# Mô tả bài toán cho cả 4 thuật toán RF, XGBoost, Ada Boost, Gradient Boost

Trong bài toán này, chúng ta sẽ được cho một bộ dataset mô tả thông tin về nhân viên trong một công ty, bao gồm các features liên quan đến nhân viên và mức lương của họ. Nhiệm vụ của chúng ta là phân tích, xử lí bộ data dưới đây và trả lời các câu hỏi yêu cầu người làm phải thực hiện coding.
employee_data.csv

## Thực hiện các yêu cầu sau đây

1. **Đọc dữ liệu**
   Sử dụng pandas, đọc file csv được cung cấp, sau đó hiển thị ra màn hình để hiểu các trường dữ liệu.
2. **Label Encoding**
   Chuyển đổi các cột dữ liệu dạng chữ (cụ thể là cột "Gender" và "Position") sang dạng số bằng cách sử dụng `LabelEncoder` từ thư viện `sklearn`.
3. **Tách dữ liệu thành bộ feature (X) và label (y)**

- Sử dụng các cột "Gender", "Experience (Years)" và "Position" làm features đầu vào (`X`).
- Sử dụng cột "Salary" làm biến đầu ra (`y`).

4. **Tách tập dữ liệu thành tập train và test**

- Chia dữ liệu thành tập huấn luyện (`X_train, y_train`) và tập kiểm tra (`X_test, y_test`) với tỷ lệ `80:20`.
- Đảm bảo rằng việc chia tách dữ liệu là ngẫu nhiên nhưng tái lập (reproducibility) được với `random_state=42`

# Câu hỏi trắc nghiệm yêu cầu coding

Câu 1: Khái niệm "bagging" trong Random Forest có ý nghĩa gì đối với sự đa dạng của các cây trong rừng?

A. Tăng độ sâu của từng cây để tăng đa dạng

B. Sử dụng các tập dữ liệu huấn luyện khác nhau cho mỗi cây để tạo ra sự đa dạng

C. Thay đổi số lượng đặc trưng tại mỗi nút phân chia

D. Áp dụng các thuật toán tối ưu hóa khác nhau cho mỗi cây

Đáp án: B

Câu 2: Hãy sắp xếp lại các bước thực hiện thuật toán Random Foresttheo thứ tự đúng để hoàn thiện quy trình.

1. Tính entropy hoặc Gini để chọn feature tốt nhất.
2. Lặp lại quá trình chọn feature tốt nhất và loại cột đã chọn cho đến khi không còn feature nào.
3. Loại bỏ feature đã được chọn ra khỏi bảng dữ liệu.
4. Xây dựng một cây quyết định cho mỗi bootstrapped dataset.
5. Chọn ngẫu nhiên 2 feature từ dataset.
6. Tạo $N$ bootstrapped dataset từ dữ liệu gốc. Các lựa chọn:

A. $6 \rightarrow 5 \rightarrow 4 \rightarrow 1 \rightarrow 3 \rightarrow 2$

B. $4 \rightarrow 6 \rightarrow 5 \rightarrow 1 \rightarrow 3 \rightarrow 2$

C. $6 \rightarrow 4 \rightarrow 5 \rightarrow 1 \rightarrow 3 \rightarrow 2$

D. $5 \rightarrow 6 \rightarrow 1 \rightarrow 3 \rightarrow 4 \rightarrow 2$

Đáp án đúng:

C. $6 \rightarrow 4 \rightarrow 5 \rightarrow 1 \rightarrow 3 \rightarrow 2$

Câu 3: Random Forest có khả năng xử lý dữ liệu thiếu (missing data) như thế nào?

A. Random Forest Không thể xử lý dữ liệu thiếu

B. Sử dụng trung bình để điền dữ liệu thiếu

C. Tự động bỏ qua các mẫu có dữ liệu thiếu

D. Sử dụng các kỹ thuật nội suy trong quá trình huấn luyện cây

Đáp án: D

Câu 4: So với các phương pháp ensemble khác như Gradient Boosting, Random Forest thường có ưu điểm gì trong việc xử lý dữ liệu lớn và đa dạng?

A. Random Forest khó bị overfitting hơn

B. Random Forest có thời gian huấn luyện nhanh hơn

C. Random Forest có thể song song hóa dễ dàng hơn do các cây độc lập

D. Random Forest đạt được độ chính xác cao hơn trong mọi trường hợp

Đáp án: C

Câu 5: Sử dụng các hàm `mean_squared_error` và `r2_score` từ thư viện `sklearn.metrics`, hãy tính toán giá trị MSE và ( R^2 ) của mô hình Random Forest trên tập kiểm tra (sử dụng các tham số `n_estimators=50` và `random_state=42`). Giá trị MSE và ( R^2 ) là bao nhiêu?

A. **MSE**: 781254527.5, $R^2$: 0.5572

B. **MSE**: 827087272.8, $R^2$: 0.6572

C. **MSE**: 781254527.5, $R^2$: 0.6572

D. **MSE**: 827087272.8, $R^2$: 0.5572

Đáp án: D

Câu 6: Hãy thử nghiệm số lượng cây của mô hình Random Forest, với các giá trị `n_estimators` khác nhau (10, 20, 50, 100), số lượng cây nào đem lại MSE (Mean Squared Error) nhỏ nhất? ( `random_state=42` )

A. 10 cây

B. 20 cây

C. 50 cây

D. 100 cây

Đáp án đúng: B

Câu 7: Hãy thử nghiệm các giá trị độ sâu của cây `max_depth` từ 1 đến 10, độ sâu nào mang lại MSE (Mean Squared Error) nhỏ nhất? ( `random_state=42` )

A. Độ sâu 1

B. Độ sâu 3

C. Độ sâu 5

D. Độ sâu 10

Đáp án đúng: C