Bootcamp: Cientista de Dados

Desafio Prático

Módulo 2: Desenvolvimento de Soluções Utilizando Spark

Objetivos de Ensino

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 da Saúde (OMS), são a segunda principal causa de morte da 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 o paciente fuma ou não, vai prever se aquele paciente terá 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 *stroke_data.csv* está disponível em:

- https://drive.google.com/file/d/163BGU_RzXR29UbVVkPpv8tYnUejlPBVr/view?usp=share_link.

Para uma descrição das colunas, veja a seção “Attribute information” em:

- https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

As questões objetivas vão guiá-lo para a análise exploratória e para 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

Atividades

Os alunos deverão desempenhar as seguintes atividades:

1. Assistir as aulas gravadas sobre os módulos Spark SQL e Spark ML;
2. Assistir a aula interativa sobre Spark ML;
3. A seguir, você estará apto a responder às perguntas. :-)