Projeto prático em equipe

Desenvolvimento e comparação de modelos de Aprendizado de Máquina

1. Descrição geral:

A equipe deverá desenvolver ao menos 3 modelos de aprendizado de máquina para um conjunto de dados específico de forma sistemática para selecionar o melhor modelo, relatando todo o processo.

A equipe deve ter no mínimo 4 e no máximo 5 integrantes. Este projeto só pode ser entregue em equipe. Todos os membros da equipe serão entrevistados individualmente e separadamente sobre os resultados apresentados pela equipe.

A equipe apresentará seus resultados com um relatório detalhado explicando e justificando suas decisões.

Esse projeto vale 10 pontos.

2. Requisitos do projeto

2.1 Requisitos gerais

1. O projeto deve ser realizado na plataforma Google Colab.
2. O conjunto de dados deverá ser escolhido em uma das seguintes fontes de dados abertas: UC Irvine Machine Learning Repository, Kaggle datasets e Amazon's AWS datasets.
3. O conjunto de dados precisa ser tabular, não podendo ser imagens, sons, ou outros tipos. O problema pode ser de regressão ou classificação.
4. O conjunto e dados precisa ter no mínimo 8 atributos, distribuídos entre atributos números e categóricos.
5. A equipe deverá modificar aleatoriamente o conjunto de dados escolhido. 10% das colunas devem ser selecionadas aleatoriamente para serem modificadas.

3% dos dados das colunas selecionadas devem aleatoriamente ser substituídos por valores aleatórios. 3% dos valores das colunas selecionadas devem ser aleatoriamente removidos.

6. O conjunto de dados modificado deve ser analisado e tratado usando pipelines para ser preparado para o treinamento de modelos de aprendizado de máquina. Esse passo deve desconsiderar qualquer informação anterior à modificação do conjunto de dados.
7. A equipe deverá treinar ao menos 3 modelos de algoritmos diferentes, justificando as escolhas.
8. Cada modelo treinado deve ser avaliado com validação cruzada. Modelos de classificação devem apresentar análise de precisão e revocação e curva ROC. Modelos de regressão devem apresentar análise de RMSE e MAE.
9. Para cada modelo treinado, devem ser explorados seus hiperparametros usando busca em grid para definir o melhor conjunto de hiperparametros.
10. Todos os modelos treinados devem ser regularizados de forma a minimizar o sobreajuste.
11. Todos os membros do grupo deverão estar identificados no arquivo Notebook Colab, sob pena do membro ter sua nota zerada por não ter sua participação confirmada pela equipe. A constituição da equipe tem que estar no início do documento.
12. Os nomes de variáveis, funções, métodos, classes e qualquer outra estrutura definida pela equipe no código devem ser autodescritivas, não permitindo dúvidas quanto a finalidade de uso daquela variável.
13. Todas as etapas de análise e desenvolvimento devem ser descritas usando estruturas de texto do Notebook Colab. As descrições devem atender à norma culta da língua, além de ser concisas, claras e corretas.

2.2 Requisitos da Análise do problema:

A. Entendimento do Problema:
    a. Especificação do objetivo geral e dos objetivos específicos
B. Enquadramento do Problema:
    a. Especificação do tipo de problema: Quanto à Supervisão, Quanto à Tarefa, Quanto ao Modo;
    b. Especificação das Medidas de Desempenho (para Classificação ou Regressão)

2.3 Requisitos da Análise de Dados:

A. Obtenção dos Dados
    a. Fontes de dados abertas: UC Irvine Machine Learning Repository, Kaggle datasets ou Amazon's AWS datasets
B. Divisão do Dados
    a. Amostragem estratificada em Conjunto de Treinamento e Conjunto de Teste
C. Exploração dos Dados
    a. Visualização dos Dados com gráficos de dispersão e histograma
    b. Levantamento de hipóteses sobre as distribuições dos dados

c. Busca de Correlações (Coeficientes de Correlação)

D. Preparação dos Dados
   a. Modificação dos Dados para gerar novos desafios: 10% das colunas devem ser selecionadas aleatoriamente para serem modificadas. 3% dos dados das colunas selecionadas devem aleatoriamente ser substituídos por valores aleatórios. 3% dos valores das colunas selecionadas devem ser aleatoriamente removidos.
   b. Limpeza dos Dados:
      i. Para dados Categóricos com mais de duas categorias: Codificação One-Hot ou Ordinal
      ii. Para dados Numéricos ausentes: escolher e justificar a estratégia que será usada (Ex: remover a coluna com valores ausentes, remover as linhas com valores ausentes, atribuição da mediana geral, atribuição mediana da subcategoria, treinar preditor para atribuir valores, etc...),
      iii. Escalonamento de Características: Normalização ou Padronização
      iv. Construção de Pipeline Transformador para automatizar o pré-processamento dos atributos categóricos e numéricos

2.4 Requisitos da Construção dos Modelos e Aprendizado de Máquina

A. Seleção de Modelos:
   a. Seleção de 3 modelos de algoritmos diferentes, justificando as escolhas.

B. Treinamento do Modelo
   a. Demonstrar treinamento com gráficos de curvas de treino e validação

2.5 Requisitos da Avaliação do Modelo

A. Avaliação dos Modelos:
   a. Demonstrar Validação Cruzada
   b. Análise de Desempenho:
      i. Para Classificação (precisão e revocação e curva ROC)
      ii. Para Regressão (Raiz do Erro Quadrático Médio - RMSE e Erro Médio Absoluto (MAE)

B. Refinamento dos Modelos
   a. Definir o melhor conjunto de hiperparametros (Busca em Grid)
   b. Minimizar o sobreajuste (Regularização)
   c. Comparação final do desempenho dos modelos

3. Entregas e prazos:

- Definição das equipes: 03/10/2022
- P1 – Entrega de resultados preliminares – 20%
  - Apresentação: 7/11/2022
  - Deve apresentar as etapas de análise do problema e análise de dados
  - Deve apresentar ao menos um modelo treinado com hiperparametros com valores padrão para servir de referência inicial para o desenvolvimento dos modelos.

