# BÁO CÁO TÌM HIỂU THUẬT TOÁN: XGBOOST CLASSIFIER

## 1. Giới thiệu và Ý nghĩa thuật toán

**XGBoost (Extreme Gradient Boosting)** là một thuật toán học máy mạnh mẽ thuộc nhóm **Ensemble Learning** (Học kết hợp), cụ thể là kỹ thuật **Boosting**.

* **Nguyên lý hoạt động:** Không giống như các thuật toán cây quyết định đơn lẻ (như ID3 hay CART), XGBoost xây dựng mô hình bằng cách kết hợp hàng loạt các "cây quyết định yếu" (weak learners) một cách tuần tự. Cây sinh ra sau sẽ tập trung sửa lỗi (residual) của các cây trước đó.
* **Ý nghĩa:** XGBoost được thiết kế để tối ưu hóa cả về **tốc độ tính toán** và **hiệu suất mô hình**. Nó tích hợp sẵn cơ chế Regularization (chống quá khớp - overfitting), giúp mô hình tổng quát hóa tốt hơn trên dữ liệu mới. Đây là thuật toán thường xuyên giành chiến thắng trong các cuộc thi dữ liệu (Kaggle).

## 2. Cách sử dụng và Các tham số quan trọng

Trong Python, XGBoost được sử dụng thông qua thư viện xgboost và có API tương thích với scikit-learn.

### Các tham số (Hyperparameters) cốt lõi:

|  |  |  |
| --- | --- | --- |
| **Tham số** | **Ý nghĩa** | **Giá trị thường dùng** |
| **n\_estimators** | Số lượng cây quyết định được tạo ra. | 100, 500, 1000 |
| **learning\_rate (eta)** | Tốc độ học. Kiểm soát mức đóng góp của mỗi cây vào kết quả chung. Giá trị nhỏ giúp mô hình bền vững nhưng cần nhiều cây hơn. | 0.01 - 0.3 |
| **max\_depth** | Độ sâu tối đa của mỗi cây. Càng sâu mô hình càng phức tạp và dễ bị overfitting. | 3 - 10 |
| **subsample** | Tỷ lệ dữ liệu được lấy mẫu ngẫu nhiên để huấn luyện từng cây. Giúp giảm overfitting. | 0.6 - 0.8 |
| **colsample\_bytree** | Tỷ lệ số lượng đặc trưng (cột) được chọn ngẫu nhiên cho mỗi cây. | 0.6 - 0.8 |
| **gamma** | Mức giảm hàm mất mát tối thiểu cần thiết để thực hiện phân chia một nút lá. | 0 - 5 |
| **objective** | Hàm mục tiêu cần tối ưu. | binary:logistic (nhị phân), multi:softmax (đa lớp) |

## 3. Ưu điểm và Nhược điểm

### Ưu điểm:

1. **Hiệu suất vượt trội:** Thường cho độ chính xác cao hơn so với ID3, CART và Gradient Boosting truyền thống.
2. **Tốc độ xử lý:** Hỗ trợ tính toán song song (parallel processing), tận dụng tốt đa luồng CPU.
3. **Xử lý dữ liệu thiếu (Missing Data):** Có cơ chế tự động học hướng đi tốt nhất cho các giá trị bị khuyết (NaN).
4. **Regularization:** Tích hợp L1 và L2 regularization giúp mô hình tránh bị overfitting tốt hơn.

### Nhược điểm:

1. **Khó tinh chỉnh (Tuning):** Có quá nhiều tham số cần điều chỉnh để đạt kết quả tối ưu.
2. **Mô hình "Hộp đen" (Black Box):** Khó giải thích trực quan về luồng ra quyết định so với cây quyết định đơn lẻ (như ID3/CART).
3. **Dễ Overfitting nếu không kiểm soát:** Nếu để max\_depth quá lớn hoặc learning\_rate quá cao trên tập dữ liệu nhỏ.

