ÁP DỤNG MỘT SỐ THUẬT TOÁN PHÂN LỚP VÀO BỘ DỮ LIỆU SOCIAL NETWORK ADS

Huỳnh Trọng Khoa - 18520918

***Tóm tắt nội dung*—Đề tài sử dụng 4 thuật toán phân lớp là Support Vector Machine, Logistic Regression, Naive Bayes và Decision Tree để dự đoán kết quả của bộ dữ liệu có tên "Social Network Ads" trên trang Kaggle. Thông qua việc sử dụng các mô hình đã được cài sẵn trên thư viện scikit-learn, chúng tôi có thể đánh giá quá trình phân tích dữ liệu của 4 thuật toán và đưa ra kết quả tốt nhất trong 4 thuật toán. Việc đánh giá được thực hiện thông qua kết quả Confusion Matrix và 4 đại lượng đặc trưng của bài toán phân lớp là Accuracy, Precision, Recall, F1-Score.**

I. Giới thiệu

Bài toán phân lớp (classification) là một trong những bài toán thuộc lĩnh vực máy học. Nội dung của bài toán này là đưa ra dự đoán kết quả thuộc 1 trong các lớp đã được định nghĩa trước thông qua một mô hình (model). Mô hình được xây dựng trên tập dữ liệu được gán nhãn trước. . Có các loại bài toán phân lớp như: phân lớp nhị phân, phân lớp đa lớp,... Với đề tài này thì chúng tôi sử dụng bài toán phân lớp nhị phân.

II. Bộ dữ liệu

Bộ dữ liệu được sử dụng ở đây có tên "Social Network Ads"được lấy trên trang Kaggle. Bao gồm 400 mẫu dữ liệu thông tin khác hàng và cho biết khách hàng có mua hàng hay không.

Các thuộc tính bao gồm:

Tên thuộc tính	Mô tả
UserID	Mã số định danh
Gender	Giới tính
Age	Độ tuổi
Estimated Salary	Mức lương ước đoán
Purchased	Dự đoán có mua hàng hay không

III. Xử lý dữ liệu

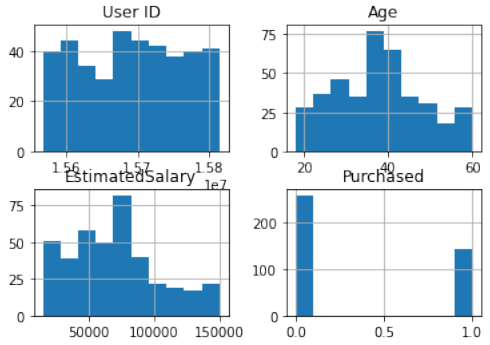
A. Tiền xử lý dữ liệu

Vì bộ dữ liêu đã đầy đủ nên ở bước này chúng tôi sẽ tiến hành lựa chọn thuộc tính để làm Input và Output.

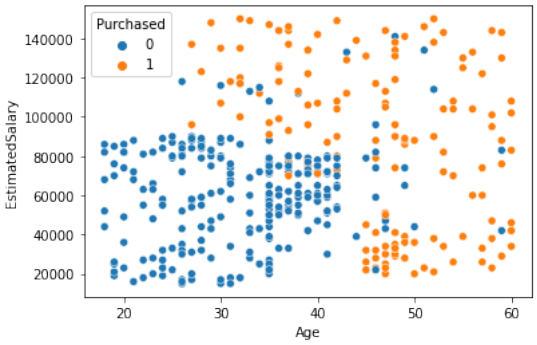
Input: Age, Estimated Salary

Output: Purchased

Và bộ dữ liệu sẽ đưa chia với tỉ lệ 3-1 cho quá trình train và test.



