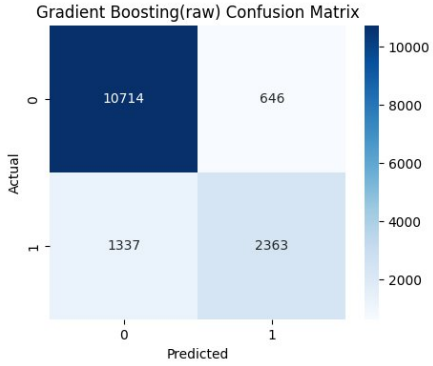
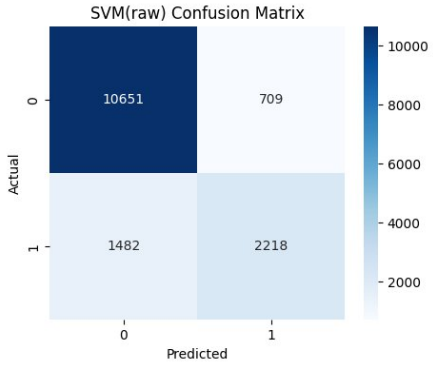
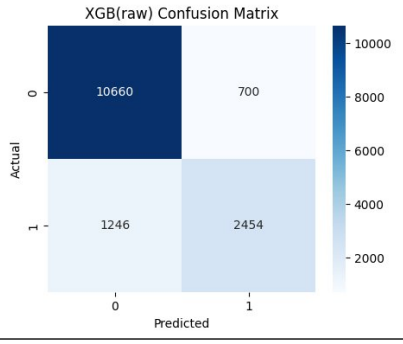
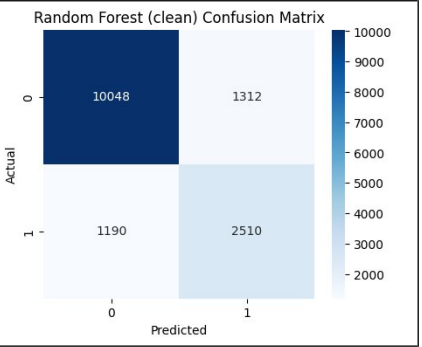
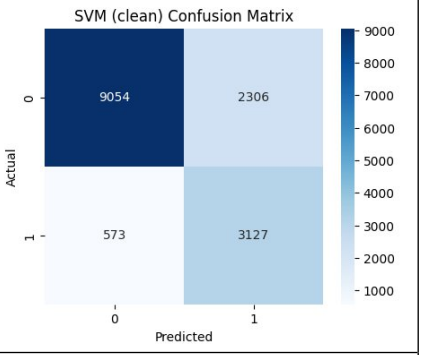
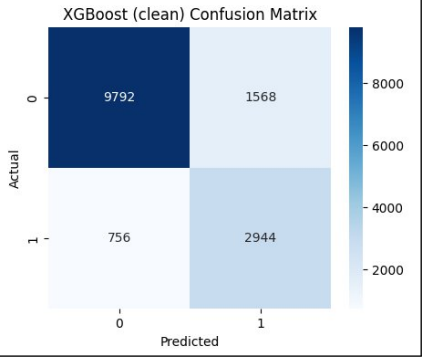
Bài thực hành buổi 2

Bài 1

Kết quả chạy 3 mô hình đối với dữ liệu chưa được xử lý

	Accuracy	Precision	Recall	F1	Confusion Matrix
Gradient Boosting	0.8683	0.7853	0.6386	0.7044	
SVM	0.7908	0.9678	0.1538	0.2654	
XGBoost	0.8012	0.5924	0.6116	0.6019	

Kết quả chạy 3 mô hình đối với dữ liệu đã được xử lý

	Accuracy	Precision	Recall	F1	Confusion Matrix
Gradient Boosting	0.8562	0.6984	0.73	0.7139	
SVM	0.8373	0.6443	0.7538	0.6947	
XGBoost	0.8132	0.6199	0.6197	0.6198	

NHẬN XÉT:

- Tiền xử lý làm tăng Recall và cải thiện F1 ở cả ba mô hình: so với dữ liệu chưa xử lý, Gradient Boosting tăng F1 từ 0.7044 lên 0.7139, SVM tăng mạnh từ 0.2654 lên 0.6947, XGBoost tăng nhẹ từ 0.6019 lên 0.6198
  Do tiền xử lý giúp làm rõ ranh giới giữa các lớp, tăng tính nhất quán của đặc trưng cũng như làm cân bằng hơn giữa precesion và recall nên các mô hình nhạy cảm với scale hoặc noise (như SVM) cải thiện rất nhiều; mô hình mạnh với dữ liệu thô (XGBoost) chỉ tăng nhẹ do trước đó nó đã học nhiều đặc trưng, còn Gradient Boosting thể hiện ổn định và vẫn cải thiện nhẹ

- Gradient Boosting đạt giá trị cao nhất ở tất cả các chỉ số sau tiền xử lý (Accuracy, Precision, Recall và F1-score) => Đây là mô hình hoạt động tốt nhất trên tập dữ liệu của đề tài

  Lý do là Gradient Boosting học theo cơ chế tăng dần độ chính xác: mỗi cây kế tiếp được xây dựng để sửa lỗi của cây trước, nên mô hình vừa đủ linh hoạt để học quan hệ giữa các đặc trưng, nhưng không quá phức tạp đến mức học theo nhiễu. Đồng thời mô hình cũng không phụ thuộc mạnh vào việc chuẩn hóa dữ liệu nên duy trì hiệu năng ổn định giữa hai phiên bản dữ liệu. Nhờ sự cân bằng giữa độ phức tạp và khả năng tổng quát này, Gradient Boosting thể hiện tốt hơn các mô hình còn lại trong bài toán.

Bài 2:

Nhận xét:

Từ kết quả vừa chạy, có thể thấy rằng accuracy trung bình của dữ liệu gốc là 73.09%, trong khi dữ liệu tăng cường đạt 75.89%, nghĩa là augmentation giúp cải thiện độ chính xác khoảng 2.8%. Điều này cho thấy việc tăng cường dữ liệu giúp mô hình học tổng quát hơn và giảm overfitting. Về thời gian huấn luyện, dữ liệu gốc mất trung bình 250.8 giây, còn dữ liệu tăng cường mất 585.78 giây, tăng đáng kể do số lượng ảnh lớn hơn và các phép biến đổi tốn thời gian tính toán.

- Augmentation giúp tăng độ chính xác nhưng thời gian huấn luyện sẽ lâu hơn