Tên: Lương Nguyễn Minh Chánh

Lớp: MSEHCM20

Môn: MLOps

**Dự án MLOps\_MSE.HCM\_SP25**

**Mô tả các bước thực hiện**

**1. Tạo dữ liệu mẫu**

* Sử dụng hàm make\_classification trong sklearn để sinh dữ liệu mẫu cho bài toán phân loại.
* Dataset gồm 2000 mẫu, mỗi mẫu có 10 đặc trưng (features).
* Thêm nhiễu (noise) vào dữ liệu để mô phỏng tính chất dữ liệu thực.
* Áp dụng StandardScaler để chuẩn hóa dữ liệu đầu vào.

**2. Chia tập dữ liệu**

Sử dụng train\_test\_split để chia dữ liệu thành 2 phần:

* Tập train: 70% dữ liệu
* Tập test: 30% dữ liệu

**3. Xây dựng và training các mô hình phân loại**

* Các mô hình được sử dụng:
  + Logistic Regression
  + Logistic Regression (GridSearchCV tuning)
  + Random Forest
  + XGBoost
  + LightGBM
  + CatBoost
* Mỗi mô hình được training trên tập train và đánh giá trên tập test.
* Accuracy của từng mô hình được ghi lại vào MLflow.

**4. Lưu mô hình tốt nhất**

* Mô hình có độ chính xác cao nhất là CatBoost.
* Model CatBoost được lưu lại thành file best\_model.pkl sử dụng joblib.dump().

**5. Tracking và quản lý với MLflow**

* Thiết lập MLflow Experiment để tracking mỗi lần train model.
* Ghi log các tham số, độ chính xác, và mô hình lên MLflow server.
* Đăng ký mô hình tốt nhất vào MLflow Model Registry với tên Best\_Model.
* Promote mô hình thành Production stage.

**6. Tạo ứng dụng Flask predict**

* Tạo file app.py sử dụng Flask framework.
* Khi server chạy:
  + Load best\_model.pkl từ local.
  + Tạo API /predict nhận dữ liệu JSON, trả về kết quả phân loại.
* Test API bằng Postman và file test\_api.py tự động gửi request predict.

**7. Tạo ứng dụng web giao diện**

Tạo thêm file app\_web.py:

* Web form nhập 10 feature từ tay.
* Submit form sẽ được predict và hiển thị kết quả trên giao diện web.

1. **Nhận xét về hiệu suất các mô hình:**

* Logistic Regression đạt accuracy ~70%, thể hiện vai trò là baseline đơn giản.
* Random Forest nâng độ chính xác lên ~85% so với Logistic Regression.
* XGBoost đạt accuracy cao hơn một chút (~86%), cho thấy khả năng tối ưu tốt hơn trên dữ liệu có nhiễu.
* LightGBM cũng cho kết quả rất cao (~85.8%), gần tương đương XGBoost.
* CatBoost có độ chính xác cao nhất (~88.1%), vượt trội so với các mô hình khác.

=> CatBoost có lợi thế tự động xử lý categorical data và regularization tốt nên đạt hiệu suất ổn định nhất.

