Atenção

Curso de Extensão em 2025.2

Introdução a Ciência de Dados

Edital previsto para ser publicado dia

28/08/2025.

Fiquem atentos os interessados.

Realização

LADE
LABORATÓRIO DE ANÁLISES DE DADOS EDUCACIONAIS E ESTATÍSTICA APLICADA
IFCE - CAMPUS FORTALEZA

INSTITUTO FEDERAL
Ceará
Campus Fortaleza

DEFIMAT
DEPARTAMENTO DE FÍSICA E MATEMÁTICA
IFCE - Campus Fortaleza

Diretoria de Extensão e Relações Empresariais – IFCE Campus Fortaleza

INSTITUTO FEDERAL
Ceará
Campus Fortaleza

CURSO DE EXTENSÃO
2025.1

DEFIMAT

Curso Introdução à Ciência de Dados

Aluno: _____ **Professor Valberto Feitosa**	**Data: _____/_____/_____**

1. (0,25) O conceito de Aprendizado de Máquina vem sendo aperfeiçoado desde meados do século XX. Com base nas definições clássicas sobre o tema, qual das opções a seguir **melhor descreve** o que é Aprendizado de Máquina?

a) Um conjunto de técnicas para substituir sistemas especialistas na tomada de decisão.
b) O processo de construir redes neurais profundas com milhares de parâmetros treináveis.
c) A capacidade de um programa melhorar seu desempenho em uma tarefa com base em dados e experiência.
d) A criação de sistemas que funcionam sem necessidade de qualquer programação.
e) Uma abordagem estatística restrita ao reconhecimento de padrões visuais em imagens.

2. (0,25) Considere um modelo de regressão que estima a **Satisfação de Vida** com base no **PIB per capita**. Dentre as métricas abaixo, qual é **mais sensível a erros grandes**, penalizando-os com mais intensidade?

a) Acurácia.
b) MAE – Erro Médio Absoluto.
c) RMSE – Raiz do Erro Quadrático Médio.
d) Recall – Sensibilidade.
e) R^2 – Coeficiente de Determinação.

3. (0,25) Sobre a ferramenta Orange Data Mining, analise as afirmativas:

I. É uma ferramenta com código aberto e foco em fluxos interativos de aprendizado de máquina.
II. Requer conhecimento avançado em Python e scripts para aplicar modelos de regressão e classificação.
III. Possui interface baseada em blocos conectáveis (drag-and-drop), ideal para prototipagem rápida.
IV. É voltada apenas para especialistas em ciência de dados.

Assinale a alternativa correta:
a) Apenas I e III estão corretas.
b) Apenas II e IV estão corretas.
c) Apenas I e IV estão corretas.
d) Apenas I, II e III estão corretas.
e) Todas estão corretas.

4. (0,25) Sobre o tratamento e caracterização de dados em projetos de Ciência de Dados, analise as afirmativas:

I. A etapa de limpeza envolve a identificação e correção de dados ausentes, duplicados e inconsistentes.
II. A imputação pela média é sempre recomendada para valores faltantes, pois mantém a média geral do conjunto.
III. A caracterização dos dados inclui entender os tipos dos atributos, como qualitativos ou quantitativos.
IV. A análise exploratória dos dados pode incluir o uso de visualizações para revelar padrões.

Assinale a alternativa correta:
a) Apenas I e II estão corretas.
b) Apenas I, III e IV estão corretas.
c) Apenas II e IV estão corretas.

d) Apenas III e IV estão corretas.
e) Todas estão corretas.

5. (0,25)

Asserção (A): O uso da padronização (z-score) é recomendado quando os atributos possuem unidades e escalas distintas.
Razão (R): A padronização transforma os dados para que tenham média zero e desvio padrão igual a um.

a) A e R são verdadeiras, e R justifica A.
b) A e R são verdadeiras, mas R não justifica A.
c) A é verdadeira, e R é falsa.
d) A é falsa, e R é verdadeira.
e) A e R são falsas.

