Báo cáo bài tập lý thuyết môn Học máy Thống kê – DS102.K21

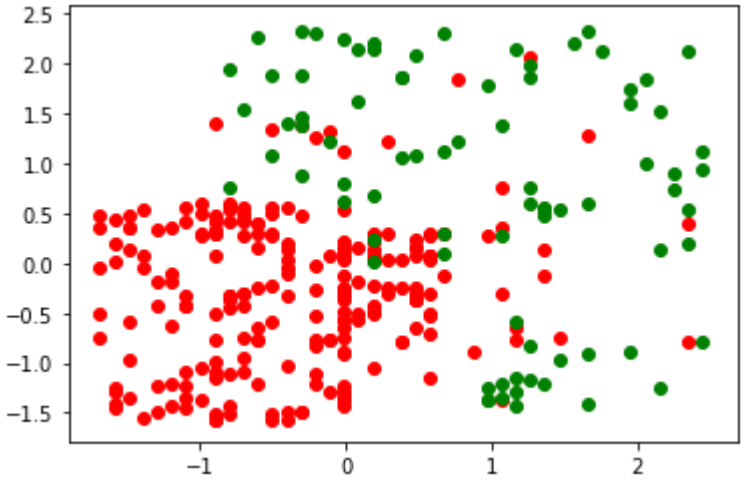
Giảng viên phụ trách môn: thầy NGUYỄN TẤN TRẦN MINH KHANG - VÕ DUY NGUYÊN
Họ và tên: Võ Hoàng Thông – MSSV: 18521462

─ Tên tập dữ liệu: Social Network Ads.
─ Nguồn: https://www.superdatascience.com/pages/machine-learning.
─ Tập dữ liệu cho biết các thông tin của khách hàng và họ có mua hàng hay không.

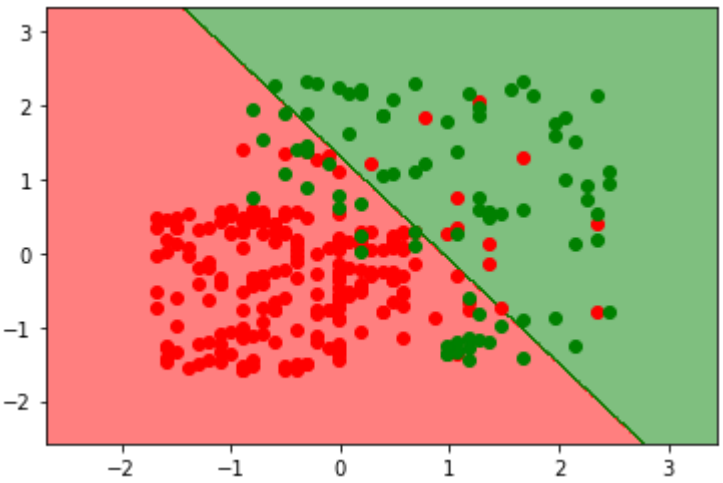
─ Tập dữ liệu chứa 400 điểm dữ liệu, mỗi điểm dữ liệu có 5 thuộc tính gồm:
+ UserID: Mã số định danh của người dùng.
+ Gender: Giới tính của người dùng.
+ Age: Độ tuổi người dùng.
+ Estimated Salary: Mức lương ước đoán của người dùng.
+ Purchased: Là một trong hai số 0 và 1. Số 0 cho biết khách hàng không mua hàng và số 1 cho biết khách hàng có mua hàng.

─ **Bài toán:** Yêu cầu dựa vào 2 thuộc tính:
+ Độ tuổi (Age).
+ Mức lương ước đoán (Estimated Salary).
Dự đoán khách hàng sẽ mua hàng hay không?

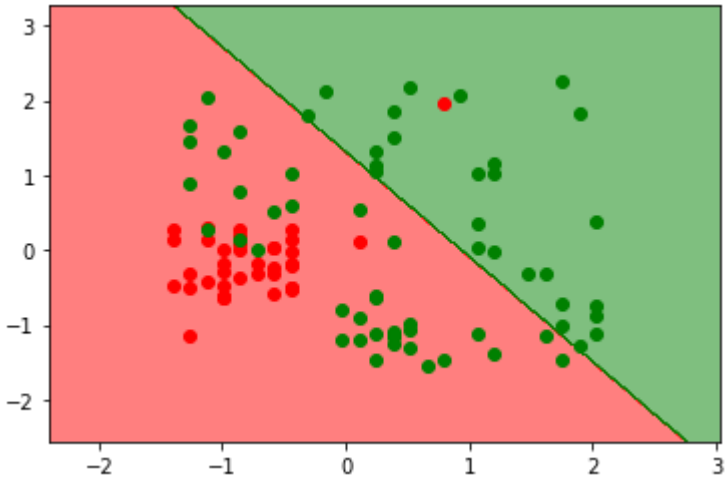
Visualize dữ liệu
Dữ liệu đã được chuẩn hóa về dạng:
+ Kỳ vọng bằng 0.
+ Phương sai bằng 1.
Trục x là độ tuổi (Age)
Trục y là Estimated Salary.



─ 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.



Trực quan hóa kết quả mô hình



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 đây cũng là một mô hình tuyến tính.

Confusion matrix của tập test:
```
[[37  1]
 [36 26]]
```
Theo ma trận trên, số lượng dữ liệu được phân loại đúng là 37+26 = 63.
Số lượng dữ liệu phân loại sai là 36+1 = 37.
Tỉ lệ điểm dữ liệu phân loại sai là 37%.

Confusion matrix của tập train:
```
[[207  12]
 [ 31  50]]
```
Tỉ lệ điểm dữ liệu phân loại sai của tập train là 14.3333%

Từ confusion matrix tập train và tập test, ta rút ra được kết luận sau:
+ Mô hình training cho tỉ lệ phân bố các điểm dữ liệu tương đối tốt với sai số là 14.3333$
+ Trong khi, mô hình testing có mức độ sai số trung bình khá, vì vậy tỉ lệ phân bố dữ liệu phân loại đúng chỉ là 63% không quá tệ.