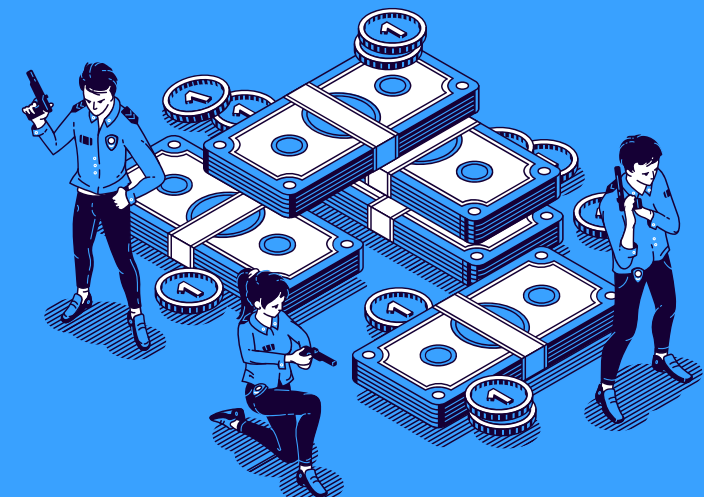


Đồ án cuối kỳ

Bài toán dự đoán khoản vay



Nhóm 7

Thành viên nhóm

Nguyễn Cao Nhân

1959024

Hồ Ngọc Thảo Trang

1959040

Lê Trần Bá Tân

1959035

Phạm Đình Chương

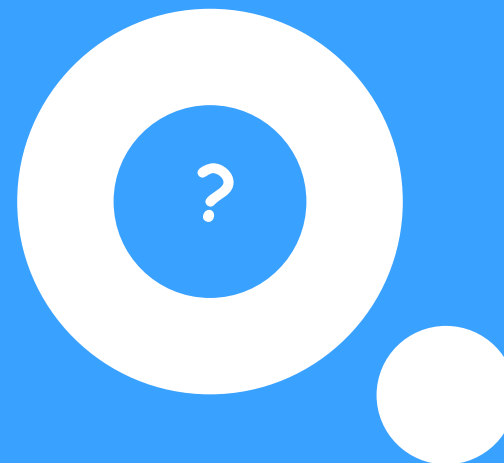
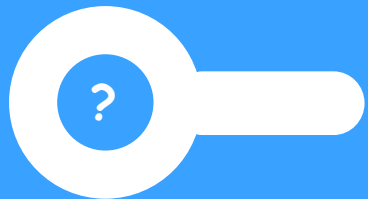
1959002



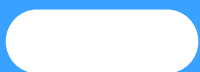
TẬP DỮ LIỆU

122607
Mẫu

67
Đặc tính



Dự đoán một khoản vay có
khả năng vỡ nợ hay không?

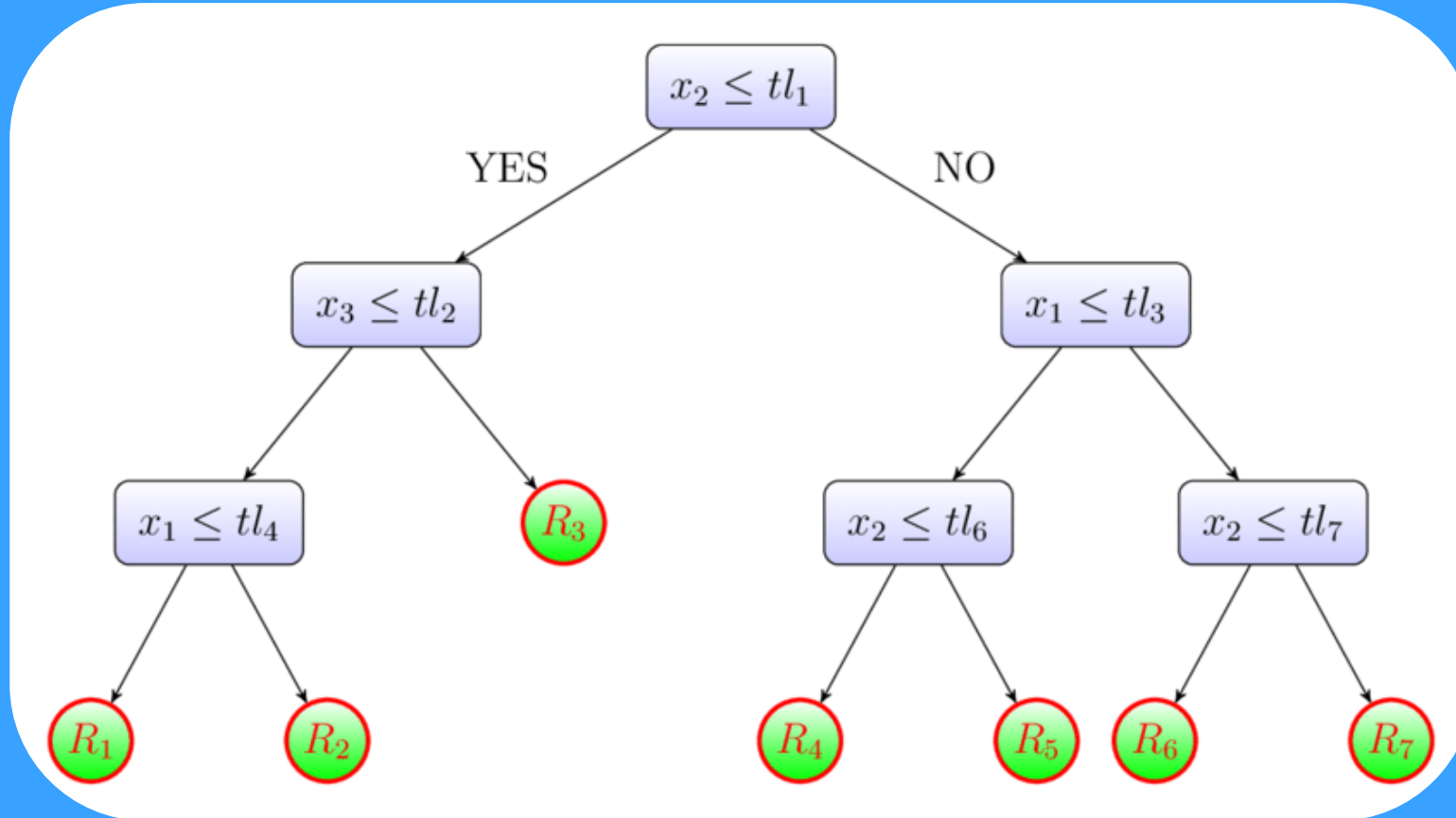


BÀI TOÁN PHÂN LOẠI NHỊ PHÂN (2 LỚP)



