Một số phương pháp trên tập dữ liệu Breast Cancer Wisconsin

Nguyễn Tuấn Anh
Khoa Khoa học Máy tính
Trường Đại học Công nghệ Thông tin
Hồ Chí Minh, Việt Nam
21520142@gm.uit.edu.vn

***Tóm tắt nội dung*—Breast Cancer Wisconsin là một tập dữ liệu phổ biến cho bài toán phân loại. Dữ liệu được thu thập từ những bệnh nhân ung thư vú tại Wisconsin, Mỹ. Bộ dữ liệu bao gồm những đặc trưng thực tế của những bệnh nhân ung thư vú lành tính (Benign) và ác tính (Malignant), dựa vào đó có thể đưa ra nhận định tình trạng của một bệnh nhân khi có những dấu hiệu nhất định. Ở đây chúng tôi xây dựng các mô hình để dự đoán, từ đó hỗ trợ phần nào cho các bác sĩ trong việc chẩn đoán tình trạng bệnh nhân. Chúng tôi thu được kết quả mô hình Multi-layer Perceptron có độ chính xác cao nhất trên tập test của hệ thống sử dụng đánh giá.**

***Từ khóa*—Classification, Data Analysis**

I. Giới thiệu

Bài toán phân loại là bài toán liên quan đến việc phân loại một đối tượng vào một trong các lớp đã được xác định trước. Ví dụ với hai lớp *chó* và *mèo*, từ các đặc trưng của một đối tượng (cân nặng, màu sắc, tập tính,...) chúng ta có thể phân loại đối tượng đó hoặc là chó hoặc là mèo.
Một cách tường minh, bài toán phân loại có thể được mô tả như sau:

- Input: Không gian đặc trưng χ, tập dữ liệu $\{(x_i, y_i)\}_{i=1}^{N}$, với $x_i \in \chi$ và $y_i \in L$ (tập các nhãn). Đặc trưng của đối tượng quan tâm x.
- Output: $y = f(x)$ với $f : \chi \mapsto L$ là mô hình ánh xạ đặc trưng trong không gian đặc trưng sang tập nhãn L.

Các phần còn lại phía dưới. Phần 2 giới thiệu về tập dữ liệu Breast Cancer Wisconsin. Phần 3 là kết quả thực nghiệm trên 4 phương pháp. Phần 4 thảo luận và phần 5 là kết luận.

II. Tập dữ liệu Breast Cancer Wisconsin

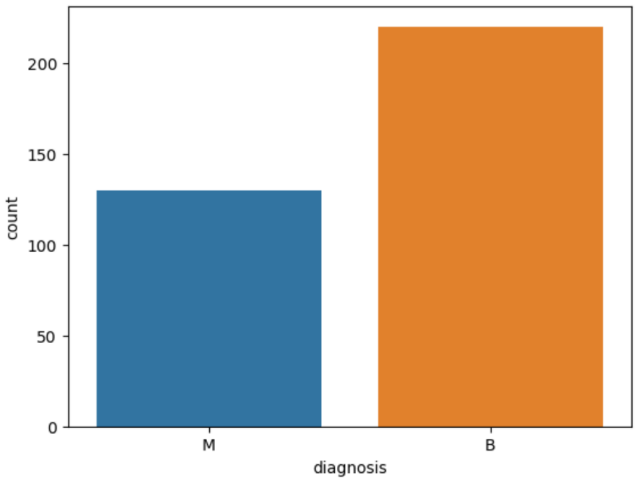
A. Tổng quan dữ liệu

Breast Cancer Wisconsin [1] là một tập dữ liệu phổ biến cho bài toán phân loại. Dữ liệu được thu thập từ những bệnh nhân ung thư vú tại Wisconsin, Mỹ. Tập dữ liệu này được tạo ra để hỗ trợ việc nghiên cứu và phát triển mô hình máy học để dự đoán liệu một khối u vú là ác tính (Malignant) hay lành tính (Benign) dựa trên các đặc trưng y học.

