

TRƯỜNG ĐẠI HỌC VĂN LANG

**KHOA CÔNG NGHỆ THÔNG TIN**

**HỌC MÁY VÀ ỨNG DỤNG BÁO CÁO ĐỒ ÁN:**

**TÌM HIỂU CÁC THUẬT TOÁN XỬ**

**LÝ DỮ LIỆU CSV**

**SVTH: Lý Minh Tài - 2274802010770**

**GVHD: Huỳnh Thái Học**

TP. Hồ Chí Minh – năm 2024

**Mục Lục**

**Nhận Xét Của Giáo Viên**

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

**Phần Mở Đầu**

Lời mở đầu:

Trong thời đại công nghệ hiện nay, việc áp dụng các thuật toán học máy (Machine Learning) để giải quyết các bài toán thực tiễn đã trở thành một xu hướng tất yếu. Những phương pháp này không chỉ giúp tăng hiệu quả trong việc xử lý dữ liệu mà còn mang lại những góc nhìn mới mẻ và chính xác hơn trong việc đưa ra quyết định.

Báo cáo này tập chung vào phân tích dữ liệu thuốc phù hợp với từng bệnh nhân thông qua các dữ liệu đầu vào bằng việc áp dụng ba thuật toán học máy. Trong đó **Naive Bayes** là một thuật toán phân loại đơn giản với giả định rằng các đặc trưng là độc lập với nhau , **Linear Regression** và **K-Nearest Neighbors (KNN)** được sử dụng trong các bài toán hồi quy

Thông qua báo cáo này, em muốn trình bày cách bài toán phân tích dữ liệu và phân tích ưu điểm và nhược điểm của từng thuật toán, từ đó đưa ra những kết luận trong bài phân tích

# Chương 1: Áp dụng thuật toán

# Dữ liệu bài toán

Dữ liệu bài toán là một bộ dữ liệu csv nhằm phân biệt được các loại thuốc ( DrugA, DrugB, DrugC,…) để phù hợp cho từng bệnh nhân. Dữ liệu bài toán rất phù hợp cho ngành y học. Bộ dữ liệu bào gồm:

Tập huấn luyện (training): Là một tập dữ liệu dùng để huấn luyện mô hình

Tập kiểm thử (testing): Là một tập dữ liệu ùng để kiểm tra mức độ chính xác của mô hình học máy nhằm đưa ra các kết quả khách quan nhất

1. **Mục tiêu bài toán**

Mục tiêu của bài toán này là dựa trên các dữ liệu được truyền vào từ đó chẩn đoán bệnh nhân phù hợp với loại thuốc nào. Mô hình học máy này phân tích từ dữ liệu có sẵn truyền vào bộ dữ liệu huấn luyện và kiểm tra thông qua bộ kiểm thử và từ đó đưa ra đánh giá về độ chính xác của mô hình

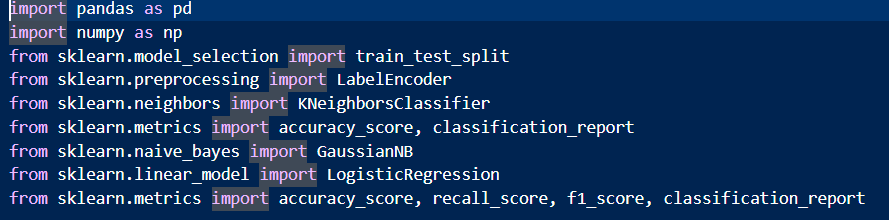
**Input và Output của bài toán:**

* **Input**

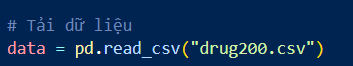
1. Age: tuổi tác
2. Sex: giới tính (F: nữ, M: nam)
3. BP: Huyết áp (HIGH, NORMAL, LOW)
4. Cholesterol: Mức cholesterol (HIGH, NORMAL)
5. Na\_to\_K: Tỉ lệ Natri/Kali trong máu

* **Output:**

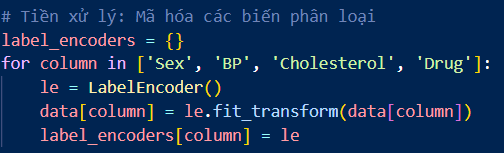
1. Kết quả dự đoán: Kết quả của tùng của bài dữ đoán bệnh nhân sẽ phù hợp với loại thuốc nào
2. Mức độ chính xác của mô hình: Accuracy, F1-score, Recall
3. **Lập trình** 
   1. **Thư viện sử dụng**



