Báo Cáo Đồ Án Cuối Kì CS116

Nguyễn Tuấn Kiệt, Hoàng Minh Hiếu, Nguyễn Nhật Nam
Khoa Khoa học máy tính, Trường Đại học Công nghệ Thông tin
Đại học Quốc gia Thành phố Hồ Chí Minh
{21521042, 21520232, 21521160}@gm.uit.edu.vn

Abstract

Báo cáo trình bày toàn diện về phương pháp, kĩ thuật để giải quyết bài toán phân loại bằng các phương pháp máy học. Sử dụng 2 tập dữ liệu Diabetes và Compas, nhóm chúng em sẽ đi qua các phần: khám phá và phân tích dữ liệu, tiền xử lí dữ liệu, thiết kế đặc trưng, lựa chọn mô hình, trình bày kết quả tốt nhất trên tập dev mà nhóm đạt được.

1 Exploratory Data Analysis

Trong dữ liệu bảng, thông thường các loại đặc trưng sẽ được phân chia thành các loại sau:

- **Dạng số học** (**numeric**): Đây là loại đặc trưng có giá trị liên tục, có thể là số thực hoặc số nguyên. Trong quá trình huấn luyện mô hình máy học, các đặc trưng số học được coi là có tính thứ tự, tức là có mối quan hệ so sánh giữa chúng. Ví dụ, số 3 gần với số 4 hơn là số 10 vì $3 < 4 < 10$.
- **Dạng phân loại** (**categorical**): Đây là loại đặc trưng có các giá trị thuộc vào một tập hữu hạn và có số lượng ít hơn so với số dòng trong tập dữ liệu. Ví dụ, trong đặc trưng **sex** của tập Compas chỉ có 2 giá trị là "**Male**" và "**Female**". Đặc trưng phân loại có thể là số học, văn bản hoặc cả hai. Ví dụ, một đặc trưng phân loại có thể chứa tập giá trị như ("**Male**", 1.0, "**Female**", 5, "**Dog**").
- **Dạng nhị phân** (**binary**): Đây là trường hợp đặc biệt của đặc trưng phân loại khi chỉ có 2 giá trị. Ví dụ, đặc trưng **sex** được mô tả ở trên là một ví dụ cho loại đặc trưng này.
- **Dạng văn bản** (**text**): Đây là loại đặc trưng chứa các giá trị không phải số, thường là ngôn ngữ tự nhiên. Trong một số trường hợp, tất cả các giá trị của đặc trưng này đều là duy nhất và không có sự trùng lặp giữa chúng.
- **Dạng thời gian ngày tháng** (**datetime**): Đây là loại đặc trưng chứa thông tin về ngày tháng và thời gian.

1.1 Compas

Train	Dev	Số cột	Cột mục tiêu	Loại bài toán
5049	721	9	two_year_recid	Phân loại nhị phân

Table 1: Thông tin tập dữ liệu Compas.

Table 1 cho thấy thông tin về tập dữ liệu Compas chứa 5049 mẫu dữ liệu cho quá trình huấn luyện và 721 mẫu dữ liệu cho quá trình kiểm định. Với 9 cột trong đó là 8 cột đặc trưng: **sex**, **age**, **race**, **juv_fel_count**, **juv_misd_count**, **juv_other_count**, **priors_count**, **c_charge_degree** và 1 cột mục tiêu: **two_year_recid**. Nhóm cũng định nghĩa tất cả các đặc trưng trong tập dữ liệu đều là **dạng phân loại**

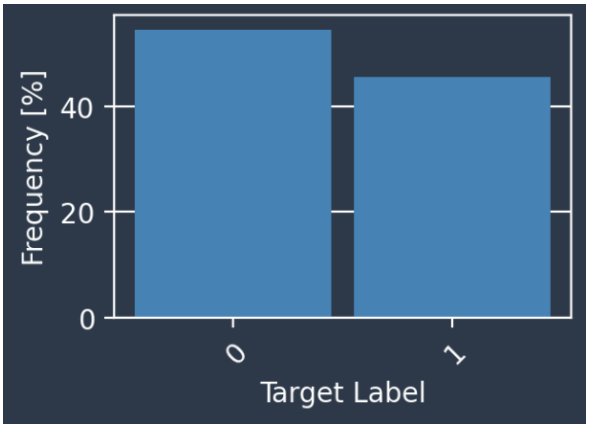

Figure 1: Phân phối nhãn của tập Compas.

