SCC0275 - Introdução à Ciência de Dados
Segundo Projeto Prático

Esse projeto tem como objetivo fazer um estudo de uma base de fraudes em compras com cartões de crédito e tentar criar um modelo que ajude a mitigar tal problema. Os dados de treinamento estão disponíveis no arquivo **`train.csv`** e os de teste no arquivo **`test.csv`**.

Utilize a biblioteca scikit-learn para criar os modelos pedidos. Você pode usar outras bibliotecas para fazer as demais análises =)

Todas as respostas devem ser justificadas com base em:

1. **Código Python mostrando a(s) análise(s) e/ou o(s) modelo(s) feitos;**
2. **O resultado da(s) análise(s) e/ou do(s) modelo(s) e**
3. **Uma explicação textual (pode ser breve) da conclusão obtida.**

Em caso de plágio (mesmo que parcial) o trabalho de todos os alunos envolvidos receberá nota ZERO.

NÃO enviar um link para o Google Colab como resposta/relatório do projeto.

Desorganização excessiva do código resultará em redução da nota do projeto. Exemplos:

- **Códigos que devem ser rodados de forma não sequencial;**
- **Projeto entregue em vários arquivos sem um README;**
- **...**

Bom projeto,
Tiago.

Questão 1 (valor 2.5 pontos)

a) Identifique a variável resposta do problema (coluna da tabela que indica se uma transação é fraude ou não) e faça um histograma mostrando a sua distribuição.
b) A base de dados conta com 31 colunas, das quais: 1 é a variável resposta, 29 são variáveis explicativas e uma é um metadado. Identifique a coluna que contém algum metadado. Justifique sua resposta.

Dica: a variável que contém metadados **NÃO** deverá ser usada nos modelos das próximas questões.

Questão 2 (valor 2.5 pontos)

a) Compare as métricas acurácia e AUC para essa base de dados simulando os seguintes modelos:
    - Modelo que classifica aleatoriamente (50% chance de dizer que é fraude e 50% de dizer que não é);
    - Modelo que classifica todos os casos como fraude;
    - Modelo que classifica todos os casos como não fraude.
b) Qual das duas métricas de avaliação deve ser usada para medir os resultados dos modelos nesta base de dados: Acurácia ou AUC?

Questão 3 (valor 2.5 pontos) - A total corretude dessa questão depende das questões anteriores

a) Usando a classe **`RandomForestClassifier`** do scikit-learn faça um 3-fold cross-validation para comparar todas as combinações dos seguintes valores de hiperparâmetros:
    - `n_estimators = [10, 50, 100, 200]`;
    - `max_depth = [2, 3, 4, 5]`.

b) Refaça os experimentos do item a) usando `class_weight='balanced'`. Explique o que esse valor do hiperparâmetro faz e analise os resultados obtidos ao usar ele.

Importante:

- Em todos os experimentos use `random_state = 42` e deixe os demais hiperparâmetros com seu valor padrão;
- Calcule o resultado do melhor modelo de cada item na base de teste. Use a métrica escolhida na questão 2.

Questão 4 (valor 2.5 pontos)

Usando o melhor modelo obtido na questão anterior, calcule o lucro que tal modelo traria no seguinte cenário:

- As top-1% transações com maior chance de fraude (de acordo com os scores do modelo) seriam impedidas de acontecer;
- Cada fraude evitada em média evita um prejuízo (gera um lucro) de R$ 100 e
- Cada não-fraude bloqueada gera em média um prejuízo de R$ 2.

Dica: Para encontrar as transações com maior chance de fraude você pode utilizar o resultado da função `predict_proba`.