Dự Đoán Bệnh Tim Bằng Máy Học

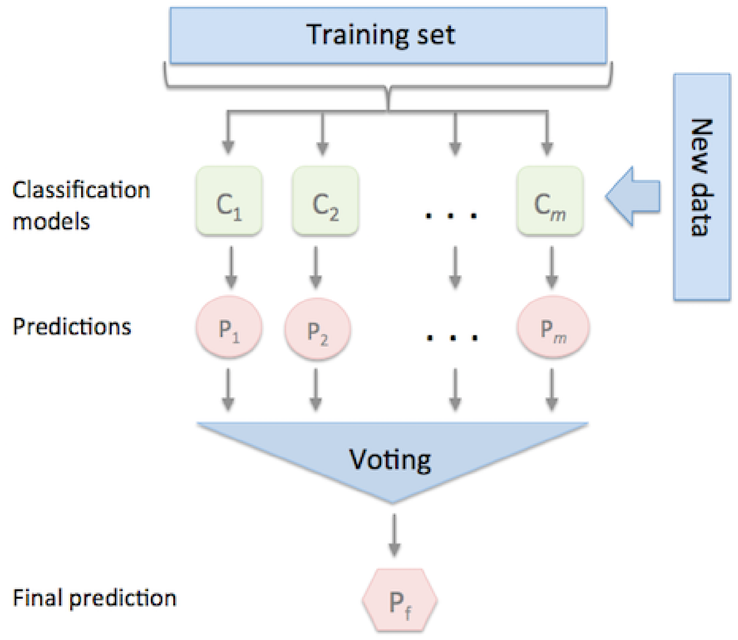
Phạm Tấn Khương, Hoàng Vũ
Huỳnh Thanh Bình, Nguyễn Minh Tú

Vấn đề và mục tiêu

Input: Thông tin người cần dự đoán bệnh tim.
Output: Có nguy cơ bị bệnh tim không?

Bệnh tim mạch là một trong những nguyên nhân gây tử vong hàng đầu trên toàn thế giới.
Mục tiêu chính của nghiên cứu là tạo ra mô hình máy học với độ chính xác cao nhất và thời gian xử lý không quá chậm.

Phương pháp



Tổng quan cho phương pháp Volting Enssemble

1 Huấn luyện và kiểm thử mô hình

Mô hình	Train Accuracy	Test Accuracy	Precision	Recall
KNN	100.00%	95.08%	94.12%	96.97%
Random Forest	90.87%	91.80%	91.18%	93.94%
Logistic Regression	86.72%	90.16%	90.91%	90.91%
Naive Bayes	83.82%	88.52%	90.63%	87.88%
SVM	94.61%	85.25%	87.50%	84.85%
Decision Tree	93.78%	81.97%	84.38%	81.82%

Kết quả kiểm thử mô hình

2 Phân tích kết quả và đánh giá

KNN đạt độ chính xác kiểm thử cao nhất (95.08%), nhưng có nguy cơ overfitting vì accuracy huấn luyện đạt tuyệt đối (100%). Điều này cho thấy KNN ghi nhớ quá mức dữ liệu huấn luyện, có thể không ổn định với dữ liệu mới.
Random Forest là mô hình cân bằng tốt giữa độ chính xác huấn luyện và kiểm thử (91.80% test accuracy), cho thấy khả năng tổng quát hóa cao, nhờ cơ chế bootstrap và bagging trong quá trình học.
Logistic Regression hoạt động ổn định và có kết quả khá tốt (90.16% accuracy), phù hợp với các bài toán tuyến tính và có khả năng giải thích mô hình.
Naive Bayes có kết quả kiểm thử khá tốt dù mô hình đơn giản (88.52%), cho thấy thuật toán này phù hợp với bài toán khi các thuộc tính độc lập tương đối.
SVM đạt độ chính xác huấn luyện cao (94.61%) nhưng kiểm thử thấp hơn (85.25%), có thể do mô hình chưa tối ưu được siêu tham số (kernel, C).
Decision Tree có mức chênh lệch lớn giữa train (93.78%) và test (81.97%), cho thấy dấu hiệu overfitting nghiêm trọng nếu không có cắt tỉa hoặc giới hạn độ sâu.

3 Chọn mô hình

Chúng tôi chọn mô hình Logistic Regression với mô hình SVM.

Kết quả

Nghiên cứu này đã ứng dụng các thuật toán học máy để dự đoán nguy cơ mắc bệnh tim dựa trên bộ dữ liệu từ UCI. Kết quả cho thấy:

- Random Forest đạt độ chính xác kiểm thử cao (91.80%), cho thấy khả năng tổng quát hóa tốt và hiệu suất ổn định.
- Logistic Regression cũng thể hiện hiệu quả với độ chính xác kiểm thử 90.16%, đồng thời dễ triển khai và giải thích, phù hợp với các ứng dụng thực tế.
- K-Nearest Neighbors (KNN) đạt độ chính xác kiểm thử cao nhất (95.08%), nhưng có dấu hiệu overfitting do độ chính xác huấn luyện đạt tuyệt đối (100%), điều này có thể ảnh hưởng đến khả năng tổng quát hóa trên dữ liệu mới.

Như vậy, nghiên cứu đã trả lời được câu hỏi đặt ra: Liệu các mô hình học máy có thể dự đoán hiệu quả nguy cơ mắc bệnh tim? Kết quả cho thấy các mô hình như Random Forest và Logistic Regression có thể được sử dụng để hỗ trợ chẩn đoán bệnh tim một cách hiệu quả.