Nguyễn Chí Thanh-64TTNT2

2251262639

**Predicting Criticality of Patients Based on Disease Symptoms Using Machine Learning**

Bài báo "Predicting Criticality of Patients Based on Disease Symptoms Using Machine Learning" nghiên cứu việc dự đoán mức độ nguy kịch (criticality) của bệnh nhân dựa trên các triệu chứng bệnh sử dụng các kỹ thuật Học máy (Machine Learning - ML)[**Abstract—Predicting disease criticality is essential for**

**immediate medical attention and optimizing healthcare facilities. In**

**this work, we propose a Machine Learning (ML) based approach to**

**predict the criticality of a patient based on disease symptoms. By**

**using Disease-Symptom Knowledge Database dataset, we created a**

**binary classification model to distinguish critical conditions of a**].-[1]

Tại sao việc này lại quan trọng?

Việc phát hiện mức độ nguy kịch của bệnh tật là cần thiết để cung cấp sự chăm sóc y tế ngay lập tức và tối ưu hóa các cơ sở chăm sóc sức khỏe[1]....

Nó giúp cải thiện kết quả điều trị cho bệnh nhân, tối ưu hóa tài nguyên y tế và đảm bảo kế hoạch điều trị hiệu quả[**For this, multiple classification**

**algorithms are employed to achieve optimal accuracy and among**

**them the Gradient Boosting (GB) classifier was found to give best**

**result with 95.71% accuracy. The study also includes Exploratory**

**Data Analysis (EDA) to understand the symptom-disease**

**relationships.**]-[2]

Dự đoán các bệnh nguy kịch dựa trên triệu chứng hỗ trợ việc phát hiện sớm bệnh nhân có nguy cơ cao và giảm tỷ lệ tử vong[2].

Nó cũng giảm thiểu áp lực tài chính lên hệ thống chăm sóc sức khỏe do tỷ lệ mắc bệnh mãn tính cao4. Năm 2016, Mỹ đã chi 3.300 tỷ USD cho chăm sóc sức khỏe, và Ấn Độ thiệt hại 236,6 tỷ USD vào năm 2015 do các tình trạng nguy kịch[**Based on the information from Medicaid services and Medicare centres, half of the US population has several chronic illnesses . As a result, in 2016 the US spent 3,300 billion dollars on healthcare, or about 10,348 dollars per person on medical care. The report of WHO and World Economic Forum revealed that India costs 236.6 billion dollars in 2015 due to critical conditions . Due to the high incidence of chronic diseases, the early identification is a priority. By detecting the criticality of diseases in their initial stages reduces the mortality rates as well as minimizes the financial pressure on healthcare systems ]**.

Hạn chế của phương pháp truyền thống: Cách truyền thống để xác định bệnh tật thường dựa vào kiến thức của các chuyên gia và kiểm tra thủ công, điều này có thể tốn thời gian và dễ mắc lỗi[**Traditional way of identifying the disease focuses on professionals’ knowledge and a manual examination, which can be time-consuming and prone to mistake. While ML solves the complicated patterns of medical data and examined in real time to make accurate predictions [7, 8]. It also revolutionizes the healthcare, particularly in the diagnosis and prediction of disease progressions. One major challenge in healthcare is the prioritization of patients with greater risk or those who are in critical condition to provide immediate attention. Symptom-based disease criticality prediction can help in prioritization for patient care, reduce hospitalization time, and further improve the survival rate [9, 10]. The Exploratory Data Analysis (EDA) helps to understand the disease severity by analysing the relationships, patterns, and distributions within data and identifies the key variables that**].-[3]

Giải pháp được đề xuất: Bài báo đề xuất một phương pháp dựa trên Học máy để giải quyết các mô hình phức tạp trong dữ liệu y tế và phân tích chúng theo thời gian thực để đưa ra dự đoán chính xác.... Phương pháp này có thể giúp ưu tiên bệnh nhân có nguy cơ cao hoặc trong tình trạng nguy kịch để được chăm sóc ngay lập tức, giảm thời gian nằm viện và cải thiện tỷ lệ sống sót[3].