Hình 1. Mô tả dữ liệu



Hình 2. Trực quan hoá dữ liệu

B. Trực quan hoá dữ liệu

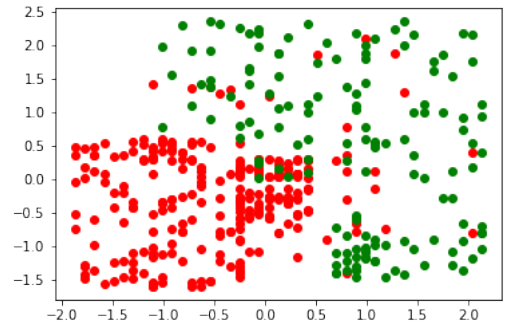
Nhận xét thấy dữ liệu 2 thuộc tính có độ chênh lệch khá lớn nên dẫn đến các điểm dữ liệu cách xa khá nhiều, và để cải thiện điều này, chúng tôi sẽ tiến hành chuẩn hoá dữ liệu.

IV. Huấn luyện mô hình

Ở đây chúng tôi sử dụng 4 thuật toán phân lớp để huấn luyện dữ liệu, đó là: Support Vector Machine, Logistic Regression, Naive Bayes, Decision Tree.

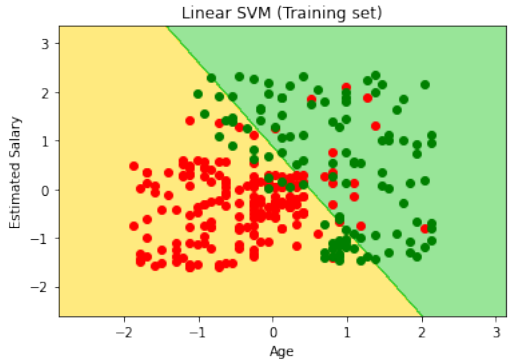
A. Support Vector Machine

Đây là thuật toán thường dùng cho việc phân loại qua việc tìm hyper-plane phân chia các lớp. Thuật toán SVM sẽ tìm một số vector hộ trợ (support vectors). Mô hình dự đoán kết quả đầu ra của dữ liệu mới dựa trên các

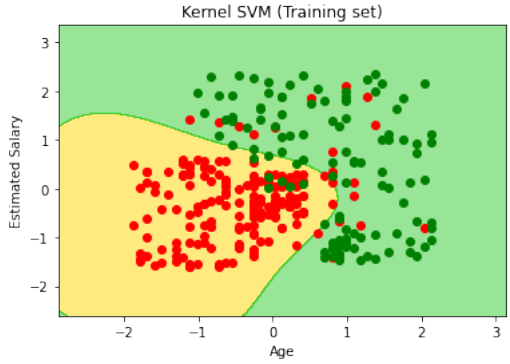


Hình 3. Dữ liệu sau khi chuẩn hoá

vector này. Sau đây là kết quả phân lớp khi sử dụng Linear SVM và Kernel SVM với kernel = 'rbf' (Radial Basic Function).



Hình 4. Trực quan phân lớp Linear SVM



Hình 5. Trực quan phân lớp Kernel SVM

B. Logistic Regression

Logistic Regression là một mô hình sử dụng cho các bài toán phân loại. Thuật toán được biến đổi từ Linear Regression bằng cách cho kết quả của Linear Regression vào hàm sigmoid, cụ thể:

$$y = sigmoid(f(x)) = sigmoid(w_0 + w_1x_1 + ... + w_nx_n)$$

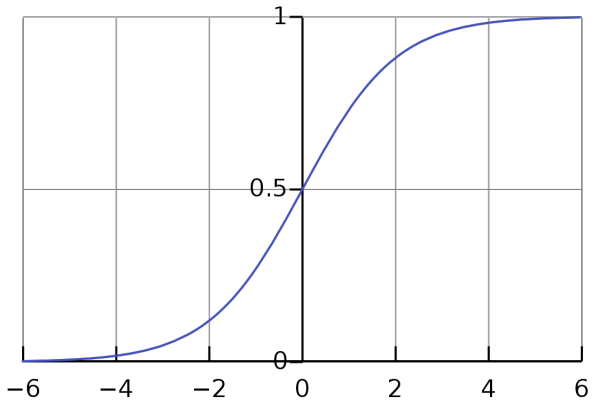
Trong đó:

$w_0, w_1, ..., w_n$ là các tham số mô hình

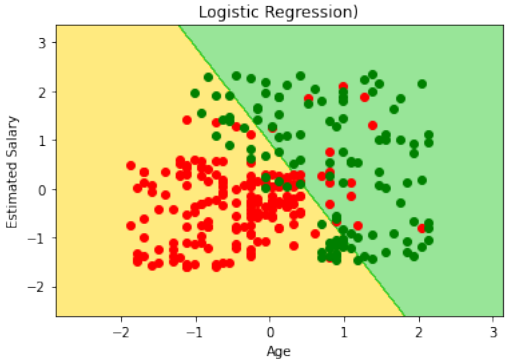
$x_0, x_1, ..., x_n$ là các biến độc lập

y là kết quả đầu ra

$$sigmoid(X) = \frac{1}{e^{-x}}$$



Hình 6. Hàm sigmoid



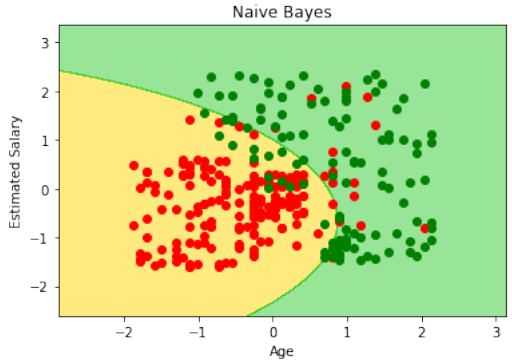
Hình 7. Trực quan phân lớp Logistic Regression

C. Naive Bayes

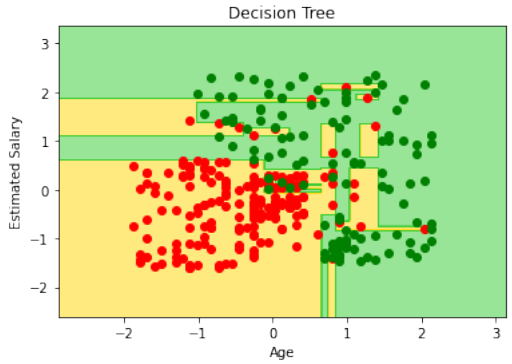
Đối với Naive Bayes Classifier, chúng tôi sử dụng mô hình Gaussian Naive Bayes. Mô hình này sử dụng chủ yếu trong các loại dữ liệu mà các thành phần là biến liên tục.

D. Decision Tree

Đối với thuật toán Decision Tree, mô hình được sử dụng theo hàm Entropy.



Hình 8. Trực quan phân lớp Naive Bayes



Hình 9. Trực quan phân lớp Decision Tree

V. Đánh giá

Mô hình phân lớp được đánh giá thông qua 4 thông số Accuracy, Precision, Recall và F1-Score được tính dựa trên Confusion Matrix.

Confusion Matrix có các tham số như sau:

- TP (true positive): điểm dữ liệu có nhãn là 1 được dự đoán chính xác là 1.
- FP (false positive): điểm dữ liệu có nhãn là 0 được dự đoán chính xác là 1.
- TN (true negative): điểm dữ liệu có nhãn là 0 được dự đoán chính xác là 0.
- FN (false negative): điểm dữ liệu có nhãn là 1 được dự đoán chính xác là 0.

Từ đó, ta có thể tính được các thông số:

- Độ chính xác (Accuracy):

$$\frac{TP+TN}{TP+TN+FP+FN}$$

- Tỉ lệ điểm dữ liệu TP trong số những điểm được phân loại là Posive (Precision):

$$\frac{TP}{TP+FP}$$

- Tỉ lệ điểm dữ liệu TP trong số những điểm được phân loại thực sự là positive (Recall):

$$\frac{TP}{TP+FN}$$

Precision và Recall đều là các số không âm hoặc nhỏ hơn bằng 1.

Precision cao đồng nghĩa với việc độ chính xác của các điểm tìm được là cao.

Recall cao đồng nghĩa tỉ lệ bỏ sót các điểm thực sự là positive thấp.