    - A equipe apresentará o projeto para o professor, demonstrando e explicando atendimento aos requisitos.
- P2 – Entrega final – 80%
    - Entrega do arquivo final: 20/11/2022
        - Deve conter todas as etapas do processo de AM.
        - Deve ser enviado via plataforma Eldman Cursos:
            - arquivo do Google Colab (ipynb)
            - arquivo .csv ou xls da fonte de dados original
            - arquivo .csv ou xls da fonte de dados modificada
        - O arquivo deve ter explicações explicitas para as decisões tomadas no desenvolvimento do código.
    - Apresentação: 21/11/2022 a 28/11/2022
        - A data exata de cada equipe será definida pelos professores. Todas equipes devem estar preparadas para apresentar no dia 21/11/2022.
        - A não apresentação do projeto incorre em multa de 50% na nota.
        - A equipe apresentará o projeto para o professor, demonstrando o atendimento aos requisitos e justificando as decisões tomadas.
        - O professor poderá fazer perguntas e solicitar alterações do código a cada membro da equipe após a apresentação, sendo esse um dos critérios a ser utilizado pelo professor para pontuar a equipe. Essa entrevista individual define a nota individual até o máximo da nota da equipe. Caso o aluno não tenha bom desempenho, a nota individual poderá ser a nota da equipe com multa de 25%.
        - Todos os membros do projeto devem ter conhecimento total sobre todos os aspectos dele, não importando que parte específica ficou responsável individualmente.

4. Observações Gerais

Durante as aulas, no momento das práticas, as equipes podem e devem fazer consultas ao professor sobre o projeto.

Similaridades totais ou parciais entre os resultados e o material de outros autores serão penalizadas com nota zero no projeto todo.

Avaliação

- Todos os membros do grupo deverão estar identificados nos documentos enviados, sob pena do membro ter sua nota zerada por não ter sua participação ratificada pela equipe. A constituição da equipe tem que estar no início do documento.
- A nota individual de cada aluno será no máximo a nota da equipe. Essa nota só será alcançada caso o aluno tenha bom desempenho na entrevista individual realizada

na apresentação final. Caso o aluno não tenha bom desempenho, a nota individual poderá ser 75% da nota da equipe.

- Rubrica:

Formulário de Avaliação de Processo de ML (RUBRICA)						
Etapa	Critério	DESEMPENHO (PESOS)				Nota por Critério
		Inexistente (0%)	Deficiente (33%)	Aceitável (66%)	Adequado (100%)	
		0,0	0,3	0,7	1,0	
1. Análise do problema	a) Identificação da equipe	0,0	0,3	0,7	1,0	0,3
	b) Entendimento do Problema: Objetivo geral e	0,0	0,3	0,7	1,0	0,3
	c) Enquadramento do Problema (Tipo de	0,0	0,3	0,7	1,0	0,3
	d) Documentação (Código e Gramática)	0,0	0,2	0,4	0,6	0,2
2. Análise de Dados	a) Obtenção dos Dados (fontes de dados abertas determinadas)	0,0	0,3	0,7	1,0	0,3
	b) Divisão do Dados (Amostragem estratificada)	0,0	0,7	1,3	2,0	0,7
	c) Exploração dos Dados (Visualização Gráfica/Busca de Correlações)	0,0	0,7	1,3	2,0	0,7
	d) Preparação dos Dados (Modificação dos Dados/Limpeza dos Dados)	0,0	0,7	1,3	2,0	0,7
	e) Construção de Pipeline Transformador	0,0	0,3	0,7	1,0	0,3
	f) Documentação (Código e Gramática)	0,0	0,5	1,1	1,6	0,5
3. Construção dos Modelos	a) Seleção e justificativa de 3 Modelos distintos	0,0	0,7	1,3	2,0	0,7
	b) Demonstração do treinamento do Modelo (curvas de treino e validação)	0,0	0,3	0,7	1,0	0,3
	c) Documentação (Código e Gramática)	0,0	0,2	0,4	0,6	0,2
4. Avaliação do Modelo	a) Treinamento com Validação Cruzada	0,0	1,0	2,0	3,0	1,0
	b) Análise de Desempenho	0,0	1,3	2,7	4,0	1,3
	c) Refinamento dos Modelos (Randomize / Grid Search)/ Regularização	0,0	1,0	2,0	3,0	1,0
	d) Comparação final do desempenho dos modelos (Conj Teste)	0,0	0,3	0,7	1,0	0,3
	e) Documentação (Código e Gramática)	0,0	0,7	1,5	2,2	0,7
		0,0	10,0	20,0	30,0	10,0

Alcance	Penalização sobre a Nota Final do Grupo
Grupo	Falta a Apresentação (-50%)
Individual	Falta de Membro na apresentação (-25%)
Individual	Resposta Deficiente ao questionamento do Professor (-25%)