Figure 1 thể hiện phân phối các nhãn của tập dữ liệu Compas với nhãn **0** chiếm 54.49% và nhãn **1** chiếm 45.51%.

Ở Figure 2, có thể thấy độ tương quan của các đặc trưng trong tập dữ liệu so với cột mục tiêu rất là thấp khi độ tương quan cao nhất có giá trị chỉ là 0.27 là cột **priors_count**, có thể thấy nếu không tiền xử lí cẩn thận và làm tốt giai đoạn thiết kế đặc

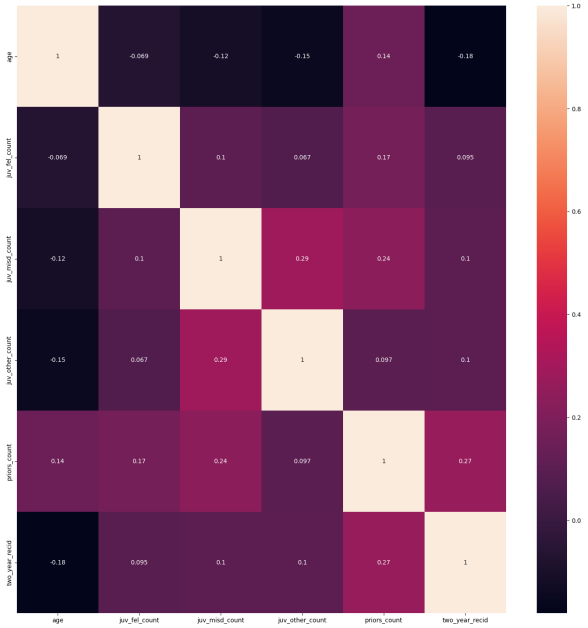

Figure 2: Độ tương quan giữa các feature trong tập Compas.

trưng, rất khó để có thể thiết lập một hiệu suất ở mức ổn cho tập dữ liệu này hoặc cải thiện hiệu suất.

Figure 5 đã trực quan hóa các miền giá trị của từng đặc trưng, thông qua đó ta có thể quyết định đặc trưng nào nên là dạng phân loại và đặc trưng nào nên là dạng số học dễ dàng hơn.

1.2 Diabetes

Train	Dev	Số cột	Cột mục tiêu	Loại bài toán
537	77	9	Outcome	Phân loại nhị phân

Table 2: Thông tin tập dữ liệu Diabetes

Table 2 cho thấy thông tin về tập dữ liệu Diabetes chứa 537 mẫu dữ liệu cho quá trình huấn luyện và 77 mẫu dữ liệu cho quá trình kiểm định. Với 9 cột trong đó là 8 cột đặc trưng: **Pregnancies**, **Glucose**, **BloodPressure**, **SkinThickness**, **Insulin**, **BMI**, **DiabetesPedigreeFunction**, **Age** và 1 cột mục tiêu: **Outcome**.

Về vấn đề loại đặc trưng, nhóm quyết định tất cả các đặc trưng trong tập dữ liệu đều là **dạng số học**.

Figure 3 thể hiện phân phối các nhãn của tập dữ liệu Compas với nhãn **0** chiếm 64.98% và nhãn **1** chiếm 35.02%.

So với Compas, tương quan giữa cột mục tiêu so với các đặc trưng còn lại khá cao, điều này giúp giảm bớt gánh nặng lên quá trình tiền xử lí và thiết kế đặc trưng trong các giai đoạn tiếp theo.

