VÕ HỒNG VIỆT - 22725461

Mô hình **Logistic Regression** cho kết quả tốt nhất với **99% độ chính xác**.Sử dụng predict\_proba() để dự đoán xác suất từng lớp thay vì nhãn cứng.So sánh giữa các mô hình thì ta thấy SVM và Logistic Regression đạt kết quả tốt nhất. Naive Bayes và k-NN kém hiệu quả hơn.

**Ưu điểm**:

Xây dựng pipeline khoa học, có đầy đủ bước tiền xử lý, chọn mô hình, đánh giá.

Sử dụng cross-validation và GridSearchCV giúp mô hình tổng quát hơn.

Dữ liệu được chuẩn hóa giúp mô hình ML hoạt động ổn định.

**Hạn chế**:

Chưa thử nghiệm các mô hình deep learning như CNN, có thể cải thiện kết quả.

Chưa có kiểm tra kỹ về các đặc trưng quan trọng nhất trong ba nhóm (shape, margin, texture).

Có thể dùng PCA hoặc feature selection để giảm số chiều.

VÕ HỒNG VIÊT - 22725461

* **Mục tiêu chính của quy trình:**
  + **Overfit tập huấn luyện:** Đảm bảo mô hình học tốt dữ liệu huấn luyện, tức là đạt được hiệu suất cơ bản (baseline) để xác định khả năng học của mô hình.
  + **Giảm overfitting:** Tăng khả năng tổng quát hóa của mô hình, giúp cải thiện hiệu suất trên tập kiểm định (validation set) và tránh việc mô hình chỉ "học vẹt" dữ liệu huấn luyện.
  + **Tinh chỉnh siêu tham số:** Điều chỉnh các siêu tham số (hyperparameters) như learning rate, batch size, hoặc kiến trúc mô hình để đạt hiệu suất tối ưu trên cả tập huấn luyện và kiểm định.
* **Quy trình thực hiện gồm 3 bước cụ thể:**
  + **Overfit tập huấn luyện:** Huấn luyện mô hình trên tập dữ liệu huấn luyện cho đến khi đạt độ chính xác cao, thậm chí có thể overfit (học quá kỹ dữ liệu huấn luyện). Điều này giúp kiểm tra khả năng học tối đa của mô hình.
  + **Giảm overfitting:** Áp dụng các kỹ thuật như regularization (L2, dropout), tăng cường dữ liệu (data augmentation), hoặc điều chỉnh kiến trúc mô hình để giảm hiện tượng overfit, từ đó cải thiện hiệu suất trên tập kiểm định và dữ liệu mới.
  + **Tinh chỉnh siêu tham số:** Sử dụng các phương pháp như grid search, random search hoặc tối ưu hóa Bayesian để tìm ra bộ siêu tham số tốt nhất, giúp mô hình đạt hiệu suất tối ưu và ổn định trên các tập dữ liệu khác nhau.