Bootcamp: Cientista de Dados

Desafio

Módulo II	Desenvolvimento de Soluções Utilizando Spark

Objetivos

Exercitar os seguintes conceitos trabalhados no Módulo:

- ✓ Exercitar o módulo Spark SQL do Apache Spark.
- ✓ Exercitar o módulo Spark MLLib do Apache Spark.

Enunciado

Doenças ligadas ao coração afetam milhões de pessoas ao redor do mundo, e segundo a Organização Mundial de Saúde (OMS) é a segunda principal causa de morte na população mundial. Como cientista de dados, você foi contratado para criar um modelo preditivo que, a partir de dados de pacientes - como idade, gênero, nível de glicose, se é fumante ou não - vai prever se eles terão um derrame cerebral ou não.

Você tem acesso a um arquivo que possui atributos de pacientes e um atributo “stroke” (derrame), que indica se aquele paciente sofreu um evento de derrame ou não.

O conjunto de dados está disponível em: https://dcc.ufmg.br/~pcalais/stroke_data.csv

Para uma descrição das colunas, veja a seção “Attribute information” em https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

As questões objetivas vão lhe guiar sobre a análise exploratória e o modelo preditivo que você criará a partir dos dados.

Links úteis:

- https://spark.apache.org/docs/latest/sql-getting-started.html

- https://spark.apache.org/docs/latest/ml-classification-regression.html#decision-tree-classifier