Mô tả bài toán

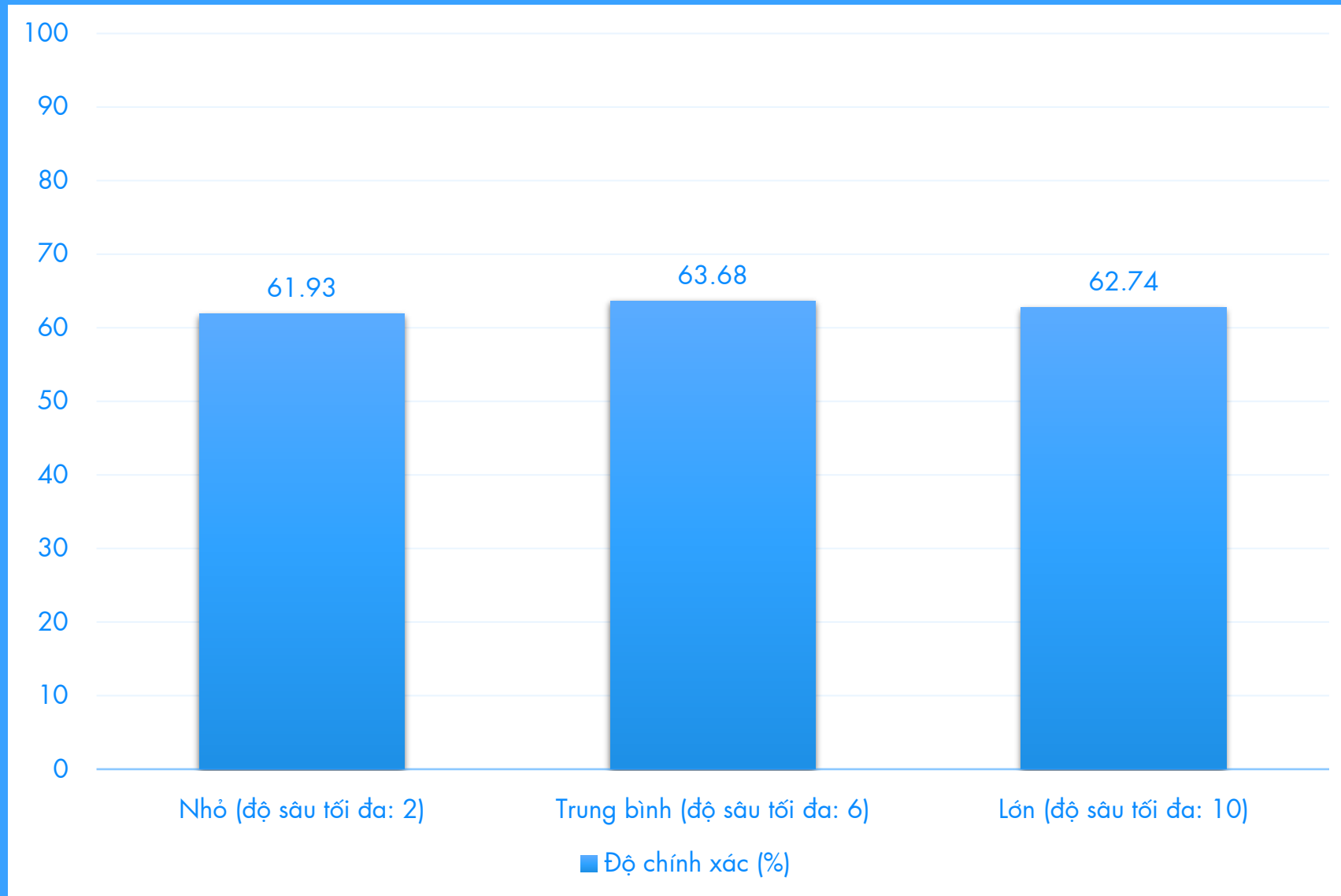
Cơ sở: **DECISION TREE**



Sử dụng các loại thuật toán khác nhau để so sánh hiệu quả

Độ chính xác cơ sở

Decision Tree – Trong Notebook được cấp sẵn



HƯỚNG TIẾP CẬN

1. EXPLORATORY DATA ANALYSIS (EDA)

Hiểu dữ liệu và tìm đặc trưng

2. RESAMPLING

Xử lý vấn đề bất cân bằng trong tập dữ liệu

3. FEATURE SELECTION

Trích xuất đặc tính

4. ONE-HOT ENCODING

Tiền xử lý đặc tính có dữ liệu dạng phân loại

5. FEATURE SCALING

Cân bằng giá trị đặc tính

6. HYPERPARAMETERS TUNING

Tìm kiếm bộ siêu tham số tối ưu

7. TRAINING

Huấn luyện mô hình

8. EVALUATION

Kiểm định hiệu quả mô hình

LỰA CHỌN MÔ HÌNH

7 MÔ HÌNH

NHÓM
THUẬT
TOÁN DỰA
TRÊN CÂY

DECISION TREE

RANDOM
FOREST

K-NEAREST
NEIGHBORS

NHÓM
THUẬT
TOÁN DỰA
TRÊN
KHOẢNG
CÁCH

NHÓM
THUẬT
TOÁN DỰA
TRÊN XÁC
SUẤT

(BERNOULLI)
NAÏVE BAYES

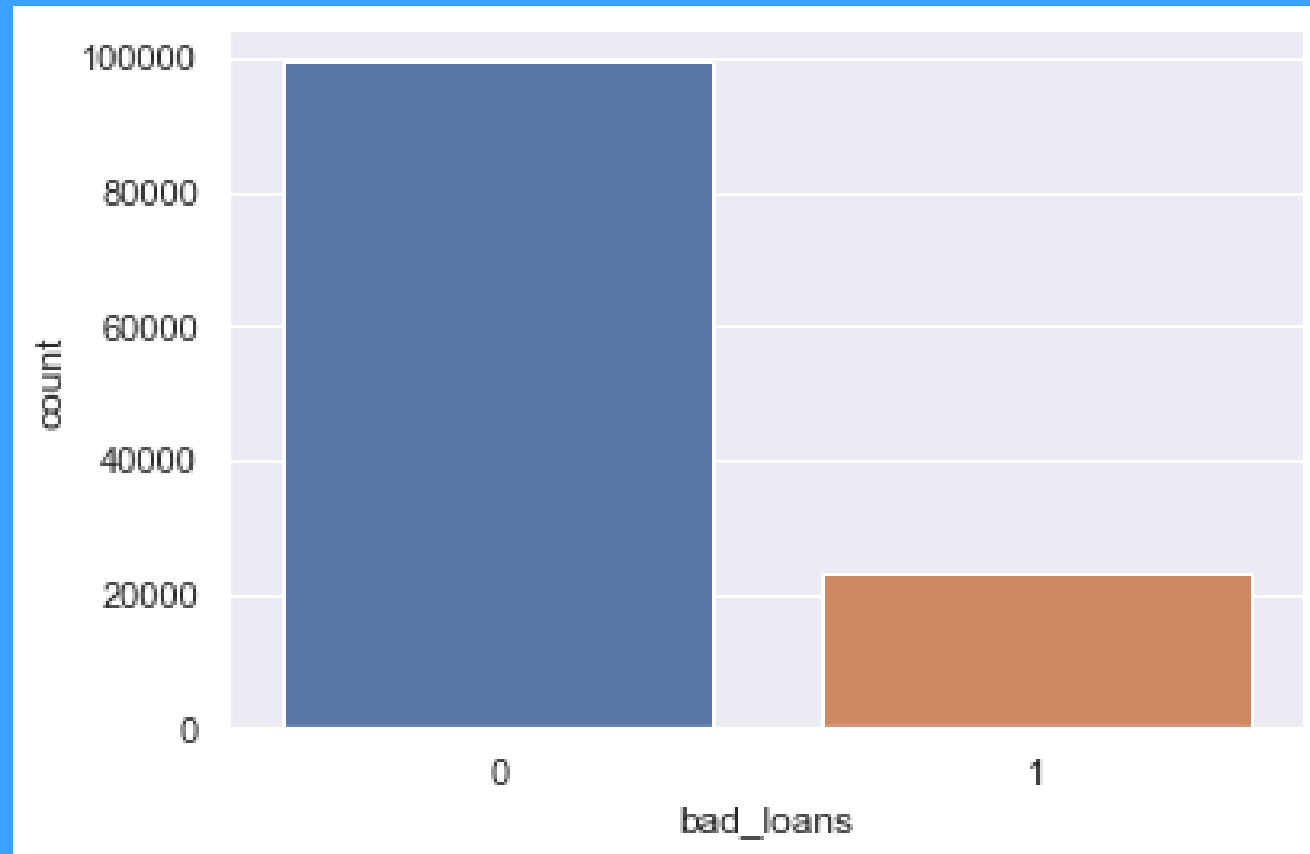
LOGISTIC
REGRESSION

SUPPORT
VECTOR
MACHINE

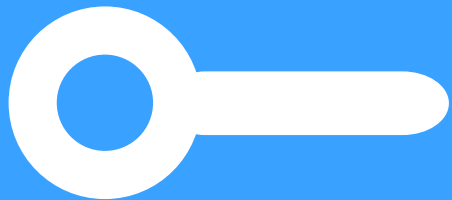
ARTIFICIAL
NEURAL
NETWORK

NHÓM
THUẬT
TOÁN
KHÁC

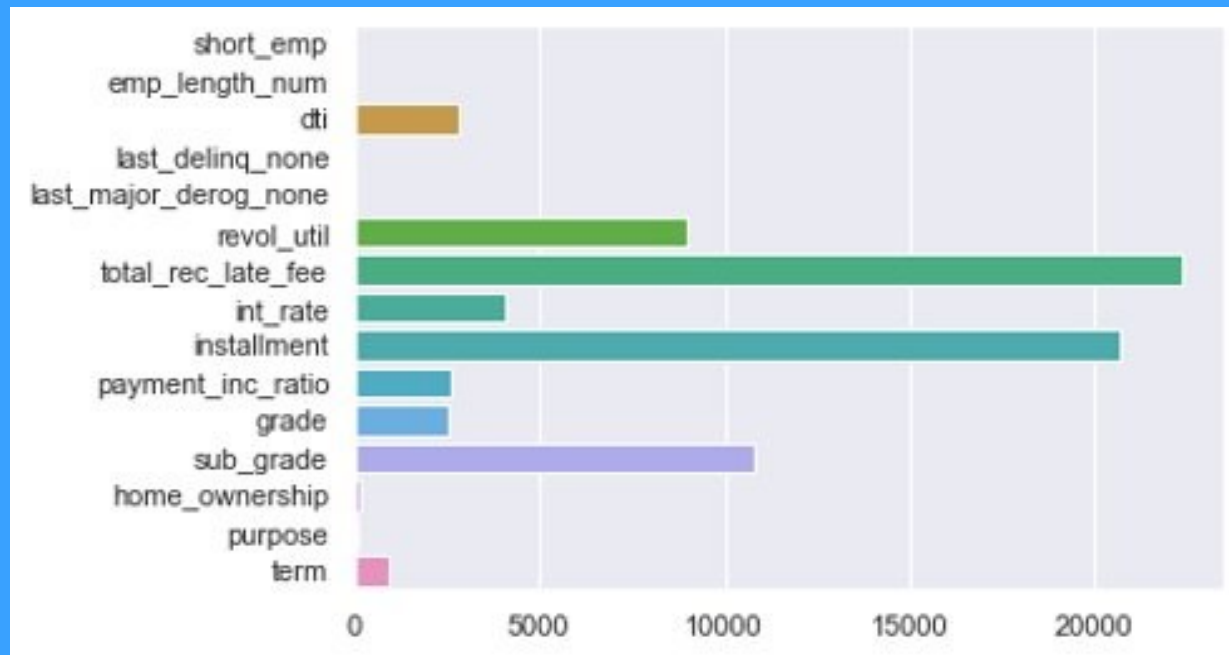
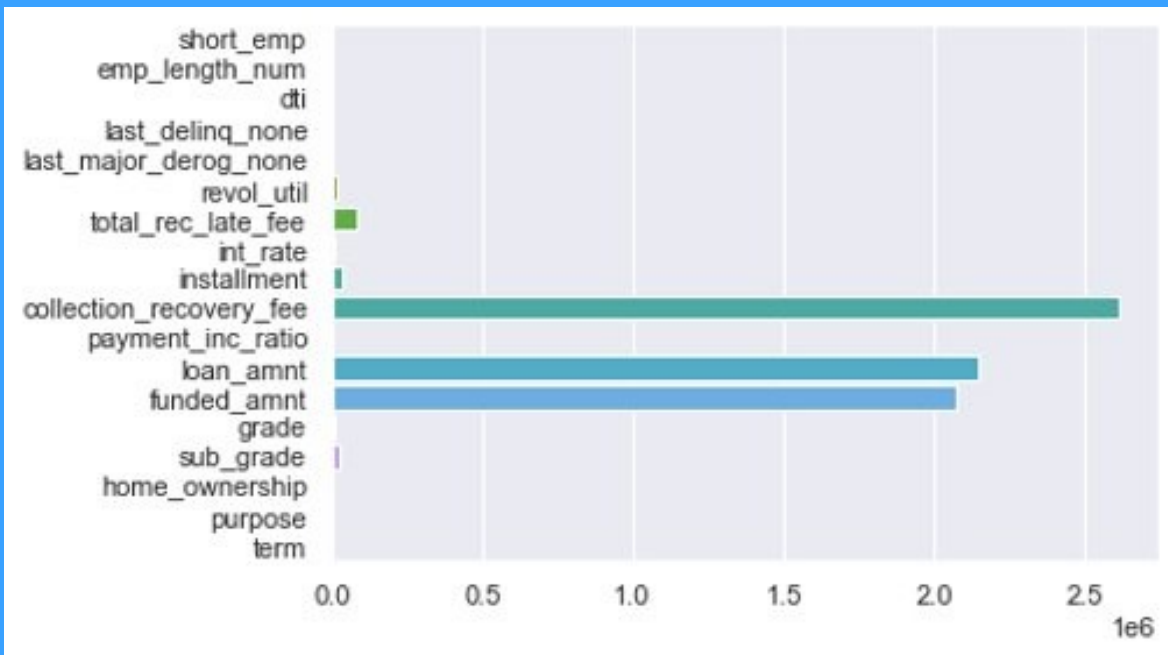
1. EXPLORATORY DATA ANALYSIS (EDA)