Tập dữ liệu được cung cấp là tập con của Breast Cancer Wisconsin, có 350 dòng và 31 cột tương ứng với 31 đặc trưng y học. Để thống nhất chúng tôi xem tập này như là tập dữ liệu gốc để thực hiện đánh giá. Hình 1 là 5 dòng đầu tiên của tập dữ liệu. Đặc biệt, cột **diagnosis** cho biết các biểu hiện khối u của bệnh nhân là lành tính (B) hay ác tính (M). Có tổng cộng 220 khối u lành tính và 130 khối u ác tính, Hình 2.

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
0	M	15.78	17.89	103.60	781.0	0.09710	0.1292	0.09954	0.06606	0.1842
1	M	19.17	24.80	132.40	1123.0	0.09740	0.2458	0.20650	0.11180	0.2397
2	M	15.85	23.95	103.70	782.7	0.08401	0.1002	0.09938	0.05364	0.1847
3	M	13.73	22.61	93.60	578.3	0.11310	0.2293	0.21280	0.08025	0.2069
4	M	14.54	27.54	96.73	658.8	0.11390	0.1595	0.16390	0.07364	0.2303

Hình 1. 5 dòng đầu tiên của tập dữ liệu Breast Cancer Wisconsin thu gọn.

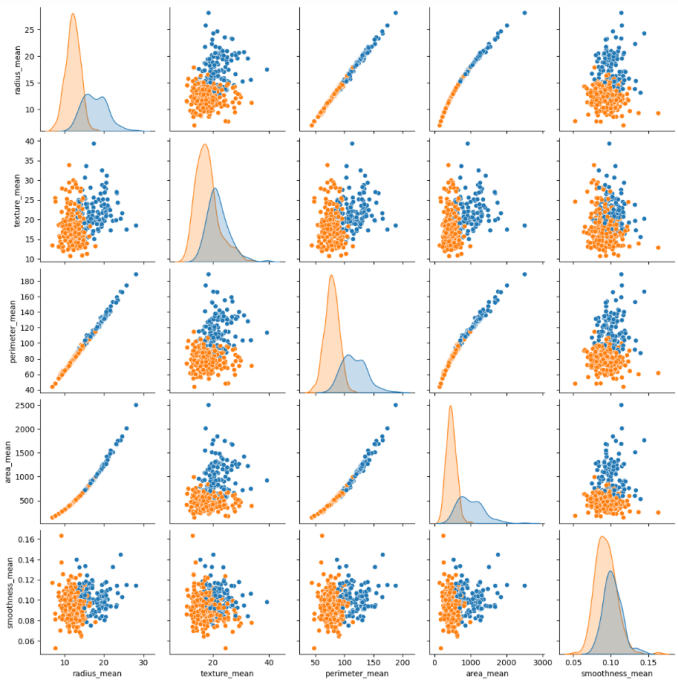


Hình 2. Số lượng phân loại cho từng khối u.

Ngoài ra, tập dữ liệu cũng được chuẩn bị rất kĩ, không có giá trị bị thiếu. Có thể dựa trên đó để làm cơ sở xây dựng mô hình máy học.

B. Trực quan và phân tích dữ liệu

Dựa trên Hình 3, một số nhận xét được rút ra như sau. *Một là*, mối tương quan giữa các đặc trưng: đặc biệt là `radius_mean`, `perimeter_mean`, và `area_mean`. Những đặc trưng này có liên quan về mặt hình học, khi bán kính khối u tăng, thì chu vi và diện tích của nó cũng tăng theo. `Texture_mean` cũng cho thấy mức độ tương quan với `radius_mean`, `perimeter_mean`, và `area_mean`, các khối u lớn hơn có thể có kết cấu khác so với khối u nhỏ. *Hai là*, phân loại dựa trên diagnosis. Đối với các đặc trưng



Hình 3. Pairplot cho 5 đặc trưng radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean. Dấu chấm xanh biểu thị biến quyết định Malignant, dấu chấm cam biểu thị Benign.