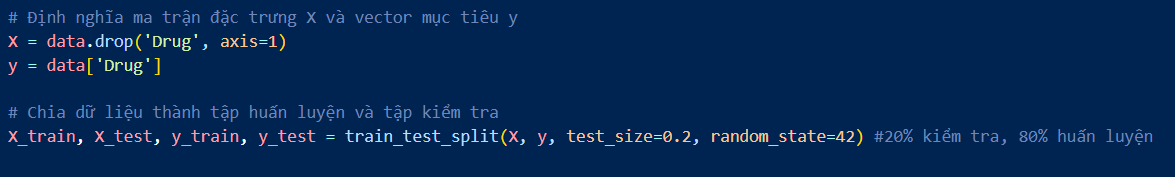
**3.2 Đưa dữ liệu vào**



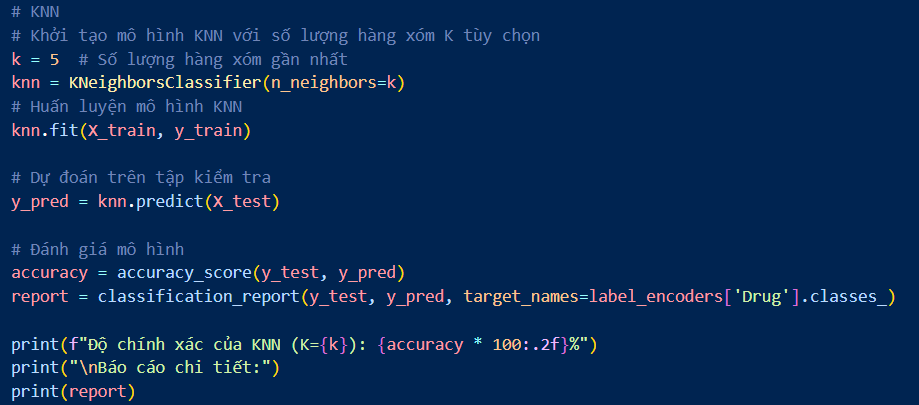
**3.3 Mã hóa dữ liệu**



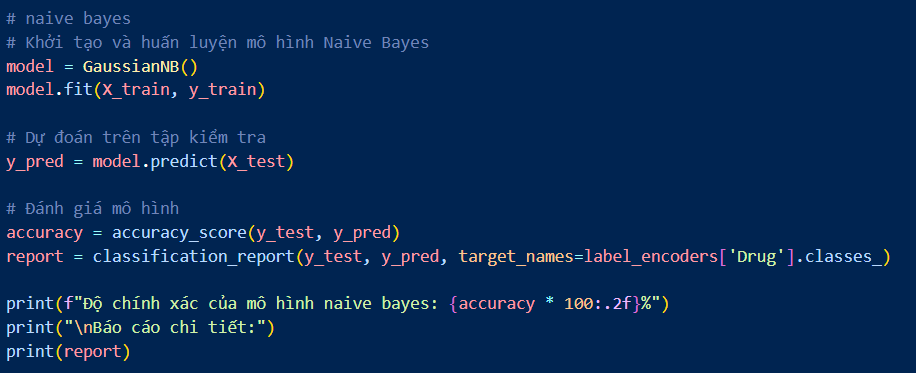
**3.4 Định nghĩa ma trận và chia dữ liệu**



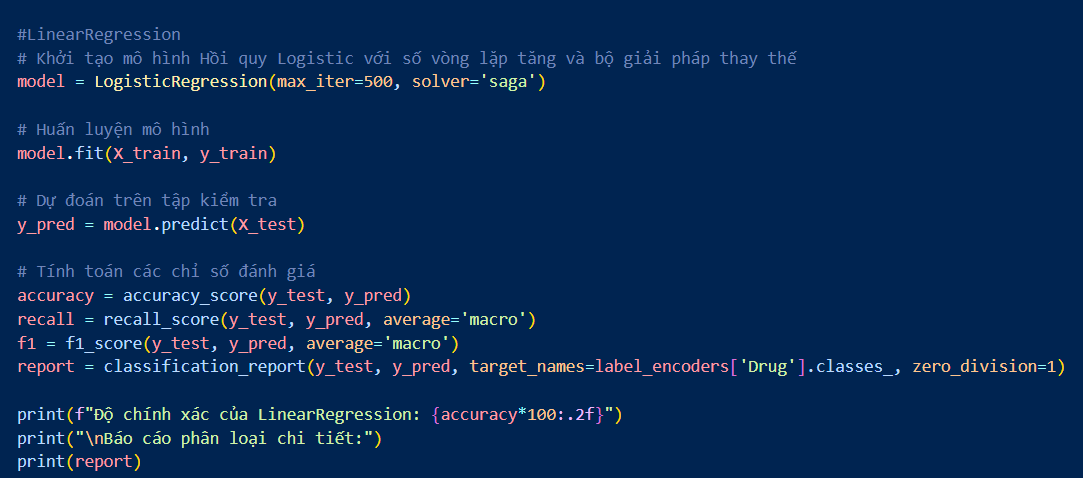
**3.5 Khởi tạo mô hình và đưa đánh giá mô hình KNN**



