Bộ giáo dục và đạo tạo

Trường Đại Học Sài Gòn

Khoa công nghệ thông tin

Môn Phương Pháp Nghiên Cứu Khoa Học

Đề Cương Chi Tiết

|  |  |
| --- | --- |
| Tên đề tài | Ứng dụng học máy để dự đoán bệnh tim (Heart Disease Prediction) và sử dụng kỹ thuật ensemble machine learning |
| Lĩnh vực nghiên cứu | Công nghệ phần mềm & Học máy (Machine Learning) |
| Loại hình nghiên cứu | Sản Phẩm |
| Người hướng dẫn | Đỗ Như Tài |
| Sinh Viên Thực Hiện | Phạm Tấn Khương – 3122410191  Hoàng Vũ - 3122560089  Huỳnh Thanh Bình - 3122410033  Nguyễn Minh Tú - 3120411167 |

**I) Giới Thiệu**

Bệnh tim là một trong những nguyên nhân gây tử vong hàng đầu trên thế giới. Việc dự đoán sớm nguy cơ bệnh tim giúp cải thiện khả năng chẩn đoán, điều trị kịp thời, và giảm tỷ lệ tử vong. Các phương pháp truyền thống như chẩn đoán lâm sàng của bác sĩ thường dựa trên triệu chứng (đau ngực, khó thở) và thang điểm nguy cơ (Framingham Risk Score), nhưng thiếu độ chính xác khi xử lý dữ liệu lớn và mẫu ẩn.

Với sự tiến bộ của học máy, các nghiên cứu gần đây đã áp dụng nhiều thuật toán như K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, Naive Bayes, Support Vector Machine (SVM), và Decision Tree để dự đoán bệnh tim dựa trên dữ liệu số (tuổi, huyết áp, cholesterol). Tập dữ liệu UCI Heart Disease (303 mẫu, 14 đặc trưng) là một nguồn phổ biến để thử nghiệm các mô hình này.

**Vấn đề cần giải quyết:**

* Độ chính xác thực tế: Cần mô hình có Test Accuracy cao để dự đoán chính xác trên dữ liệu mới.
* Phát hiện bệnh nhân: Đảm bảo Recall cao để ít bỏ sót người có bệnh (False Negatives thấp), quan trọng trong y khoa.
* Giảm báo nhầm: Đạt Precision cao để tránh chẩn đoán sai người không bệnh thành có bệnh (False Positives thấp).
* Overfitting: Train Accuracy quá cao so với Test Accuracy cho thấy mô hình học thuộc dữ liệu huấn luyện, không tổng quát hóa tốt

**II) MỤC TIÊU VÀ PHẠM VI NGHIÊN CỨU**

|  |  |
| --- | --- |
| **1 ) MỤC TIÊU** |  |

* Xây dựng hệ thống dự đoán bệnh tim hiệu quả bằng học máy, hỗ trợ bác sĩ và bệnh nhân trong chẩn đoán sớm.
* Đánh giá hiệu suất của 6 mô hình học máy qua các chỉ số:
  + Train Accuracy: Đo độ chính xác trên dữ liệu đã học (train set).
  + Test Accuracy: Đo độ chính xác trên dữ liệu mới (test set).
  + Precision: Đo tỷ lệ dự đoán "có bệnh" đúng (ít báo nhầm).
  + Recall: Đo tỷ lệ phát hiện người thực sự có bệnh (ít bỏ sót).
* So sánh hiệu suất để chọn mô hình tốt nhất, ưu tiên Recall cao (phát hiện bệnh nhân) và Test Accuracy ổn định (khả năng thực tế).

2) **PHẠM VI NGHIÊN CỨU**

* **Dữ liệu**: Tập trung vào tập dữ liệu UCI Heart Disease (303 mẫu, chia 80% train, 20% test). |
* **Đặc trưng**: Các yếu tố chính bao gồm:
  + Tuổi (age): Nguy cơ tăng theo tuổi.
  + Giới tính (sex): Nam giới có nguy cơ cao hơn.
  + Huyết áp lúc nghỉ (trestbps): Huyết áp cao là dấu hiệu nguy cơ.
  + Cholesterol (chol): Mức cao liên quan đến bệnh mạch vành.
  + Nhịp tim tối đa (thalach): Nhịp tim bất thường là chỉ số quan trọng.
* **Mô hình**: Phân tích 6 thuật toán: KNN, Random Forest, Logistic Regression, Naive Bayes, SVM, Decision Tree. |
* **Phương pháp**: Chuẩn hóa dữ liệu bằng StandardScaler, tối ưu KNN bằng GridSearchCV. |