như radius_mean, perimeter_mean và area_mean, các khối u ác tính thường có giá trị cao hơn so với các khối u lành tính. Điều này cho thấy rằng các khối u ác tính có xu hướng lớn hơn về kích thước. Có sự phân biệt không rõ ràng hơn nhưng vẫn có thể nhận thấy trong texture_mean và smoothness_mean giữa các khối u ác tính và lành tính. *Ba là*, sự khác biệt về phân phối. Các biểu đồ trên đường chéo chính thể hiện phân phối mỗi đặc trưng. Các khối u ác tính thường có phân phối dịch về các giá trị cao hơn trong các đặc trưng như radius_mean, perimeter_mean và area_mean. Điều này cho thấy rằng những đặc trưng này có thể quan trọng trong việc phân biệt giữa các khối u ác tính và lành tính. *Bốn là*, hiện tượng Đa cộng tuyến. Sự tương quan mạnh mẽ giữa một số đặc trưng cho thấy có thể có hiện tượng đa cộng tuyến. Điều này là một yếu tố quan trọng cần xem xét trong phân tích thống kê và mô hình dự đoán.

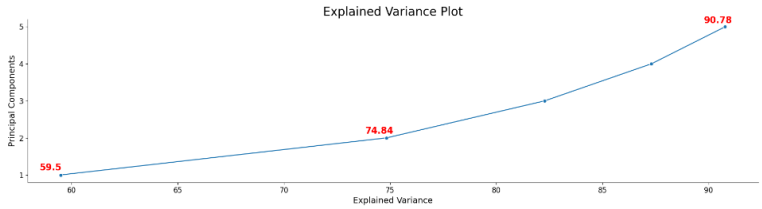
Hình 4 là kết quả kiểm định bằng Analysis of Variance [2] (ANOVA test) của các đặc trưng với biến quyết định diagnosis (ngưỡng 0.05 trong thống kê). Chúng tôi thấy có 5 đặc trưng không có mối quan hệ đáng kể với biến phân loại: fractal_dimension_mean, texture_se, smoothness_se, symmetry_se và fractal_dimension_se. Có nghĩa là, việc giá trị của các đặc trưng này có thay đổi thì cũng không ảnh hưởng lớn đến kết quả dự đoán. Chúng tôi sẽ loại bỏ những cột này khỏi dữ liệu và sau đó PCA [3] để tìm đặc trưng nào giải thích sự khác biệt nhất trong dữ liệu.

Hình 5 là biểu đồ khi thực hiện PCA. Chúng tôi nhận thấy với 5 thành phần thì phương sai được giải thích là lớn nhất

```
ANOVA statistic radius_mean and diagnosis (p-value): 9.50955780232019e-66
ANOVA statistic texture_mean and diagnosis (p-value): 4.408761134754431e-17
ANOVA statistic perimeter_mean and diagnosis (p-value): 1.2919845785668225e-68
ANOVA statistic area_mean and diagnosis (p-value): 6.916140587957887e-61
ANOVA statistic smoothness_mean and diagnosis (p-value): 1.5019437484316503e-09
ANOVA statistic compactness_mean and diagnosis (p-value): 5.343420650639344e-32
ANOVA statistic concavity_mean and diagnosis (p-value): 1.9103442135885402e-55
ANOVA statistic concave points_mean and diagnosis (p-value): 1.4132262751548634e-70
ANOVA statistic symmetry_mean and diagnosis (p-value): 3.666605973492627e-12
ANOVA statistic fractal_dimension_mean and diagnosis (p-value): 0.14942442399343042
ANOVA statistic radius_se and diagnosis (p-value): 5.95098630517521 45e-34
ANOVA statistic texture_se and diagnosis (p-value): 0.764804243515583
ANOVA statistic perimeter_se and diagnosis (p-value): 1.0842901039502082e-31
ANOVA statistic area_se and diagnosis (p-value): 1.4368044583091132e-32
ANOVA statistic smoothness_se and diagnosis (p-value): 0.23423020177274337
ANOVA statistic compactness_se and diagnosis (p-value): 2.0009772576300834e-08
ANOVA statistic concavity_se and diagnosis (p-value): 9.274590214840813e-13
ANOVA statistic concave points_se and diagnosis (p-value): 1.2334114400298187e-15
ANOVA statistic symmetry_se and diagnosis (p-value): 0.7658915983361326
ANOVA statistic fractal_dimension_se and diagnosis (p-value): 0.06379821249891209
ANOVA statistic radius_worst and diagnosis (p-value): 2.5997290293050978e-76
ANOVA statistic texture_worst and diagnosis (p-value): 2.841240031847008e-21
ANOVA statistic perimeter_worst and diagnosis (p-value): 4.496109712837722e-79
ANOVA statistic area_worst and diagnosis (p-value): 8.287036505273896e-65
ANOVA statistic smoothness_worst and diagnosis (p-value): 1.761920442217327e-13
ANOVA statistic compactness_worst and diagnosis (p-value): 9.221970891152136e-34
ANOVA statistic concavity_worst and diagnosis (p-value): 3.39709553820498e-50
ANOVA statistic concave points_worst and diagnosis (p-value): 4.606526870471364e-74
ANOVA statistic symmetry_worst and diagnosis (p-value): 3.9539836249487335e-17
ANOVA statistic fractal_dimension_worst and diagnosis (p-value): 3.653351708695353e-08
```

