Aula 7 – Exercício de Programação
Validação-cruzada e overfitting/underfitting

Nesse exercício de programação, você deve resolver um problema de classificação, incluindo validação-cruzada com conjunto de teste, assim como verificação de overfitting/underfitting.

1) Qual tipo de problema você deve resolver?

Esse exercício de programação foi dado pelo professor Ryan Baker no MOOC Big Data in Education, disponibilizado pela plataforma http://edx.org. Ryan Baker é um dos principais pesquisadores na área de Mineração de Dados Educacionais, sendo um dos percursores no uso de mineração de dados para predizer comportamentos e emoções de estudantes em ambientes de aprendizagem inteligentes.

Nesse exercício, ele disponibiliza uma versão simplificada da base de dados usada no artigo:

- GODWIN, K.E., ALMEDA, M.V., PETROCCIA, M., BAKER, R.S., & FISHER, A.V. (2013). CLASSROOM ACTIVITIES AND OFF-TASK BEHAVIOR IN ELEMENTARY SCHOOL CHILDREN. POSTER PAPER. PROCEEDINGS OF THE ANNUAL MEETING OF THE COGNITIVE SCIENCE SOCIETY, 2428-2433.

O arquivo que contém a base de dados foi disponibilizado no e-aula e se chama "a1-in.csv". Também disponibilizei o artigo no e-aula.

Neste exercício, você utilizará a base de dados para prever comportamentos *off-task* e *on-task*, indicados pela variável ONTASK (Y=Yes, N=No), configurando assim um problema de classificação binária. O comportamento *on-task*, por exemplo, resolver um exercício no ambiente de aprendizagem, ocorre quando os estudantes estão engajados e concentrados na atividade de aprendizado proposta. Por outro lado, o comportamento *off-task*, como conversar com colegas sobre assuntos não relacionados às atividades, navegar em conteúdo irrelevante na internet ou adotar uma abordagem de tentativa e erro na resolução de exercícios, refere-se a momentos em que os estudantes se distraem ou se envolvem em atividades alheias ao aprendizado. A identificação desses comportamentos é fundamental para compreender e aprimorar a eficácia dos ambientes educacionais.

2) O que deve ser realizado

2.1) Você deve usar todos os algoritmos vistos em aula para resolver o problema de classificação binária do exercício.

2.2) Primeiramente, você vai buscar no artigo quais são as variáveis usadas e descrever textualmente no seu código.

2.3) Teste seu algoritmo usando todas as variáveis que fazem sentido. Por exemplo, matrícula, escola, coder, entre outros, não fazem sentido para a generalização do modelo. Quais fariam sentido? Explique em uma célula textual no seu código a sua escolha.

2.4) Não esqueça de realizar todos os tratamentos necessários. OBS: Algoritmos como árvore de decisão não exigem pré-processamento, mesmo assim pode ser interessante testar se há perda e ganho fazendo tratamento de dados para esses algoritmos.

2.5) Antes de fazer a validação cruzada, você deve separar os dados em conjunto de treinamento e conjunto de teste (como fazíamos até aqui) e fazer a validação cruzada apenas no conjunto de treinamento.

2.6) **Utilize a validação cruzada para comparar diferentes algoritmos de classificação (você deve testar todos os vistos em aula) e identificar qual oferece o melhor desempenho em seu problema**. A validação cruzada ajuda a assegurar que a seleção do algoritmo não seja influenciada por uma divisão específica de treino e teste, proporcionando uma avaliação mais estável e confiável da capacidade do modelo de generalizar para novos dados.

2.7) **Após selecionar o melhor algoritmo com a ajuda da validação cruzada, treine o modelo final usando todo o conjunto de treinamento disponível** e então avalie sua performance utilizando um conjunto de teste separado. Isso fornece uma estimativa realista da performance do modelo em dados não vistos, crucial para entender como o modelo funcionará na prática.

2.8) Use todas as métricas de classificação vistas em aula para avaliação nos itens 2.5 e 2.6.

2.9) Em uma célula textual do notebook Jupiter da sua solução, descreva textualmente no seu código qual foi o algoritmo que encontrou o melhor resultado. Explique por que você acredita que aquele é o melhor resultado. Embora você deva mostrar os resultados para todas as métricas, busque aqui trazer uma explicação um pouco mais aprofundada, tentando elucidar quais as métricas são mais interessantes para esse tipo de problema e quais foram os melhores resultados de acordo com essa métrica.

2.10) Quais foram os valores das métricas para o conjunto de treinamento e teste? Você acredita que o seu modelo teve *overfitting* ou *underfitting*? Explique por que em uma célula textual do notebook Jupiter.

3) Como você será avaliado

Para lhe ajudar a desenvolver adequadamente o seu trabalho, abaixo segue as orientações serão dadas no formulário de avaliação por pares para essa atividade.

1. Entendimento e Preparação dos Dados (2 pontos)

- Identificação das Variáveis Relevantes (1 ponto): Clareza na identificação e justificativa das variáveis escolhidas com base no artigo fornecido, excluindo variáveis não relevantes como matrícula, escola, etc.
- Pré-processamento dos Dados (1 ponto): Eficiência e adequação dos métodos de pré-processamento aplicados, mesmo em algoritmos que não exigem estritamente esse passo, como árvores de decisão.

2. Aplicação de Algoritmos de Classificação (1 ponto)

- Diversidade e Implementação (1 ponto): Uso correto de múltiplos algoritmos de classificação vistos em aula.

3. Validação Cruzada (3 pontos)

- Separação de todos os dados em conjunto de treinamento e teste (1 ponto):
- Implementação e Uso Correto (1 ponto): Aplicação correta da validação cruzada para comparar o desempenho dos algoritmos e selecionar o modelo apenas no conjunto de treinamento.
- Justificativa da decisão (1 ponto): Texto de justificativa de qual foi o melhor algoritmo escolhido.

4. Análise de Desempenho (4 pontos)

- Treinamento do melhor modelo com todo o conjunto de treinamento (1 ponto): O melhor modelo escolhido na validação cruzada deve ser treinado com todo o conjunto de treinamento.
- Avaliação de Overfitting/Underfitting (2 pontos): Identificação e análise correta de sinais de overfitting ou underfitting baseada nas métricas de desempenho em ambos os conjuntos de treinamento e teste.
- Comparação e Seleção do Modelo (1 ponto): Discussão detalhada sobre a performance dos modelos em relação às métricas de classificação. O colega trouxe uma

explicação um pouco mais aprofundada, tentando elucidar quais as métricas são mais interessantes para esse tipo de problema e quais foram os melhores resultados de acordo com essa métrica.