Big Data Real-Time Analytics Com Python e Spark Versão 3.0

Mini-Projeto 1 Definição do Problema e Fonte de Dados

Tem aumentado de forma contínua o número de pacientes com doença hepática devido ao consumo excessivo de álcool, inalação de gases nocivos, ingestão de alimentos contaminados e uso de drogas e anabolizantes.

Neste mini-projeto vamos construir um modelo de Machine Learning capaz de prever se um paciente vai ou não desenvolver uma doença hepática com base em diversas características do paciente. Esse modelo pode ajudar médicos, hospitais ou governos a planejar melhor o orçamento de gastos de saúde ou mesmo criar políticas de prevenção.

Como nosso objetivo é prever uma classe (sim ou não), usaremos aprendizado supervisionado para classificação, criando diferentes versões do modelo com diferentes algoritmos e passaremos por todo o processo de Machine Learning de ponta a ponta. Usaremos como fonte de dados o dataset disponível no link abaixo:

https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

Não é preciso fazer o download, pois forneceremos o dataset a você ao final do capítulo. Este conjunto de dados contém registros de pacientes hepáticos e registros de pacientes não hepáticos coletados na Índia. A coluna "Dataset" é um rótulo de classe usado para dividir os grupos em paciente hepático (que tem a doença hepática) ou não (sem doença).