Hình 4. Anova testing. Dòng màu đỏ là đặc trưng không có mối quan hệ đáng kể với biến quyết định.

(90.78). Như vậy, khi tiền xử lí dữ liệu trong tương lai, chúng tôi sẽ xem xét số thành phần từ 5 trở lên.



Hình 5. Biểu đồ Explained Variance.

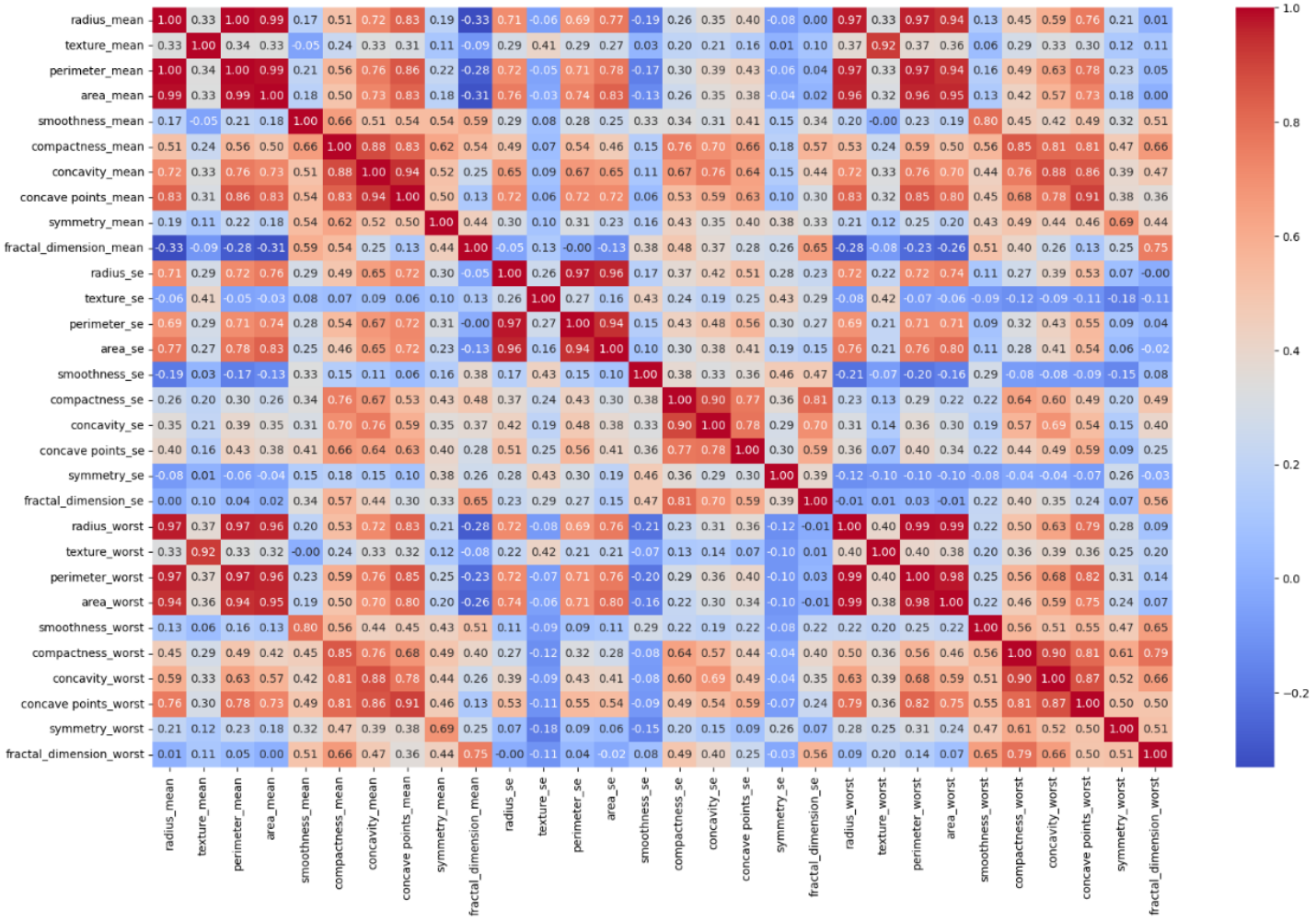
Hình 6 cho thấy một số đặc trưng, đặc biệt là những đặc trưng liên quan đến kích thước và hình dạng của khối u (bán kính, chu vi và diện tích), thể hiện sự tương quan mạnh với nhau. Điều này cho thấy khi một trong những đặc trưng này tăng, những đặc trưng khác cũng có xu hướng tăng. Các đặc trưng như radius_worst, perimeter_worst, concave points_worst có khả năng ảnh hưởng đến kết quả. Sự xuất hiện của tương quan cao giữa một số đặc trưng cho thấy khả năng xảy ra hiện tượng đa cộng tuyến.

III. Thực nghiệm

Quá trình thực nghiệm được thực hiện trên nền tảng của Website môn học. Các phương pháp được tinh chỉnh tối ưu ở Public Test trước khi submit.

A. Tiền xử lí dữ liệu

Ở bước tiền xử lí, có nhiều sự lựa chọn để có được đặc trưng và số liệu tốt. Tuy nhiên, những phương pháp riêng biệt như Principal Component Analysis, loại bỏ đặc trưng có tương đồng cao,...tác động lớn lên dữ liệu và đặc trưng bài toán. Bằng thực nghiệm và theo quan điểm chủ quan của mình, chúng tôi chỉ thực hiện Scale dữ liệu và xử lí ngoại lệ ở bước tiền xử lí.



Hình 6. Ma trận tương quan.

Breast Cancer Wisconsin là tập dữ liệu có chứa ngoại lệ, chúng tôi sử dụng Robust Scaler và phương pháp xử lí ngoại lệ Interquartile Range [4] (IQR) để giải quyết điều này. Một lí do đơn giản để chúng tôi lựa chọn chính là chúng đều dựa trên phạm vi giữa các phân vị (IQR), điều này đặc biệt hữu ích khi làm giảm sự nhạy cảm với ngoại lệ.

