleatória dos dados (com reposição) e utilizando apenas um subconjunto aleatório das variáveis explicativas em cada divisão. A predição final é obtida por meio da votação, no caso de classificação. Esta abordagem apresenta vantagens, como a seleção automática de características, onde a escolha aleatória de variáveis em cada divisão ajuda a evitar a dominância de uma única variável e a melhorar a generalização do modelo. Além disso, é eficaz mesmo em conjuntos de dados com alta dimensionalidade e uma grande quantidade de variáveis explicativas. Por outro lado, as Florestas Aleatórias têm desvantagens, como a complexidade computacional, visto que construir várias árvores de decisão e combinar suas previsões pode ser computacionalmente intensivo, especialmente para conjuntos de dados grandes. Além disso, a interpretabilidade do modelo pode ser comprometida devido à combinação de várias árvores de decisão.

c) Um método de classificação alternativo viável para o problema em questão é a regressão logística, um modelo estatístico comumente empregado na predição de eventos binários com base em variáveis explicativas. Além disso, outras alternativas incluem Máquinas de Vetores de Suporte, Redes Neurais Artificiais, *K-Nearest Neighbors, Naive Bayes*, entre outros métodos disponíveis na literatura.
   Cabe ao candidato justificar a escolha de acordo com o algoritmo escolhido.

Questão 02

a) Como todas as variáveis categóricas são dicotômicas, pode se pensar em dois testes de acordo com a distribuição de probabilidade associada. Para a distribuição normal pode-se aplicar testes paramétricos como o teste t student para amostras independentes, o qual deve-se observar a existência ou não de homogeneidade da variância. No caso de a normalidade não poder ser observada, deve-se utilizar um teste não paramétrico como o Mann Whitney, o qual avalia a distribuição da variável considerada.
   Para as variáveis numéricas pode-se utilizar correlação de Pearson, no caso de as variáveis apresentarem distribuição normal, ou spearman.

b) Como a variáveis TSH tem distribuição continua, os pesquisadores podem pensar na família de distribuição gaussiana. Outras alternativas seriam a gamma, binomial negativa, beta etc; as funções de ligação mais comuns para essas famílias de distribuição são a identidade, log, inversa etc.; a depender da família de distribuição, é necessário avaliar normalidade, colinearidade, outlier etc; a interpretação do coeficiente do modelo irá depender da família de distribuição adotada, mas no geral este fornece a variação média da variável de exposição na variação ou ocorrência do desfecho avaliado; a qualidade do ajuste pode ser avaliada por estatísticas como deviance, erro quadrático médio, verossimilhança, AIC BIC etc.
c) A multicolineareade observada entre as variáveis pode afetar a qualidade do ajuste e interpretação dos resultados, uma vez que se torna mais difícil de avaliar o efeito isolado de uma variável sobre o desfecho. As formas mais comuns de lidar com a multicolinearidade são: exclusão das variáveis, agrupamento por meio de técnicas componente principal (aqui há perda de informação das variáveis isoladas) e modelagem por equações estruturais (aqui é possível a correlação entre as variáveis colineares, isolando desta forma cada fator).