TÌNH TRẠNG BẤT CÂN BẰNG TRONG TẬP DỮ LIỆU



1. EXPLORATORY DATA ANALYSIS (EDA)



CHÊNH LỆCH VỀ ĐIỂM CHI-SQUARED Ở NHIỀU ĐẶC TÍNH

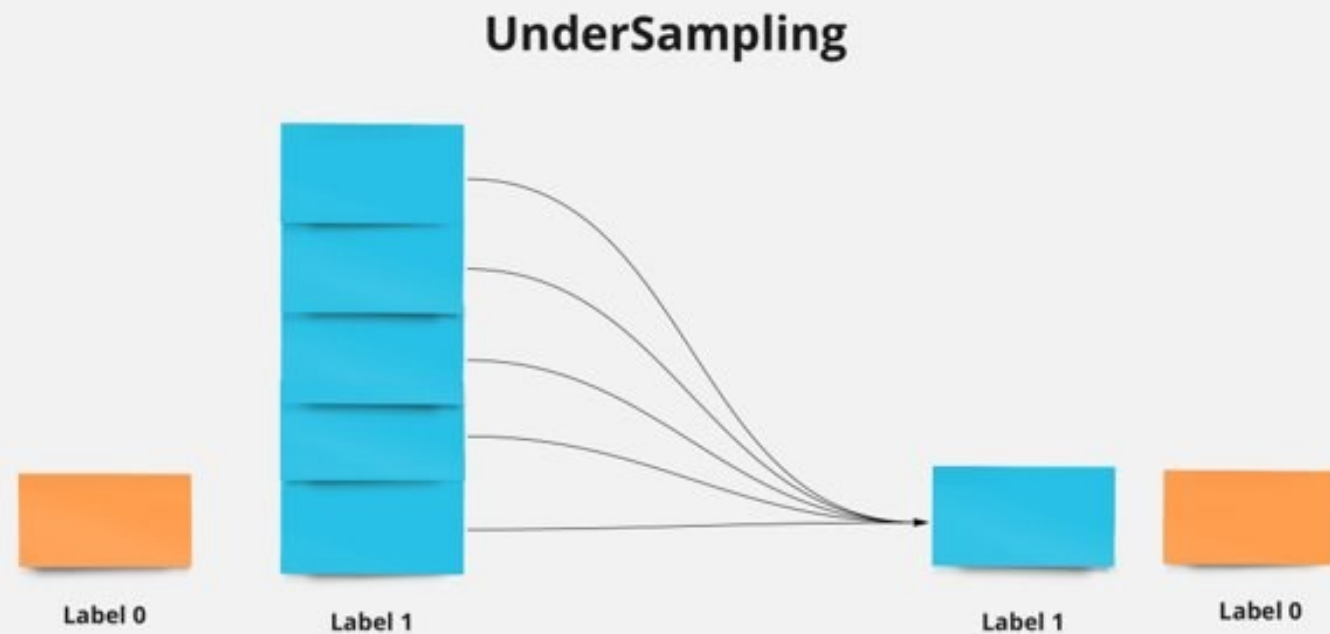


2. RESAMPLING

Kỹ thuật undersampling

Tránh trùng lặp dữ liệu

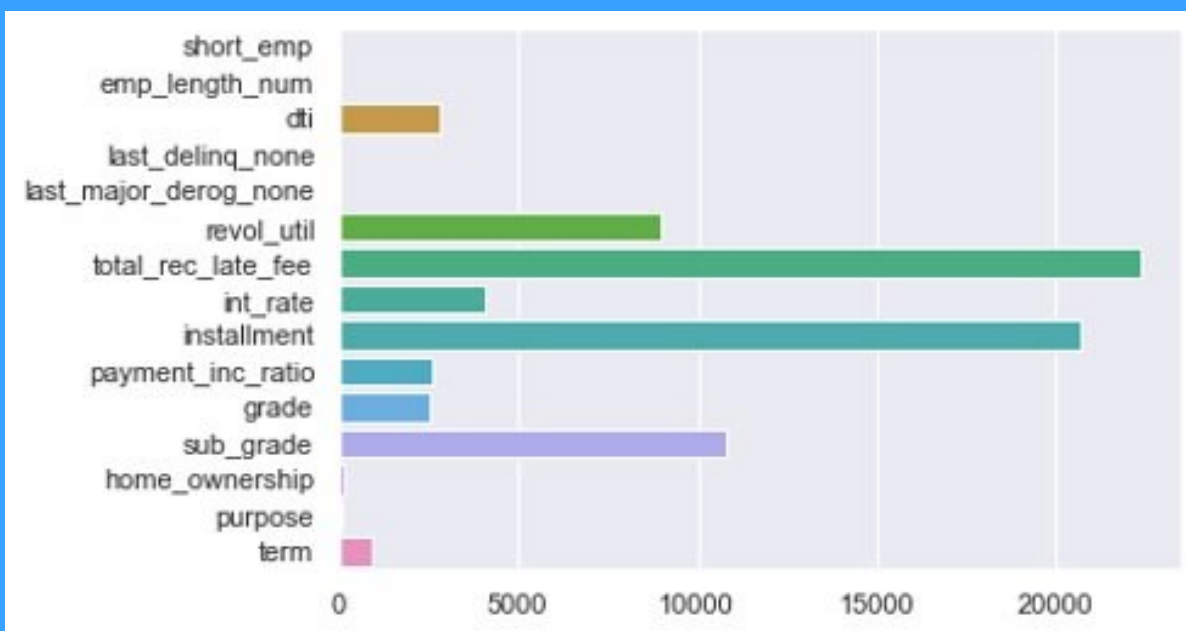
Số mẫu dữ liệu giảm xuống còn
khoảng 46000 mẫu
=> Thời gian training vừa phải



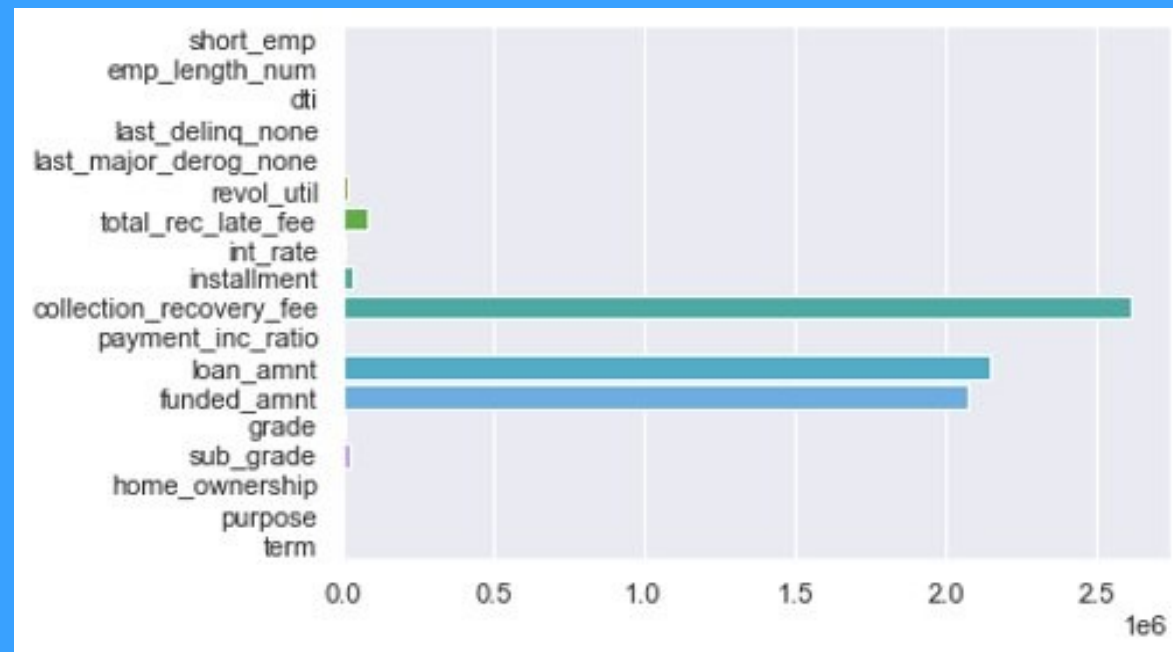
3. FEATURE SELECTION

Tiếp cận theo 2 hướng

Xử lý bài toán thông qua 2 lần chạy riêng biệt để so sánh



Run 1: loại bỏ những đặc tính có điểm chi-squared quá cao, chỉ giữ những đặc tính trong phạm vi cho phép



Run 2: giữ những đặc tính có điểm chi-squared rất cao so với tổng thể

4. ONE-HOT ENCODING

[1 , 4 , 2 , 0 , 3]



$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$

Normal array

One hot encoding

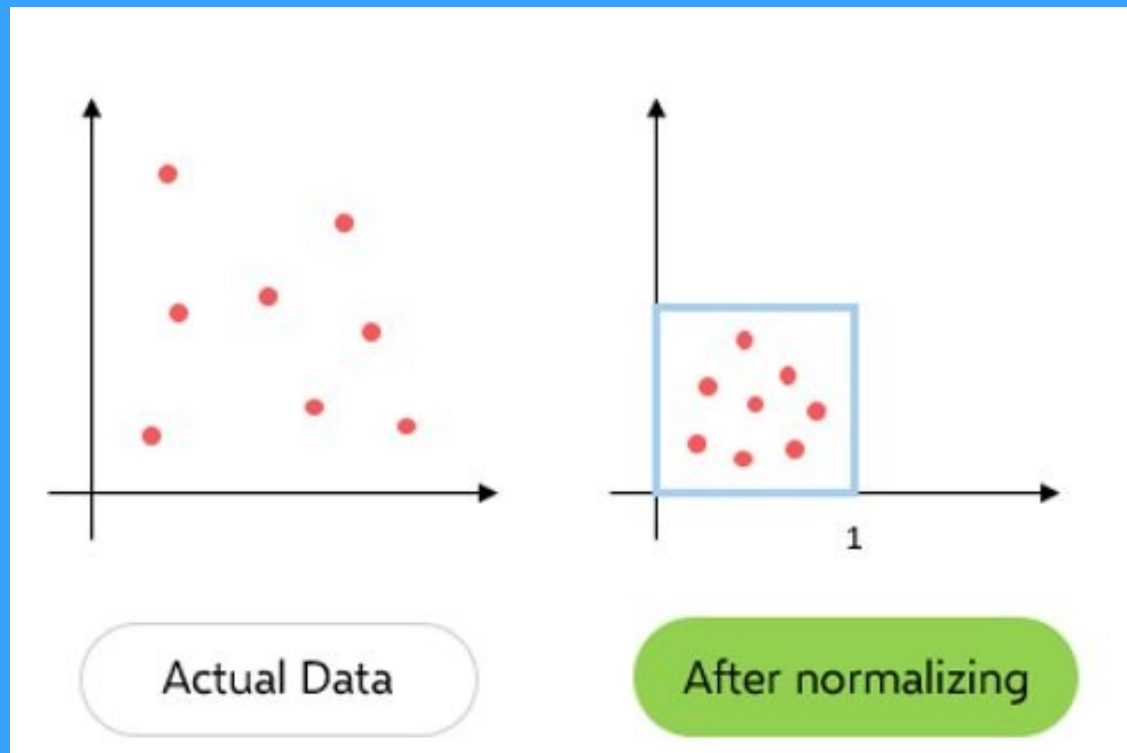
Xử lý các đặc tính dạng loại

Hiệu quả: loại bỏ khả năng mô hình nhận diện mối tương quan tuyến tính giữa nhãn cần dự đoán và đặc tính dạng loại (sai về mặt ý nghĩa)

5. FEATURE SCALING

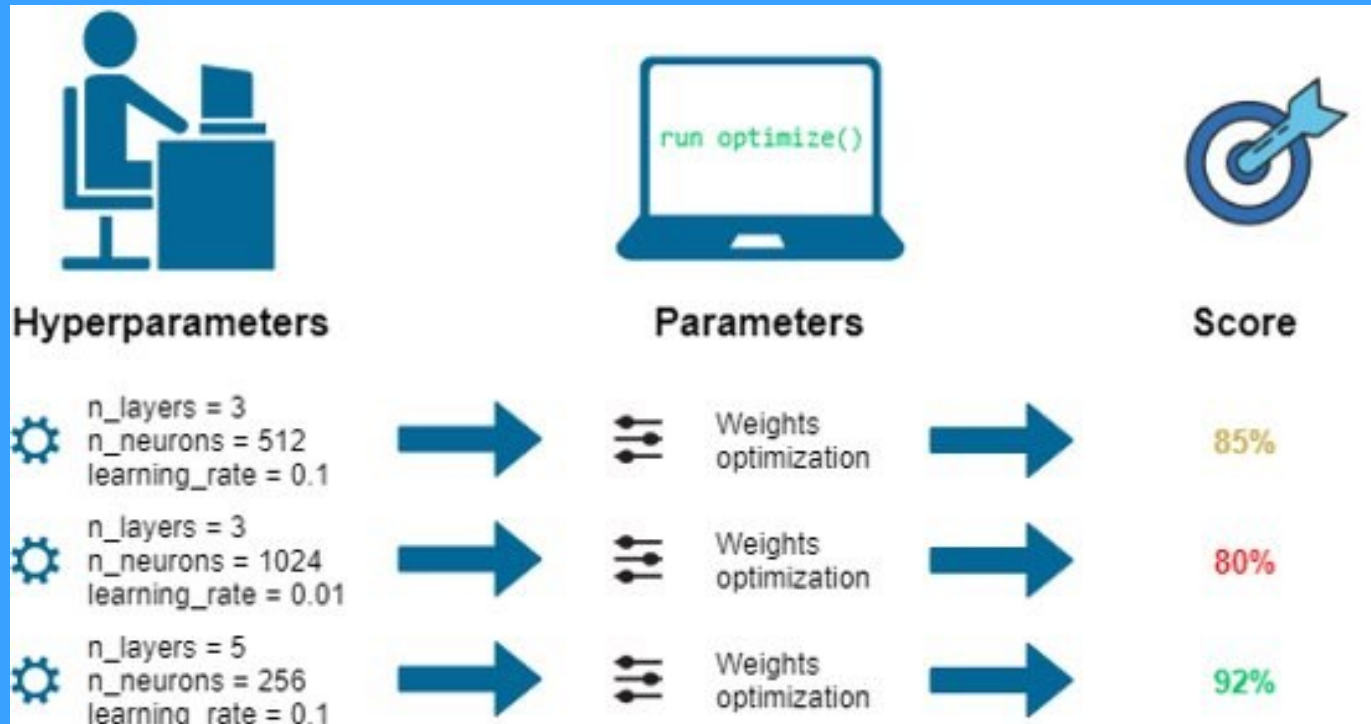
Normalization → Sử dụng min-max scaler

Hiệu quả: cân bằng vùng giá trị của các đặc tính => tăng hiệu quả tính toán và ý nghĩa toán học cho các thuật toán dựa trên khoảng cách