6. (0,25)
Asserção (A): A acurácia é uma boa métrica em problemas de classificação com classes desbalanceadas.
Razão (R): Quando há grande desbalanceamento entre as classes, métricas como precisão, recall e F1-score são mais informativas.

a) A e R são verdadeiras, e R justifica A.
b) A e R são verdadeiras, mas R não justifica A.
c) A é verdadeira, e R é falsa.
d) A é falsa, e R é verdadeira.
e) A e R são falsas.

7. (0,75) A **visualização de dados** é uma ferramenta fundamental na ciência de dados, permitindo interpretar padrões, tendências e possíveis anomalias no conjunto analisado.

a) (0,25) Explique a função dos seguintes gráficos:

- **i. Gráfico de linhas**
- **ii. Histograma**
- **iii. Boxplot**

b) (0,25) Apresente um exemplo prático de uso de cada um desses gráficos em um projeto real de ciência de dados, destacando seu papel na análise.

c) (0,25) Entre os gráficos apresentados, qual é mais indicado para identificar **outliers**? Justifique sua resposta com base em suas características visuais e interpretativas.

8. (0,75) Considere a seguinte tabela, que representa a **quantidade de horas de estudo por dia** e a **nota obtida** por 10 estudantes em uma avaliação:

Estudante	Horas de Estudo	Nota Obtida
A	1.0	55
B	2.5	63
C	3.0	66
D	4.0	71
E	2.0	60
F	5.0	75
G	4.5	73
H	3.5	69
I	2.5	63
J	1.5	58

a) (0,25) Construa um **gráfico de dispersão (scatter plot)** relacionando “Horas de Estudo” (eixo X) com “Nota Obtida” (eixo Y). Explique o padrão observado.

b) (0,25) Calcule os seguintes valores para a variável **Nota Obtida**:

- Média
- Mediana
- Moda

c) (0,25) Com base nas medidas de tendência central calculadas, o desempenho dos estudantes pode ser considerado homogêneo? Justifique.

INSTITUTO FEDERAL
Ceará
Campus Fortaleza

CURSO DE EXTENSÃO
2025.1

DEFIMAT

Curso Introdução à Ciência de Dados

Aluno: _____ **Professor Valberto Feitosa**	**Data: _____/_____/_____**

1. (0,25) Ao avaliar a performance de diversos modelos preditivos para um problema de regressão e outro de classificação, várias métricas podem ser utilizadas para determinar qual modelo oferece o melhor desempenho. Considere as métricas para regressão e classificação, bem como as técnicas de detecção de overfitting e underfitting.

Nesse contexto, quais métricas devem ser utilizadas para determinar qual modelo oferece o melhor desempenho?

a) Para avaliar um modelo de regressão, deve-se utilizar a métrica Accuracy (acurácia) para determinar a proporção de previsões corretas, enquanto, no problema de classificação, o uso do R^2 ajustado é essencial para medir a variabilidade explicada pelo modelo, ajustada pelo número de features.

b) Para modelos de classificação, é importante utilizar a métrica R^2 para entender a proporção da variância explicada pelo modelo, enquanto o F1-score deve ser utilizado em problemas de regressão em que há um equilíbrio significativo entre as classes.

c) Para a detecção de overfitting e de underfitting, pode ser realizada a observação do trade-off entre viés e variância nas curvas de aprendizagem, independentemente do tipo de modelo (regressão ou classificação), sendo as métricas Accuracy e R^2 ajustado suficientes para medir a performance em ambos os casos.

d) No problema de regressão, o RMSE (Root Mean Square Error) é ideal para avaliar a média dos erros ao quadrado das previsões, e, para problemas de classificação, a análise da matriz de confusão permite calcular métricas como Precision, Recall, e F1-score, auxiliando na detecção de overfitting e underfitting.

e) No problema de regressão, o uso do MAE (Mean Absolute Error) é preferível ao RMSE (Root Mean Square Error) quando se deseja penalizar fortemente grandes erros, e, para problemas de classificação, a matriz de confusão é suficiente para detectar overfitting e underfitting.

