Projeto de Disciplina

Big Data e Processamento Distribuído

Valor: 50 pontos
Entrega: 16/01 às 23:59
Formato: Jupyter Notebook
Individual ou em duplas.
Objetivo: realizar um ciclo de ciência de dados completo no Spark.

Nesse projeto, vocês deverão realizar um ciclo completo de ciência de dados utilizando o PySpark. Isso significa que vocês deverão explorar e preparar dados, treinar um modelo de aprendizado de máquina e fazer análise dos resultados obtidos.

Para escolher um dataset, vocês poderão visitar o Kaggle e selecionar algum que estejam com vontade de explorar. Vocês deverão optar por datasets apropriados para as seguintes tarefas:

- Classificação
- Agrupamento
- Recomendação

Todo o projeto deve ser construído em **um único Notebook**. Nele deverão conter, além do código, análises, explicações e motivações para a escolha do dataset e do algoritmo de aprendizado de máquina. **IMPORTANTE: Todos os imports utilizados deverão ser colocados no início do Notebook.**

Parte I: Exploração de Dados.

Vocês deverão utilizar as funcionalidades de RDD e/ou Dataframes para analisar e limpar os dados. Como essa tarefa é dependente de cada conjunto de dados, não há um modelo rígido a seguir. Porém, vocês deverão realizar no mínimo **2 análises** (estatísticas, análise com gráficos, etc.) e **3 transformações** (filtragem, remoção de características, remoção/troca de valores nulos, normalizações, etc). As transformações devem ser pautadas no que for descoberto ao analisar os dados. Por exemplo: normalização dos valores por discrepância de magnitude entre características.

Parte II: Criação de um Modelo e Análise de Resultados.

Nessa etapa, vocês deverão rodar um algoritmo da biblioteca MLlib do Spark para aprender um modelo de aprendizado de máquina com os dados que vocês acabaram de organizar. Vocês deverão motivar a escolha do algoritmo, que deve ser um dos disponíveis dentro da MLlib do Spark. Além disso, vocês deverão dividir os dados utilizando alguma metodologia de validação (cross-validation, 60-40, 80-20, etc), e validar a performance do seu modelo, analisando os resultados.