Rất tốt! Dưới đây là phần **giải thích chi tiết chức năng & bản chất của các thư viện** xuất hiện trong đoạn mã bạn cung cấp:

## **📚 1. Thư viện cốt lõi**

| **Thư viện** | **Chức năng** | **Bản chất** |
| --- | --- | --- |
| kagglehub | Tải datasets từ Kaggle | Công cụ hỗ trợ truy cập nhanh bộ dữ liệu trực tiếp từ Kaggle thông qua API |
| pandas (pd) | Xử lý dữ liệu dạng bảng (DataFrame) | Dùng cho thao tác lọc, gộp, xử lý dữ liệu hàng/cột |
| numpy (np) | Tính toán số học, mảng | Cung cấp các phép toán vector/matrix cực nhanh |
| matplotlib.pyplot (plt) | Vẽ biểu đồ đơn giản | Công cụ trực quan hóa dữ liệu kiểu biểu đồ đường, cột, scatter... |
| seaborn (sns) | Vẽ biểu đồ thống kê đẹp mắt | Thêm lớp biểu diễn thống kê lên matplotlib (heatmap, boxplot, v.v.) |

## **🤖 2. Thư viện Học máy (scikit-learn)**

| **Import** | **Chức năng** | **Bản chất/Khi dùng** |
| --- | --- | --- |

### **🎯 Tiền xử lý & chọn đặc trưng**

| StandardScaler | Chuẩn hóa dữ liệu (trung bình 0, std = 1) | Dùng trước khi huấn luyện mô hình nhạy cảm với scale |  
 | VarianceThreshold | Loại bỏ đặc trưng có phương sai thấp | Loại bỏ feature gần như không thay đổi, không có thông tin |  
 | SelectKBest, f\_classif, mutual\_info\_classif | Chọn K đặc trưng tốt nhất | Dựa vào thống kê (F-test) hoặc mức độ phụ thuộc (MI) |

### **🧪 Chia dữ liệu & tạo mẫu**

| train\_test\_split | Chia tập train và test | Dùng để đánh giá mô hình đúng cách |  
 | resample | Tái tạo lại dữ liệu (upsample/undersample) | Hữu ích khi cần cân bằng dữ liệu giữa các lớp |

### **🧠 Mô hình học máy**

| **Mô hình** | **Chức năng** | **Khi nào dùng?** |
| --- | --- | --- |
| RandomForestClassifier | Mô hình cây ngẫu nhiên | Mạnh, chống overfitting tốt |
| DecisionTreeClassifier | Mô hình cây quyết định | Hiểu dễ, nhưng dễ overfit |
| LogisticRegression | Hồi quy nhị phân | Bài toán 0/1, baseline tốt |
| KNeighborsClassifier | KNN | Dựa trên khoảng cách, dễ hiểu |
| GaussianNB | Naive Bayes phân phối chuẩn | Nhanh, nhẹ, phù hợp với dữ liệu độc lập |

### **📏 Đánh giá mô hình**

| **Import** | **Chức năng** | **Ý nghĩa** |
| --- | --- | --- |
| accuracy\_score, precision\_score, recall\_score, f1\_score | Các chỉ số đánh giá mô hình | So sánh giữa dự đoán & nhãn thật |
| confusion\_matrix, classification\_report | Ma trận lỗi & báo cáo tổng quan | Hiển thị đầy đủ TP, FP, FN, TN theo lớp |

### **✅ Tóm lại:**

* Bạn đang sử dụng **pipeline khá hoàn chỉnh** cho một bài toán phân loại: từ tiền xử lý → chọn đặc trưng → mô hình → đánh giá.
* Tất cả thư viện học máy chính đều đến từ **scikit-learn**, một thư viện phổ biến, chuẩn mực trong machine learning Python.

Bạn có muốn mình giúp viết **pipeline xử lý đầy đủ (bằng Python)** từ những thư viện trên không?