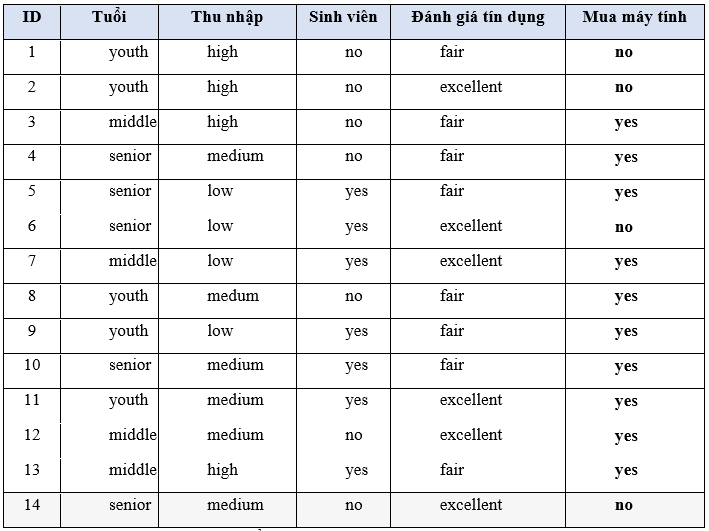
**Buổi thực hành 4: Naïve Bayes**

**Bài 1**: Cho cơ sở dữ liệu khách hàng đã thu thập được như sau:



Dùng giải thuật Naïve Bayes để đưa ra dự đoán việc các khách hàng mới có mua máy tính hay không dựa trên các thuộc tính như sau:

Khách hàng:

S1 : (age = youth, income = medium, student = yes, credit\_rating = fair)

S2 : (age = middle, income = high, student = yes, credit\_rating = fair)

S3 : (age = youth, income = low, student = no, credit\_rating = excellent)

**Bài làm**

* S1 : (age = youth, income = medium, student = yes, credit\_rating = fair)

Ta có: P(Cyes) = 10/14 = 0.714, P(Cno) = 4/14 = 0.285

Các xác suất thành phần:

P(Age = youth | Cyes) = 3/10 = 0.3

P(Age = youth | Cno) = 2/4 = 0.5

P(income = medium | Cyes) = 5/10 = 0.5

P(income = medium | Cno) = 1/4 = 0.25

P(student = yes | Cyes) = 6/10 = 0.6

P(student = yes | Cno) = 1/4 = 0.25

P(credit\_rating = fair | Cyes) = 7/10 = 0.7

P(credit\_rating = fair | Cno) = 1/4 = 0.25

Vậy ta có:

P(X|Cyes) = 0.3 \* 0.5 \* 0.6 \* 0.7 = 0.063

P(X|Cno) = 0.5 \* 0.25 \* 0.25 \* 0.25 = 0.0078

P(X|Cyes) \* P(Cyes) = 0.063 \* 0.714 = 0.045

P(X|Cno) \* P(Cno) = 0.0078125 \* 0.285 = 0.0022

🡺 Từ kết quả trên ta thấy P(X|Cyes) \* P(Cyes) có giá trị lớn nhất, do đó thuật toán Naïve Bayes sẽ kết luận rằng khách hàng X sẽ mua máy tính.

* S2 : (age = middle, income = high, student = yes, credit\_rating = fair)

Ta có: P(Cyes) = 10/14 = 0.714, P(Cno) = 4/14 = 0.285

Các xác suất thành phần:

P(Age = middle | Cyes ) = 4/10 = 0.4

P(Age = middle | Cno ) = 0/4 = 0

P(income = high | Cyes ) = 2/10 = 0.2

P(income = high | Cno) = 2/4 = 0.5

P(student = yes | Cyes ) = 6/10 = 0.6

P(student = yes | Cno) = 1/4 = 0.25

P(credit\_rating = fair | Cyes) = 7/10 = 0.7

P(credit\_rating = fair | Cno) = 1/4 = 0.25

Vậy ta có:

P(X|Cyes) = 0.4 \* 0.2 \* 0.6 \* 0.7 = 0.0336

P(X|Cno) = 0 \* 0.5 \* 0.25 \* 0.25 = 0

P(X|Cyes) \* P(Cyes) = 0.0336 \* 0.714 = 0.024

P(X|Cno) \* P(Cno) = 0 \* 0.285 = 0

🡺 Từ kết quả trên ta thấy P(X|Cyes) \* P(Cyes) có giá trị lớn nhất, do đó thuật toán Naïve Bayes sẽ kết luận rằng khách hàng X sẽ mua máy tính.

* S3 : (age = youth, income = low, student = no, credit\_rating = excellent)

Ta có: P(Cyes) = 10/14 = 0.714, P(Cno) = 4/14 = 0.285

Các xác suất thành phần:

P(Age = youth | Cyes) = 3/10 = 0.3

P(Age = youth | Cno) = 2/4 = 0.5

P(income = low | Cyes) = 3/10 = 0.3

P(income = low | Cno) = 1/4 = 0.25

P(student = no | Cyes) = 4/10 = 0.4

P(student = no | Cno) = 3/4 = 0.75

P(credit\_rating = excellent | Cyes) = 3/10 = 0.3

P(credit\_rating = excellent | Cno) = 3/4 = 0.75

Vậy ta có:

