Random Forest algorithm

Các bước chính:

1. **Lặp lại B lần** (thường B ≈ 100):
    - Tạo **tập huấn luyện mới** bằng **sampling with replacement** từ tập gốc có M ví dụ.
    - Huấn luyện một cây quyết định trên tập mới này.
2. Sau khi có **B cây**, để dự đoán:
    - Cho mỗi cây **vote** trên đầu vào mới.
    - Lấy **đa số phiếu** làm kết quả cuối cùng (cho bài toán phân loại).
    - Hoặc **trung bình dự đoán** (cho bài toán hồi quy).

Bagged Trees vs. Random Forest:

- Nếu chỉ dừng ở việc dùng `sampling with replacement` để tạo B cây → gọi là **bagged decision trees**.
- **Random forest** cải tiến thêm bằng cách **giảm số lượng đặc trưng được xem xét tại mỗi nút**.

Khác biệt chính của Random Forest:

- Tại mỗi nút của cây:
    - **Chọn ngẫu nhiên K đặc trưng (K < N)** từ tất cả N đặc trưng.
    - Chỉ xem xét các đặc trưng đó để chọn phép chia có **information gain** cao nhất.
- Với N lớn, thường chọn K = √N.

Điều này giúp:

- Tạo ra **các cây càng khác biệt nhau hơn**.
- Khi **bình chọn**, kết quả tổng hợp sẽ **ít bị lệch** do 1 cây riêng lẻ → **mô hình tổng thể trở nên ổn định, chính xác hơn.**