Practical exercises and applied

Koalas

APACHE Spark Streaming

Spark MACHINE LEARNING

Data Bootcamp
BEST DATA TRAINING

Machine Learning with PySpark

Exercise 1

In this exercise we are going to import and preprocess the data from 'data/heart.csv' that we will use to train a binary classification model with PySpark. To do this, you'll need to initialize a Spark session, load the data with the correct schema, and analyze its distribution.

That is, you must complete the import part and exploratory analysis of the data.

Solution: *Exercise_Solution_Machine Learning with Spark.ipynb*

Machine Learning with PySpark

Exercise 2

In this exercise you are going to apply minimal pre-processing on the data set from the previous exercise, so that you can use it to train a model.

Solution: *Exercise_Solution_Machine Learning with Spark.ipynb*

Machine Learning with PySpark

Exercise 3

In this exercise you will train a binary classification model with the PySpark machine learning library, with the pre-processed data set from the previous exercise.

Once the model has been trained, you will have to make a prediction with the test data set and compare the results. Next, you will need to obtain different evaluation metrics to determine if the model is suitable or not.

Solution: *Exercise_Solution_Machine Learning with Spark.ipynb*