**III) Phương pháp nghiên cứu**

|  |  |
| --- | --- |
| **1 ) NỘI DUNG** |  |

* **Quy trình thực hiện**:
  1. Thu thập dữ liệu từ tập UCI Heart Disease (303 mẫu, 14 đặc trưng).
  2. Tiền xử lý dữ liệu:
     + Xử lý giá trị thiếu (nếu có)
     + Chuẩn hóa đặc trưng bằng StandardScaler để cải thiện hiệu suất mô hình.
  3. Chia dữ liệu: 80% tập huấn luyện (242 mẫu), 20% tập kiểm tra (61 mẫu).
  4. Huấn luyện 6 mô hình:
     + KNN: Tối ưu tham số bằng GridSearchCV (n\_neighbors, weights).
     + Random Forest, Logistic Regression, Naive Bayes, SVM, Decision Tree: Sử dụng tham số mặc định hoặc điều chỉnh cơ bản.
  5. Đánh giá: Tính Train Accuracy, Test Accuracy, Precision, Recall dựa trên kết quả dự đoán.
* **Công thức đánh giá**:
  1. Train Accuracy = (Số đoán đúng / Tổng mẫu train) × 100
  2. Test Accuracy = (Số đoán đúng / Tổng mẫu test) × 100
  3. Precision = (TP / (TP + FP)) × 100
  4. Recall = (TP / (TP + FN)) × 100
  5. Trong đó:
     + TP (True Positives): Số mẫu có bệnh đoán đúng.
     + FP (False Positives): Số mẫu không bệnh đoán sai thành có bệnh.
     + FN (False Negatives): Số mẫu có bệnh đoán sai thành không bệnh.

**IV) KẾT QUẢ VÀ ĐÁNH GIÁ**

| **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% |

**Đánh giá chi tiết**:

* **KNN**:
  + Test Accuracy cao nhất (95.08%), Recall tốt nhất (96.97%) → phát hiện 97% người có bệnh.
  + Precision 94.12% → ít báo nhầm (6% sai khi nói "có bệnh").
  + Train 100% → dấu hiệu overfitting, cần kiểm tra thêm trên dữ liệu lớn hơn.
* **Random Forest**:
  + Test 91.80%, Recall 93.94%, Precision 91.18% → cân bằng tốt giữa phát hiện và độ tin cậy.
  + Train 90.87% → ít overfitting hơn KNN.
* **Logistic Regression**:
  + Test 90.16%, Precision và Recall đều 90.91% → ổn định, phù hợp dữ liệu đơn giản.
* **Naive Bayes**:
  + Test 88.52%, Recall 87.88% → bỏ sót nhiều hơn so với KNN, RANF.
* **SVM**:
  + Test 85.25%, Recall 84.85% → kém trong phát hiện bệnh nhân.
* **Decision Tree**:
  + Test thấp nhất (81.97%), Recall 81.82% → bỏ sót gần 18% người có bệnh, kém hiệu quả.

**Sản phầm cuối cùng** : Sử dụng 2 mô hình có phần trăm cao nhất để sử dụng kỹ thuật ensemble nhằm tăng khả năng chính xác hơn

**Định hướng nghiên cứu tiếp theo**:

* Thu thập dữ liệu lớn hơn (vượt 303 mẫu) để giảm overfitting và kiểm tra tổng quát hóa của KNN.
* Tối ưu tham số cho Random Forest và SVM (số cây, kernel) để tăng Test Accuracy và Recall.
* Thử nghiệm mô hình sâu (Deep Learning) với dữ liệu bổ sung như tín hiệu ECG nếu có.

**V) Tài liệu tham khảo**

[**Github chương trình**](https://github.com/BodduSriPavan-111/Heart-Disease-Prediction-Using-Machine-Learning)

[Heart disease prediction using machine learning algorithms](https://www.matec-conferences.org/articles/matecconf/pdf/2024/04/matecconf_icmed2024_01122.pdf) của các tác giả Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain và Preeti Nagrath

[Heart disease prediction using machine learning algorithms](file:///C:\Users\somna\Downloads\iopscience.iop.org\article\10.1088\1757-899X\1022\1\012072\pdf) của các tác giả Govardhan Logabiraman, D.Ganesh, M. Sunil Kumar, A. Vinay Kumar, Nitin Bhardwaj

[Youtube Hướng dẫn](https://www.youtube.com/watch?v=Cokm4izzHe4)