Universidade Federal do Ceará
Centro de Ciências
Departamento de Computação

CKP9011 e CK0268 – Introdução à Ciência de Dados

2024.1

Lista 7

Exercício: Classificação
Objetivos: Exercitar os conceitos referente à classificação.
Data da Entrega: 22/09/2024

OBS 1: Exercício Individual.
OBS 2: A entrega da lista deverá ser executada utilizando-se o SIGAA.

Questão 1

Utilizando os dados referente a postagens no WhatsApp, crie um modelo preditivo (classificador binário) para, dado uma determinada mensagem (postagem) classificá-la como "viral" (classe positiva) ou "não viral" (classe negativa).

Para rotular as mensagens únicas (agrupadas) utilize a seguinte estratégia:
Calcule um limiar (threshold). Por exemplo, mediana do número de compartilhamentos mais dois desvios padrões.
As mensagens com quantidade de compartilhamentos maiores ou iguais ao limiar definido devem ser rotuladas como "virais". As demais mensagens devem ser rotuladas como "não virais".

A avaliação experimental deverá considerar:

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