Cách tiếp cận:

Nghiên cứu này sử dụng kết hợp Phân tích Dữ liệu Khám phá (Exploratory Data Analysis - EDA) và các thuật toán ML để phân tích và dự đoán mức độ nguy kịch của bệnh nhân dựa trên triệu chứng[**In this work, we apply EDA with ML algorithms to analyse as well as predict the criticality of patients based on the disease symptoms. The 'Disease-Symptoms Knowledge Database' [17] is used in this work. We classified the diseases into either critical or non-critical by developing models that can estimate criticality. This will help in speeding up of diagnostic processes, and even save lives through early diagnosis of critically ill patients. It also supports healthcare professionals to take instant decision.**].... EDA giúp hiểu rõ mối quan hệ, mẫu hình và phân phối trong dữ liệu, xác định các biến chính để xây dựng mô hình dự đoán[3]....

Tập dữ liệu được sử dụng là "Disease-Symptoms Knowledge Database"2..., chứa mối liên hệ giữa triệu chứng và bệnh tật từ bản tóm tắt xuất viện của Bệnh viện New York Presbyterian năm 2004 [**After pre-processing, the target and source columns were renamed as Symptoms and Disease respectively. There are a total of 150 distinct diseases with 405 symptoms in the dataset. The Criticality column was added as a target column where 1 indicates a critical condition and 0 indicates a non-critical condition. The criticality of the disease was determined by taking the help from a Physician from a Government Hospital “Capital Hospital, Bhubaneswar” to assess if the disease requires immediate medical attention or not based on the disease symptoms.**].-[4]

Tập dữ liệu ban đầu có 150 bệnh phổ biến nhất và 405 triệu chứng[4]. Sau khi tiền xử lý, cột triệu chứng được đổi tên thành "Symptoms" và cột bệnh thành "Disease"[4].

Một cột mục tiêu mới là "Criticality" đã được thêm vào, với giá trị 1 cho tình trạng nguy kịch và 0 cho tình trạng không nguy kịch.... Mức độ nguy kịch này được xác định với sự trợ giúp của một Bác sĩ từ Bệnh viện Chính phủ Capital Hospital, Bhubaneswar[4].

Mô hình được xây dựng là mô hình phân loại nhị phân (critical hoặc non-critical)2....

Các thuật toán ML đã áp dụng: Bốn thuật toán phân loại ML đã được sử dụng và so sánh:

Gradient Boosting (GB)[2]...

Random Forest (RF)[ **Exploratory Data Analysis (EDA) results**

**For analyzing the dataset, EDA was performed on the pre-**

**processed dataset. Figure 1 illustrates the distribution of**

**critical and non-critical cases by symptoms and diseases. It**

**helps us to figure out if there is a significant pattern or overlap**

**between the two cases. It represents the overlap between**

**critical and non-critical situations. The overlapping data**

**points in the figure indicate the risk of misclassifying the**

cases]-[5]

AdaBoost[5]...

Extra Tree (ET)[5]...

Tập dữ liệu được chia thành 80% để huấn luyện và 20% để kiểm tra[**For configuration of each considered algorithm, we used train test split of the dataset as 80% train and 20% test.**]

Kết quả và Phân tích:

Kết quả EDA:

Phân phối các trường hợp nguy kịch và không nguy kịch theo triệu chứng và bệnh tật cho thấy có sự phủ lấp (overlap), điều này chỉ ra nguy cơ phân loại sai[5].

Số lượng trường hợp nguy kịch và không nguy kịch hơi mất cân bằng.

