Tree Ensemble

Vấn đề của một cây quyết định duy nhất

- **Nhạy cảm cao với dữ liệu**:

  Chỉ cần **thay đổi một mẫu dữ liệu nhỏ**, cây quyết định có thể:

  - Chọn **feature khác** để chia ở nút gốc.
  - Tạo ra **cấu trúc cây hoàn toàn khác**.

→ Dẫn đến **kết quả không ổn định**, dễ **overfit** dữ liệu huấn luyện.

Giải pháp: **Tree Ensemble (Tập hợp nhiều cây)**

- **Ý tưởng chính**: Thay vì xây 1 cây → **xây nhiều cây**.
- Mỗi cây là một "giả thiết hợp lý" để phân loại (cat / not cat).
- Khi dự đoán:
  - Chạy **tất cả các cây** trên ví dụ cần dự đoán.
  - **Lấy biểu quyết đa số (majority vote)** để đưa ra dự đoán cuối cùng.

Sampling with replacement

- **Sampling with replacement** nghĩa là:
  - **Lấy một mẫu (ví dụ huấn luyện)** từ tập dữ liệu.
  - **Sau đó bỏ lại mẫu đó** vào trong tập (để có thể được chọn lại).
- Do đó: **Một mẫu có thể xuất hiện nhiều lần**, trong khi một số khác **không xuất hiện** trong tập mẫu mới.
- Mục đích là tạo ra nhiều cây quyết định khác nhau bằng cách:
  - **Huấn luyện mỗi cây** trên **một tập mẫu có hoàn lại** khác nhau.
  - Sau đó cho các cây **biểu quyết** để đưa ra dự đoán chính xác và ổn định hơn.