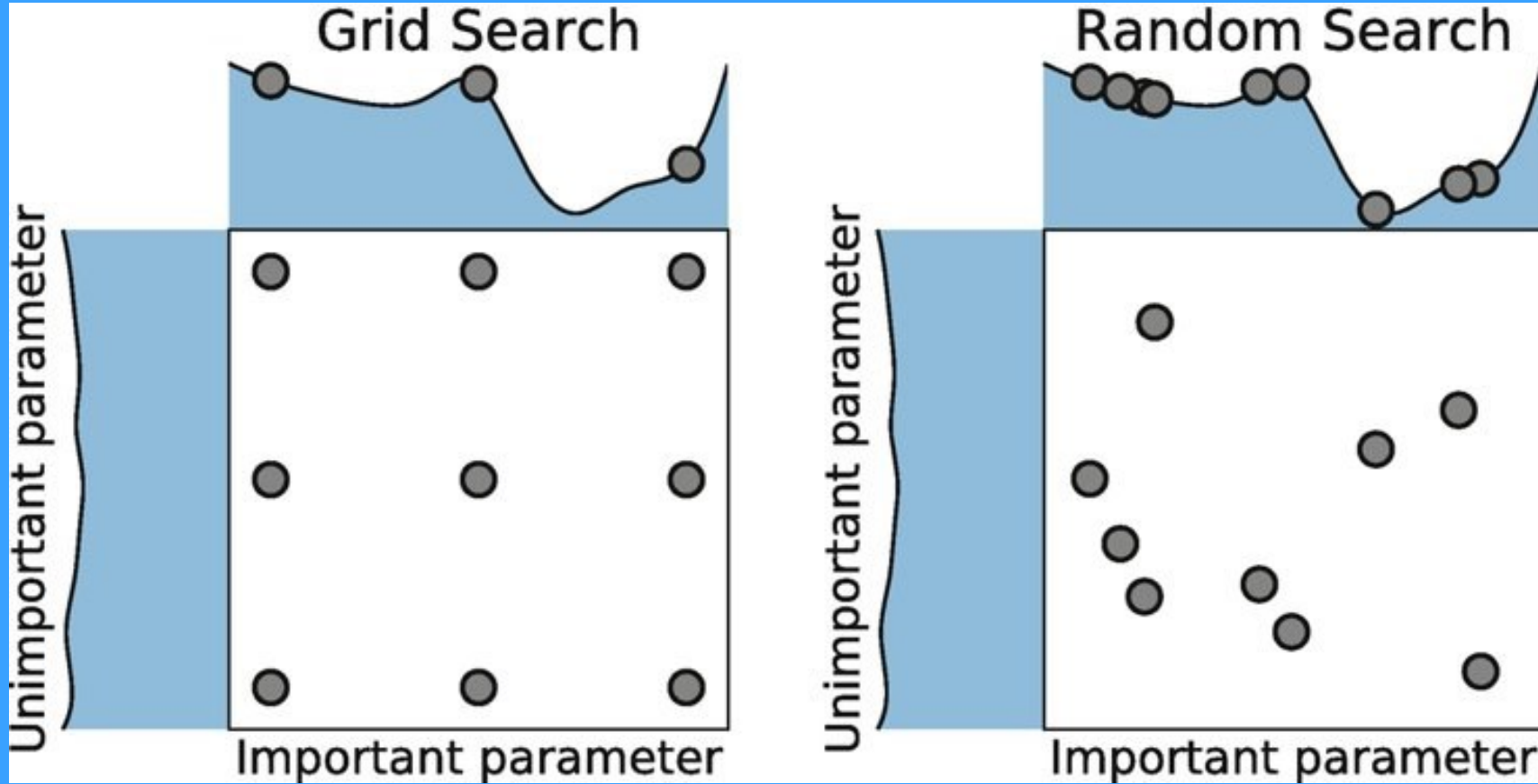
6. HYPERPARAMETERS TUNING

Các siêu tham số có ảnh hưởng không nhỏ đến hiệu quả huấn luyện và dự đoán của các mô hình



