**MACHINE LEARNING**

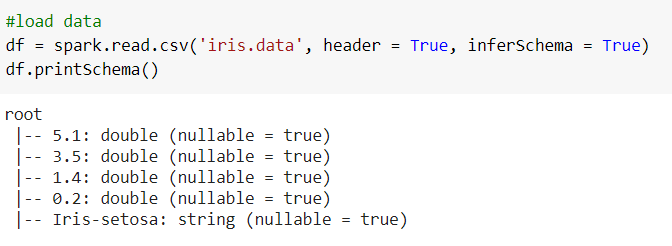
Apache Spark là một trong những công cụ mã nguồn mở được sử dụng rộng rãi nhất và được hỗ trợ cho việc học máy và dữ liệu lớn. [1]

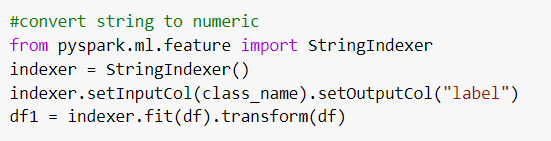
Các bước để xây dựng một chương trình học máy với PySpark có thể gồm các bước sau:

Bước 1: Tiền xử lý dữ liệu



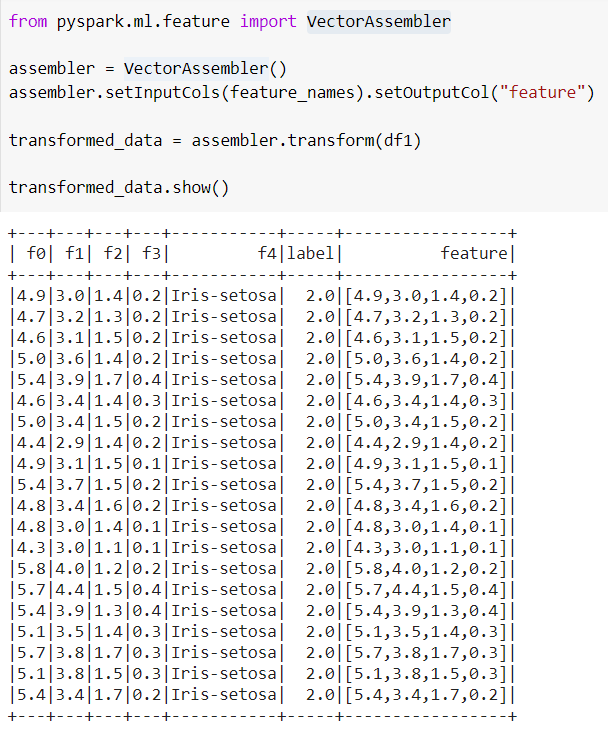
Trước tiên, ta khởi tạo Spark.





Ta nhận thấy ở dữ liệu có cột với các giá trị kiểu String, do đó ta sử dụng StringIndexer chuyển về dạng số để thực hiện việc xử lý dữ liệu ở bước sau.

Tiếp theo, ta chuyển các cột dữ liệu thành dạng vector bằng VectorAssembler.



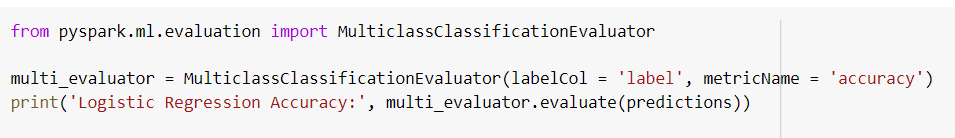
Bước 2: Xây dựng mô hình học máy (phân loại, dự đoán)

Ở đây ta tách dataset thành 2 phần: dữ liệu train và dữ liệu test. Sau đó xây dựng mô hình học máy với dữ liệu train được xử lý để tính toán tỉ lệ phân loại, dữ liệu test sẽ sử dụng kết quả tính toán tỉ lệ trước đó để phân loại dữ liệu.



Bước 3: Đánh giá mô hình

Tùy theo dữ liệu và yêu cầu của bài toán mà ta có những cách xây dựng mô hình học máy khác nhau, do đó, độ chính xác của các mô hình là một dữ kiện quan trọng, là thước đo để đánh giá mô hình học máy tốt hay xấu. Vì vậy, ta có thể đánh giá một mô hình học máy như sau:



**TÀI LIỆU THAM KHẢO**

1. https://tek4.vn/apache-spark-machine-learning-voi-pyspark-va-mllib/