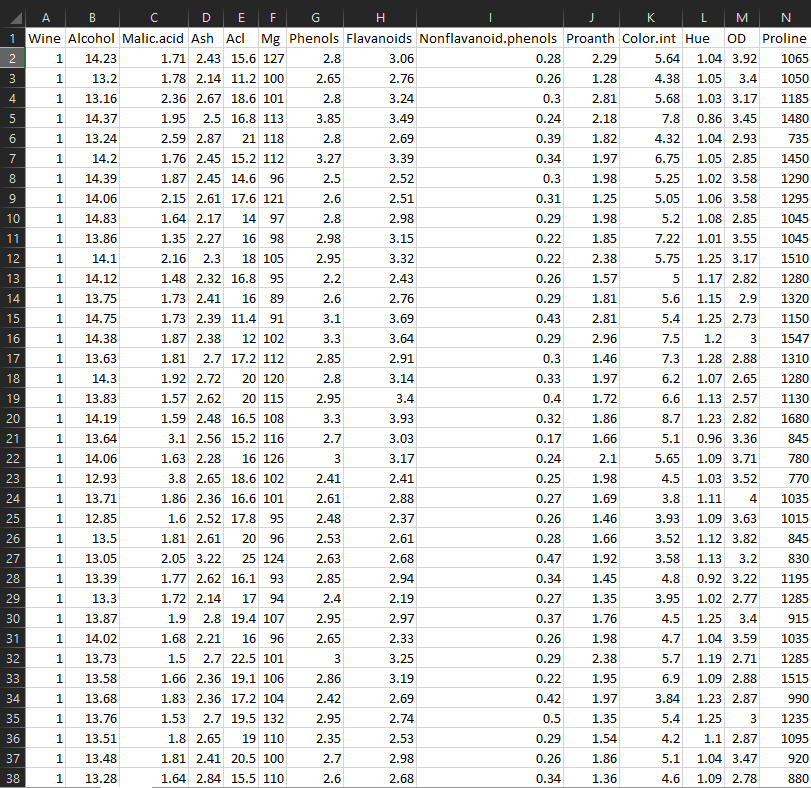
# **BÀI TOÁN PHÂN LỚP RƯỢU**

**by Trần Anh Duy – 19810310499**

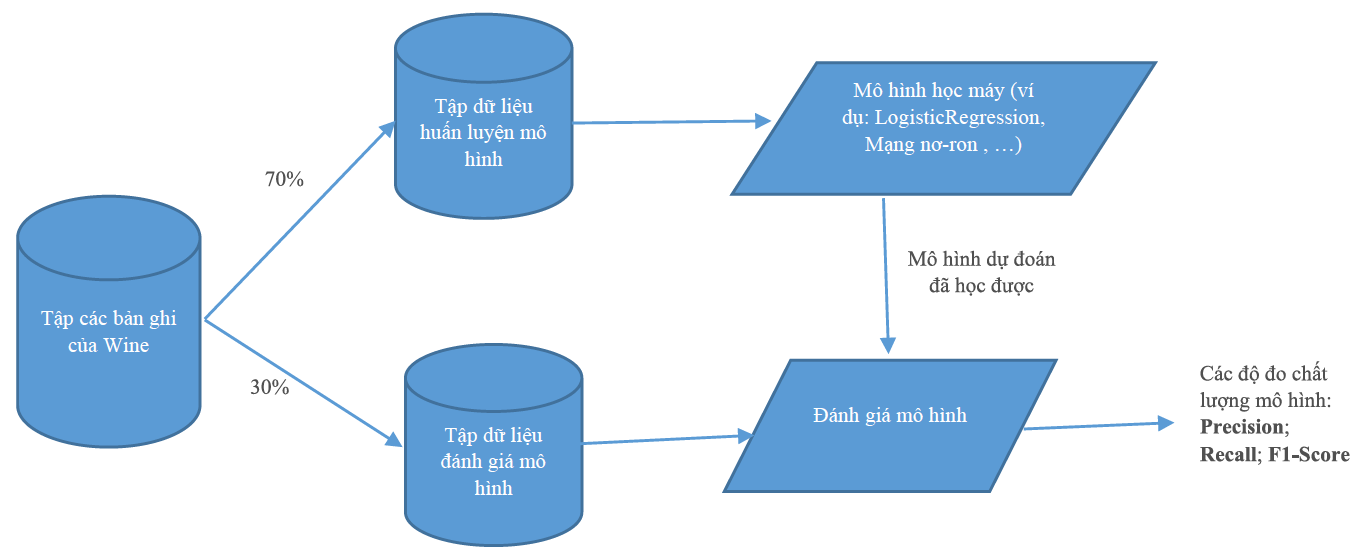
## **Mô tả bài toán**

Chúng ta có file wine.csv là bảng dữ liệu rượu có cấu trúc ở dưới:



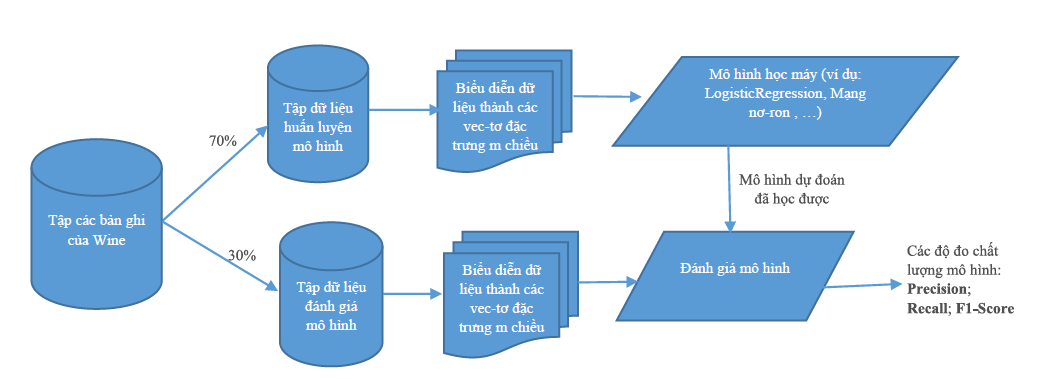
Theo đó, trường Wine để chỉ thuộc tính lable, đánh số thứ tự từ 1 đến 3. Các trường còn lại là các thuộc tính chất lượng của rượu. Trong bài này, chúng ta sẽ không dùng trường Proline.

## **Mô hình hệ thống học máy**

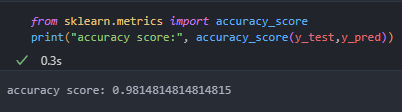


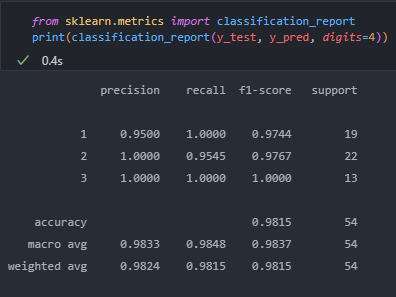
## **Giải bài toán phân lớp rượu không sử dụng phương pháp giảm chiều dữ liệu (PCA)**

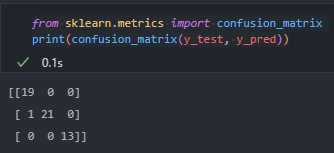
* Chúng ta sẽ cần chia dataset thành 2 tập dữ liệu: \_train và \_test, trong đó: \_train chứa 70% số bản ghi trong dataset, \_test chứa 30% số bản ghi trong dataset. Sau đó chúng ta chọn một mô hình học máy để học một mô hình phân lớp rượu, ở đây là Logistic Regression.
* **Mô hình hệ thống học máy:**