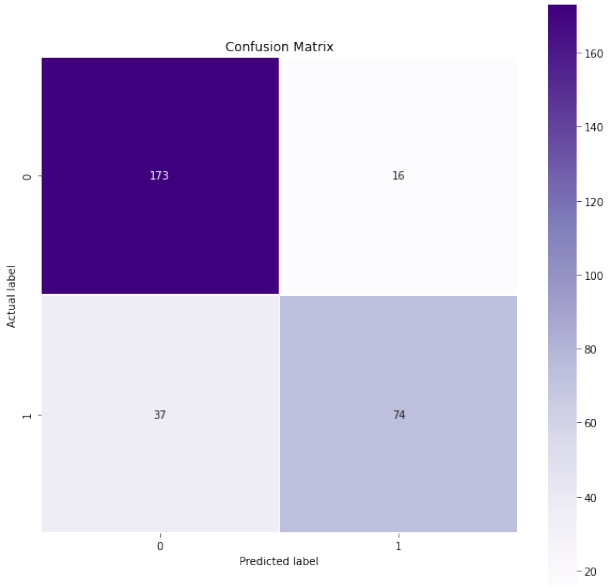
Một mô hình tốt thì cả Precision và Recall đều cao.

- F1-Score là harmonic mean của Precision và Recall (giả sử 2 đại lượng khác 0):

$$2\frac{precisionxrecall}{precision+recall}$$

	SVM	LR	Naive Bayes	D Tree
Accuracy	0.93	0.89	0.9	0.91
Precision	0.88	0.89	0.89	0.83
Recall	0.91	0.75	0.78	0.91
F1-Score	0.89	0.81	0.83	0.87

