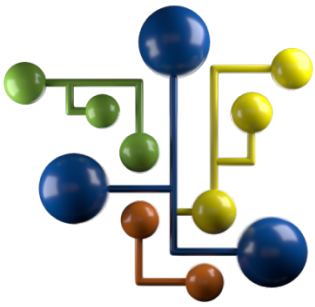


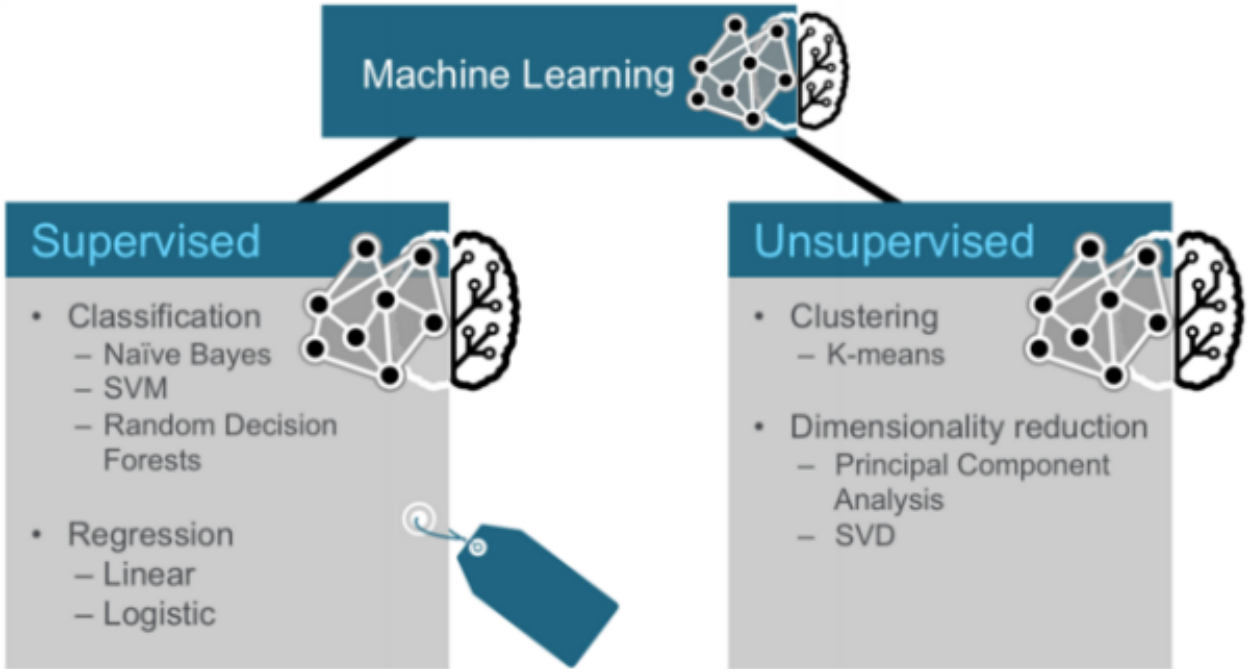
Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Algoritmos de Machine Learning Suportados Pelo Apache Spark MLLib

O MLLib não suporta todos os algoritmos de Machine Learning. Alguns algoritmos não foram construídos para execução em paralelo e de forma distribuída e por isso não foram implementados no MLLib. Porém muitos outros algoritmos podem ser usados, sejam algoritmos de aprendizagem supervisionada ou aprendizagem não supervisionada.



Vejamos a descrição dos principais algoritmos de Machine Learning suportados pelo MLLib.

A classificação é uma família de algoritmos de aprendizagem supervisionada que designam valores de entrada como pertencendo a uma de várias classes pré-definidas. Alguns casos de uso comum de classificação incluem: detecção de fraude com cartão de crédito e detecção de spam. Os dados de classificação são rotulados, por exemplo, como spam / não-spam ou fraude / não-fraude. O algoritmo de Machine Learning atribui um rótulo ou classe para novos dados, classificando esses novos dados com base em características pré-determinadas.

As árvores de decisão criam um modelo que prevê a classe ou o rótulo com base em várias características de entrada. As árvores de decisão funcionam avaliando uma expressão contendo uma característica em cada nó e selecionando uma alternativa para o próximo nó com base na resposta.

No clustering, o algoritmo cria grupos de objetos em categorias, analisando semelhanças entre exemplos de entrada. Esses grupos são chamados de clusters (não confundir com cluster de computador, ok? O termo é igual, mas são coisas diferentes). Utilizações de clustering incluem: resultados da pesquisa de agrupamento, agrupamento de clientes, detecção de anomalias e categorização de texto. Clustering utiliza algoritmos não supervisionados, ou seja, não são expostos as possíveis saída.

Algoritmos de filtragem colaborativa recomendam itens (esta é a parte do filtro) com base na preferência de informações de muitos usuários (esta é a parte colaborativa). A abordagem de filtragem colaborativa baseia-se na semelhança; as pessoas que gostavam de itens semelhantes no passado vão gostar itens semelhantes no futuro. O objetivo de um algoritmo de filtragem colaborativa é tomar preferências de usuários, e criar um modelo que pode ser usado para recomendações ou previsões.

Uma das características importantes sobre MLLib é que ele contém apenas algoritmos de Machine Learning que podem ser executados em paralelo através de um cluster. Alguns algoritmos clássicos não estão incluídos no MLLib, pois não foram criados para processamento em paralelo. Em compensação, o MLLib permite criar modelos com alguns importantes algoritmos como Árvores de Decisão e K-Means. Os algoritmos do MLLib são otimizados para computação em clusters e com grandes conjuntos de dados. Se o seu objetivo for aplicar Machine Learning em conjuntos de dados pequenos ou médios, com certeza o MLLib não será a melhor opção para isso e nesse caso podemos usar o Scikit-Learn ou mesmo o TensorFlow.

O MLLib oferece duas bibliotecas para a construção dos modelos: