**CHƯƠNG 2: PHÂN TÍCH NAÏVE BAYES CLASSIFIER**

**2.1. Phân tích bài toán**

**Input:**

* File .csv chứa dữ liệu huấn luyện cho bài toán phân lớp.
* Dữ liệu gồm:
  + Các cột điều kiện: đặc trưng (attributes)
  + Cột cuối: là nhãn lớp (class label)
* Ví dụ: Tập dữ liệu chọn laptop theo tuổi, nghề nghiệp, mục đích sử dụng…

**Output:**

* Dự đoán **lớp (nhãn)** cho một mẫu mới dựa trên thuật toán Naïve Bayes.
* Tính xác suất **P(C|X)** với từng lớp và chọn lớp có xác suất cao nhất.

**🔹 2.2. Các giải thuật**

**2.2.1. Huấn luyện mô hình Naïve Bayes**

1. **Tính xác suất tiên nghiệm cho mỗi lớp**:
2. **Tính xác suất có điều kiện cho từng thuộc tính theo từng lớp**:

**2.2.2. Dự đoán lớp cho một mẫu mới**

1. **Tính xác suất hậu nghiệm cho mỗi lớp** (áp dụng định lý Bayes):
2. **Chọn lớp có xác suất cao nhất**:

**2.3. Tạo GUI và chạy thử chương trình**

**2.3.1. Giao diện chương trình**

* Cho phép:
  + Chọn file .csv dữ liệu huấn luyện.
  + Nhập dữ liệu của mẫu cần phân lớp.
  + Nút “Chạy dự đoán” để tính toán.

**2.3.2. Thử nghiệm chương trình**

* **Dataset 1**: Laptop Choice
  + Gồm các cột: tuổi, nghề nghiệp, mục đích → nhãn: laptop đã chọn
  + Mẫu đầu vào: Tuổi = “18–22”, Nghề = “Sinh viên”, Mục đích = “Học tập”
  + Dự đoán: Samsung
* **Dataset 2**: Acute Inflammation (UCI)
  + 6 thuộc tính, nhãn phân loại bệnh → chạy cho kết quả chính xác
* **Dataset 3**: Credit Approval (UCI)
  + 15 thuộc tính hỗn hợp, có dữ liệu nhiễu → độ chính xác thấp hơn

**🔹 2.4. Nhận xét**

* **Dữ liệu sạch, chuẩn hóa tốt** → mô hình Naïve Bayes cho kết quả rất tốt.
* **Dữ liệu nhiễu hoặc mất mát** → cần xử lý trước khi áp dụng thuật toán.
* Phù hợp với các bài toán có dữ liệu rời rạc (categorical), ít thuộc tính liên tục.

**🔧 Gợi ý mở rộng:**

* Cho phép chọn thuộc tính cần xét (tự động hoặc thủ công)
* Thêm chức năng xử lý dữ liệu thiếu (missing values)
* Hỗ trợ lưu và tải mô hình huấn luyện để sử dụng lại