2. (0,25) O teorema de Bayes é um mecanismo formal para atua- lizar probabilidades. Considere o caso de um analista de mercado que, após o encerramento de um pregão, pretende divulgar informações sobre a probabilidade de que- da de determinada ação. O analista tinha uma previsão inicial de queda dessa ação de 10% e recebeu novas in- formações sobre a economia, no que diz respeito a um aumento da taxa de juros. O analista tem registros de que, quando houve queda nessa ação, em 20% das vezes essa queda foi precedida pelo aumento dos juros e de que, nos dias em que a ação esteve em alta, apenas em 5% das vezes elas foram precedidas pela notícia de aumento da taxa de juros. Levando-se em conta esse cenário, e com base no teo- rema de Bayes, a nova probabilidade de queda da ação será de

a) 45% b) 40% c) 33% d) 31% e) 29 %

3. (0,25) Sobre Árvores de Decisão, considere:

I. O critério de divisão mais comum é o ganho de informação, que pode ser calculado a partir de medidas como entropia ou índice Gini.
II. São suscetíveis a overfitting e instabilidade quando pequenas alterações nos dados geram árvores diferentes.
III. Exigem escalonamento prévio das variáveis para funcionarem corretamente.
Assinale a alternativa correta:

a) Apenas I está correta.
b) Apenas I e II estão corretas.
c) Apenas II e III estão corretas.
d) Apenas III está correta.
e) Todas estão corretas.

4. (0,25) Sobre a Regressão Logística, considere:

I. É utilizada para modelar variáveis dependentes binárias, estimando a probabilidade de ocorrência de um evento.
II. Garante que as previsões fiquem no intervalo [0,1] por meio da função sigmoide.
III. Utiliza o método dos mínimos quadrados para estimar os parâmetros, garantindo uma função de custo convexa.
Assinale a alternativa correta:

a) Apenas I está correta.
b) Apenas II está correta.
c) Apenas I e II estão corretas.
d) Apenas III está correta.
e) Todas estão corretas.

5. (0,25)

Asserção (A): A regressão linear é utilizada para prever valores contínuos e requer que a relação entre variáveis seja linear.
Razão (R): Isso ocorre porque a regressão linear estima parâmetros que minimizam a soma dos erros quadráticos, assumindo linearidade nos coeficientes.

a) A e R são verdadeiras, e R justifica A.
b) A e R são verdadeiras, mas R não justifica A.
c) A é verdadeira, e R é falsa.
d) A é falsa, e R é verdadeira.
e) A e R são falsas.

6. (0,25)

Asserção (A): No algoritmo K-Vizinhos Mais Próximos (KNN), utilizar valores ímpares de K em problemas de classificação binária ajuda a evitar empates.
Razão (R): Isso ocorre porque o KNN classifica uma instância com base na classe mais frequente entre seus vizinhos mais próximos, e um número ímpar reduz a chance de empate.

a) A e R são verdadeiras, e R justifica A.
b) A e R são verdadeiras, mas R não justifica A.
c) A é verdadeira, e R é falsa.
d) A é falsa, e R é verdadeira.
e) A e R são falsas.

