# Nội dung đề tài

Hiện nay việc áp dụng các công nghệ tiên tiến vào trong y học đang được các nhà khoa học, nhà nghiên cứu đặc biệt chú ý. Hơn thế nữa tính trạng bệnh tim hiện nay có rất nhiều những rủi ro cho người mắc bệnh. Để giúp y bác sỹ chuẩn đoán cũng như phòng ngừa bệnh được tốt nhất, nhóm mình đề suất đề tài áp dụng học máy trong chuẩn đoán bệnh tim.

# Phương pháp tiếp cận

Ta có thể thấy giữa Machine Learning và lý thuyết xác suất có một sự liên hệ rất khăng khít. Các phương pháp phân loại dựa trên lý thuyết xác suất về cơ bản có thể hiểu là việc tính xem xác suất một sự việc của chúng ta sẽ xảy ra theo hướng như thế nào. Xác suất của hướng nào càng cao thì khả năng sự việc xảy ra theo hướng đó càng nhiều. Điều này đặc biết có ý nghĩa trong bài toán dự đoán và phân lớp của lĩnh vực Machine Learning. Theo thống kê học hiện đại thì tương ứng với mỗi bài toán giải quyết theo phương pháp xác suất thường đi kèm theo một phân phối xác suất phù hợp với bài toán đó. Tương ứng với mỗi phân phối xác suất chúng ta có một cách tính riêng các đại lượng cần thiết cho quá trình chạy các thuật toán như kỳ vọng, độ lệch chuẩn ...

Đề tài này chúng em chọn sử dụng thuật toán **Naive Bayes** và phân phối **Gaussian** để tiến hành triển khai bài toán

## Phân lớp Naïve Bayes

Là các phương pháp học phân lớp có giám sát và dựa trên xác suất

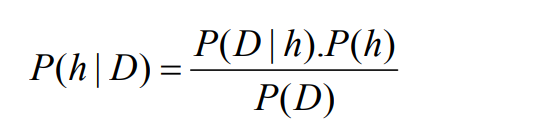
Dựa trên một mô hình (hàm) xác suất

Việc phân loại dựa trên các giá trị xác suất của các khả năng xảy ra của giả thuyết

Là một trong các phương pháp học máy thường xuyên được sử dụng trong các bài toán thực tế

Dựa trên định lý Bayes

Định lý Bayes:



*P(h)*: Xác suất trước rằng h là đúng

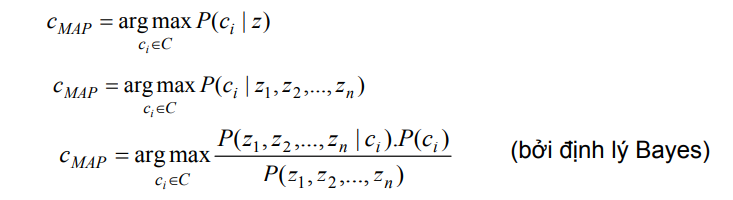
*P(D)*: Xác suất của tập dữ liệu được quan sát

*P(D|h)*: Xác suất của tập dữ liệu được quan sát với điều kiện h đúng

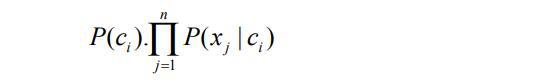
*P(h|D)*: Xác suất của h đúng với điều kiện D được quan sát

Phân loại Naïve Bayes

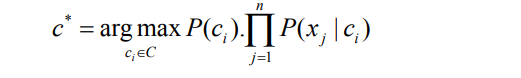
* Biều diễn bài toán:
* Một tập học *D\_train*, trong đó mỗi ví dụ học *x* được biểu diễn là một vecto n chiều: *(x­1, x2, x3, …, xn)*
* Một tập xác định các phân lớp: C = {*c1, c2, c3, …, cm*}
* Với một ví dụ z mới thì z sẽ thuộc phân lớp nào?
* Mục tiêu
* Xác định phân lớp có thể (phù hợp) nhất đối với z

****

* Giải thuật
* Giai đoạn học (traning phase), sử dụng một tập học:
* Đối với mỗi phân lớp có thể *ci* thuộc *C,*
* Tính giá trị xác suất trước *P(ci)*
* Tính giá trị xác suất xảy ra của mỗi thuộc tính đối với mỗi phân lớp *P(xi|ci)*
* Giai đoạn phân lớp (classification phase):
* Đối với mỗi phân lớp, tính giá trị

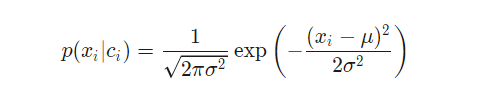


* Xác định phân lớp có thể nhất, tính giá trị:



## Phân phối Gaussian

Với mỗi một dữ liệu *xi,* thuộc class *ci*, chúng ta thấy *xi* tuân theo phân phối chuẩn với kì vọng và độ lệch chuẩn. Khi đó ta có công thức sau:



**Note**: Ta có thể sử dụng thư viện *sklearn* để hỗ trợ cho việc tính toán và áp dụng vào bài toán chuẩn đoán bệnh

## Thiết kế giao diện web bằng SpringBoot

# Tập dữ liệu

Tập dữ liệu bao gồm 76 trường (chỉ sử dụng 12 trường):

**1. age**

**2. sex**

**3. cp -** kiểu đau ngực

* 1: đau thắt ngực điển hình
* 2: đau thắt ngực không điển hình
* 3: không đau
* 4: không có triệu chứng

**4. trestbps -** huyết áp

**5. chol** - lượng mỡ trong máu

**6. fbs** (lượng đường trong máu > 120 mg/dl)

* 1 = true
* 0 = false)

**7. restecg -** điện tâm đồ

* 0
* 1
* 2

**8. thalach** - nhịp tim tối đa

**9. exang -** tập thể dục có gây đau thắt ngực

* 0: không
* 1: có

**10. oldpeak**

**11. thal**

* 3
* 6
* 7

**12. num -** the predicted attribute

* 0: < 50%
* 1: > 50%