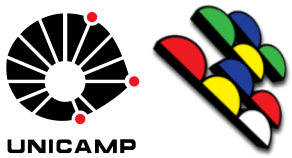

UNIVERSIDADE ESTADUAL DE CAMPINAS
ENGENHARIA E ADMINISTRAÇÃO DE SISTEMAS DE BANCO DE DADOS
CT0611 - MINERAÇÃO DE DADOS

Docente Responsável: Prof. Dr. Julio C. dos Reis [dosreis@unicamp.br]

Monitor: Eryck Pedro da Silva [eryck@unicamp.br]

Atividade 01: Classificação

Objetivo

Nesta atividade, vocês aplicarão os conceitos de mineração de dados para realizar uma classificação utilizando árvores de decisão. O objetivo é prever a sobrevivência dos passageiros do Titanic com base em informações como idade, gênero e classe do bilhete. Serão avaliadas habilidades de leitura e tratamento de dados, visualização, seleção de atributos e avaliação de modelos.

Os principais aspectos avaliados incluem:

- Leitura e tratamento básico de dados
- Aplicação de técnicas de visualização
- Identificação e tratamento de atributos relevantes para classificação
- Avaliação das classificações geradas

Cenário

O naufrágio do RMS Titanic é um dos desastres marítimos mais famosos da história. Muitos fatores influenciaram a sobrevivência dos passageiros, como classe social, gênero e idade. Nesta atividade, você utilizará um conjunto de dados real para prever se um passageiro sobreviveu ou não com base nessas informações.

Tarefas

1. Leitura e Tratamento de Dados [2,0 pontos]

- Carregar o dataset "titanic.csv" utilizando a biblioteca Pandas.
- Exibir informações gerais do dataset (quantidade de linhas e colunas, tipos de dados, valores ausentes).
- Tratar valores ausentes, substituindo-os ou removendo-os de forma justificada.
- Codificar variáveis categóricas relevantes para facilitar a análise.

2. Visualização dos Dados [2,0 pontos]

- Criar um histograma ou boxplot para visualizar a distribuição de idades.
- Criar um gráfico de barras para analisar a relação entre gênero e sobrevivência.
- Criar um gráfico de dispersão ou violin plot para verificar a influência da tarifa paga na sobrevivência.

3. Tratamento e Seleção de Dados para Classificação [3,0 pontos]

- Selecionar atributos relevantes para a classificação (evitando características/colunas que não contribuem para a previsão).
- Dividir os dados em conjuntos de treino (70%) e teste (30%).
- Implementar árvores de decisão utilizando a biblioteca Scikit-Learn:
    i. Testar diferentes critérios de geração
    ii. Testar diferentes níveis de profundidade
    iii. Averiguar quais configurações apresentam melhor ou pior desempenho, a ser analisado na Avaliação

4. Avaliação das Classificações Obtidas [3,0 pontos]

- Gerar a matriz de confusão de cada modelo gerado
- Calcular e interpretar as métricas Accuracy, Precision, Recall, F1-Score e ROC/AUC para cada modelo gerado.
- Discutir os resultados obtidos e sugerir melhorias nos modelos.
    i. Qual foi o melhor? Com quais parâmetros? De forma análoga, qual foi o melhor?

Arquivos e Ferramentas

- Arquivo de dados: "titanic.csv"
- Ferramentas: Python 3.10+, Pandas, NumPy, Matplotlib, Scikit-Learn

Submissão

- Esta tarefa pode ser realizada em dupla.
- Apenas um arquivo por dupla deve ser submetido.
- Você deve entregar um arquivo Jupyter Notebook (.ipynb), feito e baixado pelo Google Colab, subdividido de forma correspondente às Tarefas propostas nesta atividade.

- Apenas um integrante da equipe deve submeter o arquivo com a solução documentada.
- O arquivo deve ser nomeado da seguinte forma:

atividade01-classificacao-<nome_dos_integrantes>.ipynb

- [Exemplo: `atividade01-classificacao-rafael-juliana.ipynb`];
- Esta entrega tem peso de **40%** da nota final desta disciplina.
- A entrega deve ser feita até **29/03/2025 (Sábado)** às 23:59 via Moodle.

Critérios de Avaliação

- **Leitura e tratamento de dados (20%)**: Correta carga e tratamento dos dados, justificando as decisões.
- **Visualização dos dados (20%)**: Uso adequado de gráficos para explorar relações entre atributos.
- **Classificação com árvores de decisão (30%)**: Correta seleção de atributos, implementação e otimização dos modelos.
- **Avaliação do modelo (30%)**: Cálculo correto das métricas e discussão dos resultados obtidos com a comparação dos modelos gerados.