7. (0,75)

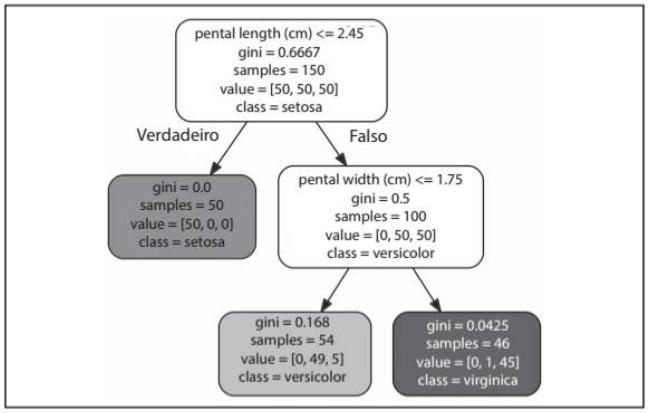
a) (0,25) Quais são as vantagens e desvantagens do algoritmo KNN ?

b) (0,50) Utilize a base de dados abaixo para classificar o ponto (9,5). Para isso, considere $k = 5$ e apresente todos os cálculos necessários.

Instância	Classe	x_1	x_2
Teste	**?**	**9**	**5**
1	1	4	9
2	1	4	3
3	2	5	12
4	2	14	7
5	2	11	3
6	1	8	6
7	2	9	8
8	1	6	8
9	1	8	3
10	1	2	8
11	1	2	6
12	2	12	8
13	2	10	8
14	2	11	12

8. (0,75)

a) (0,25) A figura abaixo mostra uma árvore de decisão treinada para classificar flores Iris em três espécies: setosa, versicolor e virginica. Os atributos utilizados foram medidas das pétalas (petal length e petal width).



Explique o significado de cada informação presente nos nós da árvore:

I- Condição (ex.: petal length (cm) <= 2.45)
II- gini
III- samples
IV- value
V- class

b) (0,50) Implemente o código em Python que treina o modelo de árvore de decisão para a base **Iris** e gera a árvore de decisão mostrada na figura, incluindo a plotagem. Utilize o trecho inicial abaixo e **complete o restante** .

```
from sklearn.datasets import load_iris
from sklearn.tree import DecisionTreeClassifier, plot_tree
import matplotlib.pyplot as plt

# Carregar dataset Iris

iris = load_iris()
X, y = iris.data[:, 2:], iris.target  # apenas petal length e petal width
```

Desafio Técnico – Cientista de Dados - 01

Cenário

Evasão de clientes em uma instituição financeira (Exited).

Objetivo

Você recebeu um conjunto de dados contendo informações sobre 10.000 clientes de um banco. O conjunto inclui atributos relacionados ao perfil, comportamento e situação financeira de cada cliente. A variável de interesse indica se o cliente permaneceu ou deixou a instituição.

Tarefa

Neste desafio, espera-se que você:

1- Elabore uma linha de trabalho coerente e bem estruturada, com base no problema apresentado;

2- Execute essa linha de ação até o ponto que julgar apropriado, de acordo com seu próprio

planejamento;

3- Utilize esse processo para explorar o comportamento dos dados e levantar hipóteses iniciais

sobre os fatores mais relevantes relacionados à variável de interesse (selecionar as variáveis mais importantes para prever o comportamento observado);

4- Documente, no próprio código, o planejamento adotado e um resumo das conclusões preliminares que surgirem ao longo da execução.

Sobre os dados

A base disponibilizada é composta por dados estruturados, abrangendo variáveis demográficas, financeiras e operacionais. Trata-se de um conjunto apropriado para abordagens voltadas à compreensão de padrões, levantamento de indicadores e identificação de perfis.

Dataset fornecido: Churn_Modelling.csv, com as seguintes variáveis:

Variável	Tipo	Descrição
`RowNumber`	Identificador	Índice da linha no dataset.
`CustomerId`	Identificador	ID único do cliente.
`Surname`	Categórica	Sobrenome do cliente.
`CreditScore`	Numérica	Pontuação de crédito (quanto maior, melhor).
`Geography`	Categórica	País de residência (`France` , `Spain` , `Germany`).
`Gender`	Categórica	Sexo do cliente (`Male` , `Female`).
`Age`	Numérica	Idade do cliente.
`Tenure`	Numérica	Tempo (em anos) como cliente.
`Balance`	Numérica	Saldo da conta.
`NumOfProducts`	Numérica	Número de produtos bancários que o cliente possui.
`HasCrCard`	Binária	Possui cartão de crédito (1 = sim, 0 = não).
`IsActiveMember`	Binária	Cliente é ativo (1 = sim, 0 = não).
`EstimatedSalary`	Numérica	Salário estimado.
`Exited`	Binária (Alvo)	1 = cliente saiu do banco, 0 = permaneceu.

