Báo cáo đánh giá Metric các thuật toán ML-Classfication

Lê Văn Thức

Ngày 30 tháng 3 năm 2025

1 Giới thiệu

Machine Learning (Học máy) là một lớp các thuật toán và mô hình cơ bản giúp máy tính học hỏi từ dữ liệu. ML được sử dụng rộng rãi trong nhiều lĩnh vực nhờ khả năng tự động tìm ra các mẫu ẩn và đưa ra dự đoán. Để đánh giá hiệu suất của các mô hình ML, chúng ta sử dụng các metric như Accuracy, Precision, Recall, F1-Score và ROC-AUC.

Ở báo cáo thực nghiệm này, chúng ta sẽ thử nghiệm các mô hình ML cơ bản sau: Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, và Random Forest của scikit trên tập dữ liệu chẩn đoán tiểu đường nhằm đánh giá các mô hình thông qua các metric đã nêu và hiểu sâu hơn về vai trò của từng metric trong việc đo lường hiệu quả dự đoán.

2 Tập dữ liệu

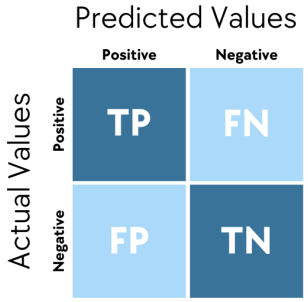
Tập dữ liệu được sử dụng là tập dữ liệu chẩn đoán bệnh tiểu đường gồm 768 mẫu, mỗi mẫu gồm 8 cột tiêu chí đánh giá là: "Pregnancies", "Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI", "DiabetesPedigreeFunction", "Age"và một cột kết quả: "Outcome"tương ứng.

Thông qua đánh giá tập dữ liệu, ta nhận thấy có nhiều ô có giá trị là 0(ví dụ như cột SkinThickness có tận 227 giá trị là 0, mặc dù về mặc logic thì các ô này là 0 thì không hợp lý. Vì số lượng dữ liệu thiếu quá lớn nên nếu bỏ đi các hàng này thì dữ liệu sẽ bị mất mát nhiều, thay vào đó ta sẽ impute các ô trống bằng giá trị trung vị của các cột.

Tập dữ liệu được tách thành 2 phần train và test riêng biệt, tập train gồm 614 mẫu với 213 mẫu là 1, tập test gồm 154 mẫu với 55 mẫu là bị.

3 Metric

Trong bài toán classification 0-1 thì khái niệm confusion matrix và các ô trên đó là trọng tâm của các metric dưới đây.



Hình 1: Confusion Matrix

3.1 Accuracy Score

Accuracy Score đo lường tỷ lệ dự đoán đúng trên tổng số dự đoán, được tính bằng công thức:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+TN+FP}}$$

Metric này cung cấp một cái nhìn tổng quát về hiệu suất của mô hình, đặc biệt hiệu quả khi phân bố các lớp được cân bằng.

3.2 Precision Score

Độ chính xác dương cho biết tỷ lệ các trường hợp dương tính đúng trong tổng số các dự đoán dương tính mà mô hình đưa ra:

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}}$$

Chỉ số này rất quan trọng khi chi phí của dự đoán dương tính sai là cao.

3.3 Recall Score

Độ nhạy, còn được gọi là sensitivity, đo lường tỷ lệ các trường hợp dương tính thực sự mà mô hình nhận diện đúng:

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}}$$

Metric này đặc biệt quan trọng khi việc bỏ sót các trường hợp dương tính có thể gây hậu quả nghiêm trọng. Cụ thể trong trường hợp này, khi chẩn đoán các bệnh nghiệm trọng như tiểu đường, bạn sẽ muốn Recall Score cao nhất có thể.

3.4 F1-Score

F1-Score là trung bình điều hòa của Precision và Recall, cung cấp một sự cân bằng giữa hai chỉ số này:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Chỉ số này hữu ích khi làm việc với dữ liệu mất cân bằng, giúp đánh giá toàn diện hơn về hiệu suất của mô hình.