B. Mô hình sử dụng

Chúng tôi đang giải quyết bài toán phân loại bao gồm 2 lớp cần dự đoán. Chúng tôi sử dụng 4 phương pháp điển hình cho nhiệm vụ phân loại, chính là Logistic Regression [5], K-Nearest Neighbors base classify [6], Support Vector Machine [7] và Multi-layer Perceptron [8]. Chúng tôi nhắc lại mục tiêu chính của 4 phương pháp này trong bài toán phân loại:

- Logistic Regression: Dự đoán xác suất một bệnh nhân có khối u lành tính hay ác tính.
- K-Nearest Neighbors: Phân loại một bệnh nhân dựa trên nhãn của K bệnh nhân gần nhất trong không gian đặc trưng.
- Support Vector Machine: Tìm ranh giới phân loại tốt nhất giữa các mẫu.
- Multi-layer Perceptron: Xây dựng một mô hình Neural Network có khả năng học được các mối quan hệ phức tạp giữa các đặc trưng.

C. Optimization

Các bộ optimizer được sử dụng đều là mặc định của Scikit-learn: Logistic Regression với L-BFGS, K-Nearest Neighbors với auto Algorithm,...Khi thực nghiệm, một số tham số được điều chỉnh cụ thể như sau:

- Logistic Regression: random_state=42, max_iter=1000.
- K-Nearest Neighbors: n_neighbors=7.
- Support Vector Machine: random_state=42.
- Multi-layer Perceptron: hidden_layer_sizes=(150,), max_iter=1000, random_state=42.

D. Đánh giá mô hình

Phần tiền xử lí như trước đó đã đề cập. Đối với 3 phương pháp đầu tiên, chúng tôi thực hiện thêm một số bước riêng

biệt để cải thiện độ chính xác ước tính. Cụ thể, chúng tôi thực hiện giảm chiều đối với mô hình Support Vector Machine, loại bỏ đặc trưng có tương đồng cao đối với mô hình Logistic Regression, K-Nearest Neighbors. Cuối cùng là cân bằng dữ liệu đối với mô hình K-Nearest Neighbors, Support Vector Machine. Sau đó các mô hình được đánh giá trước thông qua phương pháp KFold với n_splits=10 trên training set. Bảng I là kết quả đánh giá.

Chúng tôi thấy rằng gần như 4 phương pháp đều khác biệt rất ít. Chúng tôi đặt ra nghi ngờ rằng, đối với training set, cả 4 phương pháp đều hội tụ. Có nghĩa là, các số liệu về độ chính xác trên training set không mang nhiều ý nghĩa về mặt thống kê. Mặc dù vậy, chúng tôi vẫn dựa trên độ chính xác cao đó để chọn ra bộ tham số tối ưu và tiền xử lí hiệu quả. Làm tiền đề kiểm định tiếp tục trên Private Test.

E. Kết quả trên Private Test

Sau khi kiểm định trên training set, chúng tôi chọn tham số tốt nhất cho từng phương pháp cũng như phần tiền xử lí tối ưu. Chi tiết mỗi phương pháp như bảng II. Cột Score là kết quả thực nghiệm trên Private Test.

Chúng tôi nhận thấy rằng, phương pháp Logistic Regression và Multi-layer Perceptron cho ra kết quả đầy hứa hẹn. Trong khi K-Nearest Neighbors và Support Vector Machine thực sự chưa khả quan. Chúng tôi sẽ phân tích điều này ở phần IV-E.

IV. Thảo luận

A. Hiệu quả của việc xử lí ngoại lệ

Ngoại lệ là những giá trị nhiễu có thể làm cho mô hình dự đoán sai. Bảng III cho thấy tác động của việc xử lí ngoại lệ đến kết quả dự đoán.

Bảng III
Trước và sau khi sử lí ngoại lệ.

Phương pháp	non - IQR	IQR
Logistic Regression	84.52	86.71
K-Nearest Neighbors	72.44	79.02
Support Vector Machine	76.83	77.93
Multi-layer Perceptron	85.61	88.9

B. Hiệu quả của việc scale dữ liệu

Đối với dữ liệu có nhiễu hoặc có giá trị quá khác so với còn lại, thì việc scale dữ liệu sẽ góp phần làm giảm sự nhạy cảm với nhiễu và ngoại lệ. Đặc biệt, scale còn giúp đưa dữ liệu về phạm vi nhất định, hoặc đưa về cùng phân phối, điều này đặc biệt hữu ích cho các mô hình máy học. Bảng IV cho thấy sự khác biệt khi sử dụng Robust Scaler cho mỗi phương pháp.