Tương tự với Compas, Figure 6 cũng giúp trực

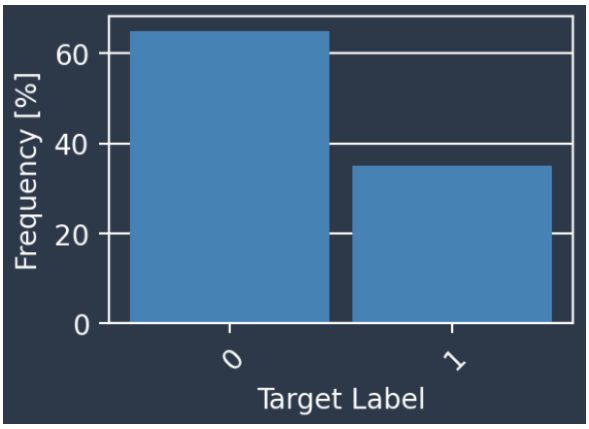

Figure 3: Phân phối nhãn của tập Diabetes.

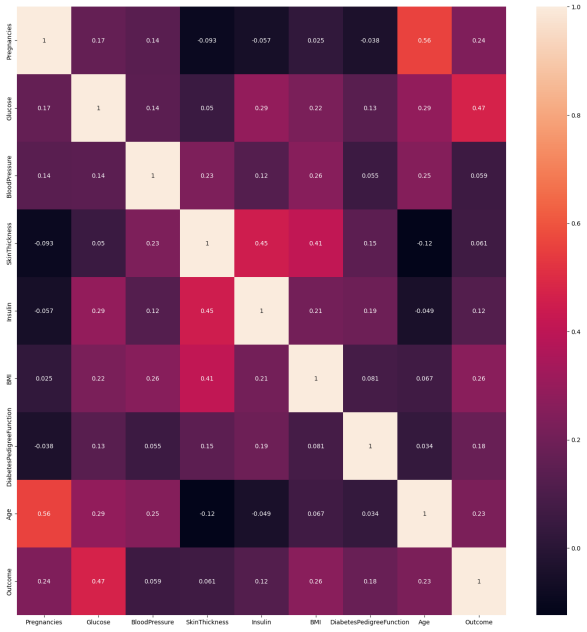

Figure 4: Độ tương quan giữa các feature trong tập Diabetes.

quan hóa miền giá trị của các đặc trưng và có cùng mục đích với Figure 5

2 Data Preprocessing

Trong quá trình tiền xử lý dữ liệu của cả hai tập Compas và Diabetes, nhóm chúng em đã sử dụng SimpleImputer từ thư viện scikit-learn để xử lý dữ liệu thiếu theo hai chiến lược chính. Đối với đặc trưng số học, nhóm đã sử dụng phương pháp điền giá trị trung vị (median). Trong khi đó, đối với các đặc trưng phân loại, nhóm đã chọn giá trị có tần suất xuất hiện cao nhất để điền vào các giá trị thiếu.

Đặc biệt, khi xử lí tập Diabetes và quan sát biểu đồ phân phối đặc trưng như Figure 6, nhóm nhận thấy rằng các đặc trưng như **BloodPressure, BMI, Glucose, Insulin, SkinThickness** thường có nhiều

