**Tóm Tắt Bài Báo**

**Analyzing the impact of feature selection on the**

**accuracy of heart disease prediction**

**Muhammad Salman Pathan, Avishek Nag, Muhammad Mohisn Pathan, Soumyabrata Dev**

Dưới đây là tóm tắt chi tiết toàn bộ nội dung của bài báo “Analyzing the impact of feature selection on the accuracy of heart disease prediction”:

1. Bối Cảnh và Mục Tiêu

Bài báo khởi đầu bằng việc nhấn mạnh tính cấp bách trong việc chẩn đoán sớm bệnh tim mạch – nguyên nhân gây tử vong hàng đầu trên toàn cầu. Với số liệu thống kê báo cáo của WHO cho thấy tỷ lệ tử vong từ bệnh tim rất cao, việc cải thiện độ chính xác của dự đoán qua các kỹ thuật học máy (machine learning – ML) được xem là rất cần thiết. Tuy nhiên, dữ liệu y tế thường có số lượng đặc trưng (features) quá lớn và không phải tất cả đều có ý nghĩa trong việc dự đoán, gây ra hiện tượng "curse of dimensionality". Do đó, bài báo đặt mục tiêu xác định các đặc trưng quan trọng nhất để tối ưu hóa hiệu năng của các mô hình dự đoán bệnh tim.

2. Tổng Quan Các Nghiên Cứu Liên Quan

Bài báo điểm qua những công trình nghiên cứu trước đó đã áp dụng các thuật toán ML như Logistic Regression, Decision Tree, SVM, Random Forest, MLP… kết hợp với nhiều kỹ thuật chọn lọc đặc trưng như Genetic Algorithm, PCA, LASSO, mRMR,… Mặc dù nhiều nghiên cứu đã đạt được kết quả khả quan, nhiều phương pháp vẫn còn thiếu sự hướng dẫn hệ thống trong việc lựa chọn các đặc trưng đầu vào. Do đó, nghiên cứu này nhấn mạnh tầm quan trọng của phân tích mối tương quan giữa các biến và lựa chọn các đặc trưng phù hợp dựa trên bằng chứng thống kê.

3. Phương Pháp Nghiên Cứu Đề Xuất

a. Bộ Dữ Liệu

- CVD Dataset:

Ban đầu gồm 29.072 bản ghi với 12 đặc trưng chính (trong đó 11 đặc trưng là các triệu chứng lâm sàng và đặc trưng mục tiêu “stroke”) do McKinsey & Company cung cấp. Sau đó, do dữ liệu bị mất cân bằng (chỉ khoảng 548 trường hợp bị stroke) đã được cân bằng lại thành 1.096 mẫu thông qua random down sampling.

- Framingham Dataset:

Gồm 4.240 bản ghi với 15 đặc trưng (14 đặc trưng đầu vào và 1 đặc trưng cho rủi ro CHD trong 10 năm). Tương tự, quá trình cân bằng dữ liệu đã chọn ra 1.114 mẫu để đảm bảo tỷ lệ giữa các lớp.

b. Tiền Xử Lý Dữ Liệu

- Xử lý giá trị thiếu: Các bản ghi có giá trị null bị loại bỏ nhằm tránh sai lệch trong quá trình huấn luyện mô hình.

- Giảm tính mất cân bằng: Kỹ thuật random down sampling được áp dụng để tạo ra các bộ dữ liệu cân bằng giữa nhóm bệnh nhân có và không có bệnh.

c. Phân Tích Tương Quan

Bước này nhằm xác định mối quan hệ giữa các đặc trưng và biến mục tiêu.

- Với CVD dataset, các đặc trưng như tuổi (age), tăng huyết áp (hypertension), tiền sử bệnh tim (heart\_disease) và mức đường trung bình (avg\_glucose\_lvl) cho thấy mối tương quan rõ rệt với biến “stroke”.

- Tương tự, với Framingham dataset, các yếu tố như tuổi, huyết áp tâm thu (sysBP), tình trạng tăng huyết áp (prevalentHyp) và mức đường (glucose) liên quan mật thiết đến biến “TenYearCHD”.

d. Chọn Lọc Đặc Trưng

Để giảm số lượng biến không cần thiết và tăng hiệu quả dự đoán, bài báo sử dụng kỹ thuật lọc (filter-based) dựa trên ANOVA-F test.

- Thông qua hàm `f\_classif()` trong thư viện scikit-learn và phương pháp SelectKBest, các đặc trưng được xếp hạng theo mức độ ảnh hưởng đến biến mục tiêu.

- Kết quả của ANOVA-F test khẳng định các đặc trưng được lựa chọn (ví dụ: tuổi, hypertension, glucose, huyết áp) trùng khớp với kết quả phân tích tương quan từ bước trước.

e. Đánh Giá Hiệu Năng Mô Hình

Các mô hình ML được huấn luyện và đánh giá dựa trên các chỉ số:

- Accuracy: Tỷ lệ dự đoán đúng trên tổng số mẫu.

- F1-score: Trung bình điều hòa giữa độ chính xác (precision) và khả năng thu hồi (recall).

- ROC Curve: Đánh giá khả năng phân biệt của mô hình dựa trên tỷ lệ đúng (TPR) và tỷ lệ sai (FPR).

Các mô hình được so sánh khi huấn luyện với tập đầy đủ các đặc trưng và tập đặc trưng đã được chọn lọc.

4. Kết Quả và Thảo Luận

- Trên tập đầy đủ đặc trưng:

Các mô hình như MLP, Logistic Regression, SVC, Random Forest đều cho kết quả khả quan; tuy nhiên, mô hình MLP đạt được độ chính xác cao nhất khoảng 73% (đi kèm ROC và F1-score tương đương). Trong khi đó, mô hình Dummy (dự đoán ngẫu nhiên) cho kết quả chỉ khoảng 46%.

- Ảnh hưởng của việc chọn lọc đặc trưng:

Sau khi áp dụng ANOVA-F để lựa chọn các đặc trưng quan trọng, hiệu năng của các mô hình được cải thiện đáng kể. Việc giảm số lượng đặc trưng không chỉ giúp nâng cao độ chính xác dự đoán mà còn giảm thời gian tính toán trong quá trình huấn luyện.

5. Kết Luận và Hướng Phát Triển

Bài báo khẳng định rằng việc chọn lọc đặc trưng là yếu tố quan trọng trong việc cải thiện hiệu quả dự đoán của các mô hình ML trong lĩnh vực chẩn đoán bệnh tim.

- Ưu điểm của phương pháp:

Tăng độ chính xác, giảm thời gian huấn luyện và giảm độ phức tạp của mô hình bằng cách loại bỏ các đặc trưng dư thừa.

- Hướng phát triển:

Các tác giả đề xuất mở rộng nghiên cứu theo hướng áp dụng các phương pháp kết hợp (ensemble methods) và tinh chỉnh thêm các kỹ thuật chọn lọc đặc trưng để đạt hiệu quả cao hơn trong chẩn đoán bệnh tim mạch. Ngoài ra, mã nguồn của nghiên cứu cũng được công khai trên GitHub với mục đích tái sinh (reproducible research).