BÁO CÁO BÀI TẬP LÝ THUYẾT MÔN HỌC MÁY THỐNG KÊ (DS102.K21)

Giảng viên phụ trách môn : Thầy NGUYỄN TẤN TRẦN MINH KHANG – VÕ DUY NGUYÊN

Họ và tên sinh viên: Võ Hoàng Thông - MSSV: 18521462

- **Bài toán:**

Yêu cầu dựa vào 2 thuộc tính: Độ tuổi và mức lương ước đoán.

Dự đoán khách hàng có mua hàng hay không ?

- **Phương pháp:** Sử dụng thuật toán phân lớp Support Vector Machine.
- **Source File: File đính kèm (Social-Network-Ads.ipynb)**
- **Dataset:** Bao gồm 400 điểm dữ liệu.

+ 300 điểm dữ liệu đầu tiên để huấn luyện.

+100 điểm dữ liệu cuối cùng để đánh giá.

	User ID	Gender	Age	EstimatedSalary
0	15624510	Male	19	19000
1	15810944	Male	35	20000
2	15668575	Female	26	43000
3	15603246	Female	27	57000
4	15804002	Male	19	76000

Mô tả bộ dữ liệu

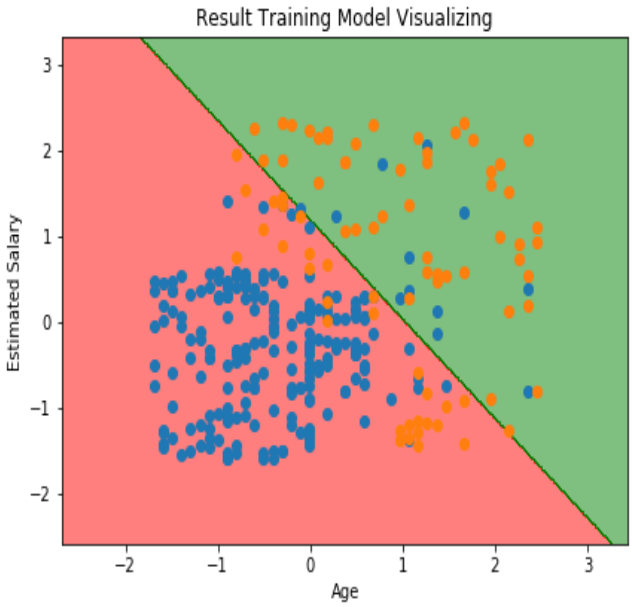
Trực quan hóa dữ liệu

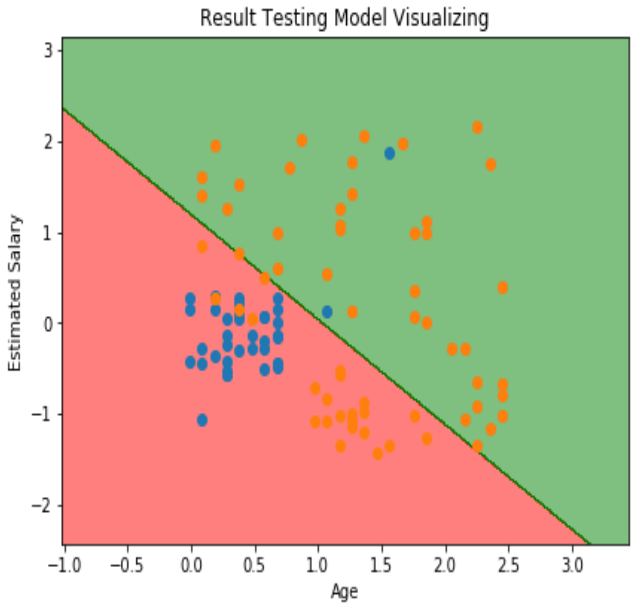
— Theo hình vẽ, ta thấy các điểm có sự phân bố thành 2 mảng.

+ Mảng dưới trái phần lớn có màu xanh, tức khách hàng không mua hàng.

+ Mảng bên phải và mảng bên trên phần lớn có màu cam, tức khách hàng có mua hàng.

Trực quan hóa kết quả mô hình





─ Ta trực quan hóa kết quả mô hình trên mặt phẳng tọa độ bằng cách vẽ 2 vùng phân chia mà mô hình thu được sau quá trình huấn luyện.

─ Nhận xét:

+ Mô hình có độ chính xác chấp nhận được, vẫn có nhiều điểm phân chia nhầm.

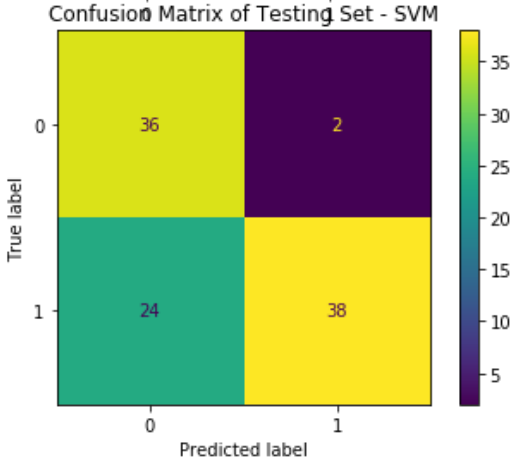
+ Mô hình phân chia theo một đường thẳng, vì SVM là một mô hình phân loại tuyến tính.

Confusion matrix

_ Theo ma trận trên, số lượng dữ liệu được phân loại đúng là 208+50 = 258.

_ Số lượng dữ liệu phân loại sai là 11+31 = 42.

_ Tỉ lệ điểm dữ liệu phân loại sai là 42/300 ~ 14%.



_ Số lượng dữ liệu được phân loại đúng là 36+38 = 74.
_ Số lượng dữ liệu phân loại sai là 26.

_ Tỉ lệ điểm dữ liệu phân loại sai là 26/100 ~ 26%.

Đánh giá mô hình

+ Training Dataset cho độ chính xác là 86%.

+ Testing Dataset cho độ chính xác là 74%.

Nhận xét:

+ Mô hình cho kết quả huấn luyện tốt.

+ Đánh giá kết quả mô hình huấn luyện trên tập kiểm thử cho độ chính xác là 74% tương đối khá, phù hợp với kết quả tập training dataset trước đó cho thấy sự phân bố dữ liệu khá đồng đều của hai tập kiểm thử và huấn luyện.