BÁO CÁO BÀI TẬP THỰC HÀNH 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 3 thuộc tính: giới tính, tuổi và mức lương.

Dự đoán khách hàng có mua điện thoại hay không ?

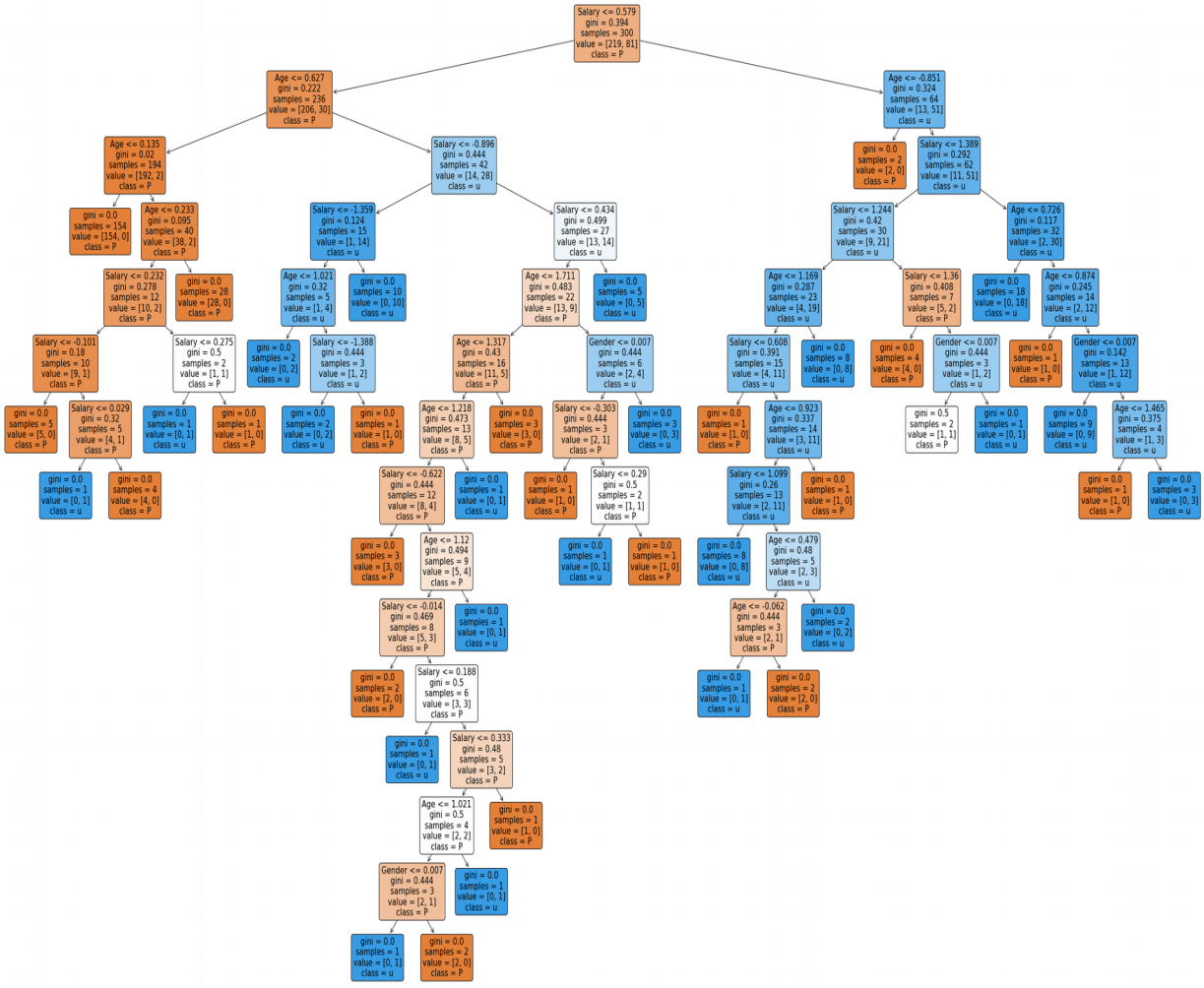
- **Phương pháp:** Sử dụng thuật toán phân lớp Decision Tree và Logistic Regression.
- **Source File: File đính kèm (iphone_purchase_records.ipynb)**
- **Dataset:** Bao gồm 400 điểm dữ liệu.

+ 300 điểm dữ liệu đầu tiên để huấn luyện.

+100 điểm dữ liệu cuối cùng để đánh giá.

	Gender	Age	Salary	Purchase Iphone
0	Male	19	19000	0
1	Male	35	20000	0
2	Female	26	43000	0

Mô tả bộ dữ liệu



Sơ đồ cây quyết định

Confusion matrix

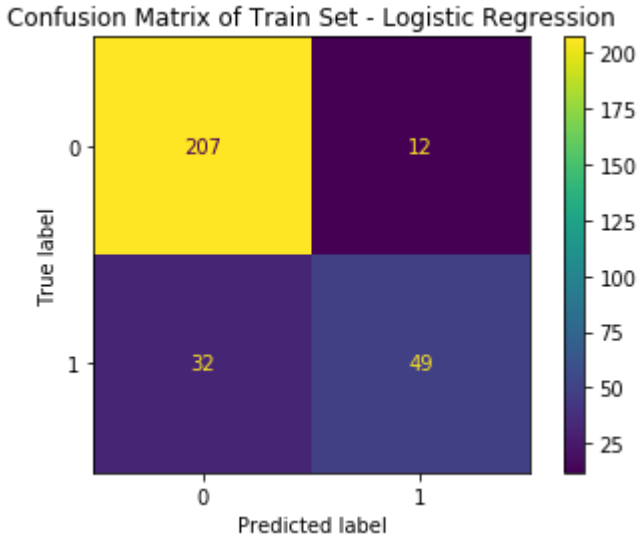
Ở trong bài toán này, tôi sẽ sử dụng hai phương pháp để giải:

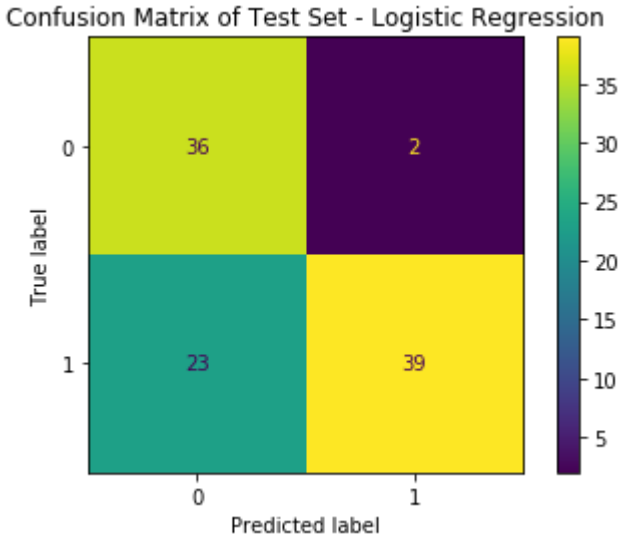
+ Logistic Regression

+ Decision Tree

Bên dưới đây, tôi sẽ trình bày confusion matrix và ý nghĩa của loại ma trận này trên hai thuật toán khác nhau cho cùng một bài toán.

Logistic Regression





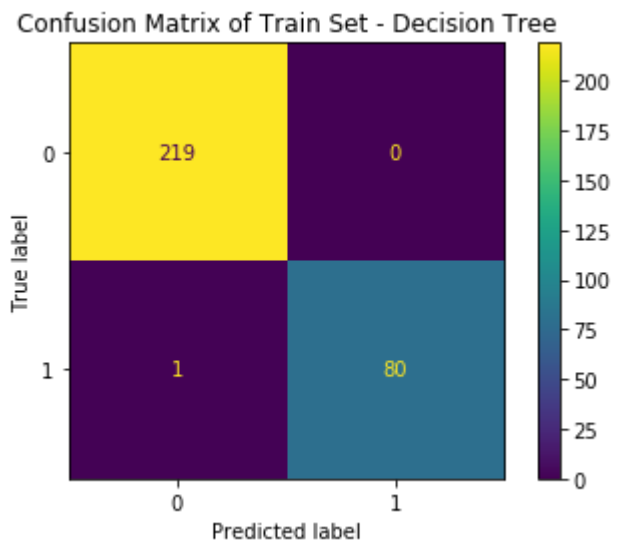
+ Đối với train set: Theo ma trận trên, số lượng dữ liệu phân loại đúng là 207+49 = 256 điểm dữ liệu. Số điểm dữ liệu phân loại sai là 12+32 = 44.

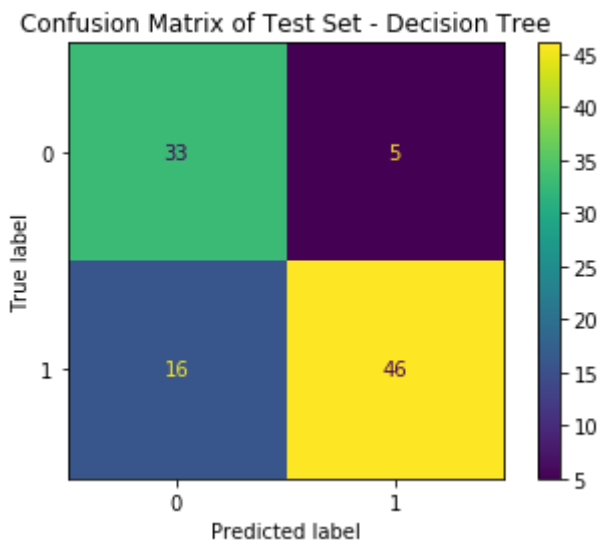
+ Đối với test set: Số lượng dữ liệu phân loại đúng là 36+39 = 75 điểm dữ liệu, số điểm dữ liệu phân loại sai là 25.

+ Tỉ lệ điểm dữ liệu phân loại sai của test set là 25/100 = 0.25

+ Tỉ lệ điểm dữ liệu phân loại sai của train set là 44/300 = 0.14

Decision Tree





+ Đối với train set: Theo ma trận trên, số lượng dữ liệu phân loại đúng là 219+80 = 299 điểm dữ liệu. Số điểm dữ liệu phân loại sai là 1.

+ Đối với test set: Số lượng dữ liệu phân loại đúng là 33+46 = 79 điểm dữ liệu, số điểm dữ liệu phân loại sai là 21

+ Tỉ lệ điểm dữ liệu phân loại sai trên train set là 1/300

+ Tỉ lệ điểm dữ liệu phân loại sai trên test set là 21/100

Đánh giá mô hình

Logistic Regression:

+ Train set: Cho độ chính xác là 0.85

+ Test set: Cho độ chính xác là 0.75

Decision Tree

+ Train set: Cho độ chính xác là 0.996666

+ Test set: Cho độ chính xác là 0.79

Nhận xét:

+ Trên cùng một bộ training set với hai thuật toán khác nhau, thuật toán Decision Tree cho kết quả tốt nhất với tỉ lệ gần như tuyệt đối.

+ Cũng vậy, trên bộ testing set, Decision Tree cho kết quả tốt hơn hẳn Logistic Regression (0.79 > 0.75).

+ Có thể nói mô hình thuật toán Decision Tree phù hợp với bài toán này hơn.