Bảng IV
Trước và sau khi scale dữ liệu.

Phương pháp	non - Robust Scaler	Robust Scaler
Logistic Regression	81.02	86.71
K-Nearest Neighbors	75.73	79.02
Support Vector Machine	74.63	77.93
Multi-layer Perceptron	87.74	88.9

C. Hiệu quả của việc loại bỏ đặc trưng có tương đồng cao

Những đặc trưng có tương đồng cao thường không đóng góp nhiều cho quá trình dự đoán. Trong nhiều trường hợp, chúng còn làm mô hình trở nên phức tạp hơn, mất nhiều chi phí tính toán hơn. Ngoài ra, những đặc trưng đó cũng có khả năng làm xảy ra hiện tượng đa cộng tuyến. Bảng V cho thấy sự khác biệt khi loại bỏ các đặc trưng radius_mean, perimeter_mean, area_mean, concavity_mean, radius_se, perimeter_se, radius_worst, perimeter_worst (được phát hiện trước).

Bảng V
Trước và sau khi loại bỏ đặc trưng.

Phương pháp	non - Remove	Remove
Logistic Regression	84.51	86.71
K-Nearest Neighbors	76.83	79.02

D. Hiệu quả của việc giảm chiều dữ liệu

Mục tiêu chính của PCA là giảm số chiều của dữ liệu trong không gian để giảm độ phức tạp và chi phí tính toán, đồng thời giữ lại sự biến động chính trong dữ liệu. Về bản chất, PCA sẽ tìm hướng (vector riêng của Covariance Matrix) để cực đại Rayleigh Quotients [9]. Nghĩa là, PCA thực hiện việc chuyển đổi dữ liệu từ các biến ban đầu sang các thành phần chính mới, sao cho các thành phần này giữ lại phần lớn sự biến động của dữ liệu. Bảng VI là sự khác biệt khi giảm chiều dữ liệu.

Bảng VI
Trước và sau khi giảm chiều.

Phương pháp	non - PCA	PCA
Support Vector Machine	75.73	77.93

E. Hiệu quả của các phương pháp

Có thể thấy, kết quả trên Private Test của các phương pháp khác biệt so với Public Test. Điều này hoàn toàn có thể được giải thích một cách rõ ràng. Đầu tiên, với K-Nearest Neighbors, do sự đơn giản của thuật toán nên việc xác định chính xác nhãn (thuộc về cụm nào) cho tất cả dữ liệu rất khó khăn. Tiếp theo, chúng tôi xem xét cả 2 mô hình Logistic Regression và Support Vector Machine. Mô hình Support Vector Machine sẽ có gắng tối ưu hóa khoảng cách margin giữa các lớp và hiệu quả với dữ liệu có kích thước lớn. Còn mô hình Logistic Regression sẽ cố gắng tìm ra một siêu phẳng tối ưu và dùng một hàm để ánh xạ xác suất một đối tượng thuộc về. Lưu ý đây là một mô hình tuyến tính. Trong nhiều trường hợp, các đặc trưng trong tập dữ liệu Breast Cancer Wisconsin có mối quan hệ tương đối tuyến tính với nhãn phân loại. Điều này có nghĩa là các mô hình tuyến tính như Logistic Regression có thể hoạt động tốt trong việc phân loại. Mặc dù có thể có một số mức độ gần như tuyến tính, nhưng cũng cần lưu ý rằng mối quan hệ giữa các đặc trưng và nhãn phân loại có thể phức tạp hơn một mối quan hệ tuyến tính đơn giản. Và đây chính là lí do mô hình Multi-layer Perceptron cho kết quả tốt nhất. Nghĩa

Bảng I
ĐÁNH GIÁ TRÊN TRAINING SET.

