

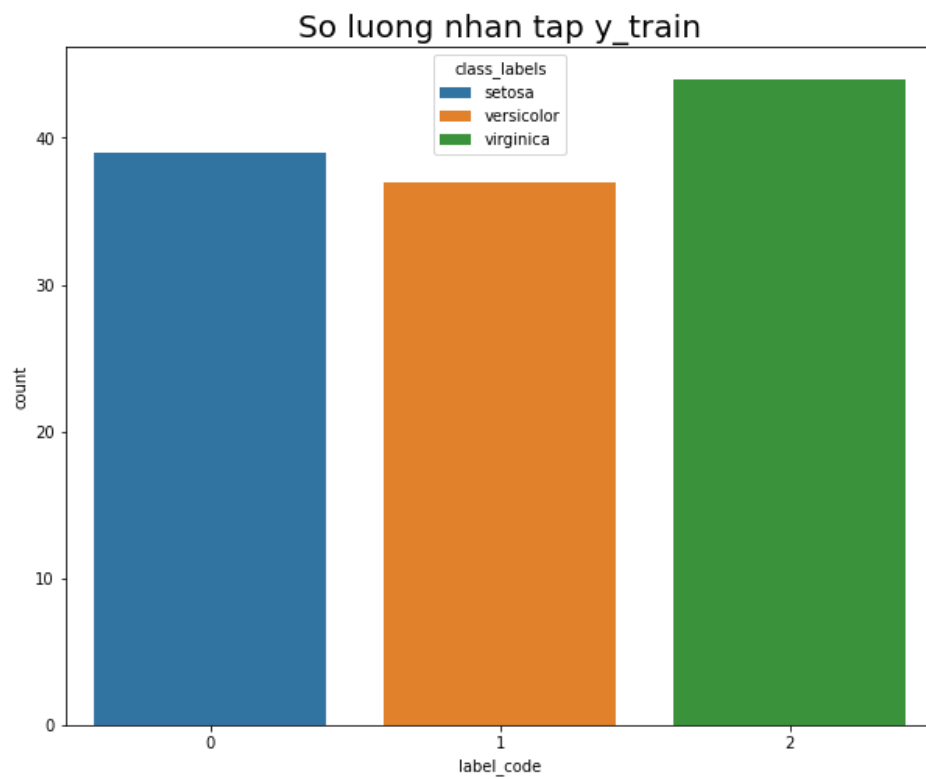
BÁO CÁO THỰC HÀNH 2

MÔ HÌNH PHÂN LỚP

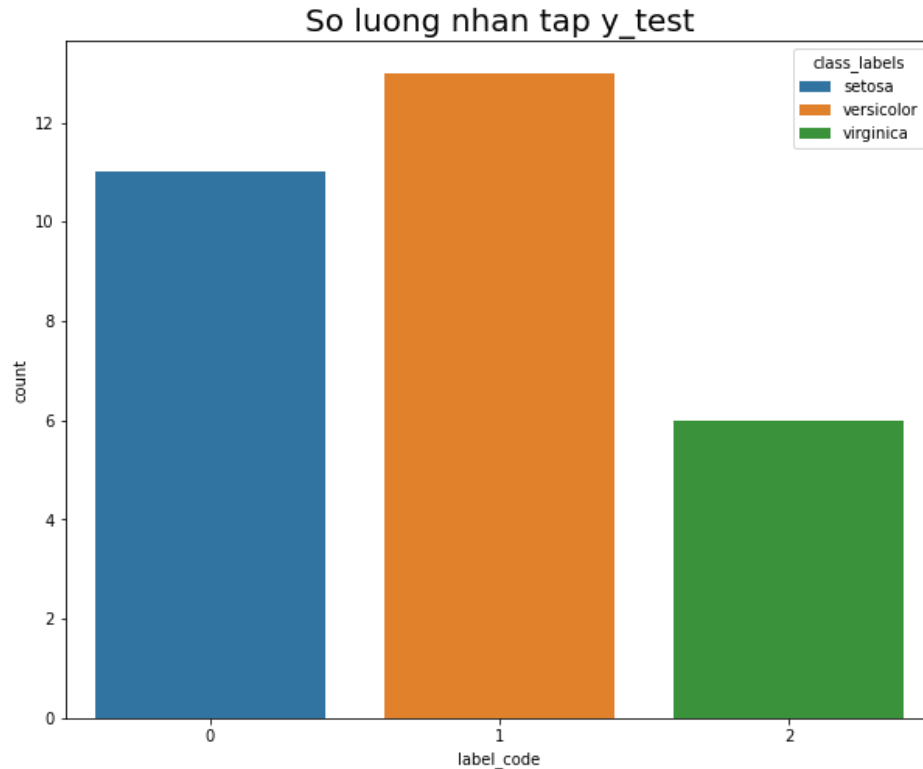
Sinh viên: Ngô Huỳnh Trương – 20522085

Môn học: Học máy thống kê – DS102

Bài 1: Hãy thống kê số lượng nhãn (label) trên tập training và tập test.



{0: 39, 1: 37, 2: 44}



{0: 11, 1: 13, 2: 6}

Bài 3: Thực hiện huấn luyện mô hình K láng giềng gần nhất (KNN) trên bộ dữ liệu, sau đó so sánh độ chính xác (Accuracy) với mô hình LogisticRegression

Độ chính xác (Accuracy) của mô hình LogisticRegression cao hơn KNN trong bài toán này, 73.33% so với 66.67%

Bài 4: Đánh giá 2 mô hình vừa xây dựng trên 3 độ đo sau: precision_score, recall_score và f1_score sử dụng macro average.

	Logistic Regression	K Nearest Neighbor (KNN)
Precision (macro avg)	0.70	0.68
Recall (macro avg)	0.71	0.68
f1-score (macro avg)	0.70	0.65
support	30	30

Accuracy của Logistic Regression (LR) = 0.73, của KNN là 0.67

precision của LR với macro avg = 0.70, 0.68 là của KNN

recall của LR với macro avg = 0.71, của KNN là 0.68

f1-score của LR sử dụng macro avg = 0.70 trong khi đó của KNN là 0.65

Ở bài toán với dataset trên đưa vào các model với hai thuộc tính thì LR cho kết quả tốt hơn KNN ở mọi độ đo

Nếu như f1-score sử dụng macro avg của LR có khoảng chênh lệch ít so với precision và recall thì f1-score (macro avg) của KNN lại thấp xuống khá đáng kể, như vậy trung bình điều hoà bị lệch như vậy chứng tỏ có kết quả thực sự thấp trong mô hình

⇒ Mô hình tối ưu hiện tại là LR

Bài 5*: Hãy sử dụng chiến lược tìm kiếm tham số GridSearchCV để tìm ra bộ tham số tốt nhất cho mô hình Logistic Regression. So sánh kết quả với mô hình gốc.

```
# View best hyperparameters

print("tuned hpyerparameters :(best parameters) ",best_model.best_params_)
print("accuracy :",best_model.best_score_)

[39] ✓ 0.1s

... tuned hpyerparameters :(best parameters) {'C': 1.0, 'penalty': 'l1'}
accuracy : 0.8166666666666668
```

Áp dụng GridSearchCV cho 2 tham số C và penalty cho mô hình LogisticRegression

Bộ tham số tốt nhất tìm được là C=1.0 và penalty = 'l1'

Kết quả tốt nhất có thể đạt được là 81.6% cao hơn 73.3% mô hình ban đầu