Bootcamp: Cientista de Dados

Desafio Prático

Módulo 5: Desafio Final

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

- ✓ Pré-processamento dos dados.
- ✓ Detecção de anomalias.
- ✓ Processamento dos dados.
- ✓ Correlações.
- ✓ Spark MlLib.
- ✓ Interpretação dos dados.

Enunciado

O derrame é uma das doenças que mais acometem a população mundial. Segundo a World Health Organization (WHO), o Acidente Vascular Cerebral (AVC) foi a segunda maior causa de morte na população mundial no ano de 2016 (https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death).

Neste desafio, vamos realizar uma análise sobre um banco de dados composto por uma pesquisa realizada com diferentes pacientes. Nesta análise, vamos tentar prever, nos baseando em algumas características específicas, se um determinado indivíduo irá ou não sofrer um AVC. Para isso, vamos aplicar o pré-processamento dos dados e a aplicar o modelo de Regressão Logística e o SVM, para indicar se um indivíduo possui ou não maior probabilidade de desenvolver o AVC.

Atividades

Os alunos deverão desempenhar as seguintes atividades:

1. Acessar o site https://community.cloud.databricks.com/ e utilizar a conta **gratuita**. O tutorial de criação da conta **gratuita** está presente na plataforma Canvas, logo abaixo da apostila, no item "**Arquivos complementares**" com o nome "**tutorial_databricks_TPD.pdf**".

2. Acessar e baixar os arquivos "**healthcare_dataset_stroke_data.csv**" e o "**desafio_CID.ipynb**" presentes na pasta:

   - https://drive.google.com/drive/folders/1alRfzr1WbOVlfvffFo8zrXSlwniW1Nd8?usp=sharing.

3. Responda às perguntas do desafio.