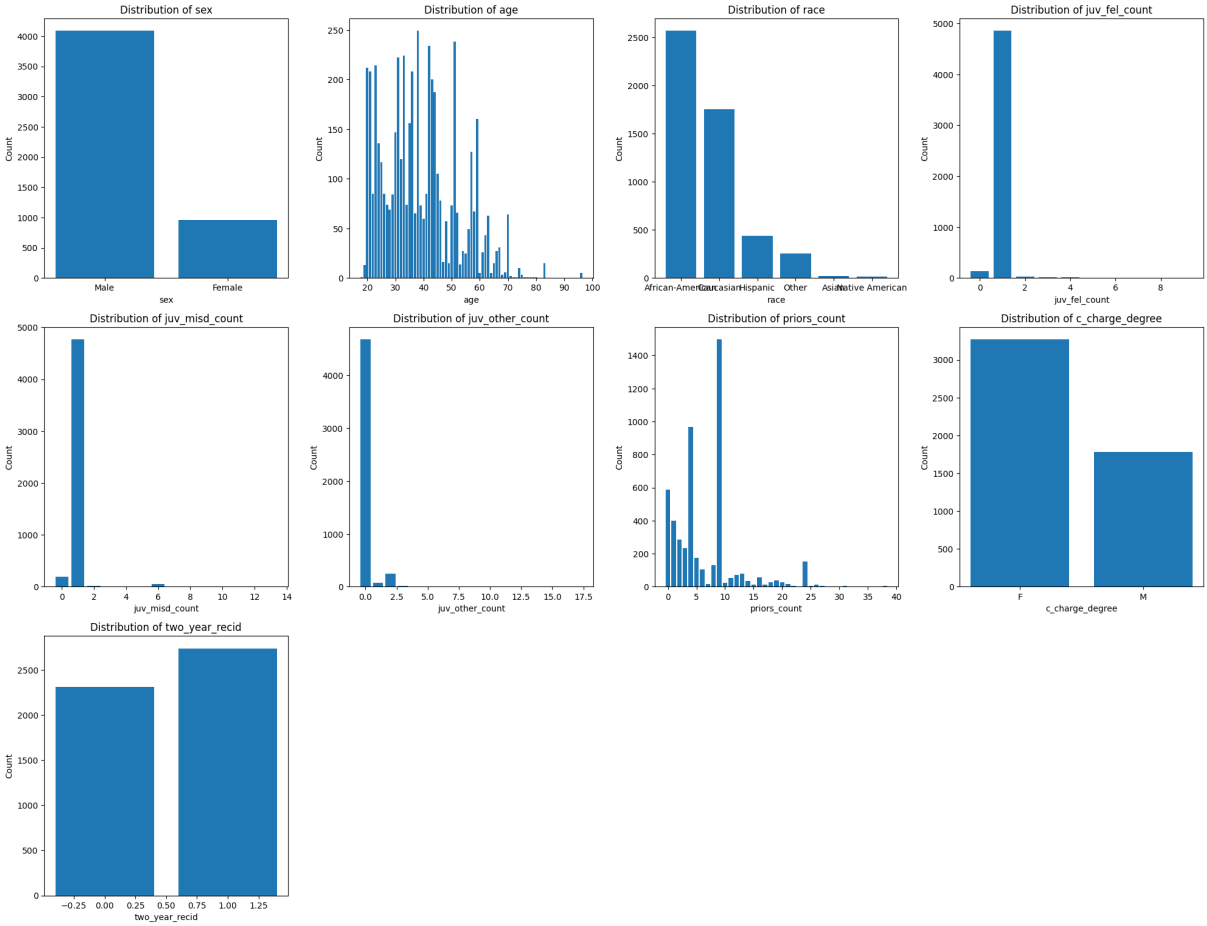

Figure 5: Phân phối của các đặc trưng trong tập dữ liệu Compas.

giá trị 0. Sau khi tìm hiểu, nhóm quyết định xem xét các giá trị 0 này là missing value và thay thế chúng bằng giá trị trung vị.

Để xử lý các outlier, nhóm đã sử dụng z-score để quyết định liệu có nên loại bỏ các mẫu dữ liệu outlier hay không, và quá trình này được thực hiện trên tập huấn luyện.

Nhóm cũng đã thực hiện chuẩn hóa dữ liệu bằng cách sử dụng StandardScaler từ thư viện scikit-learn cho các đặc trưng số học.

Cuối cùng, để cân bằng phân phối của các nhãn, nhóm đã sử dụng phương pháp SMOTE.

3 Feature & Model Selection

3.1 Feature Engineering

Nhóm chúng em chủ yếu sử dụng các kĩ thuật decomposition và feature selection như Factor Analysis, mutual information, Variance Threshold từ thư viện scikit-learn để tạo thêm đặc trưng mới.