Acessando o Dataset

```
[ ] # 📦 ETAPA 1 – Importação de Bibliotecas e Leitura dos Dados

    # Bibliotecas essenciais para análise de dados
    import numpy as np
    import pandas as pd

    # (Opcional) Biblioteca para baixar datasets do Kaggle programaticamente
    import kagglehub  # Certifique-se de que esta biblioteca está instalada no ambiente

    # 🔽 Download do dataset hospedado no Kaggle
    # Isso fará o download do dataset para o diretório local do ambiente
    path = kagglehub.dataset_download("kartiksaini18/churn-bank-customer")

    # Mostra o caminho em que os arquivos foram salvos
    print("Caminho para os arquivos do dataset:", path)

    # 📄 Leitura do arquivo CSV após download
    # Substitua o caminho abaixo se necessário, de acordo com onde o arquivo foi salvo no seu ambiente
    df = pd.read_csv("/kaggle/input/churn-bank-customer/Churn_Modelling.csv")

    # Visualização das 5 primeiras linhas da base de dados
    df.head()
```

Desafio Técnico — Classificação no Iris (offline) - 02

Objetivo

Treinar, comparar e justificar modelos de classificação no dataset Iris, explorando subconjuntos de atributos (features), pré-processamento, validação e métricas. Entregar código bem documentado e um relato curto com conclusões.

Regras

1- Sem internet e sem uso de IA.
2- Use apenas as bibliotecas já importadas no esqueleto abaixo.
3- Organize o código, com comentários claros e funções quando fizer sentido.
4- Fixe sementes aleatórias para reprodutibilidade.

Tarefas

1- Carregar e inspecionar o Iris.
2- Definir três cenários de features (ex.: só pétalas; só sépalas; todas as 4 medições).
3- Preparar pipelines de pré-processamento e modelos (ex.: StandardScaler quando necessário).
4- Treinar e comparar pelo menos 3 modelos (sugestões: LogisticRegression, KNeighborsClassifier, DecisionTreeClassifier, Bayes).
5- Avaliar com validação estratificada (ex.: StratifiedKFold, cross_val_score) e métricas (acurácia, macro F1). Mostre também matriz de confusão do melhor modelo em holdout.
6- Escrever um mini-relatório (10–15 linhas) com: melhor combinação features+modelo, métricas, e justificativa técnica.

Entregáveis

1 arquivo .py ou .ipynb com o código executável e comentado.
1 mini-relatório ao final do notebook/arquivo (markdown ou bloco de texto).

```python
# =========================
# DESAFIO TÉCNICO - IRIS
# =========================

# Imports (não altere esta seção)
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split, StratifiedKFold, cross_val_score
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.metrics import accuracy_score, f1_score, confusion_matrix, ConfusionMatrixDisplay, classification_report

# Modelos sugeridos (use pelo menos 3)
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB

RANDOM_STATE = 42
np.random.seed(RANDOM_STATE)

# ==============
# 1) Carregar dados
# ==============
iris = load_iris()
X_full = pd.DataFrame(iris.data, columns=iris.feature_names)
y = pd.Series(iris.target, name="target")

# Nome amigável das classes
target_names = iris.target_names
```

Trabalho em Equipe

Curso de Extensão: Introdução à Ciência de Dados

Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) – Campus Fortaleza

Nome do Aluno: _____
Data: _____
Professor: Valberto Feitosa

Para finalizar a **primeira avaliação**, que é composta por **atividade escrita, trabalho em equipe e atividade individual**, seguem as orientações para o trabalho em equipe:

Sua equipe deverá:

1- Escolher um **dataset** que envolva um problema de **regressão** ou **classificação**.
2- Descrever o **dataset**, suas **variáveis** e o **problema** que desejam resolver.
3- Realizar o **pré-processamento dos dados**, tratando valores ausentes, codificando variáveis, normalizando, entre outros procedimentos necessários.
4- Aplicar a **estatística descritiva** para uma compreensão inicial dos dados.
5- Selecionar as **variáveis preditoras** que entrarão no modelo com base na **teoria abordada em sala de aula**.
6- Repetir o processo de seleção de variáveis utilizando o método **SelectKBest**.

Entrega

Elaborar uma **apresentação** clara e objetiva com os principais resultados obtidos. A apresentação deverá ser realizada no dia **19/07/2025**, com **tempo máximo de 20 minutos** por equipe.

Atenção: No dia **12/07/2025** faremos o atendimento às equipes, que deverão **apresentar uma prévia** do que irão apresentar.

Trabalho Prático – Aplicação dos Algoritmos de Machine Learning

Curso de Extensão: Introdução à Ciência de Dados

Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) – Campus Fortaleza

Nome do Aluno: _____

Data: _____

Professor: Valberto Feitosa

---

1. Objetivo

Os alunos deverão:

1. Escolher uma base de dados adequada para treinar um modelo de classificação ou regressão, de acordo com os algoritmos estudados em sala.
2. Aplicar pelo menos dois dos algoritmos vistos em aula (ex.: Regressão Logística, Árvore de Decisão, Regressão Linear, Random Forest, ...).
3. Avaliar o desempenho de cada algoritmo utilizando métricas apropriadas.
4. Salvar o modelo com melhor desempenho para uso posterior.
5. Desenvolver uma aplicação web simples (front-end e back-end) que permita ao usuário inserir novos dados e obter previsões utilizando o modelo salvo.

---

2. Etapas do Trabalho

2.1 Parte 1: Escolha da Base de Dados

- A base de dados deve ser compatível com problemas de classificação ou regressão.
- Justifique a escolha, descrevendo as variáveis (features) e a variável-alvo (label).
- Exemplos:
  - Previsão de doenças
  - Risco de crédito
  - Classificação de clientes (churn)
  - Previsão de preços (ex.: imóveis, veículos)

2.2 Parte 2: Treinamento dos Modelos

- Carregar a base de dados e aplicar pré-processamento (se necessário).
- Dividir em dados de treinamento e teste.
- Treinar pelo menos dois algoritmos diferentes estudados em aula.
- Avaliar o desempenho utilizando métricas apropriadas (acurácia, precisão, recall, F1, RMSE, R^2, conforme o tipo do problema).
- Escolher o melhor modelo e salvá-lo com `joblib` ou `pickle`.

2.3 Parte 3: Desenvolvimento da Aplicação Web

Front-End

- Criar um formulário HTML que permita ao usuário inserir valores das variáveis preditoras.
- Incluir um botão "Prever" para enviar os dados.

Back-End

- Utilizar Flask ou Django.
- Carregar o modelo salvo.
- Receber os dados do formulário, processar e retornar a previsão.

Exibição do Resultado

- Mostrar ao usuário o resultado da previsão na própria página web.

---

3. Ferramentas e Tecnologias

- Linguagem: Python
- Bibliotecas ML: scikit-learn, pandas, numpy
- Salvar modelo: joblib ou pickle
- Framework web: Flask ou Django
- Front-End: HTML, CSS, JavaScript
- Comunicação front-end/back-end: HTTP POST

---

4. Entregáveis

1. Código-fonte completo (notebook ou scripts Python).
2. Aplicação web funcional.
3. Relatório PDF contendo:
    - Descrição da base de dados
    - Algoritmos utilizados
    - Comparação de métricas
    - Prints de tela da aplicação web em funcionamento