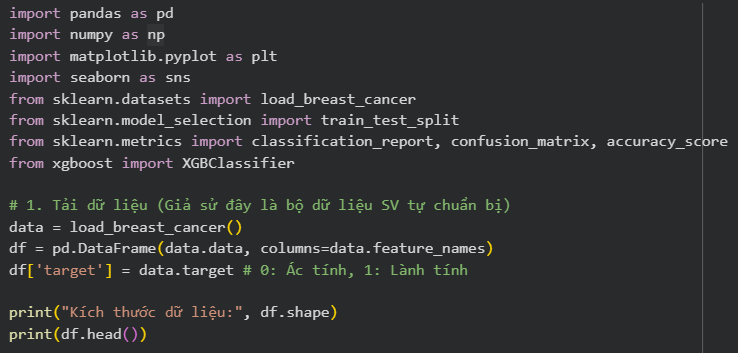
## 4. So sánh với các thuật toán khác trong nhóm

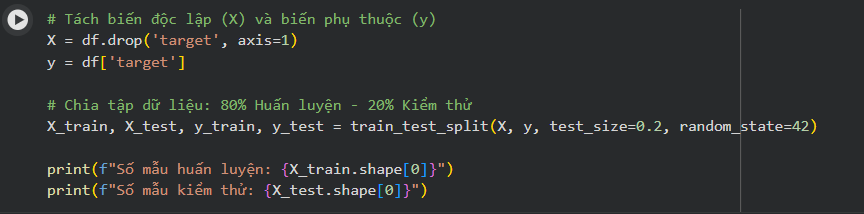
|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí** | **ID3 / CART** | **Gradient Boosting (GB)** | **XGBoost** |
| **Cấu trúc** | Cây đơn lẻ (Single Tree). | Tập hợp cây (Ensemble) tuần tự. | Tập hợp cây tuần tự + Tối ưu hóa hệ thống. |
| **Cơ chế học** | Chia nút dựa trên tin lượng (Entropy/Gini). | Cây sau sửa lỗi cây trước. | Cây sau sửa lỗi + Regularization + Pruning thông minh. |
| **Tốc độ** | Rất nhanh. | Chậm (do tuần tự). | Nhanh hơn GB nhiều lần (nhờ song song hóa). |
| **Độ chính xác** | Thấp/Trung bình. Dễ Overfit. | Cao. | Rất cao. |
| **Xử lý NaN** | Cần tiền xử lý kỹ. | Cần tiền xử lý kỹ. | Tự động xử lý bên trong thuật toán. |

## 5. Quy trình thực hiện (Ví dụ)

1. **Chuẩn bị và Tiền xử lý dữ liệu**

Em sử dụng bộ dữ liệu có sẵn từ thư viện sklearn. Bộ dữ liệu bao gồm 569 mẫu với 30 đặc trưng (features) như bán kính trung bình, độ lồi, diện tích... và cột nhãn target (0: Ác tính, 1: Lành tính).

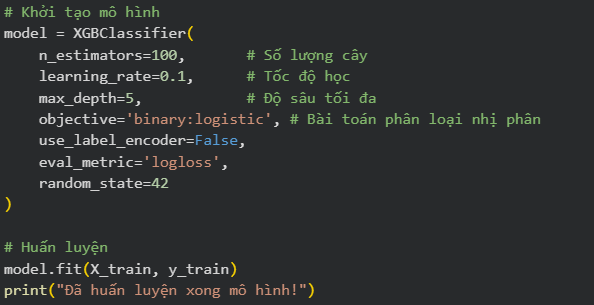




Kết quả: tập huấn luyện có 455 mẫu, tập kiểm thử có 114 mẫu.

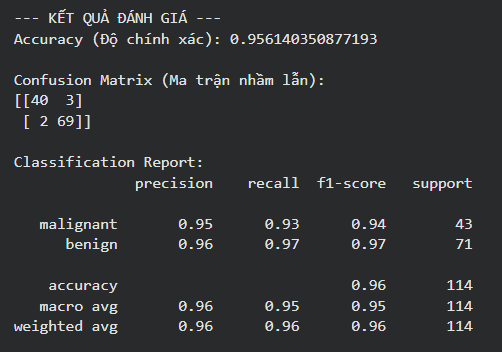
1. **Huấn luyện mô hình:**

Thiết lập mô hình XGBoost với các tham số cơ bản để cân bằng giữa tốc độ và độ chính xác.