Các triệu chứng phổ biến nhất ở các trường hợp nguy kịch là "Shortness of breath" (Khó thở), "unresponsiveness" (Không phản ứng) và "fever" (Sốt)[**Also, it shows that the number of critical and non-**

**critical cases is slightly imbalanced.**

**Figure 2 shows the frequency of the top ten common**

**symptoms observed in critical cases. The symptom frequency**

**distribution helps to identify the most common symptoms**

**that are found in critical diseases. From Figure 2, it was found**

**that 'Shortness of breath', 'unresponsiveness', and 'fever' are**

**the most frequent symptoms associated with critical**

**conditions present in the considered dataset. Figure 3 displays the correlation between critical disease and**].

Mối tương quan giữa bệnh nguy kịch và mười triệu chứng hàng đầu chủ yếu là yếu (gần 0), mặc dù có một số triệu chứng tương quan với một số bệnh nguy kịch nhất định[**Here, the correlation**

**values range from -0.10 to 0.30. The value 1, 0, and -1**

**denotes perfect positive, weak, and negative correlation**

**respectively. The visualization indicates majority of the**

**correlations are quite weak (closer to 0). But, there are also**

**certain symptoms which are correlated with some critical**

**diseases.**].

Kết quả Phân loại:

Các thuật toán đạt được độ chính xác như sau[**In this section, the results of four considered classification**

**algorithms such as RF, GB, AdaBoost and ET on the**

**considered dataset are presented. After applying the**

**classification algorithms, it was found that RF, GB,**

**AdaBoost, and ET gives 90.62%, 95.71%, 88.47%, and**

**85.52% accuracy respectively**]:

RF: 90.62%

GB: 95.71%3...

AdaBoost: 88.47%

ET: 85.52%

Gradient Boosting (GB) cho kết quả chính xác nhất trong số các thuật toán được xem xét3....

Mặc dù có những thách thức được tìm thấy trong EDA (như dữ liệu trùng lặp), việc sử dụng các phương pháp ensemble như GB và RF đã cho kết quả tốt nhất vì chúng mạnh mẽ và có thể xử lý nhiễu, điểm dữ liệu trùng lặp và các mối quan hệ phi tuyến phức tạp[**It shows that GB gives more**

**accurate result than other considered algorithms. Despite all**

**the challenges found during EDA analysis, we used ensemble**

**methods like GB, and RF which are robust and can handle**

**noise, overlapping data points and are better than simpler**

**models. They are also capable of capturing complex, non-**

**linear relationships within the data. Therefore, the model**

**gives best result by using GB classifier.**].

Báo cáo phân loại (Classification Report) của GB cho thấy độ chính xác (precision) là 0.95 cho lớp 0 (không nguy kịch) và 0.97 cho lớp 1 (nguy kịch); độ thu hồi (recall) là 0.98 cho lớp 0 và 0.93 cho lớp 1.... Điều này cho thấy GB có khả năng phát hiện chính xác cả hai lớp.

Kết luận và Hướng phát triển tương lai: Nghiên cứu kết luận rằng họ đã thành công trong việc xây dựng một mô hình sử dụng GB với độ chính xác 95% và độ chính xác (precision) 97% để phát hiện mức độ nguy kịch.... Hiệu suất này đạt được mặc dù có những thách thức về dữ liệu được xác định trong EDA, có thể là do tính mạnh mẽ của phương pháp ensemble như GB.

Trong tương lai, nhóm nghiên cứu dự định cải thiện mô hình bằng cách kết hợp thêm nhiều đặc trưng như thông tin nhân khẩu học của bệnh nhân và các bệnh đi kèm (comorbidities). Mở rộng tập dữ liệu và áp dụng các kỹ thuật học sâu (deep learning) cũng có thể mang lại hiểu biết sâu sắc hơn về tiến trình và mức độ nghiêm trọng của bệnh[**Hence to avoidthese, in future we will try to focus on improving the proposed model by incorporating more features, such as patient**

**demographics and comorbidities. Expanding the dataset and**

**applying deep learning techniques may also provide deeper**

**understanding into the disease progression and severity.**

**Additionally, the techniques like SMOTE or cost-sensitive**

**learning can help in solving class imbalance issues.**]. Các kỹ thuật như SMOTE hoặc học chi phí nhạy cảm (cost-sensitive learning) có thể giúp giải quyết vấn đề mất cân bằng lớp.

# Link code:[thuchanh2.ipynb](https://colab.research.google.com/drive/1oKyIAl9SezC_Y9MTEvwiAXbq3RYLcjUL?usp=sharing)