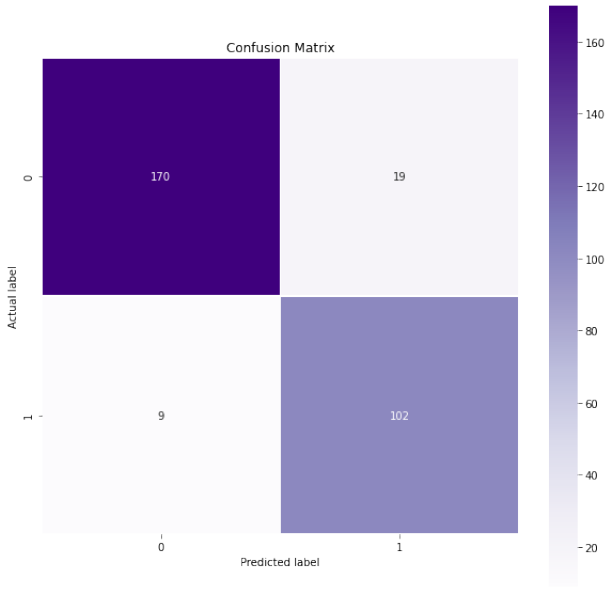
Confusion Matrix



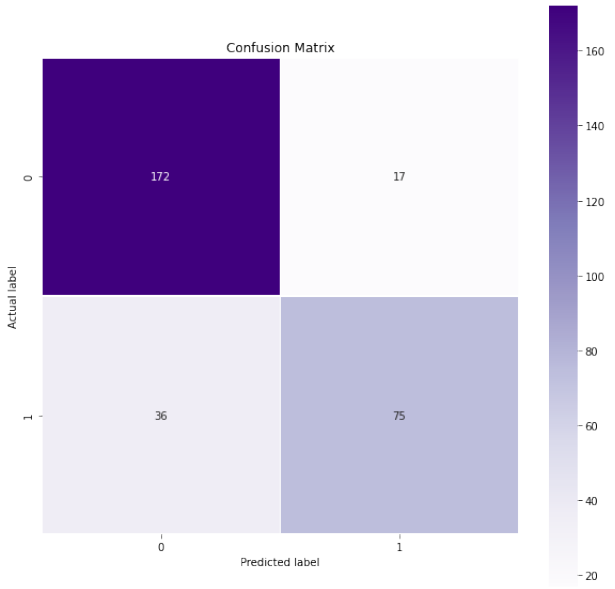
Hình 10. Confusion Matrix Linear SVM

VI. Kết luận

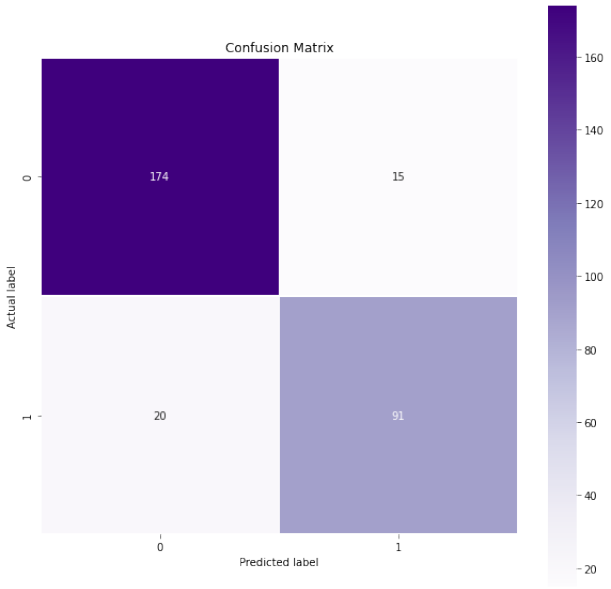
Trong 4 mô hình thì Kernel SVM cho ra kết quả tốt nhất với độ chính xác là 93. Tiếp theo là Naive Bayes với 90. Mô hình Logistic Regression mặc dù có độ chính xác khá cao nhưng điểm F1-Score thấp do số lượng điểm dự đoán sai trong quá trình huấn luyện nhiều. Mô hình



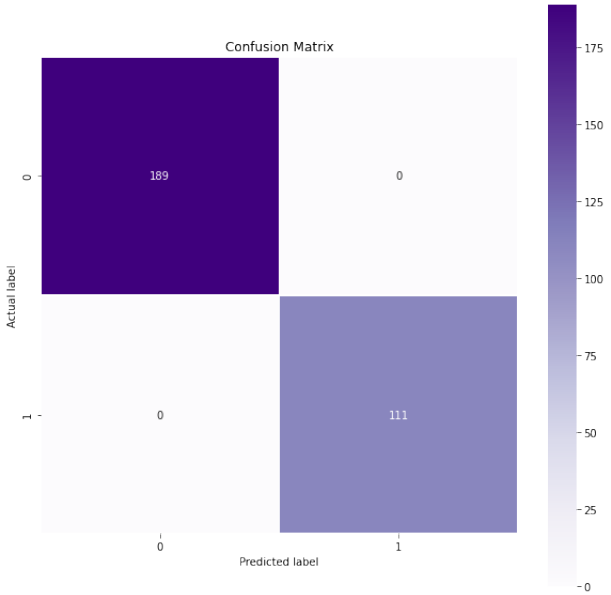
Hình 11. Confusion Matrix Kernel SVM



Hình 12. Confusion Matrix Logistic Regression



Hình 13. Confusion Matrix Naive Bayes



Hình 14. Confusion Matrix Decision Tree

Decision Tree cho ra kết quả dự đoán theo tập dự liệu tốt nhưng xấu với tập dữ liệu test vì với tính chất của mô hình thì một số dữ liệu chưa được train thì tỉ lệ dự đoán sai khá cao.

TÀI LIỆU

[1] Rakesh Raushan, "Social Network Ads - Kaggle". Available: https://www.kaggle.com/rakeshrau/social-network-ads

[2] Nguyễn Nghĩa, "Bài toán phân lớp trong Machine Learing". Available:https://eitguide.net/bai-toan-phan-lop-trong-machine-learning-classification-machine-learning/

[3] Nguyễn Nghĩa, "Bài toán phân lớp trong Machine Learing". Available:https://eitguide.net/bai-toan-phan-lop-trong-machine-learning-classification-machine-learning/

[4] Machine Learning cơ bản, "Cách đánh giá một hệ thống phân lớp". Available:https://machinelearningcoban.com/2017/08/31/evaluation/