MỤC ĐÍCH: XÁC ĐỊNH BỘ SIÊU THAM SỐ TỐI ƯU CHO HIỆU QUẢ CỦA MÔ HÌNH

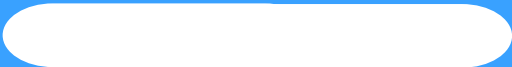
6. HYPERPARAMETERS TUNING



SỬ DỤNG KỸ THUẬT RANDOM SEARCH ĐỂ GIẢM THỜI GIAN TÌM KIẾM

7. TRAINING

HUẤN LUYỆN VỚI SỐ LƯỢNG ĐẶC TÍNH GIẢM DẦN



Run 1: trích xuất 15
đặc tính

Huấn luyện từng
mô hình với 11,
12, 13, 14, 15
đặc tính



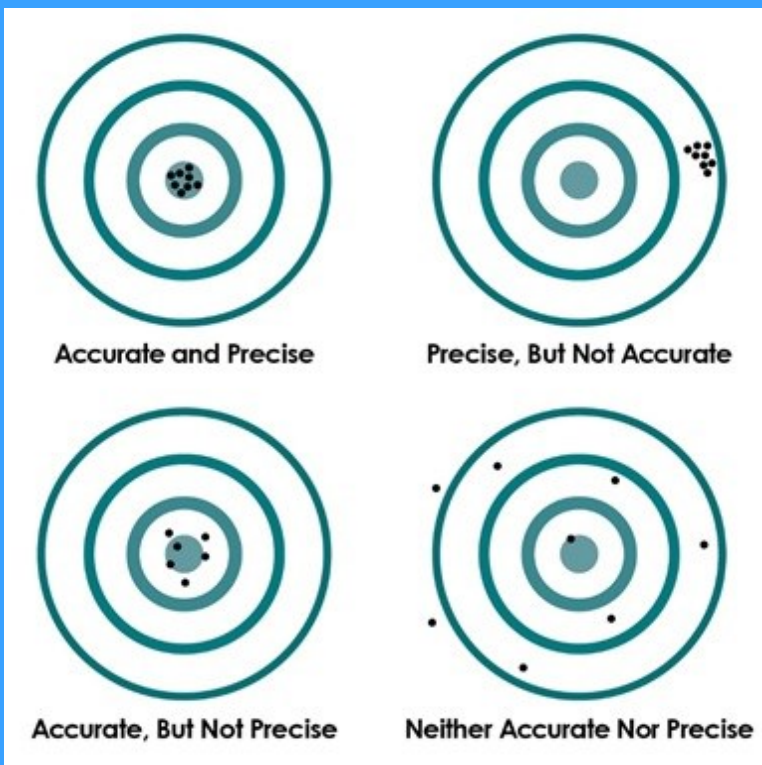
Run 2: trích xuất 18
đặc tính

Huấn luyện từng mô
hình với 14, 15, 16,
17, 18 đặc tính

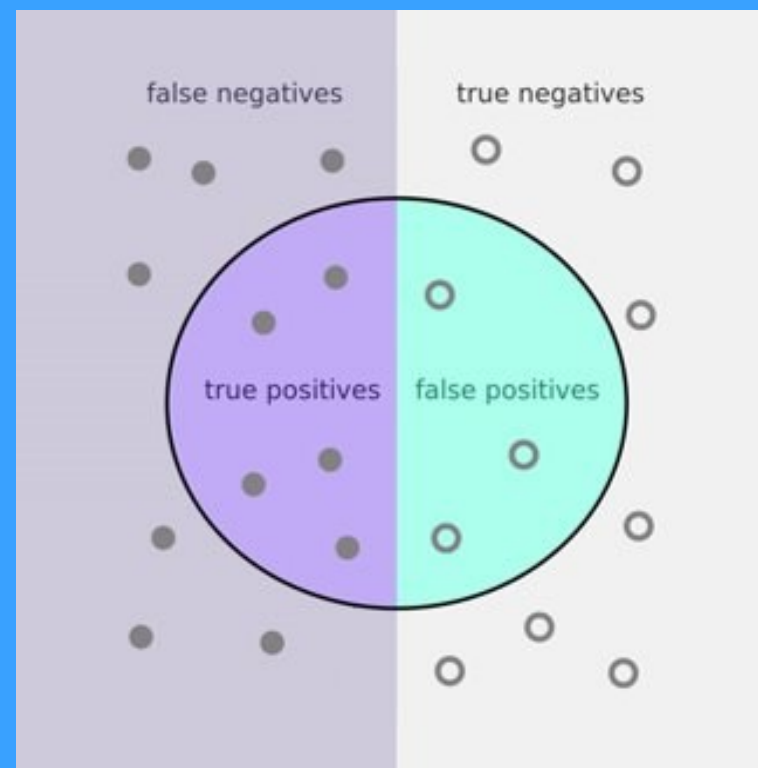
8. EVALUATION

KIỂM ĐỊNH HIỆU QUẢ MÔ HÌNH DỰA TRÊN 2 TIÊU CHÍ

ĐỘ CHÍNH XÁC

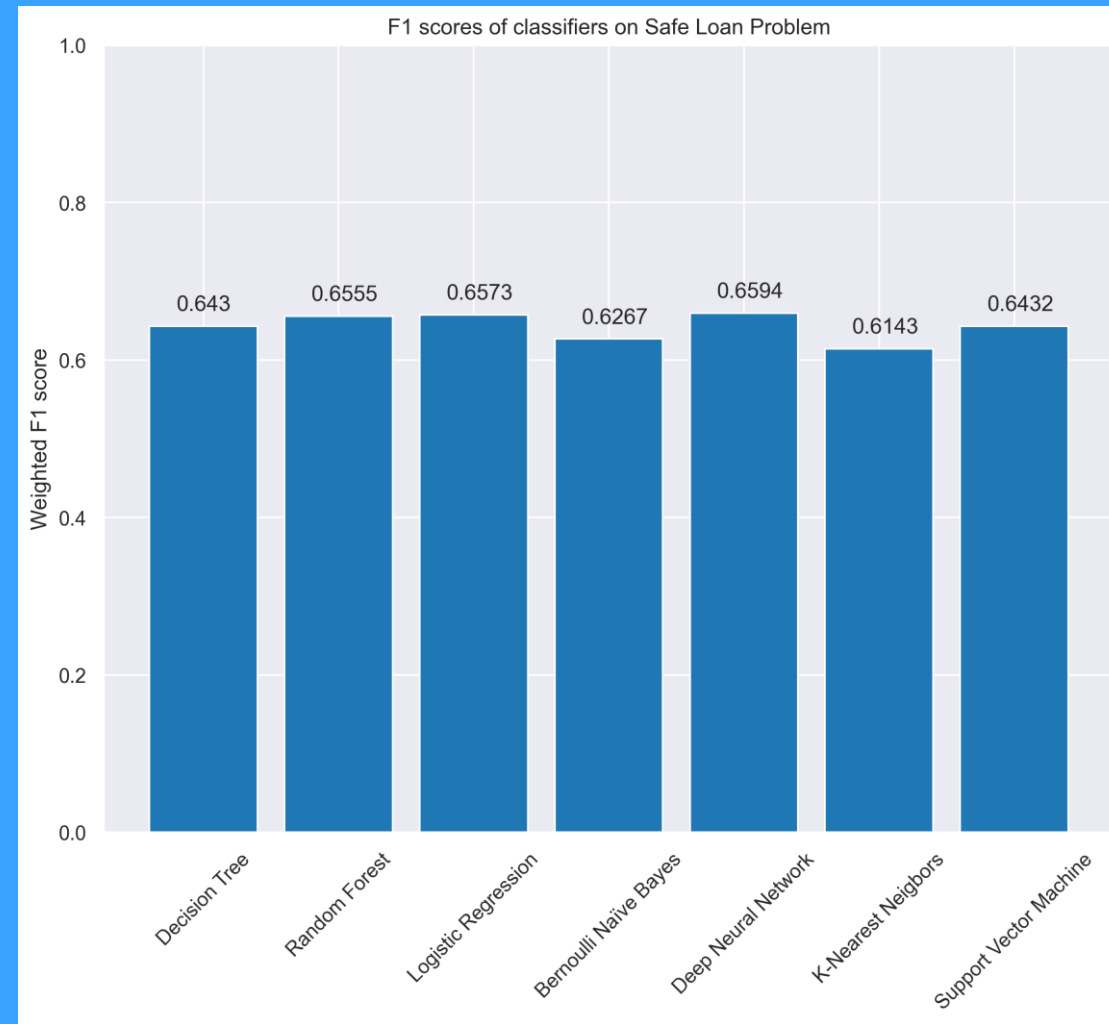
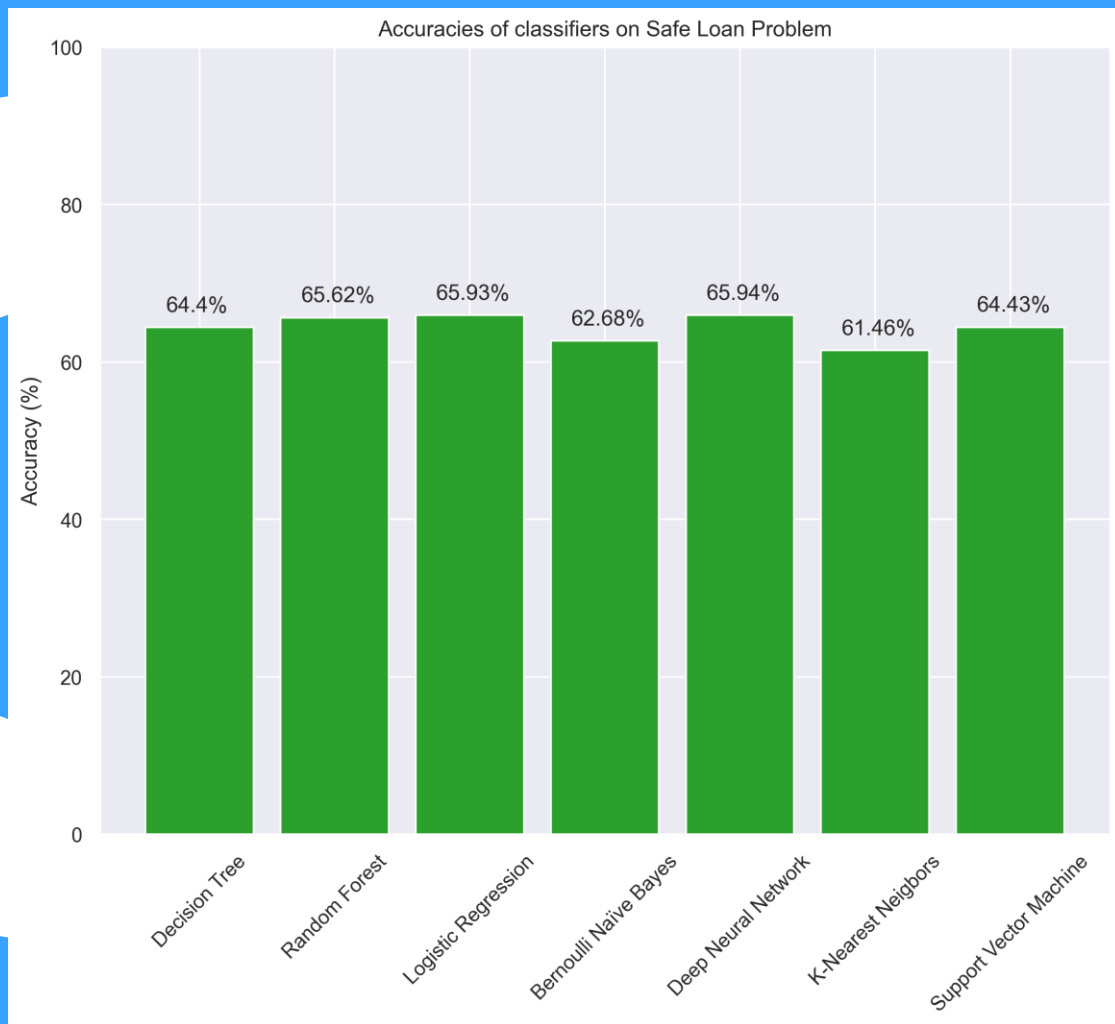


ĐIỂM F1 TRUNG BÌNH



KẾT QUẢ

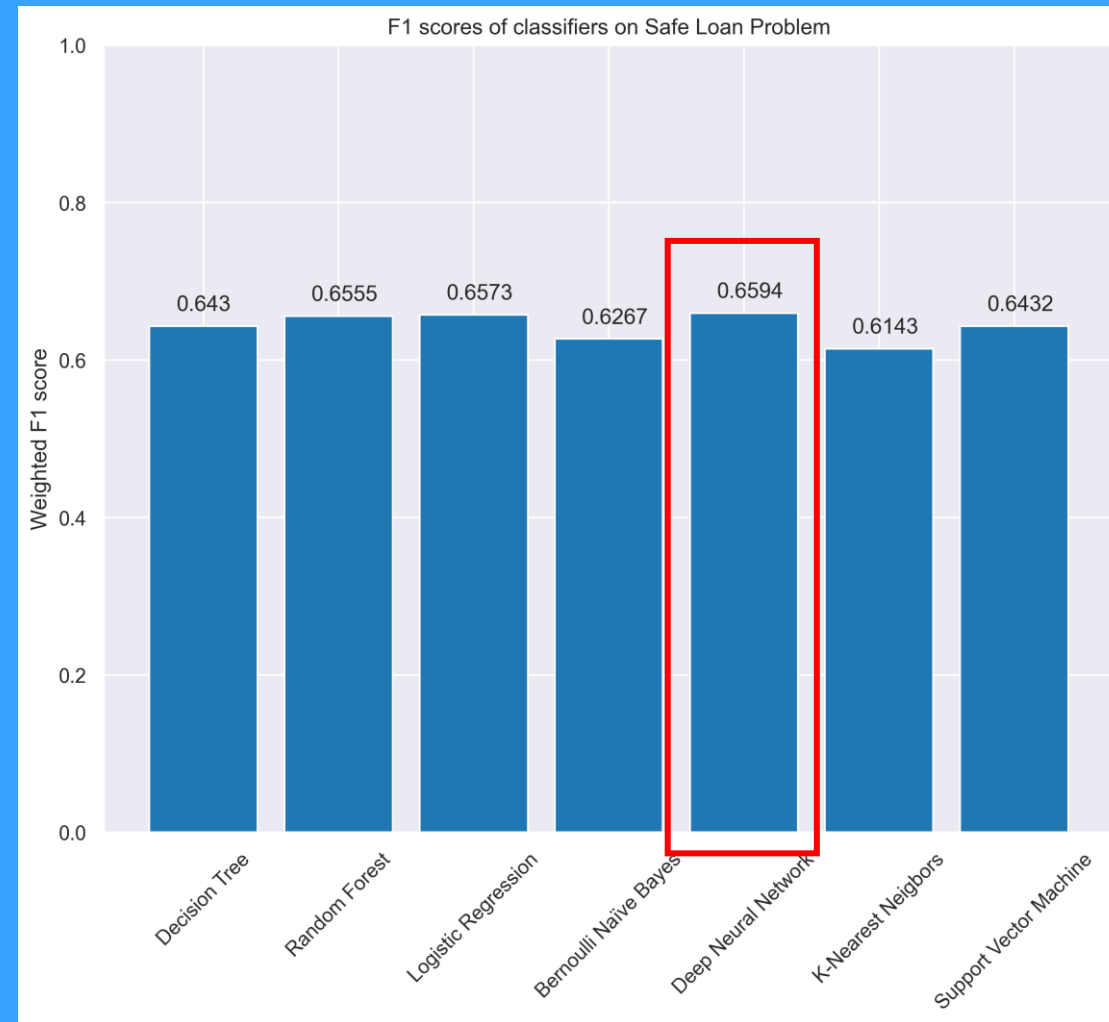
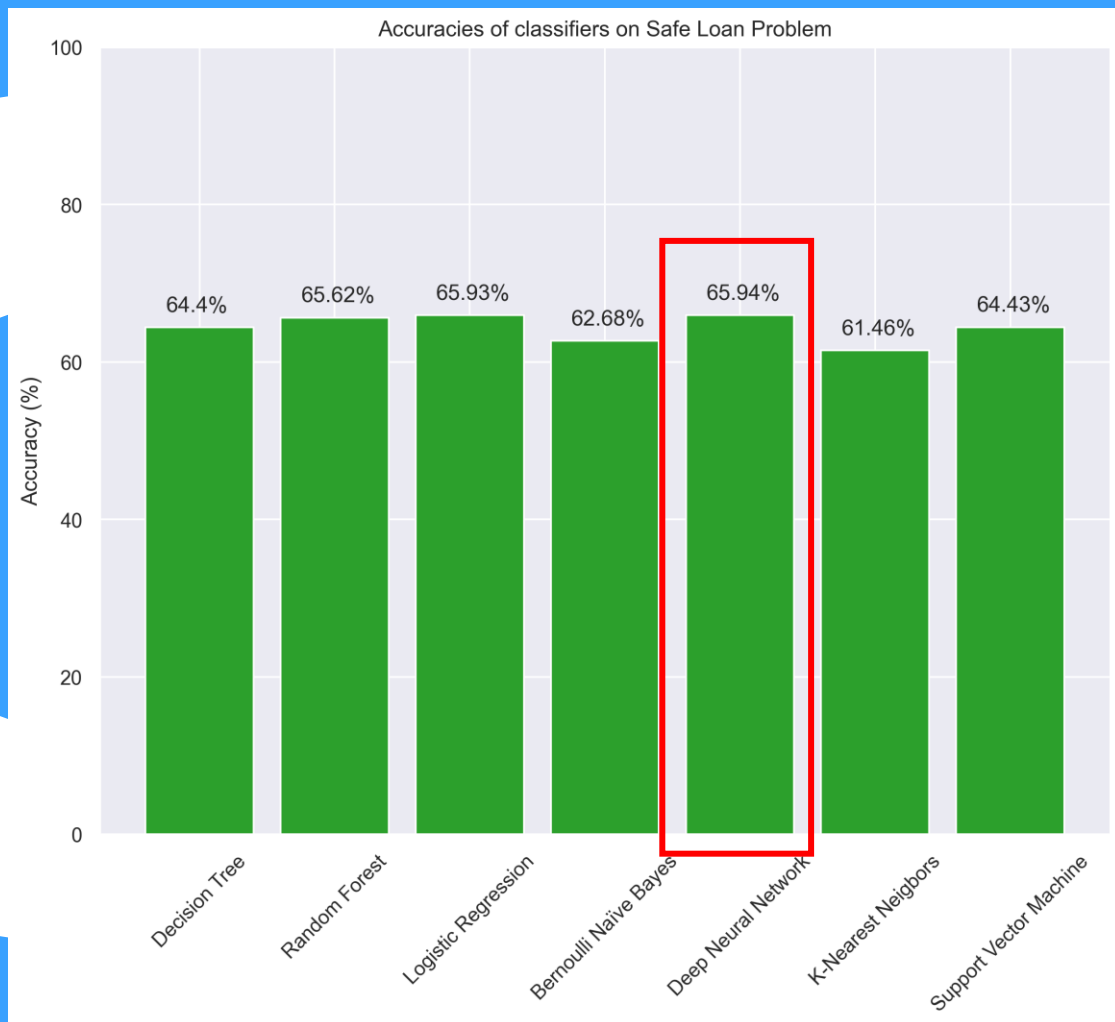
RUN 1



