KIỂM TRA THỰC HÀNH

Dữ liệu: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

- *gender – giới tính*
- *race/ethnicity – nhóm chủng tộc/xã hội*
- *parental level of education – trình độ học vấn của cha mẹ*
- *lunch – loại suất ăn (chuẩn hay miễn phí)*
- *test preparation course – có tham gia khóa luyện thi không*
- *math score, reading score, writing score – điểm các bài kiểm tra*
- *performance – nhãn đầu ra: "low", "medium", "high"*

Câu 1 – Tiền xử lý dữ liệu (2 điểm)

a) Đọc dữ liệu, kiểm tra thông tin cơ bản (số dòng, kiểu dữ liệu, giá trị thiếu).
b) Mã hóa (encoding) các biến phân loại và chuẩn hóa các cột điểm số.
c) Chia tập dữ liệu thành tập huấn luyện (80%) và kiểm tra (20%).

Câu 2 – Huấn luyện mô hình (2 điểm)

a) Huấn luyện một mô hình phân loại RandomForestClassifier để dự đoán performance.
b) In ra ma trận nhầm lẫn (confusion matrix) và classification report.

Câu 3 – Đánh giá mô hình (3 điểm)

a) Trình bày các độ đo: Accuracy, Precision, Recall, F1-score từ classification report.
b) Vẽ biểu đồ confusion matrix bằng heatmap.
c) Nếu mô hình có độ chính xác cao nhưng F1 thấp, phân tích nguyên nhân.

Câu 4 – Cải thiện mô hình (2 điểm)

a) Thử nghiệm mô hình khác (ví dụ: XGBoostClassifier hoặc LogisticRegression) và so sánh độ đo.
b) Trình bày mô hình nào tốt hơn và vì sao, dựa trên F1-score tổng quát hoặc macro.

Câu 5 – Trực quan hóa và phân tích thêm (1 điểm)

a) Vẽ biểu đồ hộp (boxplot) so sánh phân phối điểm math score giữa các nhóm performance.
b) Trình bày nhận xét về mối quan hệ giữa điểm số và nhãn dự đoán.

✅ **Yêu cầu nộp bài**

Sinh viên nộp file .ipynb kèm theo bản .html hoặc .pdf xuất từ notebook có chạy toàn bộ kết quả.