Ngoài ra, chúng em cũng áp dụng việc tạo ra các đặc trưng mới bằng cách tính tỉ lệ giữa hai đặc trưng gốc. Ví dụ, trong tập Diabetes, chúng em tạo ra đặc trưng mới **Glucose_to_BMI** bằng cách tính tỉ lệ giữa **Glucose** và **BMI**.

Trong quá trình xử lí tập Diabetes, chúng em thiết kế các ngưỡng cho các đặc trưng **Glucose, BloodPressure, BMI** để tạo ra các đặc trưng phân loại. Các ngưỡng này được xác định dựa trên kiến thức khoa học và nghiên cứu trên mạng. Tuy nhiên, đối với tập Compas, do thiếu thông tin và không thể dự đoán được ý nghĩa của các đặc trưng, chúng em quyết định không áp dụng phương pháp này.

Để đa dạng thêm các đặc trưng, nhóm đã tận dụng 2 công cụ của scikit-learn là QuantileTransformer và KBinsDiscretizer. Các tác dụng của chúng như sau:

- QuantileTransformer: giúp chuẩn hóa dữ liệu sao cho phân phối của nó gần với phân phối chuẩn, giảm thiểu ảnh hưởng của nhiễu và giúp mô hình dễ dàng học được các mối quan hệ phức tạp trong dữ liệu.

- KBinsDiscretizer: giúp chuyển đổi đặc trưng số học thành dạng phân loại, làm giảm ảnh hưởng của nhiễu và giúp mô hình dễ dàng học được các mối quan hệ non-linear trong dữ liệu.

