XGBoost

Vấn đề với cây quyết định đơn lẻ hoặc bagged trees

- Các cây riêng lẻ (hoặc bagged trees như trong random forest) đều **đối xử công bằng với mọi ví dụ huấn luyện**, nên:
  - Có thể **tốn nhiều công** học đi học lại cả tập dữ liệu, **kể cả những phần đã học tốt**.
  - Không tập trung vào những điểm mà mô hình **đang làm chưa tốt**.

Ý tưởng chính của Boosting: Học tập có trọng tâm

- Lấy cảm hứng từ **"deliberate practice"** trong học tập (luyện tập tập trung vào phần chưa giỏi).
- Khi xây dựng cây thứ 2, 3,...:
  - Tăng **xác suất chọn** các ví dụ mà cây trước **dự đoán sai**.
  - Từ đó, các cây sau **học tập trung vào lỗi của cây trước**.

XGBoost – Extreme Gradient Boosting

- Một biến thể **cực kỳ tối ưu** của boosting:
  - **Không dùng sampling với replacement**.
  - **Gán trọng số (weights)** cho từng ví dụ huấn luyện.
  - Tăng trọng số cho ví dụ bị sai → trọng số hướng dẫn cây tiếp theo tập trung học phần đó.
- Có nhiều ưu điểm:
  - Rất **nhanh và hiệu quả**.
  - **Tự động hóa** các quy trình chia, dừng, và **regularization (chống overfitting)**.
  - Được dùng rộng rãi trong các **cuộc thi như Kaggle** và nhiều ứng dụng thương mại.