* **Kết quả:**







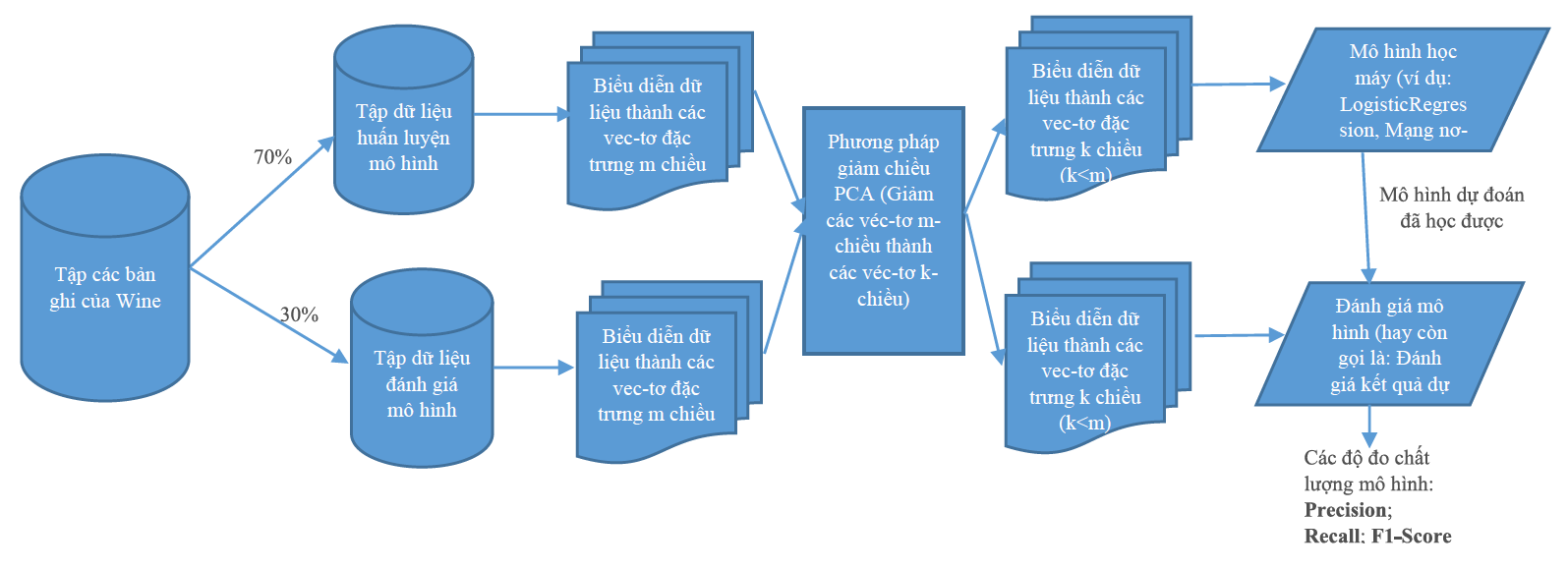
Có thể thấy, accuracy score của cách giải này là 0.98, các điểm số precision score, recall score, f1-score cũng lần lượt đạt 0.95, 1, 0.97 trong lần thử đầu tiên, còn confusion matrix trả về là [ [19 0 0]

[1 21 0]

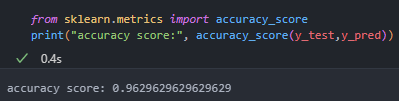
[0 0 13] ].

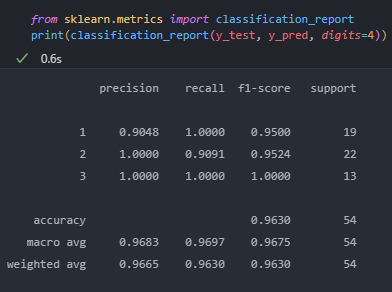
## **Giải bài toán phân lớp rượu sử dụng phương pháp giảm chiều dữ liệu (PCA)**

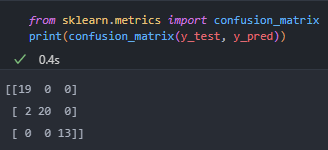
* Chúng ta sử dụng phương pháp giảm chiều PCA để giảm chiều các véc tơ biểu diễn đặc trưng trong tập D\_train và D\_test. Sau đó, chúng ta chọn một mô hình học máy để học một mô hình phân lớp rượu, ở đây là Logistic Regression.
* **Mô hình hệ thống học máy:**



* **Kết quả:**







Có thể thấy, accuracy score của cách giải này là 0.96, các điểm số precision score, recall score, f1-score cũng lần lượt đạt 0.9048, 1, 0.95 trong lần thử đầu tiên, còn confusion matrix trả về là [ [19 0 0]

[2 20 0]

[0 0 13] ].

## **So sánh 2 cách giải bài toán ở trên**

* Dễ thấy, điểm khác nhau lớn nhất là ở các điểm số accuracy score, precision score, recall score, f1-score. Các chỉ số này đều chỉ ra rằng, phương pháp giải bằng việc không sử dụng phương pháp giảm chiều dữ liệu (PCA) cho ra điểm số cao hơn khi chúng ta sử dụng phương pháp giảm chiều dữ liệu (PCA).
* Tuy nhiên, thời gian giải bài toán sử dụng phương pháp giảm chiều dữ liệu (PCA) lại ngắn hơn so với việc chúng ta không sử dụng phương pháp giảm chiều dữ liệu (PCA).

**Hết**