University of Abdelhamid Mehri-Constantine2
Faculty: NTIC
Department: IFA
Master's second year SDIA 2024/2025

Practical test

Duration : 1h30

We want to predict the 10-year risk of future coronary heart disease (CHD) from a dataset of a cardiovascular study containing 4000+ records and 15 attributes.
Write the PySpark code that:

1. Prepare the data for classification using Vector assembler
2. Extract the model using the logistic regression
3. Plot the validation results (Graphical and numerical metrics)