P(X|Cyes) = 0.3 \* 0.3 \* 0.4 \* 0.3 = 0.0108

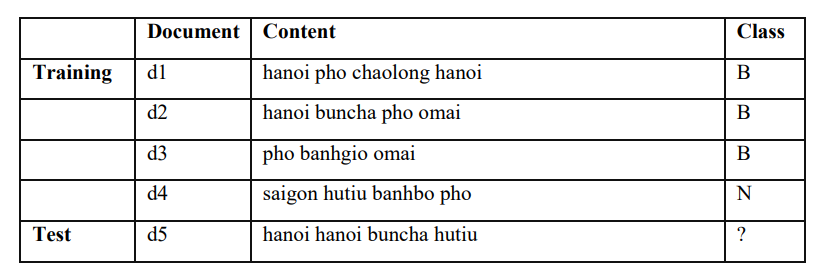
P(X|Cno) = 0.5 \* 0.25 \* 0.75 \* 0.75 = 0.0703

P(X|Cyes) \* P(Cyes) = 0.0108 \* 0.714 = 0.0077

P(X|Cno) \* P(Cno) = 0.0703 \* 0.285 = 0.020

🡺 Từ kết quả trên ta thấy P(X|Cno) \* P(Cno) có giá trị lớn nhất, do đó thuật toán Naïve Bayes sẽ kết luận rằng khách hàng X sẽ không mua máy tính.

**Bài 2:** Giả sử trong tập training có các văn bản d1, d2, d3, d4 như trong bảng dưới đây. Mỗi văn bản này thuộc vào 1 trong 2 classes: B (Bắc) hoặc N (Nam). Hãy xác định class của văn bản d5



Ta có thể dự đoán rằng d5 thuộc class Bắc.

Bài toán này có thể được giải quyết bởi hai mô hình: Multinomial Naive Bayes và Bernoulli Naive Bayes. Việc mô hình nào tốt hơn phụ thuộc vào mỗi bài toán. Chúng ta có thể thử cả hai để chọn ra mô hình tốt hơn.

Nhận thấy rằng ở đây có 2 class B và N, ta cần đi tìm p(B) và p(N) và dựa trên tần số xuất hiện của mỗi class trong tập training. Ta sẽ có:

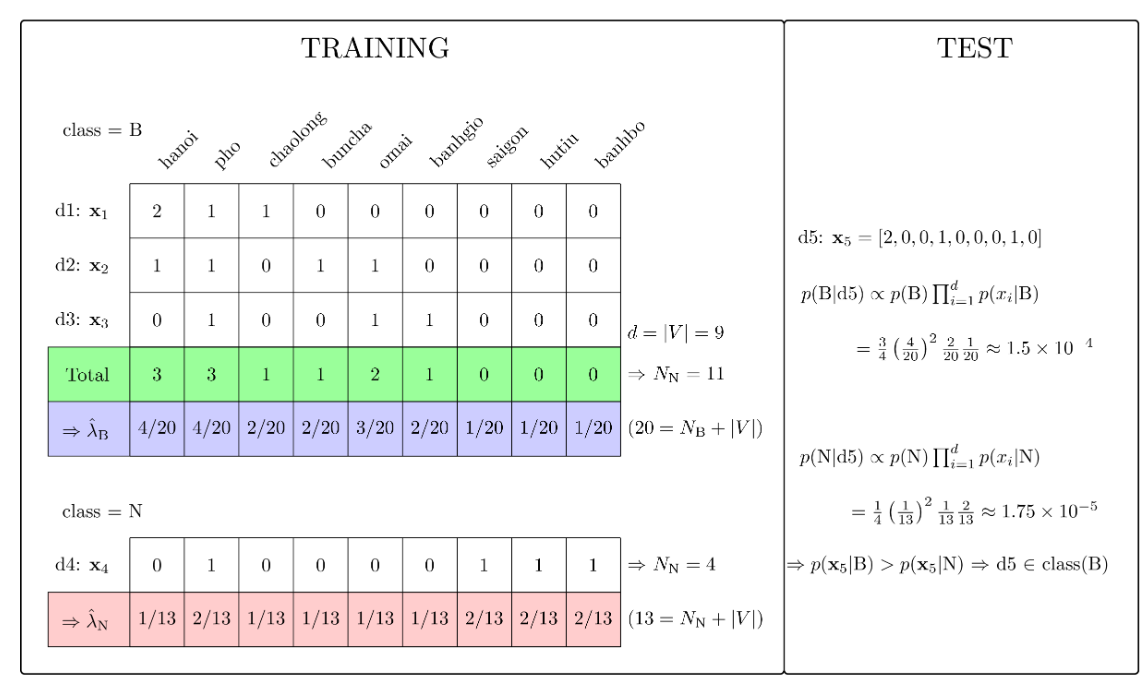
p(B)=3/4,     p(N)=1/4      (8)

Tập hợp toàn bộ các từ trong văn bản, hay còn gọi là từ điển, là:

V={hanoi, pho, chaolong, buncha, omai, banhgio, saigon, hutiu, banhbo}

Tổng cộng số phần tử trong từ điển là |V|=9|

Hình dưới đây minh hoạ quá trình Training và Test cho bài toán này khi sử dụng Multinomial Naive Bayes, trong đó có sử dụng Laplace smoothing với α=1



Chú ý, hai giá trị tìm được 1.5×10^−4 và 1.75×10^−5 không phải là hai xác suất cần tìm mà chỉ là hai đại lượng **tỉ lệ thuận** với hai xác suất đó. Để tính cụ thể, ta có thể làm như sau:

p(B|d5) = (1.5×10^−4) / (1.5×10^−4+1.75×10^−5) ≈ 0.8955,

p(N|d5) = 1 − p(B|d5) ≈ 0.1045

