Aula 8 – Exercício de Programação
Hiperparametrização e GridSearch

Nesse exercício de programação, você deve resolver o mesmo problema de classificação da aula passada, incluindo Hiperparametrização e GridSearch.

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) Use pipelines para realizar as diferentes etapas necessárias de tratamento.

2.5) Antes de fazer a validação cruzada com GridSearchCV, 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) Use GridSearchCV para hiper-parametrização com validação cruzada. Busque na documentação dos modelos qual os hiper-parâmetros mais indicados para cada

algoritmo e faixa de valores para testar. Explique qual foi a métrica que usou para escolha do melhor modelo no GridSearchCV e porque acha que ela é importante.

2.7) 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.

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 (3 pontos)**
    - **Identificação das Variáveis Relevantes (1 ponto):** Clareza na escolha e justificativa das variáveis selecionadas com base no artigo, eliminando variáveis irrelevantes para a generalização do modelo, como matrícula e escola.
    - **Pré-processamento dos Dados (2 pontos):** Uso eficaz de pipelines para o pré-processamento dos dados, garantindo uma preparação consistente e adequada para os algoritmos de classificação.
2. **Aplicação de Algoritmos de Classificação (4 pontos)**
    - **Separação de todos os dados em conjunto de treinamento e teste (1 ponto):** Execução correta da separação dos dados, preparando adequadamente os conjuntos para a validação cruzada.
    - **Diversidade e Implementação (3 pontos):** Aplicação correta e bem documentada de múltiplos algoritmos de classificação ensinados em aula, incluindo uma descrição detalhada do processo de hiperparametrização com GridSearchCV, enfocando na escolha dos parâmetros mais adequados para cada algoritmo.
3. **Análise de Desempenho (3 pontos)**
    - **Comparação e Seleção do Modelo (3 pontos):** Discussão detalhada sobre a performance dos modelos no conjunto de teste em relação às métricas de classificação e os efeitos da hiperparametrização. O aluno avaliado deve oferecer uma explicação aprofundada, elucidando quais foram os melhores resultados, enfatizando como os ajustes dos hiperparâmetros influenciaram os resultados finais, comparando com os resultados do exercício da aula anterior.