KẾT QUẢ

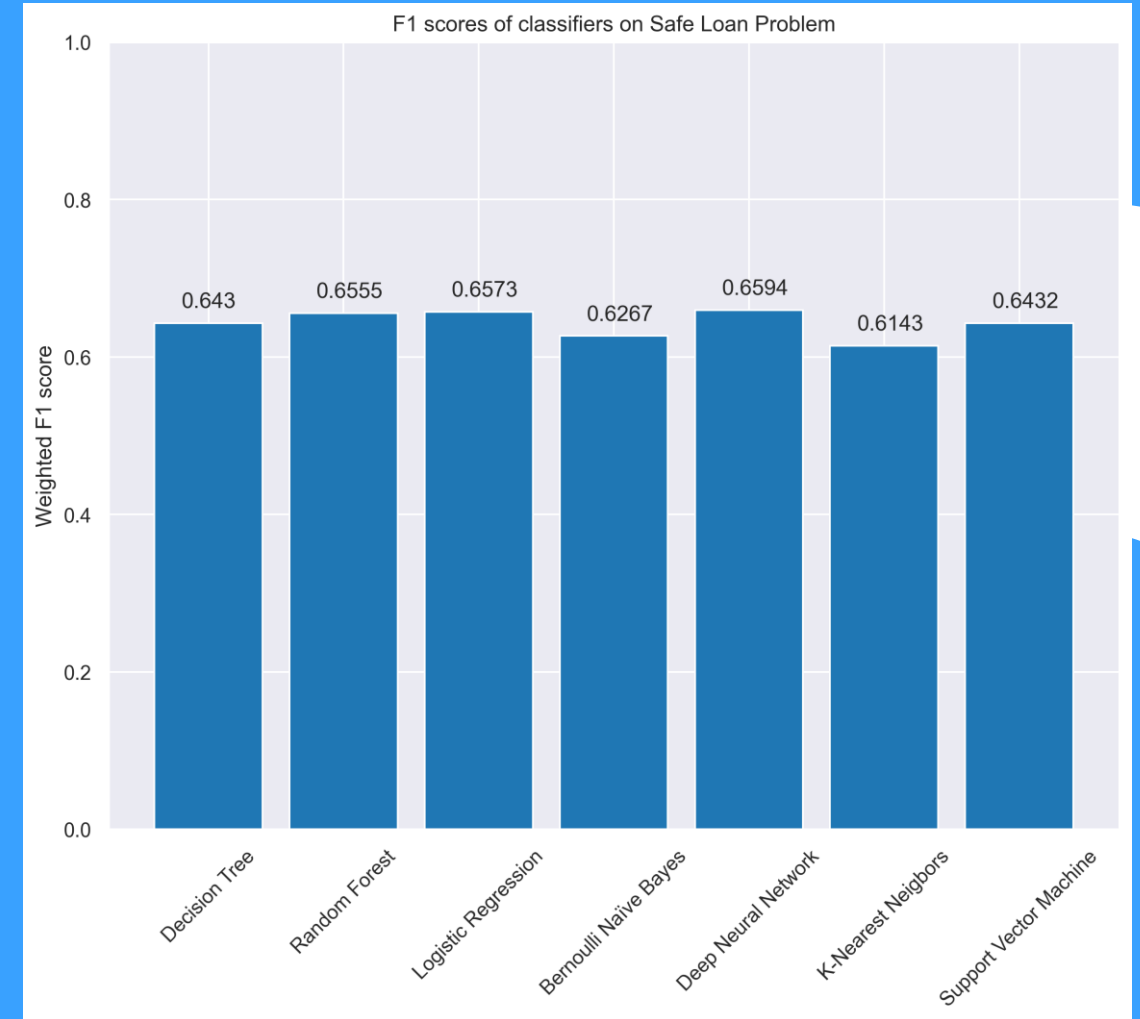
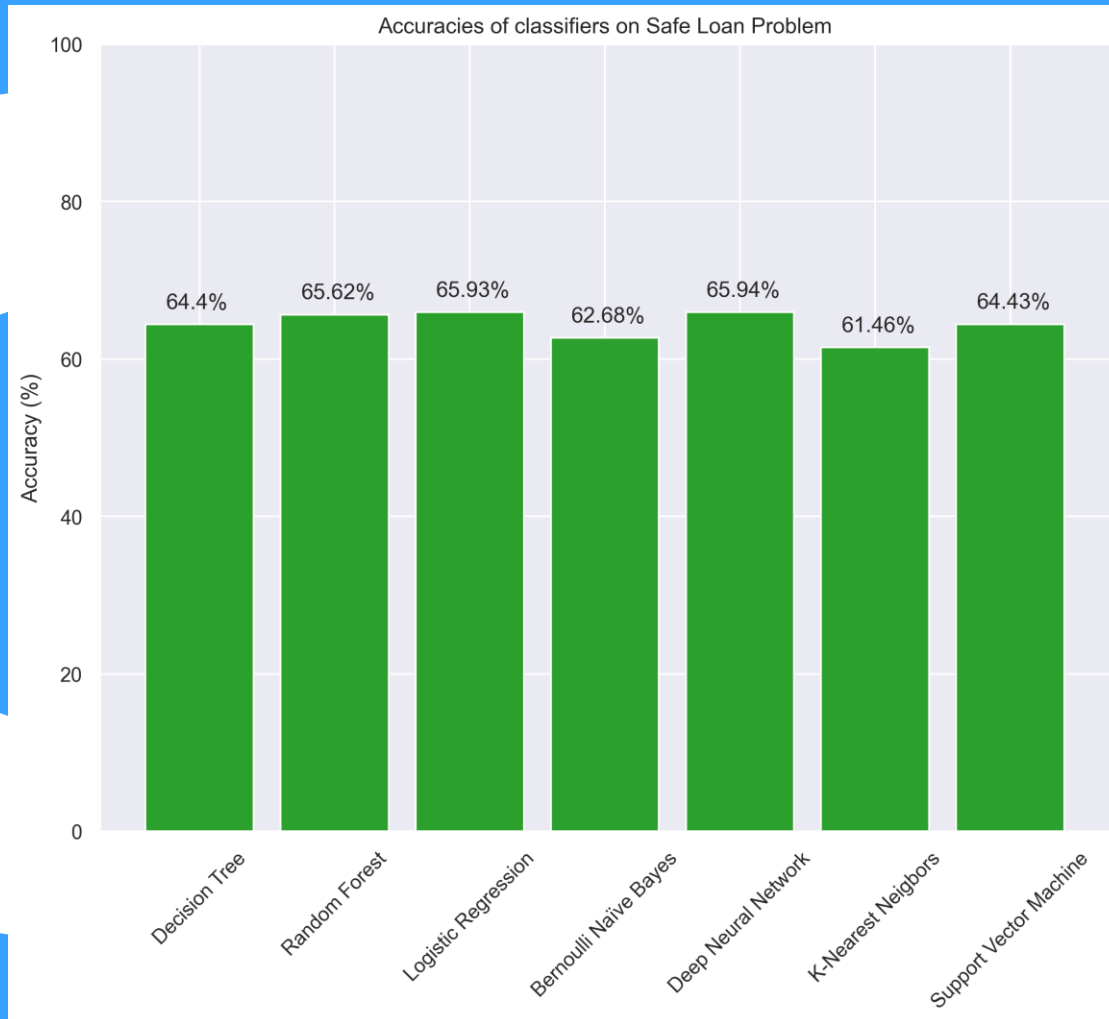
RUN 1

TỐT NHẤT: **ARTIFICIAL NEURAL NETWORK**



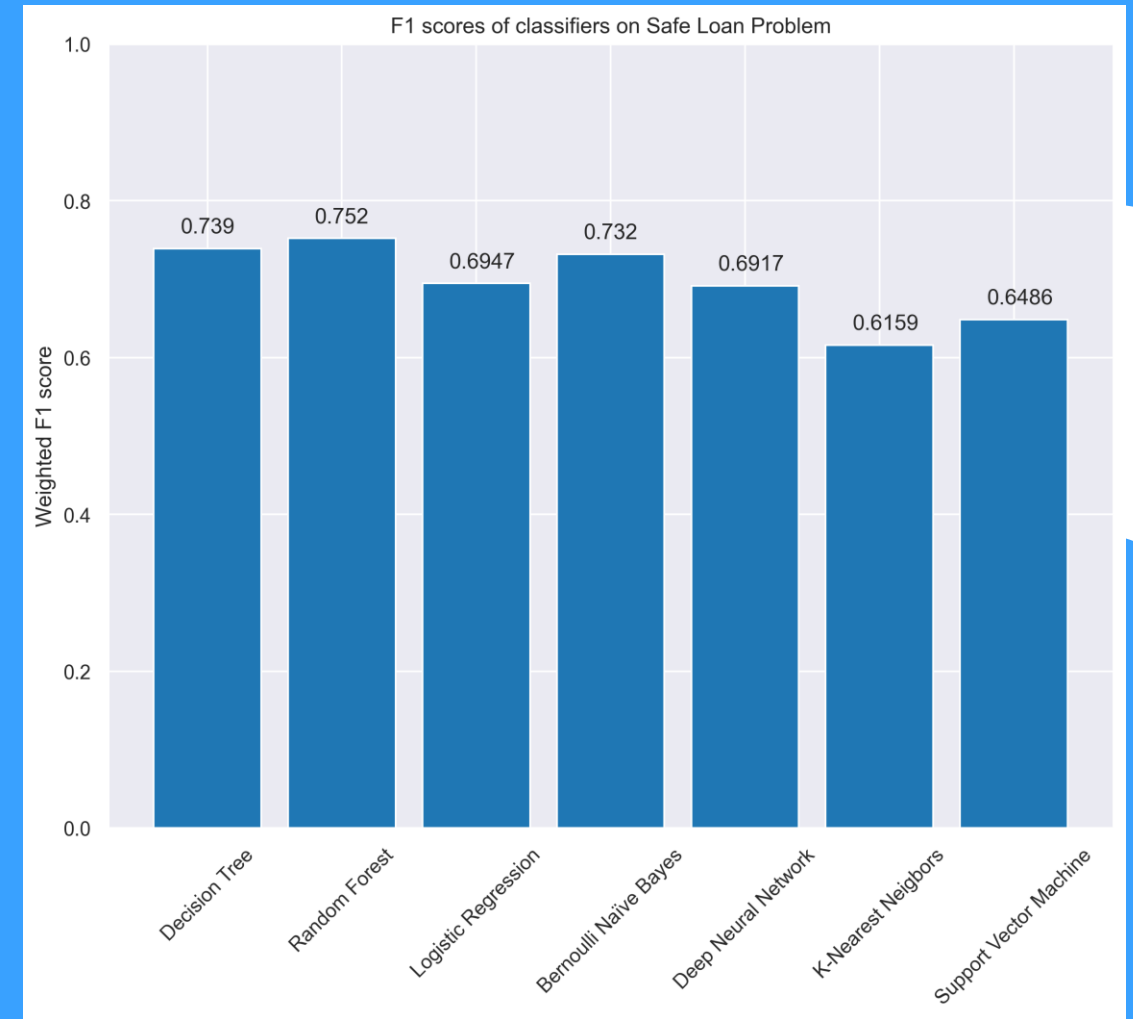
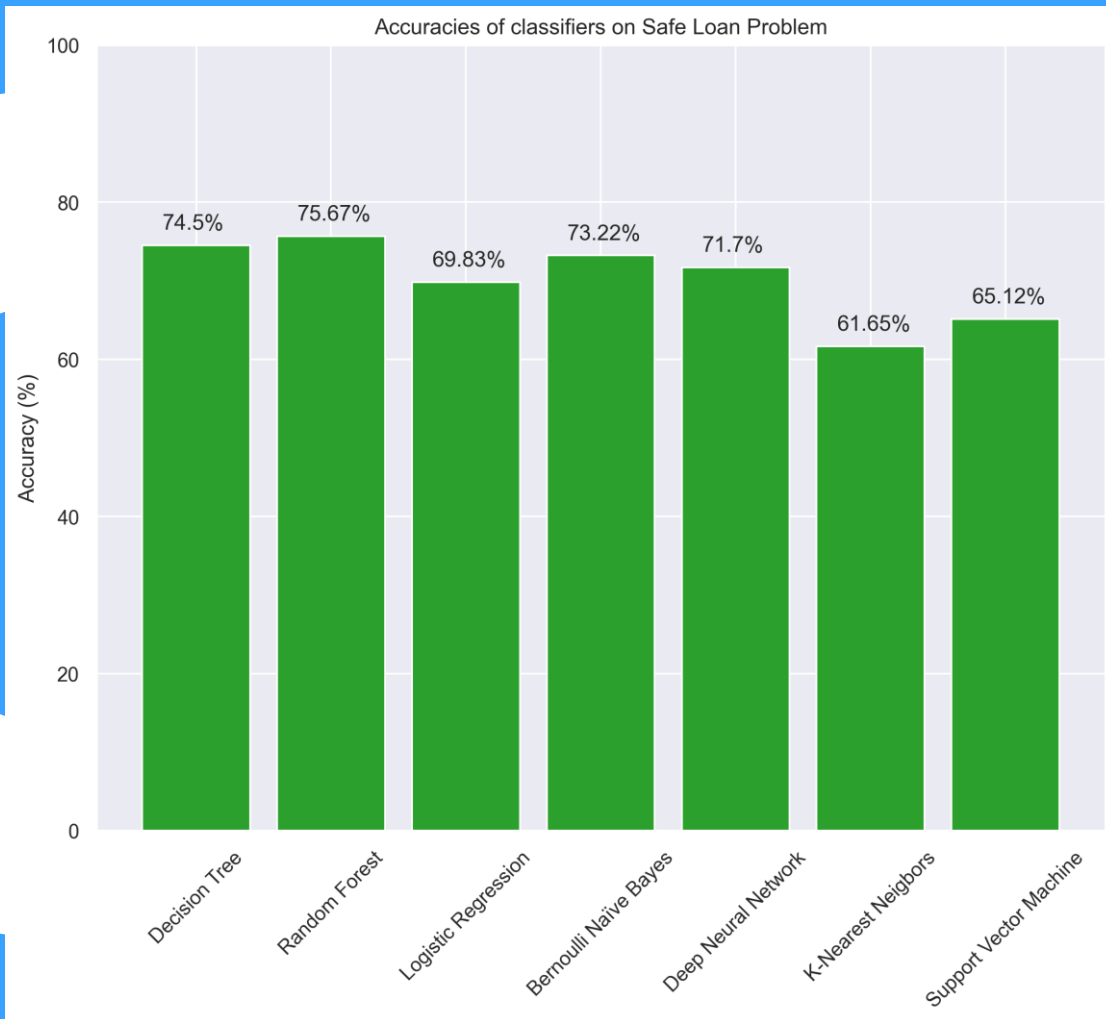
RUN 1