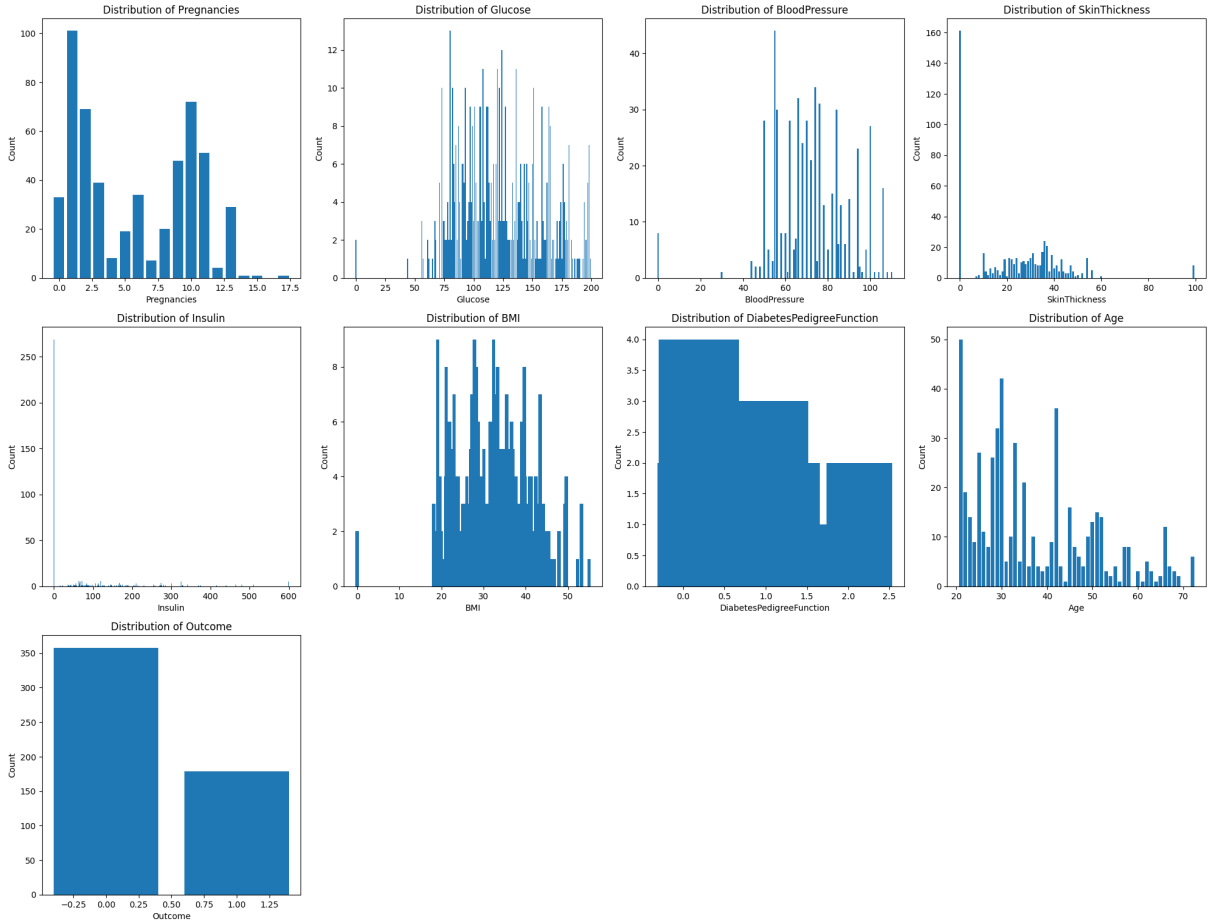

Figure 6: Phân phối của các đặc trưng trong tập dữ liệu Diabetes.

3.2 Model Selection

Trong giai đoạn xây dựng và phát triển mô hình, nhóm chủ yếu sử dụng các mô hình ensemble của cây như CatBoost, Gradient Boosting và XGBoost như hướng tiếp cận chính để giải quyết các bài toán trên 2 tập dữ liệu.

- Gradient Boosting: là một kĩ thuật ensemble với ý tưởng sử dụng các mô hình underfit để tạo ra một mô hình hiệu suất cao.

- CatBoost: một thuật toán gradient boosting được thiết kế để chủ yếu tận dụng tốt các đặc trưng phân loại một cách hiệu quả.

- XGBoost: bản cải tiến của gradient boosting cường hóa về tốc độ, hiệu suất (tính toán và lưu trữ).

3.3 Hyperparameters Optimization

Trong giai đoạn tiếp theo, nhóm sử dụng optuna - một framework sử dụng để tìm kiếm bộ siêu tham số của mô hình sao cho tối ưu nhất trên tập kiểm định theo thang đo F1.

4 Result & Discussion

Dataset	Set	F1	Accuracy
Compas	Dev	69.0565	69.3481
Compas	Test	67.7601	68.0055
Diabetes	Dev	85.7046	85.7143
Diabetes	Test	73.7249	75.3247

Table 3: Bảng kết quả.

5 Conclusion

Theo như kết quả ở Table 3, kết quả trên tập kiểm định và tập kiểm thử của tập dữ liệu Compas tuy không quá cao nhưng cũng không quá chênh lệch cho thấy mô hình mà nhóm sử đụng đã đạt được một mức khái quát hóa nhất định. Còn về tập dữ liệu Diabetes, kết quả rất chênh lệch giữa tập kiểm định và tập kiểm thử, theo phán đoán của nhóm, có rất nhiều nguyên nhân dẫn đến trường hợp này:

- Khả năng khái quát hóa thấp: có thể trong quá trình huấn luyện, đã có sự overfit trên tập kiểm định

dẫn đến sự giảm hiệu suất trên tập kiểm thử.

- Chiến lược xử lí dữ liệu thiếu: do dữ liệu thiếu chỉ xuất hiện ở tập kiểm thử mà không có ở tập kiểm định nên chiến lược xử lí mà nhóm đưa ra đã có phần tác động lên hiệu suất trên tập kiểm thử.

- Sự khác nhau với phân phối dữ liệu: có thể tập kiểm định và kiểm thử có sự chênh lệch nhau về phân phối dữ liệu, trong quá trình huấn luyện, ngoài tối ưu trên tập huấn luyện, mô hình cũng được tối ưu siêu tham số trên tập kiểm thử nên nếu trường hợp này xảy ra thì không khó đoán khi mà có sự giảm hiệu suất đáng kể trên tập kiểm thử.