Nome: _____Matrícula:_____
Data:__/__/____

EXERCÍCIO PRÁTICO DE APRENDIZADO SUPERVISIONADO
CLASSIFICAÇÃO

Considere a seguinte situação:

- Uma agência de venda de veículos pretende estimar se um determinado veículo será vendido ou não com base em 3 características: - Preço, - Quilometragem, - Ano de Fabricação.

Os dados para treino e testes do modelo encontram-se no arquivo CSV: ***car-prices.csv***

mileage_per_year	**model_year**	**price**	**sold**
21801	2000	30941.02	yes

- **mileage_per_year:** indica a quantidade de milhas que o carro rodou por ano desde sua fabricação.
- **model_year:** indica o ano de fabricação.
- **price**: indica o preço.
- **sold**: Indica se o carro foi vendido ou não (yes / no).

ATIVIDADES:

1) Crie um programa em Python de aprendizado de máquinaque possa ler os dados do arquivo CSV, treinar e aprender com base nos registros e testar possíveis classificações de novos carros que poderão ser vendidos ou não.

O QUE DEVE SER FEITO:

- **ETAPA 1: PRÉ-PROCESSAMENTO DOS DADOS - TRANSFORMAÇÃO – PARTE 1**
  1. Renomeie as colunas para português utilizando o *rename* do *Pandas*.
  2. *Sold*: Mapeie a coluna *sold* para uma nova coluna com valores 1(yes) ou 0(no).
  3. *Model_year*: Ao invés de usar a dimensão ano do modelo, crie uma nova coluna que armazene a idade do veículo (Ano atual - ano do modelo).
  4. Deixe o dataframe somente com as colunas (milhas_por_ano, idade, preco, vendido). Delete as demais colunas.
- **ETAPA 2: VISUALIZAÇÃO DOS DADOS**

1. Plote um gráfico de dispersão dos dados – verifique uma forma de exibir 3 dimensões.
    1. **Responda: É possível um modelo linear resolver o problema?**
    2. **Responda: Há ruídos?**
2. Plote um gráfico de pizza para a contagem de exemplos das classes.
    1. **Responda: Há desbalanceamento de classes?**
3. Plote histogramas e gráfico de caixas para analisar os atributos numéricos.
    1. **Responda: Há atributos com outliers?**
    2. **Responda: Há diferenças de escalas entre os atributos?**
4. Plote um gráfico de calor para verificar a correlação dos atributos.
    1. **Responda: É possível remover algum atributo?**

- **ETAPA 3 – DIVISÃO DE AMOSTRAS – PARTE 1**
    1. Divida o dataframe em X(atributos) e Y(classes).
- **ETAPA 4 – PRÉ-PROCESSAMENTO DOS DADOS – P TRANSFORMAÇÃO - PARTE 2**
    1. Se houver atributos alfanuméricos, efetue a transformação utilizando a técnica – *One Hot Encoding* ou *Label Encoder*.
    2. Normalize os dados dos atributos utilizando a abordagem – padrão 0 / 1. Armazene em um atributo específico.
    3. Normalize os dados dos atributos utilizando a abordagem – Z score. Armazene em um atributo específico.
- **ETAPA 5 – DIVISÃO DE AMOSTRAS**
    1. Utilizando a técnica *Hold Out*, efetue a divisão dos dados em uma amostra para TREINO o outra para TESTE. Utilize a proporção 70 / 30. Estratifique as classes.
        1. **Responda: Com base nas respostas da ETAPA 2, você vai utilizar todo o conjunto de dados para treinar modelos de aprendizado de máquina?**
- **ETAPA 6 – TREINAMENTO DO MODELO**
    1. Efetue treino com um modelo linear
    2. Efetue treino com o KNN
    3. Efetue treino com árvore de decisão
- **ETAPA 7 – AVALIAÇÃO DOS MODELOS**
    - Para cada modelo treinado acima, efetue testes com a massa de testes e avalie:
    1. Acurácia
    2. Matriz de Confusão
    3. F1 – Score
    4. Plote a Curva ROC
        1. **Responda: Qual foi o melhor modelo para se colocar em produção?**

```
from sklearn.dummy import DummyClassifier
from sklearn.svm import LinearSVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
```