CHÊNH LỆCH GIỮA ACCURACY VÀ F1 KHÔNG ĐÁNG KỂ



KẾT QUẢ

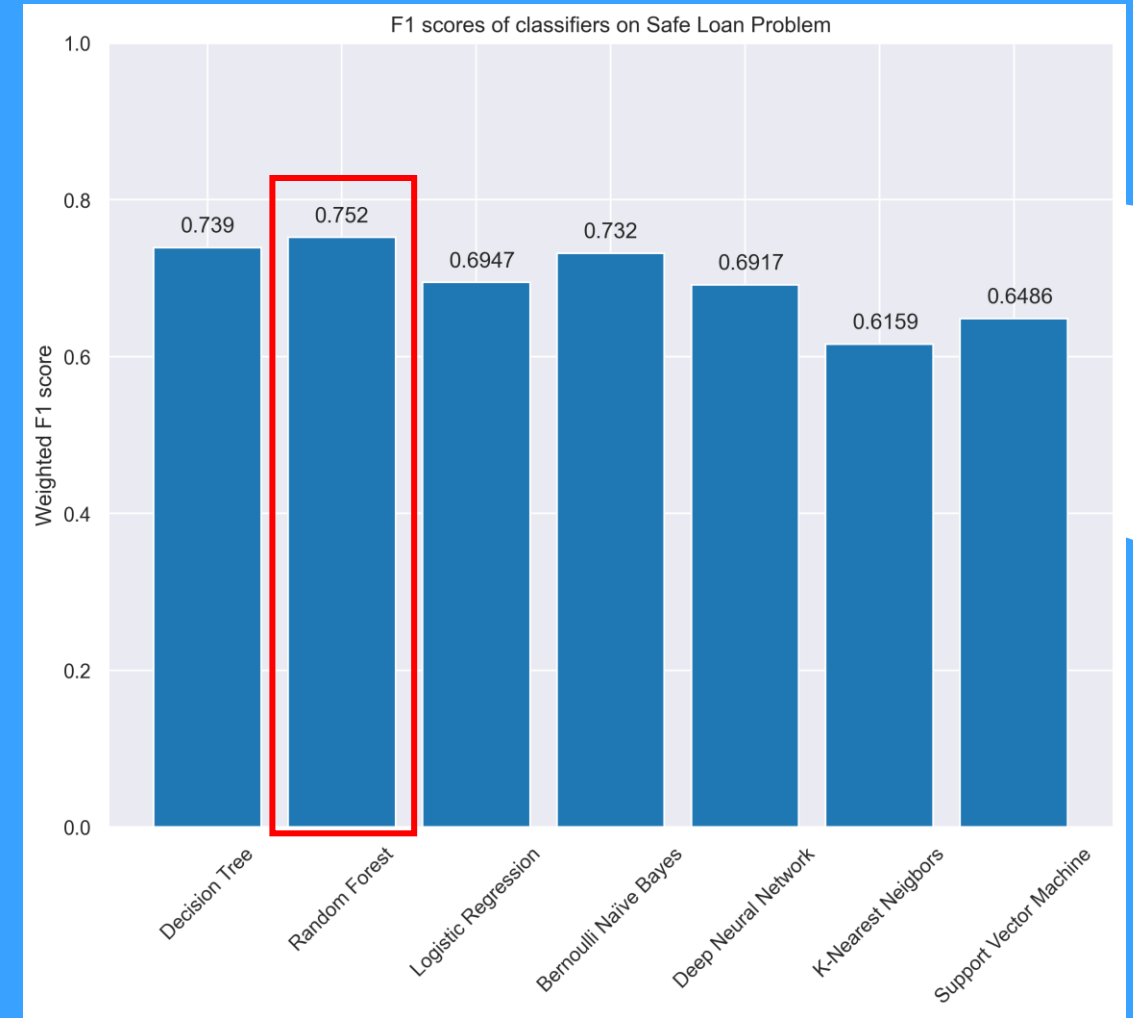
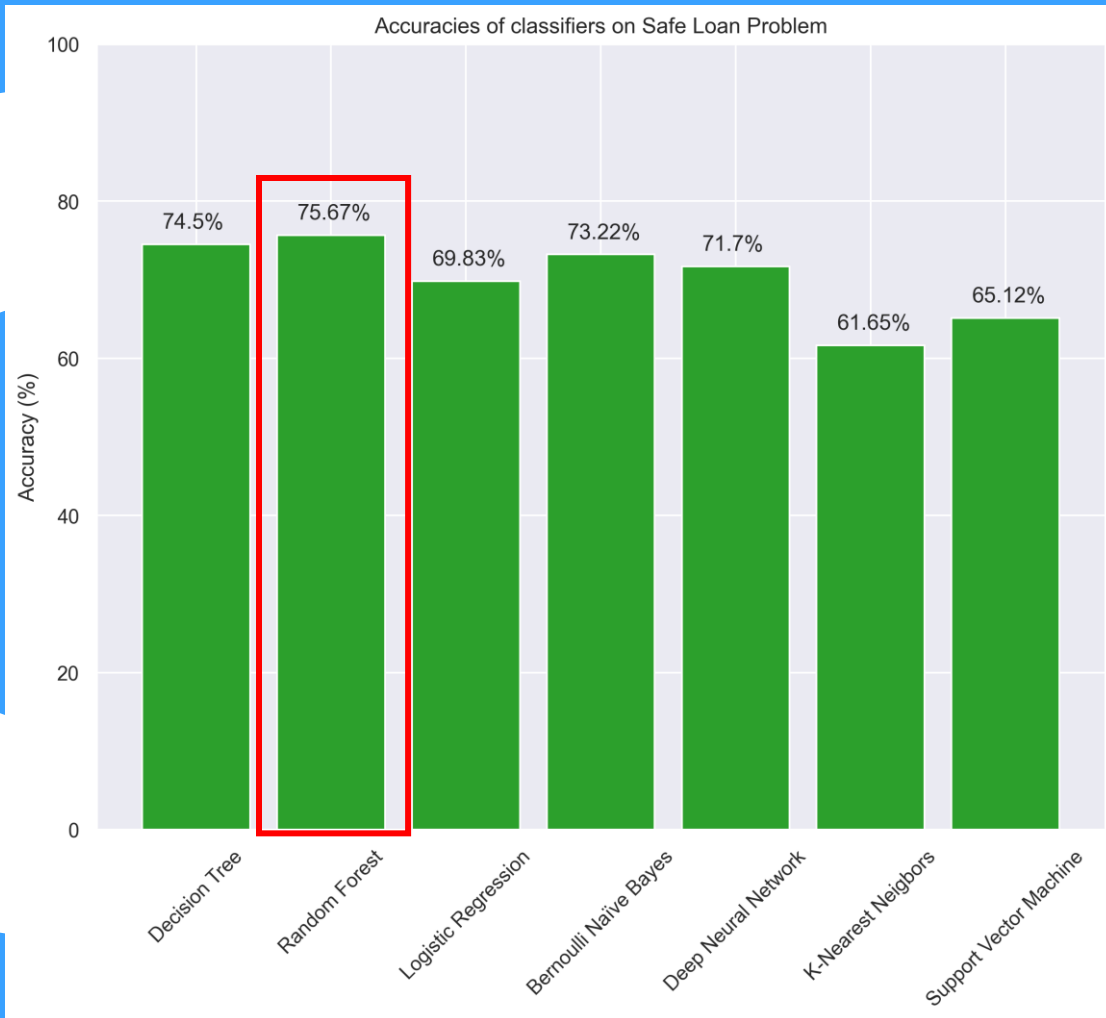
RUN 2



