Universidade Federal do Ceará
Centro de Ciências
Departamento de Computação

# CKP9011 – Introdução à Ciência de Dados
# CK0223 - Mineração de Dados
**2025.1**

**Lista 5**

Exercício: Regressão
Objetivos: Exercitar os conceitos referente à regressão.
Data da Entrega: 02/06/2025

## 1. Tarefa

Crie um arquivo Jupyter Notebook e realize as seguintes operações:

a) Ler o dataset fakeTelegram.BR_2022.csv, o qual está disponível no link a seguir: https://drive.google.com/file/d/1c_hLzk85pYw-huHSnFYZM_gn-dUsYRDm/view?usp=drive_link
b) Remova os trava-zaps, as linhas repetidas (duplicadas) e textos com menos de 5 palavras.
c) Agrupe as linhas com postagens iguais ou extremamente semelhantes. Aqui você pode utilizar uma métrica de semelhança de textos. Crie uma variável para representar a quantidade de vezes que a mensagem foi compartilhada. Observe que ao agrupar linhas que possuem a "mesma" postagem (texto), você deve escolher como valor para as variáveis data e hora da postagem, os valores da cópia mais antiga.
d) Você pode criar novos atributos numéricos, tais como: quantidade de palavras, quantidade de caracteres etc.

Utilizando os dados referente a postagens no Telegram, crie um modelo preditivo (regressor) para, dado os dados de uma postagem, prever a quantidade de compartilhamentos dessa mensagem, o que é denominado potencial de "viralização".

A avaliação experimental deverá considerar:

e) O modelo de regressão: regressão linear multivariada, regressão polinomial multivariada;
f) Explorar funções: exponencial, seno, cosseno (OPCIONAL);
g) Regularização: Com regularização (Ridge, Lasso ou ElasticNet) e sem regularização;
h) Normalização dos dados: sem normalização, Z-Score, Min-Max (OPCIONAL);
i) Pré-processamento de dados: sem pré-processamento e com pré-processamento;
j) Embedding: BOW, TF-IDF, Word2Vec;
k) N-Gramas: unigramas, bigramas, trigramas;
l) Treinamento, Validação e Teste: Outer K-Fold Cross-Validation;

## 2. Avaliação

Espera-se com a realização deste trabalho que cada estudante elabore e entregue (de forma digital) os seguintes documentos:

- Jupyter Notebook contendo o código utilizado na implementação das tarefas.
- Vídeo (disponibilizado no Youtube) apresentando e descrevendo as atividades desenvolvidas.

A avaliação deste trabalho se dará em duas etapas:

1ª. Vídeo de Apresentação do Dataset: Cada estudante irá disponibilizar um vídeo (no Youtube) apresentando o código desenvolvido para implementação das tarefas. O estudante pode utilizar slides e notebooks.

2ª. Avaliação do Notebook: O professor da disciplina irá avaliar a qualidade do notebook gerado pelo estudante, bem como dos códigos implementados e análises realizadas.

A avaliação do trabalho irá envolver os seguintes quesitos:

- Abrangência e Organização do Notebook
- Qualidade dos Códigos Utilizados
- Clareza do Texto Utilizado para Descrever as Atividades Realizadas e os Resultados Obtidos
- Domínio do Tema

**3. Data da Entrega:** 02/06/2025

- PS. O trabalho é individual.
- PS. Não serão aceitos trabalhos que não forem apresentados (por meio de vídeo disponibilizado no Youtube).
- PS. Cada estudante será responsável pela disponibilização do ambiente (software e hardware) necessário para a gravação da apresentação do seu trabalho.
- Os Notebooks deverão ser disponibilizados, em formato .ZIP, no SIGAA ou em um repositório público (GitHub ou GitLab).

"A Educação, qualquer que seja ela, é sempre uma teoria do conhecimento posta em prática".

**Paulo Freire**