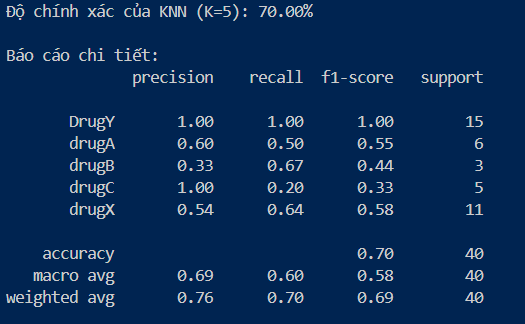
**3.6 Khởi tạo mô hình và đánh giá mô hình Naive Bayes**



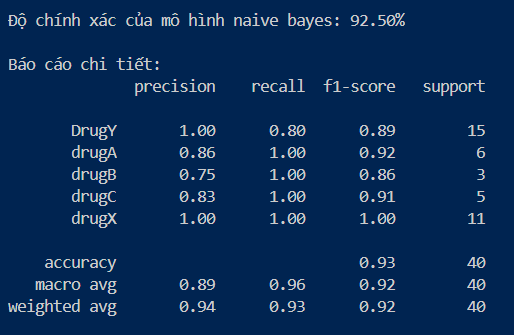
3.7 Khởi tạo và đánh giá mô hình Linear Regression



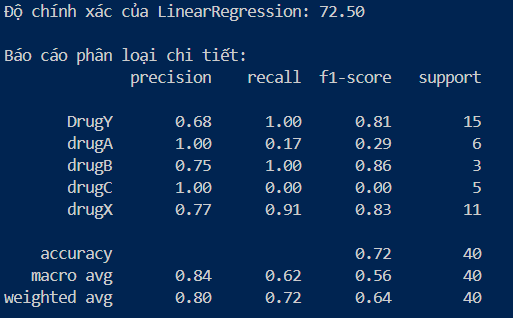
1. **Kết quả**
   1. **KNN**



* 1. **Naive Bayes**



* 1. Linear Regression



**CHƯƠNG 3: KẾT LUẬN**

1. Nearest Neighbors (KNN) , Naive Bayes, và Linear Regression là các thuật toán học máy phổ biến nhưng mỗi thuật toán đều có những đặc điểm riêng và đưa ra các kết quả khác nhau. Trong đó Naive Bayes là mô hình thuật toán đưa ra mức độ chính xác cao nhất. KNN và Linear regression cũng có độ chính xác khá cao , tuy nhiên vẫn có sự cách biệt khá lớn so với thuật toán Naive Bayes
2. **Bảng so sánh**

|  |  |  |  |
| --- | --- | --- | --- |
| Mô Hình | Accuracy | F1-Score | Recall |
| KNN (K=5) | 70% | 58% | 60% |
| Naive Bayes | 92,5% | 92% | 96% |
| Linear Regression | 72,5% | 61,6% | 55,8% |

1. **Kết luận**

Dựa vào các chỉ số như độ chính xác (Accuracy), recall và F1-score, ta có thể thấy rằng

* **Linear Regression:** Mức độ chính xác của linear regression khá cao nhưng độ nhạy (recall) và F1\_Score không được quá cao cho thấy mô hình chưa nhận diện dữ liệu tốt như mong muốn
* **Naive Bayes**:Với mức độ chính xác (92,5%), recall (96%) và F1-Score (92%) khá ấn tượng . Naive bayes là lựa chọn tốt nếu cần một mô hình đơn giản, nhanh chóng và tiết kiệm tài nguyên
* **KNN**: có độ chính xác không quá cao (70%), recall và F1-Score có tỉ lệ phần trăm khá thấp (58% và 60%). Một phần do giá trị K=5 chưa phù hợp với mô hình

Thuật toán Naive Bayes có mức độ chính xác cao nhất là vì:

* Naive Bayes: giả định rằng các đặc trưng (features) độc lập với nhau, và mô hình tính xác suất dựa trên từng đặc trưng riêng lẻ. Nếu các đặc trưng trong tập dữ liệu (tuổi, giới tính, huyết áp, cholesterol, tỷ lệ Na/K) có tính độc lập tương đối, mô hình sẽ hoạt động tốt hơn.
* KNN và Linear regression: thường sẽ phù hợp với các bài toán Hồi quy

-> Naive Bayes vượt trội hơn vì có những ưu điểm: tập dữ liệu nhỏ và không quá phức tạp, Phân loại với nhiều biến đầu vào nhưng ít phụ thuộc

1. **Những hạn chế và hướng phát triển trong tương lai**
   1. **Những hạn chế**

Trong bài phân tích này còn những hạn chế như chưa đưa ra được mô hình và các dữ liệu đầu vào chưa được đa dạng

* 1. **Hướng phát triển trong tương lai**

**Để cải thiện hơn nữa mô hình phân loại:**

**Nâng cấp thêm các kỹ thuật nâng cao như học sâu ( Deep Learning )**

**Cải thiện bộ dữ liệu đa dạng với nhiều dữ liệu đầu và**