Phương pháp	Độ chính xác trung bình	Độ lệch chuẩn
Logistic Regression	0.969	0.021
K-Nearest Neighbors	0.960	0.020
Support Vector Machine	0.969	0.025
Multi-layer Perceptron	0.960	0.036

Bảng II
CÁC PHƯƠNG PHÁP VÀ KẾT QUẢ TRÊN PRIVATE TEST.

Phương pháp	Tiền xử lý					Score
	Scale	*PCA*	*High corr*	*Outlier*	*Balance*	
Logistic Regression	✓		✓	✓		86.71
K-Nearest Neighbors	✓		✓	✓	✓	79.02
Support Vector Machine	✓	✓		✓	✓	77.93
Multi-layer Perceptron	✓			✓		88.9

là, mô hình không bị "bias" với các mối quan hệ gần tuyến tính, mà mô hình còn học được các mối quan hệ phi tuyến.

Nếu trong tương lai, tập dữ liệu được phát triển đủ lớn, các biến động về dữ liệu ít đi, thì các mô hình tuyến tính sẽ là sự lựa chọn tốt. Tuy nhiên, việc giả định sự tuyến tính như vậy là bất khả thi. Còn nếu dữ liệu phát triển theo hướng di chuyển của phân phối như hiện tại, thì việc chọn các mô hình trên cơ sở Neural Network như Multi-layer Perceptron là tất yếu.

V. KẾT LUẬN

Trong bài báo cáo, chúng tôi đã giải quyết bài toán phân loại trên tập dữ liệu Breast Cancer Wisconsin với 4 phương pháp là Logistic Regression, K-Nearest Neighbors, Support Vector Machine và Multi-layer Perceptron. Phương pháp cuối cùng, Multi-layer Perceptron, cho kết quả đầy hứa hẹn. Còn với K-Nearest Neighbors, do tính đơn giản của thuật toán nên kết quả sau cùng chưa khả quan. Ngoài ra, chúng tôi cũng áp dụng một số phương pháp tiền xử lý như scale dữ liệu, xử lý ngoại lệ, loại bỏ đặc trưng ít quan trọng, Principal Component Analysis để trích xuất đặc trưng tốt hơn, giảm nhiễu và tăng tốc quá trình thực nghiệm. Trong tương lai, chúng tôi sẽ mở rộng phạm vi ứng dụng của bài toán sang thế giới thực, hỗ trợ cho ngành y học. Chúng tôi cũng có kế hoạch áp dụng thuật toán di truyền (Genetic Algorithm) để tìm ra tập các đặc trưng tối ưu để cải thiện hiệu xuất mô hình.

TÀI LIỆU

[1] Wolberg, William, Mangasarian, Olvi, Street, Nick, and Street, "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository.

[2] F. J. Anscombe, "The Validity of Comparative Experiments," Journal of the Royal Statistical Society. Series A (General), vol. 111, no. 3, 1948, pp. 181–211.

[3] Ian T. Jolliffe, Jorge Cadima, "Principal component analysis: a review and recent developments," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.

[4] Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, Ludolf Erwin Meester, "A Modern Introduction to Probability and Statistics," Springer, London.

[5] Peng, Chao-Ying Joanne, "An Introduction to Logistic Regression Analysis and Reporting," The Journal of Educational Research, vol. 96, no. 1, 2002, pp. 3–14.

[6] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, Kieran Greer , "KNN Model-Based Approach in Classification," Lecture Notes in Computer Science, vol 2888. Springer, Berlin, Heidelberg.

[7] Corinna Cortes, Vladimir Vapnik, "Support-vector networks," Machine Learning 20, 273–297 (1995).

[8] Hassan Ramchoun, Mohammed Amine Janati Idrissi, Youssef Ghanou, Mohamed Ettaouil, "Multilayer Perceptron: Architecture Optimization and Training" International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 4.

[9] Marc Peter Deisenroth, "Mathematics for Machine Learning," Cambridge University Press; 1st edition (April 23, 2020).