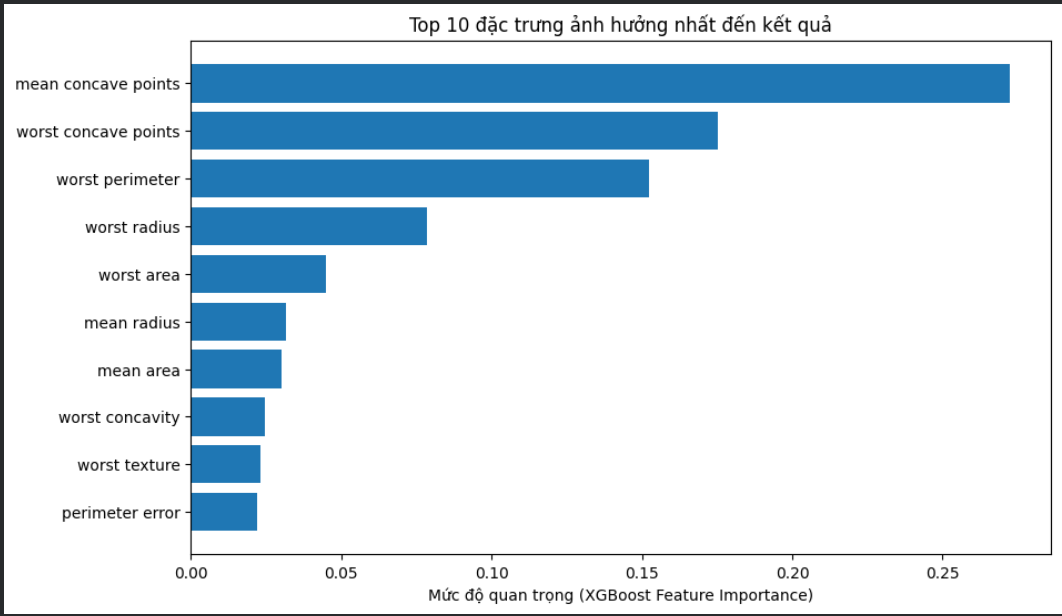
1. **Đánh giá mô hình:**

Sau khi huấn luyện, mô hình được sử dụng để dự đoán trên tập Test (dữ liệu chưa từng gặp).



* **Độ chính xác (Accuracy):** Mô hình đạt khoảng **95.61%**.
* **Classification Report:** Chỉ số F1-score cho cả hai lớp đều rất cao (0.94 và 0.97), cho thấy mô hình hoạt động rất tốt và cân bằng.
* **Ma trận nhầm lẫn (Confusion Matrix):**
  + Dự đoán đúng Ác tính (True Negative): 40 trường hợp.
  + Dự đoán đúng Lành tính (True Positive): 69 trường hợp.
  + Dự đoán sai (nhầm lẫn): Chỉ có 5 trường hợp bị phân loại sai (3 Ác tính nhầm thành Lành tính, 2 Lành tính nhầm thành Ác tính).

1. **Trực quan hóa mức độ quan trọng của đặc trưng:**



XGBoost cung cấp khả năng đánh giá đặc trưng nào ảnh hưởng nhất đến quyết định. Thông qua biểu đồ Feature Importance, chúng ta có thể thấy các yếu tố như "worst concave points", "worst perimeter" hay "mean texture" đóng vai trò quan trọng nhất trong việc xác định loại ung thư.

**Kết luận:** Với XGBoost, dù không cần tiền xử lý dữ liệu quá phức tạp (như chuẩn hóa Scale), mô hình vẫn đạt độ chính xác rất cao và thời gian huấn luyện cực nhanh, chứng minh ưu thế vượt trội so với các cây quyết định đơn lẻ.