Aula 9 – Exercício de Programação
Feature Engineering

Neste exercício de programação, você deve modificar o exercício desenvolvido na **Aula 8** para incluir Automated Feature Engineering.

1) Problema a ser resolvido:

1.1) Melhore a solução que desenvolveu na Aula 8, incorporando feature engineering. Baseando-se no exercício da Aula 8, o objetivo é prever se o aluno está engajado na atividade (on-task) ou não, usando a variável "ON-TASK" como variável alvo. Trata-se de um problema de classificação binária.

2) Tarefas a serem realizadas:

2.1) Faça uma cópia do notebook criado na Aula 8. Escolha o modelo de melhor desempenho desse exercício para aprimorá-lo.

2.2) Crie pelo menos uma nova feature usando uma técnica de geração de feature engineering automatizada discutida na videoaula. Explique por que gerou essa feature.

2.3) Com a nova feature, realize as fases de treinamento e avaliação usando GridSearchCV e pipeline.

2.4) Crie uma caixa de texto no seu notebook e comente sobre os resultados obtidos. Eles foram melhores ou piores que os da Aula 8? Explique o porquê.

2.5) Acrescente todas as features presentes na base de dados (arquivo original csv), além da nova feature criada, e implemente um processo de seleção de features usando o método Recursive Feature Elimination (RFE).

2.6) Após a seleção de features, repita as fases de treinamento e avaliação com as modificações realizadas.

2.7) Crie uma caixa de texto e comente sobre os resultados. Eles melhoraram em relação ao que encontrado no item 2.3? Explique sua análise.

Explicação das Etapas do Exercício de Machine Learning com Foco em Pipeline

Cada etapa do processo tem um propósito específico para garantir a criação de um modelo robusto. Aqui, destacamos a importância de repetir o uso do GridSearchCV e do Pipeline após a adição de novas features e após o Recursive Feature Elimination (RFE), e explicamos o papel do Pipeline no pré-processamento de dados:

1. **Automatização do Pré-processamento com Pipeline**: Uma vantagem crucial do uso do Pipeline é a integração do pré-processamento dos dados diretamente no fluxo de modelagem. Isso significa que todas as etapas de transformação de dados, como normalização, codificação e imputação de dados faltantes (algo que não estamos fazendo muito na disciplina), são realizadas dentro do Pipeline. Essa abordagem garante que o pré-

processamento seja aplicado consistentemente durante tanto a fase de treinamento quanto de validação do modelo, evitando vazamento de dados e inconsistências.

2. **Necessidade de Limpeza de Dados Antecipada**: Antes de alimentar os dados ao Pipeline, é essencial realizar uma limpeza preliminar. Esta limpeza deve incluir a remoção de dados corrompidos ou irrelevantes (por exemplo, utilizando `dropna`), a correção de formatos e a eliminação de outliers. Tais ações são fundamentais porque o Pipeline é projetado para automatizar apenas transformações consistentes e não para tomar decisões ad hoc sobre a integridade dos dados. Portanto, toda a limpeza de dados deve ser realizada antes de inserir os dados no Pipeline para uso no GridSearchCV, enquanto outras etapas, como a normalização, são apropriadamente realizadas dentro do Pipeline.

3. **Reavaliação com GridSearchCV**: Após a introdução de novas features e a aplicação do RFE, o GridSearchCV deve ser usado novamente para ajustar os hiperparâmetros do modelo. Isso é necessário porque cada modificação nas features pode alterar a dinâmica do modelo, exigindo um reajuste para maximizar o desempenho com o novo conjunto de dados.

3) Orientações para Avaliação por Pares

Obs: Pontuação final é 12, mas o Moodle vai automaticamente ajustar a nota final a 80 (que é a nota máxima do exercício).

3.1) Criação de Cópia e Escolha do Modelo (1 ponto)

- Não realizado (0 ponto): Não há evidência de que uma cópia do notebook foi feita ou que um modelo foi selecionado.
- Realizado (1 ponto): Cópia do notebook feita.

3.2) Criação de Nova Feature (2 pontos)

- Não realizado (0 ponto): Falha em criar qualquer nova feature.
- Realizado de forma insuficiente (1 ponto): Nova feature criada, mas a técnica usada ou a relevância para o problema não é claramente explicada.
- Realizado completamente (2 pontos): Nova feature bem criada usando uma técnica automatizada apropriada, com explicação clara de seu impacto potencial.

3.3) Treinamento e Avaliação com GridSearchCV e Pipeline (2 pontos)

- Não realizado (0 ponto): Não há uso de GridSearchCV ou Pipeline na avaliação e treinamento.
- Realizado de forma insuficiente (1 ponto): Uso de GridSearchCV ou Pipeline, mas aplicação incompleta ou inadequada.
- Realizado completamente (2 pontos): GridSearchCV e Pipeline são usados eficazmente para treinar e avaliar o modelo com a nova feature.

3.4) Comentário sobre Resultados no Notebook (1 ponto)

- Não realizado (0 ponto): Falta de comentário sobre os resultados.
- Realizado completamente (1 ponto): Comentários detalhados e claros sobre os resultados, com comparação efetiva aos da Aula 8 e explicação dos motivos das mudanças observadas.

3.5) Implementação do RFE (2 pontos)

- Não realizado (0 ponto): Não implementação do método RFE.
- Realizado de forma insuficiente (1 ponto): RFE implementado, mas de forma inadequada ou incompleta.
- Realizado completamente (2 pontos): Implementação correta do RFE.

3.6) Reavaliação Pós-RFE (1 ponto)

- Não realizado (0 ponto): Falta de reavaliação após o RFE.
- Realizado completamente (1 ponto): Reavaliação completa do modelo pós-RFE, usando GridSearchCV e Pipeline.

3.7) Análise Crítica dos Resultados Finais (2 pontos)

- Não realizado (0 ponto): Não há análise crítica dos resultados finais.
- Realizado de forma insuficiente (1 ponto): Análise crítica presente, mas superficial ou não abrangente.
- Realizado completamente (2 pontos): Análise crítica detalhada e profunda dos resultados finais, incluindo a eficácia das mudanças feitas.