Universidade Federal do Ceará
Centro de Ciências
Departamento de Computação

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

Lista 7

Exercício: Classificação Multiclasse
Objetivos: Exercitar os conceitos referente à classificação binária.
Data da Entrega: 30/06/2025

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.

Questão 1

Utilizando os dados referente a postagens no Telegram, crie um modelo preditivo (classificador multiclasse) para classificar uma mensagem em níveis de viralidade. Escolha uma estratégia para definir o número de níveis de viralidade (classes). Por exemplo, você definir quatro níveis de viralidade a partir dos quartis da quantidade de compartilhamentos.

A avaliação experimental deverá considerar:

e) O algoritmo de classificação: regressão logística, árvore de decisão e uma estratégia baseada em "ensemble";
f) Regularização: Com regularização (Ridge, Lasso ou ElasticNet) e sem regularização;
g) Normalização dos dados: sem normalização, Z-Score, Min-Max (OPCIONAL);
h) Pré-processamento de dados: sem pré-processamento e com pré-processamento;
i) Embedding: BOW, TF-IDF, Word2Vec;

j) N-Gramas: unigramas, bigramas, trigramas;
k) Treinamento, Validação e Teste: Outer K-Fold Cross-Validation;

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