Avaliação técnica

Instruções

Objetivo

Classificar a coluna SMS do dataset **validation_data.csv** como “ok” ou “blocked”.

Método

Você tem liberdade total para escolher o método e procedimentos de classificação (tratamento dos dados, análises, algoritmos, etc.).

A única **exigência** é que o desafio seja resolvido em **Python** (preferencial) ou **R**.

Material

Você receberá por email acesso a uma pasta do Google Drive com 2 datasets: **train_data.csv** e **validation_data.csv**.

O arquivo **train_data.csv** possui 2 colunas: **SMS e LABEL**. Você pode usar estes dados para treinar seu(s) modelo(s). Você é livre para usar os dados conforme achar mais adequado. Pode usá-los da forma como estão, pode criar *features*, extrair metadados, analisar somente os textos, analisar as URLs contidas nas mensagens, etc. Faça o que achar melhor.

O objetivo é gerar modelos que sejam capazes de reconhecer se uma mensagem (SMS) deve ser bloqueada (“blocked”) ou não (“ok”).

O arquivo **validation_data.csv** possui 1 coluna: **SMS**. Você deve usar o(s) modelo(s) que foram treinado(s) a partir do dataset train_data.csv para aplicar as labels “blocked” ou “ok” a cada um dos SMSs deste dataset.

Entregas

Você deverá entregar o material descrito abaixo, colocando-os na pasta do Google Drive onde estão os datasets utilizados neste desafio:

1. O(s) script(s) utilizado(s).
2. Relatório contendo a descrição dos procedimentos executados e resultados das análises. Em relação aos resultados das análises, não queremos somente o resultado final da classificação, mas também as informações obtidas a partir das análises dos dados: estatísticas relevantes; quais dados/features foram escolhidos para treinar os modelos e por quê; testes executados para a seleção das features; evolução das análises; fontes de erro das classificações; possíveis estratégias para melhorar níveis de acerto (mesmo que não tenham sido aplicadas/testadas no desafio) e outras informações que considerar relevantes.
3. Resultado da classificação: incluir uma coluna no dataset **validation_data.csv** com as labels atribuídas pelo(s) modelo(s).

Expectativa

O que vamos avaliar:

1. Qualidade do código.
2. Procedimentos (como os dados foram preparados/tratados, algoritmos utilizados, fundamentação das escolhas tomadas).
3. O nível de acerto da classificação também será avaliado, porém não será um critério fundamental nesta análise.

Em resumo, os critérios com maior importância em nossa avaliação são **o caminho** escolhido para resolver o problema, o **raciocínio** utilizado, as **decisões tomadas** e o **porquê**.

Se os itens 1 e 2 atenderem às nossas expectativas, você será chamado para uma conversa, onde terá a oportunidade de apresentar suas ideias e explicar de forma mais detalhada o trabalho realizado.