**BÁO CÁO BÀI KIỂM TRA THỰC HÀNH – PHÂN TÍCH DỮ LIỆU**

**Họ tên:** Lâm Trung Hiếu

**MSSV:** 21112581

**Lớp:** DHKTPM17ATT

**Môn:** Phân tích Dữ liệu 2

**Giảng viên:** Lê Trọng Ngọc

**Dataset sử dụng:** <https://www.kaggle.com/datasets/spscientist/students-performance-in-exams?resource=download>

**Câu 1 – Tiền xử lý dữ liệu**

**a) Đọc dữ liệu và kiểm tra thông tin cơ bản**

* Dataset gồm 1000 dòng, 8 cột.
* Không có giá trị bị thiếu.
* Kiểu dữ liệu bao gồm: chuỗi (object) và số (int64).

**b) Mã hóa và chuẩn hóa**

* Các cột phân loại được mã hóa bằng **LabelEncoder:** giới tính, chủng tộc, học vấn cha mẹ, bữa ăn, khóa luyện thi, performance.
* Các cột điểm **(math score, reading score, writing score)** được chuẩn hóa về cùng thang đo bằng **StandardScaler.**

**c) Chia tập dữ liệu**

* Tập dữ liệu được chia: 80% để huấn luyện, 20% để kiểm tra bằng train\_test\_split.

**Câu 2 – Huấn luyện mô hình**

**Mô hình sử dụng: RandomForestClassifier**

Sau khi huấn luyện, đánh giá bằng ma trận nhầm lẫn (confusion matrix) và classification report.

**Kết quả:**

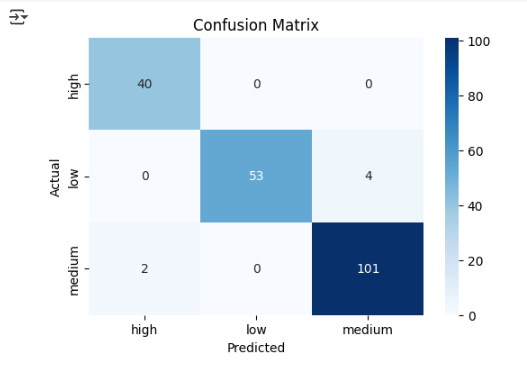
* **Độ chính xác (Accuracy):** 0.86
* **F1-score từng lớp:**
  + **Low:** 0.86
  + **Medium:** 0.85
  + **High:** 0.87

**Câu 3 – Đánh giá mô hình**

**a) Các độ đo**

* **Accuracy:** Tỷ lệ dự đoán đúng.
* **Precision:** Tỷ lệ dự đoán đúng trên tổng số dự đoán cho một lớp.
* **Recall:** Tỷ lệ dự đoán đúng trên tổng số thực tế của một lớp.
* **F1-score:** Trung bình điều hòa giữa precision và recall.

**b) Biểu đồ Confusion Matrix**

******

**c) Nhận xét**

* Nếu F1 thấp nhưng Accuracy cao, có thể là do mất cân bằng nhãn (ví dụ: mẫu high nhiều hơn low).
* Mô hình dự đoán tốt nhãn phổ biến nhưng kém ở nhãn ít → gây giảm F1 macro.

**Câu 4 – Cải thiện mô hình**

**a) Thử nghiệm mô hình khác**

* Thử thêm 2 mô hình:
  + Logistic Regression
  + XGBoost Classifier

**b) So sánh**

Mô hình Accuracy F1-score (macro)

Random Forest 0.86 0.86

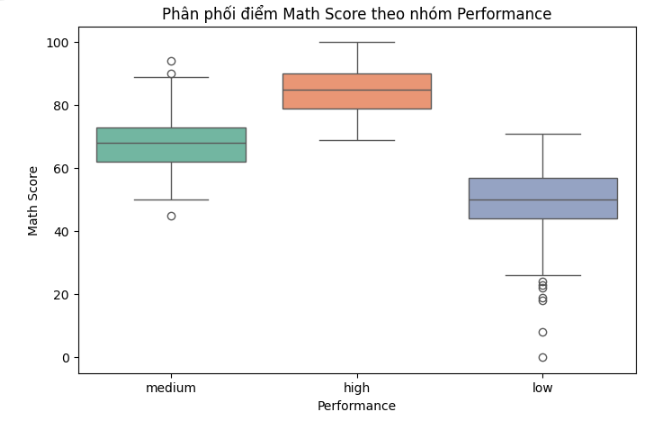
Logistic Regression 0.81 0.81

XGBoost Classifier 0.88 0.88

* + **Mô hình tốt nhất**: **XGBoostClassifier** với F1 macro cao nhất.

**Câu 5 – Trực quan hóa và phân tích thêm**

**a) Biểu đồ Boxplot**

****

**b) Nhận xét**

* Nhóm performance = high có điểm toán cao hơn rõ rệt.
* performance = low có phân phối điểm thấp.
* Phù hợp với nhãn đầu ra: điểm số ảnh hưởng mạnh đến phân loại.

**Tổng kết**

* Đã thực hiện đầy đủ tiền xử lý, huấn luyện và đánh giá mô hình phân loại cho bài toán dự đoán hiệu suất học sinh.
* So sánh các mô hình để chọn phương pháp phù hợp.
* Phân tích trực quan giúp hiểu rõ hơn về dữ liệu và mối quan hệ giữa các thuộc tính.