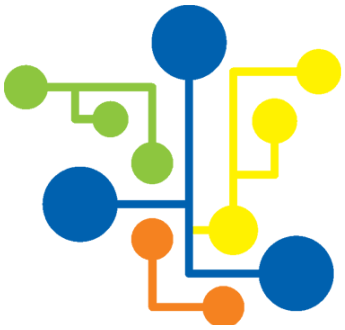

Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Mini-Projeto 2

Prevendo a Ocorrência de Doenças Cardíacas

Solução

Um grande hospital ou uma rede de serviços de saúde pode ser capaz de coletar grandes quantidades de dados sobre seus pacientes e um cluster Hadoop pode ser a solução ideal para armazenar e processar esse “Big Data da Saúde”. Nossa solução, portanto, vai utilizar um cluster Hadoop e as ferramentas:

Hive – como os dados estão em formato estruturado, usaremos o Hive para armazenar os dados no HDFS e realizar consultas interativas através da linguagem HQL.

Pig – será usado para transformação e pré-processamento nos dados.

Mahout – será usado para construção do modelo preditivo.

A solução contempla 5 etapas:

- Etapa 1 - Carregando o dataset no Hive e visualizando os dados com SQL
- Etapa 2 - Analise Exploratória e Pré-processamento nos dados com Pig
- Etapa 3 - Transformação de Dados com o Pig
- Etapa 4 - Criação do Modelo Preditivo de Classificação
- Etapa 5 - Otimização do Modelo Preditivo de Classificação

Na etapa 1, carregamos os dados em uma tabela criada com o Hive. O Hive é a solução ideal para dados estruturados e permite utilizar a HQL, uma variação da linguagem SQL, que nos permite consultar os dados de forma rápida e eficiente.

Na etapa 2, usaremos o Pig para compreender como os dados se relacionam e realizar análises estatísticas.

Na etapa 3, o Pig será usado para transformar os dados de forma a facilitar o trabalho de construção do modelo preditivo.

As etapas 4 e 5 são a criação do modelo preditivo com algoritmo Random Forest. Inicialmente criamos um modelo e avaliamos a Confusion Matrix e na sequência otimizamos o modelo aumentando o número de árvores de decisão.

Em anexo a este arquivo, você encontra o script completo, com todos os comandos comentados linha a linha.