KẾT QUẢ

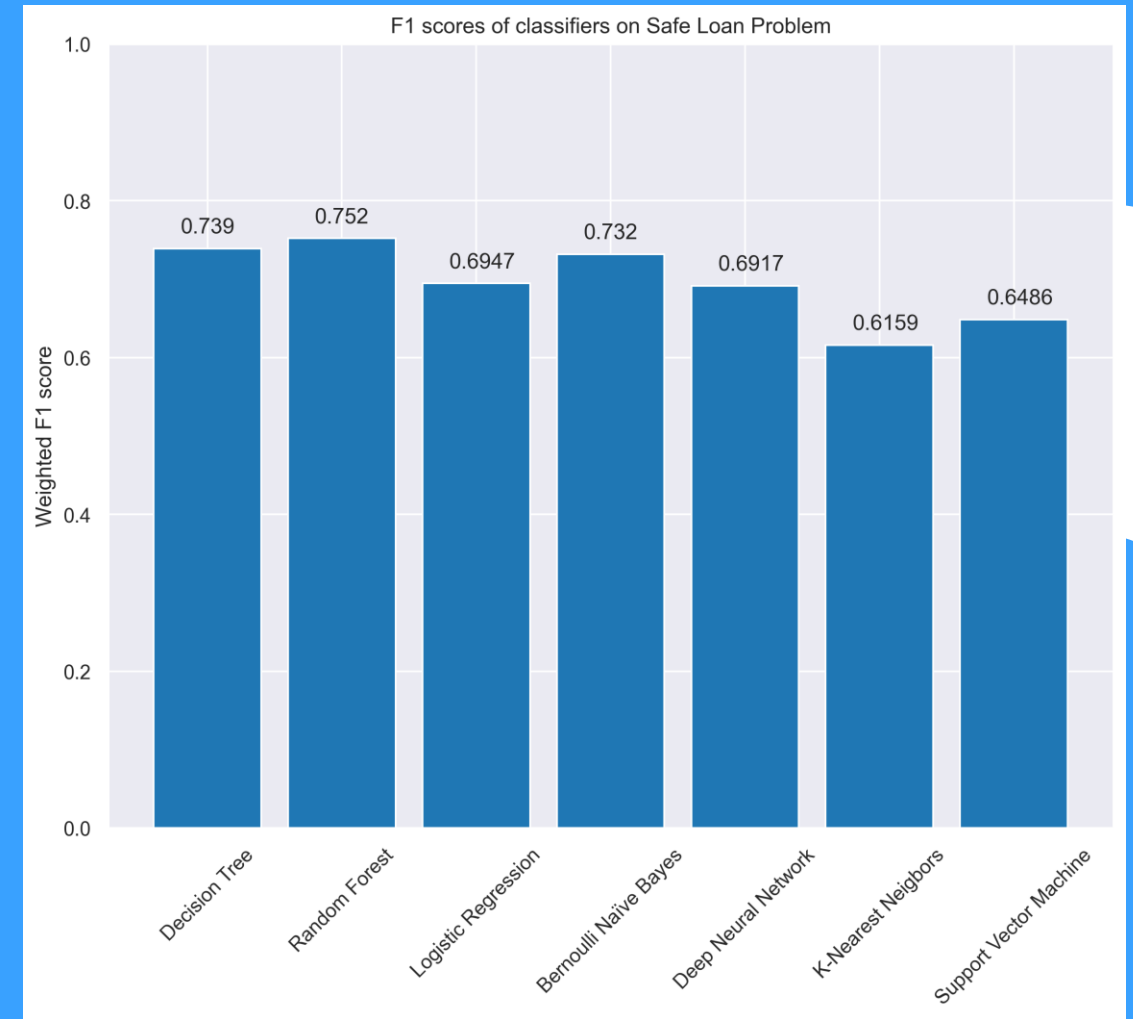
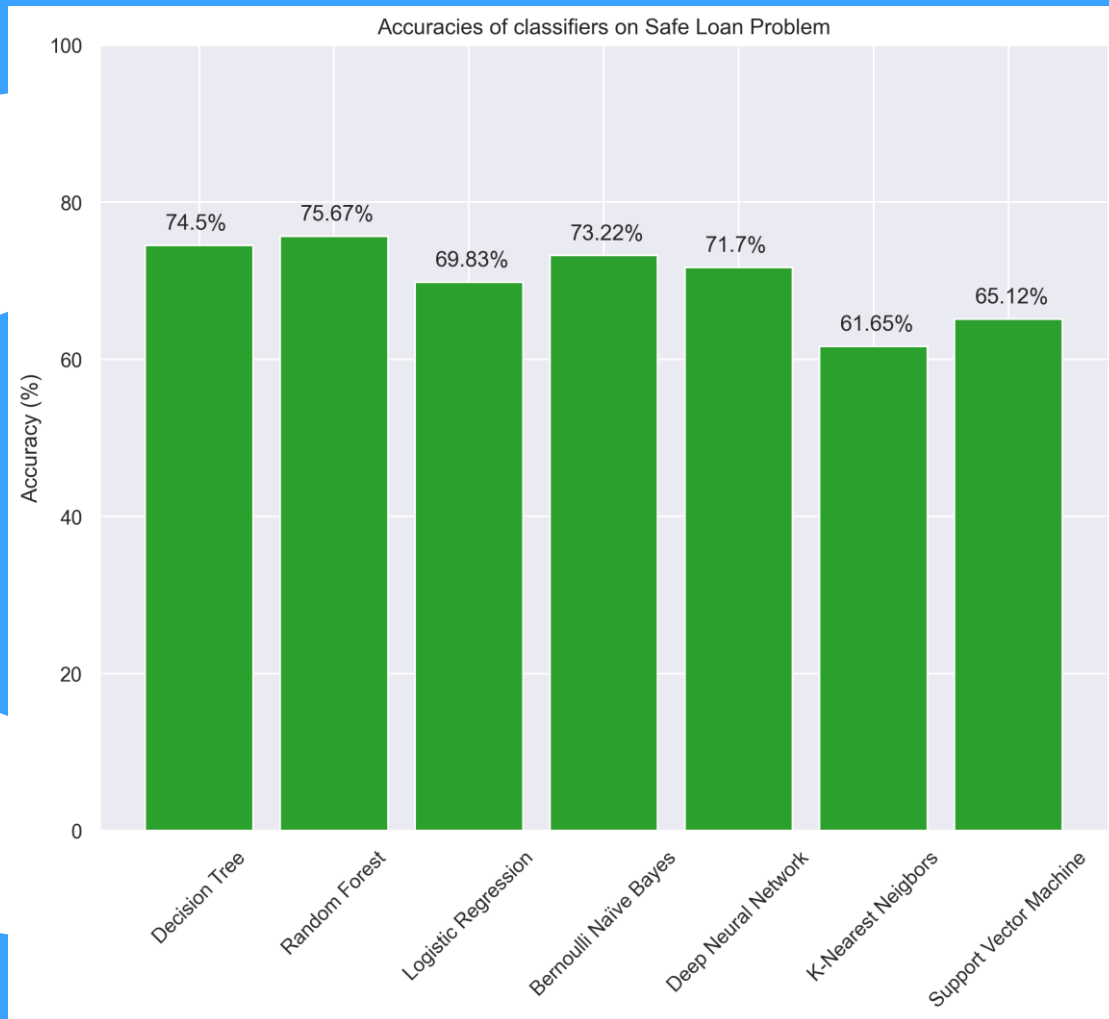
RUN 2

TỐT NHẤT: **RANDOM FOREST**



RUN 2

HIỆU QUẢ HUẤN LUYỆN TĂNG ĐÁNG KỂ VỚI CÁC THUẬT TOÁN **DẠNG CÂY** VÀ **XÁC SUẤT**



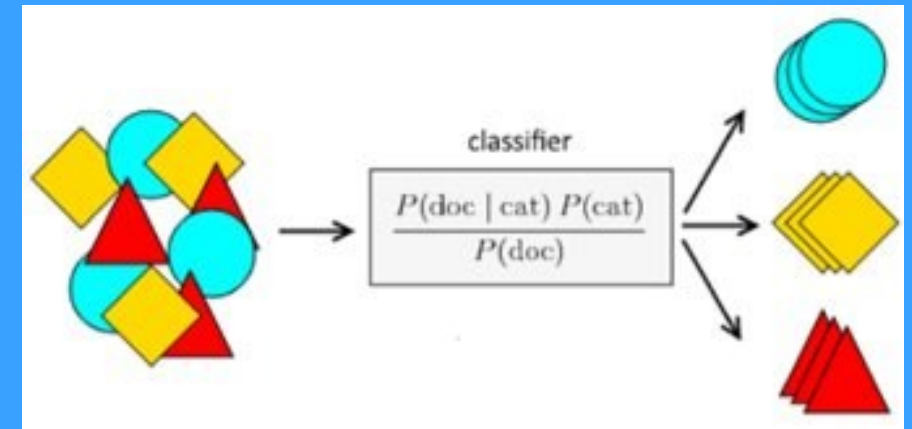
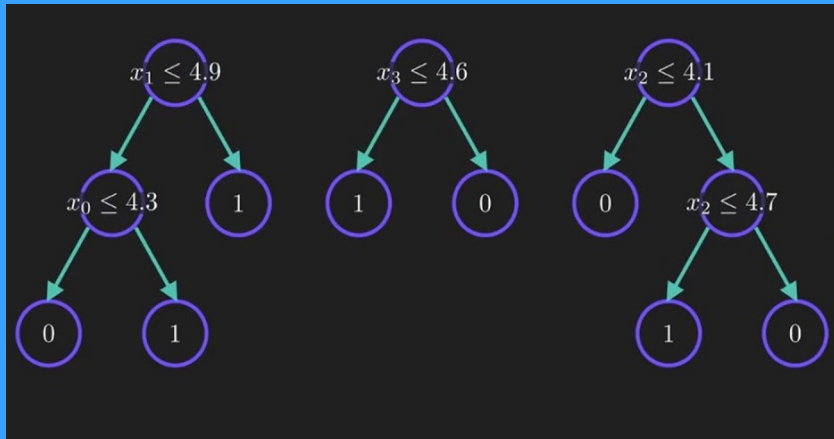


KẾT LUẬN

KẾT LUẬN

Với số lượng đặc tính hạn chế, các thuật toán có hiệu quả tương đối ngang nhau, dao động trong phạm vi từ 61-66% về độ chính xác

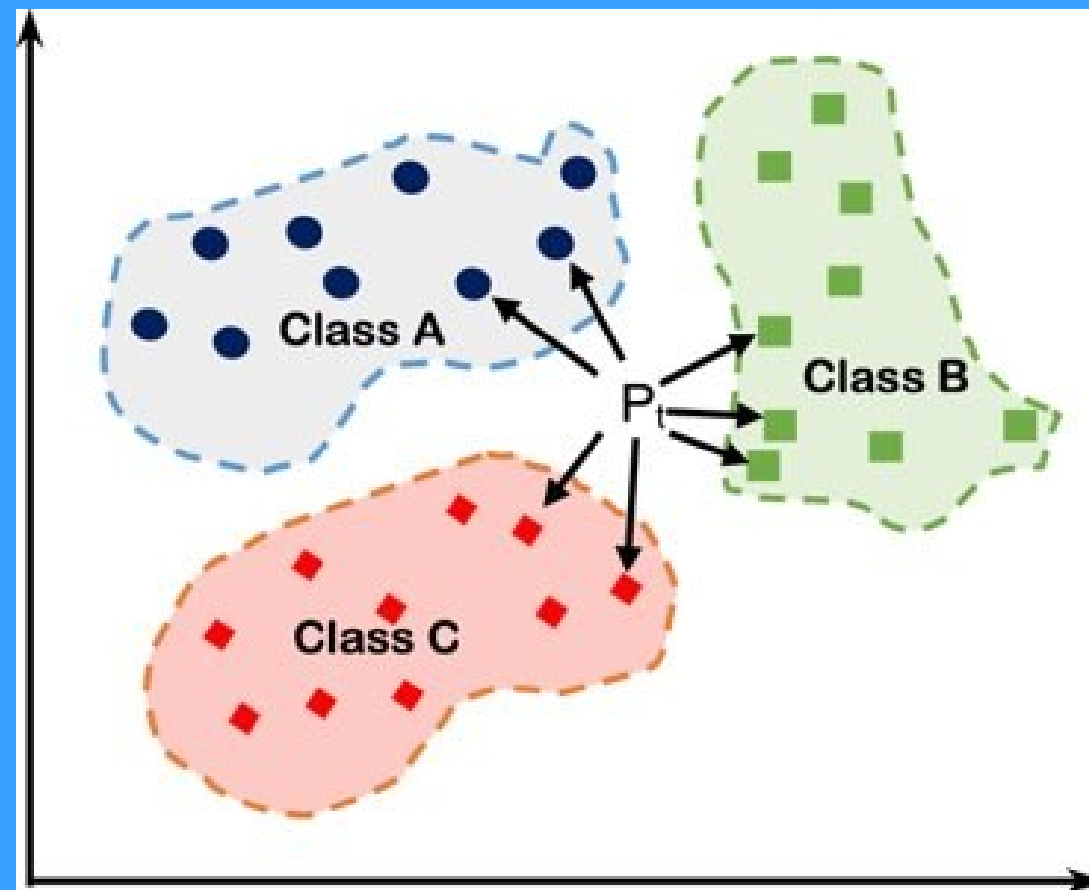
Với số lượng đặc tính cao, bao hàm những đặc tính có độ tương quan cao với nhãn, các thuật toán dạng cây hoặc xác suất đạt được nhiều hiệu quả đáng kể



KẾT LUẬN

Với cả hai lần kiểm nghiệm, thuật toán k-nearest neighbors đều có độ chính xác thấp nhất đối với bài toán

=> Tập dữ liệu không cho hiệu quả cao với thuật toán dựa trên khoảng cách



KẾT LUẬN VÀ ĐỀ XUẤT

Phân tích và trích xuất đặc tính tập dữ liệu có thể cho hiệu quả cao hơn đối với các mô hình phân loại

Giải pháp kiểm định nhiều mô hình phân loại để tìm mô hình tốt nhất cho bài toán là một hướng tiếp cận có hiệu quả

Cảm ơn thầy và các bạn
đã lắng nghe