3.5 ROC-AUC

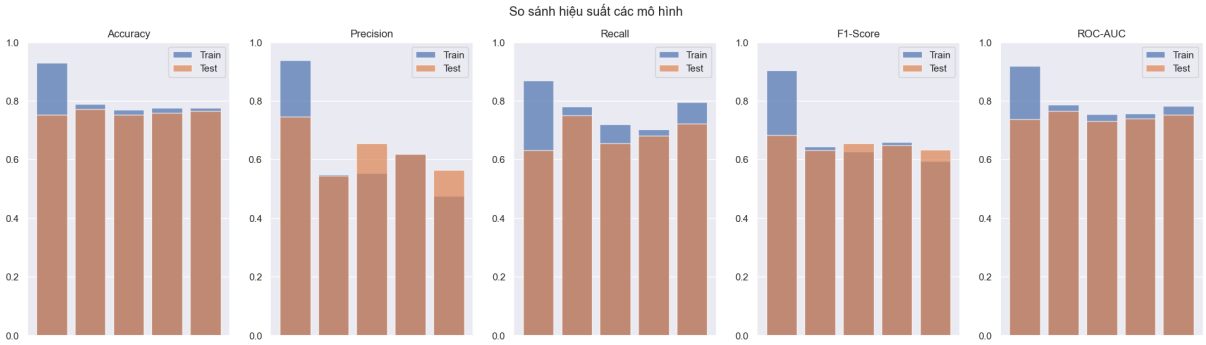
ROC-AUC (Receiver Operating Characteristic - Area Under Curve) đánh giá khả năng phân biệt giữa các lớp của mô hình. Nó được tính bằng diện tích dưới đường cong ROC, trong đó đường cong này vẽ tỷ lệ dương tính thật (Recall) so với tỷ lệ dương tính giả ở các ngưỡng khác nhau. Giá trị ROC-AUC càng cao thì mô hình càng có khả năng phân biệt tốt giữa các lớp.

4 Kết quả thực nghiệm

Vì các mô hình cơ bản nên các mô hình chạy tương đối nhanh. Chúng ta sẽ không so sánh thời gian chạy mà tập trung vào các metric là chính. Bên dưới là kết quả đạt được sau khi chạy 5 mô hình kể trên.

Bảng 1: Kết quả thực nghiệm

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logres	Train: 0.774 Test: 0.747	0.563 0.673	0.723 0.638	0.633 0.655	0.758 0.725
SVC	Train: 0.769 Test: 0.766	0.474 0.564	0.771 0.721	0.587 0.633	0.770 0.752
KNN	Train: 0.798 Test: 0.662	0.638 0.582	0.743 0.525	0.687 0.552	0.782 0.639
DT	Train: 1.000 Test: 0.740	1.000 0.691	1.000 0.623	1.000 0.655	1.000 0.720
RF	Train: 1.000 Test: 0.753	1.000 0.655	1.000 0.655	1.000 0.655	1.000 0.731



Hình 2: Confusion Matrix

Để đảm bảo tính ổn định cho các mô hình, ta sẽ tiến hành Scaling Data bằng RobustScaler và Tuning các mô hình bằng GridSearch.

Bảng 2: Kết quả thực nghiệm sau khi tuning và scaling

Model	Accuracy	Precision	Recall	F1	ROC-AUC
RF	Train: 0.930	0.939	0.870	0.903	0.918
	Test: 0.753	0.745	0.631	0.683	0.737
KNN	Train: 0.795	0.526	0.818	0.640	0.803
	Test: 0.740	0.473	0.703	0.565	0.727
Logres	Train: 0.772	0.568	0.716	0.634	0.755
	Test: 0.753	0.673	0.649	0.661	0.732
DT	Train: 0.777	0.620	0.702	0.658	0.756
	Test: 0.760	0.618	0.680	0.648	0.739
SVC	Train: 0.805	0.493	0.897	0.636	0.840
	Test: 0.779	0.527	0.784	0.630	0.781

5 Kết luận

Accuracy của các mô hình ở mức ổn tuy nhiên vì dữ liệu là chẩn đoán bệnh và bên Outcome No nhiều hơn Yes nên đánh giá bằng Recall, F1, ROC-AUC là các lựa chọn tốt hơn.

Kết quả thu được từ thực nghiệm trước khi tuning và scaling cho thấy Random Forest và Decision Tree bị overfit đáng kể khi mọi score trên tập train luôn là 1 trong khi kết quả test thì chưa tốt lắm.

Tổng thể trước khi tuning và scaling thì SVC cho kết quả tốt nhất, Logres cho hiệu suất chỉ tương đương với 2 mô hình bị overfit là Decision Tree và Random Forest trong khi KNN cho kết quả kém hơn cả.

Sau khi tuning và scaling, Metric của các mô hình đều được thiện đáng kể, đặc biệt là với KNN, với điểm ROC-AUC và Recall được cải thiện rất đáng kể, trở nên phù hợp hơn với việc dự đoán bệnh, tuy nhiên Precision lại rất thấp.

Bên cạnh đó, SVC có điểm ROC-AUC cực cao, với điểm recall vượt trội so với các mô hình còn lại, cực kỳ phù hợp cho việc dự đoán bệnh tiểu đường.

2 mô hình cây là Decision Tree và Random Forest đã ít bị overfit hơn, đặc biệt là Random Forest với điểm cao nhất theo Metric F1 bởi precision lớn hơn đáng kể so với các mô hình khác, tuy vậy điểm Recall lại thấp hơn hẳn với các mô hình còn lại, nên tính ứng dụng thực tế chưa cao.