Pyspark - MLLib Part 2 - Classification & Clustering

Agenda

This is the 13th project in the Pyspark series. The twelfth project mainly focuses on Regression in Spark MLlib. This project mainly focuses on Classification and Clustering in Spark MLlib. Apache Spark uses Spark MLlib to do machine learning. This project also includes implementation of Decision tree classifier, Random forest classifier, and K-Means clustering algorithms.

Tech stack:

➔Language: Python
➔Package: Pyspark
➔Services: Spark

Key Takeaways:

- Understanding the project overview
- Understanding the concept of Machine Learning
- Introduction to PySpark MLlib
- Understanding the Unsupervised learning
- Different types of Classification algorithms
- Implementation of Decision tree classifier
- Implementation of Random forest classifier
- Implementation of K-Means clustering
- Change datatype of column in PySpark
- Drop NA values in PySpark
- Data preprocessing in PySpark
- Splitting data into train data and test data in PySpark
- Validate the model in PySpark