

Báo cáo Bài tập lớn

Hiệu suất và ứng dụng của Decision Tree, Naive Bayes và Linear Regression trong phân loại khách hàng về thẻ tín dụng

GVHD: Th.S Hà Lê Hoài Trung

Lớp: IS252.N21 - Data Mining

Nhóm SVTH:

- *Lê Nguyễn Minh Trung - 19521061*
- *Võ Nữ Diễm Trang - 20521013*
- *Trần Gia Phong - 20521748*

NỘI DUNG TRÌNH BÀY

Giới thiệu đề tài

- Lý do; Mục tiêu; Bài toán giải quyết

Cơ sở lý thuyết

- Decision Tree, Naive Bayes & Linear Regression

Mô tả tập dữ liệu

- Mô tả; Phân bổ dữ liệu; Lựa chọn thuộc tính để huấn luyện

Thực nghiệm

- Mô hình; Tiền xử lý dữ liệu; Chạy mô hình và dự đoán

Kết quả & So sánh

Kết luận

01

GIỚI THIỆU ĐỀ TÀI

Tổng quan về đề tài

- *Đề tài tập trung vào nghiên cứu và so sánh hiệu suất của ba giải thuật quan trọng: Decision Tree, Naive Bayes và Linear Regression.*
- *Đây là một đề tài mang tính ứng dụng, nhằm tạo ra sự thực nghiệm và hiểu rõ hơn về khả năng áp dụng của các giải thuật trong lĩnh vực Data Mining.*

Lý do làm đề tài

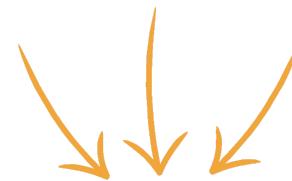
- Xuất phát từ sự tò mò và nhu cầu hiểu biết sâu hơn về Data Mining, đặc biệt là việc áp dụng các giải thuật phân loại lên các tập dữ liệu thực tế.
- Cơ hội để thực nghiệm các phương pháp và công cụ trong lĩnh vực này, và giúp xây dựng một nền tảng kiến thức rộng và sâu hơn về Data Mining cho tương lai.

Mục tiêu khi thực hiện đề tài

- *Nghiên cứu và so sánh hiệu suất của ba giải thuật quan trọng trong việc phân loại trên tập dữ liệu.*
- *Kết luận về sự hiệu quả và ứng dụng của từng giải thuật & định hướng lựa chọn giải thuật phù hợp trong các tình huống và dataset khác nhau.*



*Để đạt được mục tiêu: Chọn một bài toán phân lớp cụ thể
là "Dự đoán trạng thái của khách hàng về việc rời bỏ hoặc
tiếp tục sử dụng dịch vụ thẻ tín dụng".*



Bài toán giải quyết trong đề tài

02

CƠ SỞ LÝ THUYẾT

Decision Tree

Là một thuật toán học máy không giám sát, được sử dụng cho phân loại và hồi quy

3 công việc chính khi triển khai thuật toán Decision Tree:

- Xây dựng cây quyết định
- Cắt tỉa cây quyết định
- Đánh giá cây quyết định

Naive Bayes

- Là một giải thuật học máy có giám sát, được sử dụng cho các công việc phân lớp như phân lớp văn bản, ...
- Nó cũng là một phần của generative learning algorithms, điều đó đồng nghĩa nó tìm kiếm để mô hình hóa sự phân tán của dữ liệu đầu vào của một lớp hoặc loại được cho.
- Là một thuật đơn giản, dễ tiếp cận đối với những người mới bắt đầu với học máy

Linear Regression

- Là một trong những thuật toán học máy, dựa trên kỹ thuật học tập có giám sát.
- Là một trong những thuật toán được biết đến rộng rãi và dễ hiểu ngay cả bởi người không quen thuộc với thuật toán máy học.
- Thuật toán thực hiện hồi quy, xác định mối quan hệ giữa hai biến bằng cách khớp đường hồi quy với dữ liệu.
- Một trong hai biến là biến phụ thuộc và phụ thuộc vào biến còn lại, biến đó được gọi là biến độc lập

03

MÔ TẢ TẬP DỮ LIỆU

Mô tả tập dữ liệu

- *Tên: Predicting Credit Card Customer Segmentation*
- *Nguồn: Kaggle*
- *Số dòng dữ liệu: 10127*
- *Số cột thuộc tính: 23*

The screenshot shows the Kaggle platform interface. On the left, there's a sidebar with navigation links: Home, Competitions, Datasets (which is selected), Models, Code, Discussions, Learn, and More. The main content area displays the dataset 'Predicting Credit Card Customer Segmentation'. At the top, there's a search bar, a sign-in button, a register button, and a file download button labeled 'Download (388 kB)'. Below the download button is a large blue credit card icon with a yellow letter 'E' on it. The dataset title 'Predicting Credit Card Customer Segmentation' is prominently displayed, along with the subtitle 'Exploring Key Customer Characteristics'. There are three tabs at the bottom: 'Data Card' (selected), 'Code (16)', and 'Discussion (0)'. To the right of the tabs, there are sections for 'Usability' (rating 10.00), 'License' (CC0: Public Domain), and 'Expected update frequency' (Never). The overall background features abstract purple and blue shapes.

Bảng mô tả các thuộc tính

STT	Thuộc tính	Kiểu dữ liệu	Mô tả
1	Clientnum	Int	Số định danh của khách hàng
2	Attrition_Flag	Boolean	Cho biết khách hàng có rời bỏ hay không.
3	Customer_Age	Int	Tuổi của khách hàng
4	Gender	String	Giới tính của khách hàng
5	Dependent_count	Int	Số người phụ thuộc mà khách hàng
6	Education_level	String	Trình độ học vấn của khách hàng
7	Marital_Status	String	Tình trạng hôn nhân của khách hàng
8	Income_Category	String	Loại thu nhập của khách hàng
9	Card_Category	String	Loại thẻ của khách hàng
10	Months_on_book	Int	Số tháng khách hàng đã sử dụng dịch vụ thẻ tín dụng tính từ khi khách hàng mở tài khoản

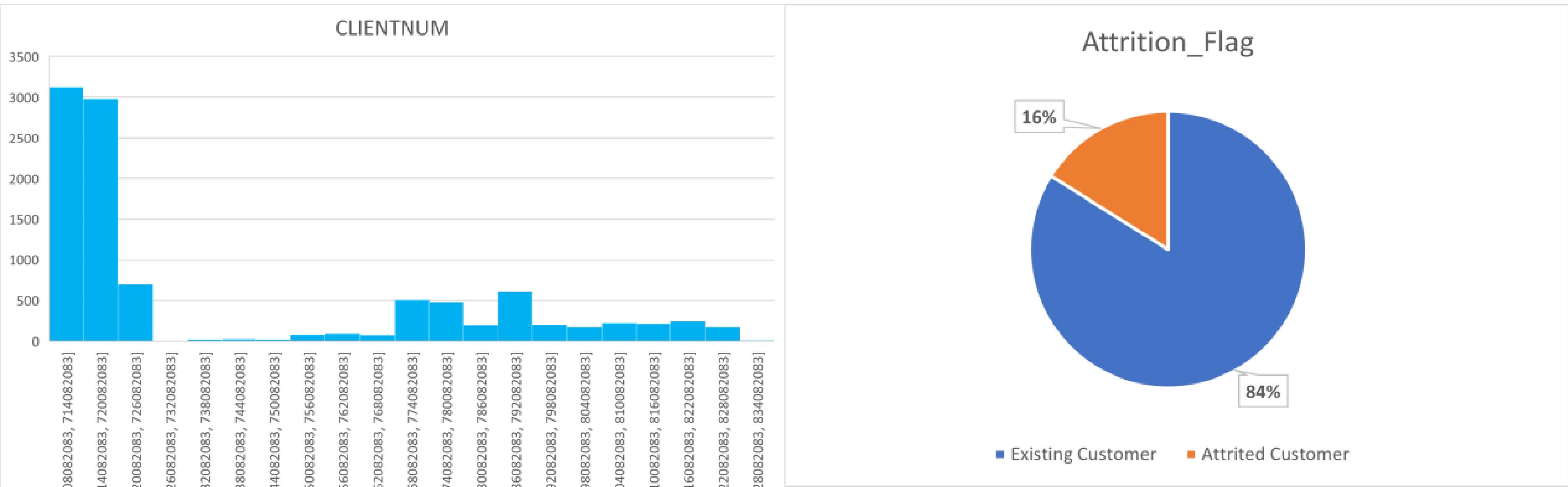
Bảng mô tả các thuộc tính

10	Months_on_book	Int	Số tháng khách hàng đã sử dụng dịch vụ thẻ tín dụng tính từ khi khách hàng mở tài khoản
11	Total_Relationship_Count	Int	Tổng số mối quan hệ giữa khách hàng và nhà cung cấp thẻ tín dụng
12	Months_Inactive_12_mon	Int	Số tháng khách hàng không hoạt động trong 12 tháng qua
13	Contacts_Count_12_mon	Int	Số lượng liên lạc của khách hàng trong 12 tháng qua
14	Credit_Limit	Int	Hạn mức tín dụng của khách hàng
15	Total_Revolving_Bal	Int	Tổng số dư quay vòng của khách hàng
16	Avg_Open_To_Buy	Int	Tỷ lệ mua trung bình của khách hàng
17	Total_Amt_Chng_Q4_Q1	Int	Tổng số tiền thay đổi từ quý 4 sang quý 1

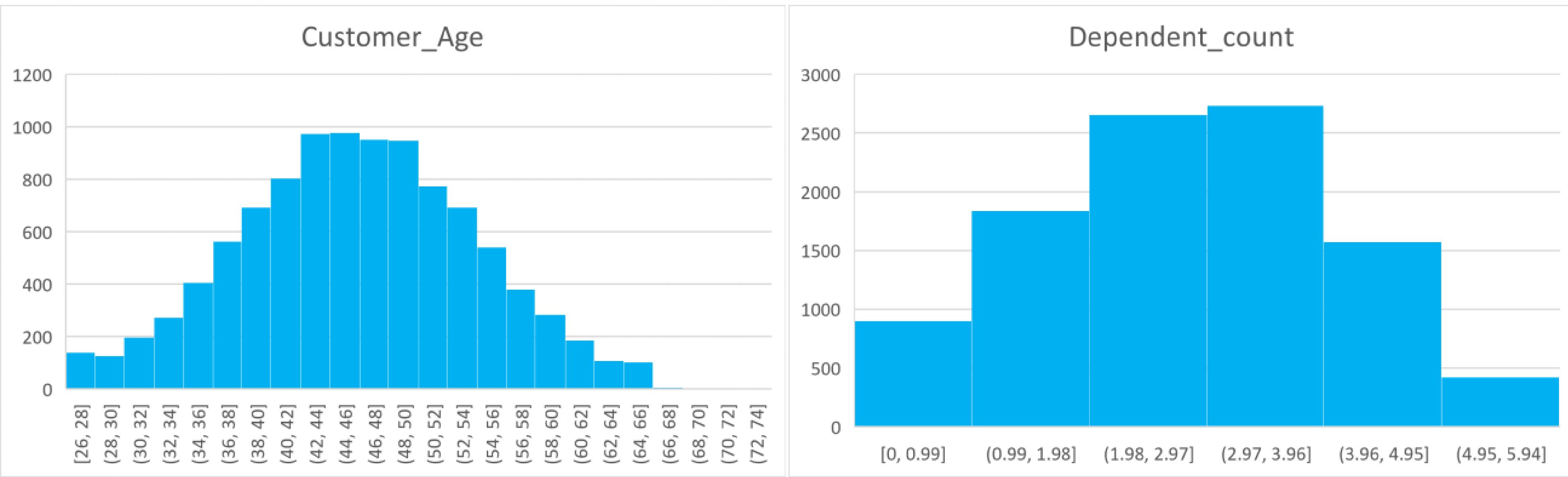
Bảng mô tả các thuộc tính

18	Total_Trans_Amt	Int	Tổng số tiền giao dịch
19	Total_Trans_Ct	Int	Tổng số lượng giao dịch
20	Total_Ct_Chng_Q4_Q1	Int	Tổng số thay đổi từ quý 4 sang quý 1
21	Avg_Utilization_Ratio	Int	Tỷ lệ sử dụng trung bình của khách hàng
22	Naive_Bayes_Classifier_Attrition _Flag_Card_Category_Contacts_ Count_12_mon_Dependent_count _Education_Level_Months_Inacti ve_12_mon_1	Float	Sử dụng Naïve Bayes để dự đoán liệu ai đó có rời đi hay không dựa trên các đặc điểm.
23	Naive_Bayes_Classifier_Attrition _Flag_Card_Category_Contacts_ Count_12_mon_Dependent_count _Education_Level_Months_Inacti ve_12_mon_2	Float	Sử dụng Naïve Bayes phân loại loại thẻ của khách hàng dựa vào số lượng liên lạc của khách hàng trong 12 tháng qua và trình độ học vấn.

Phân bố dữ liệu

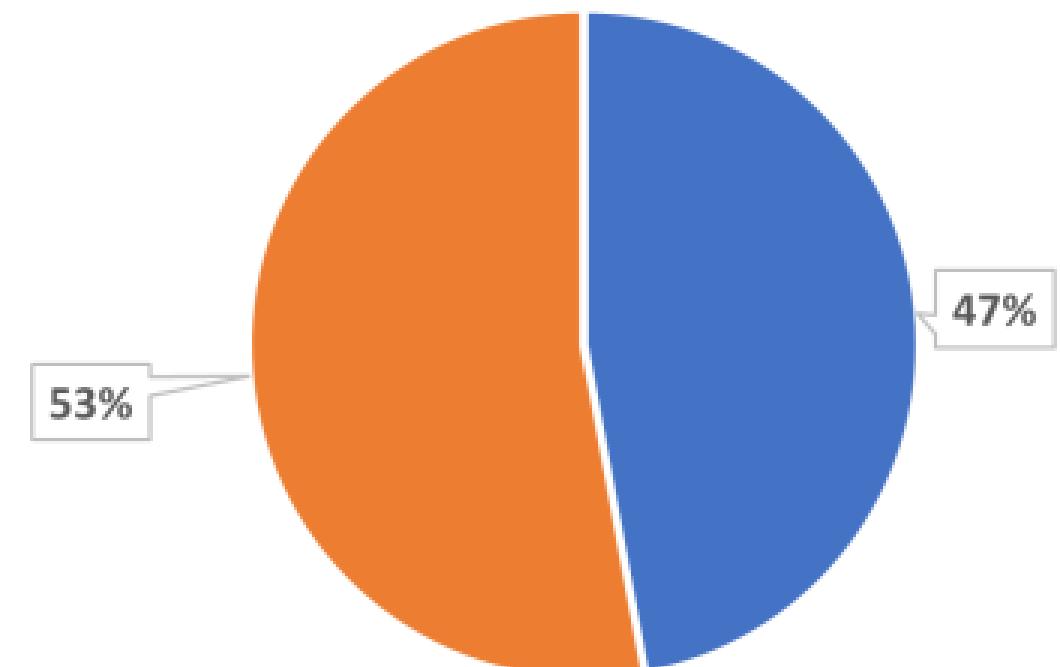


Phân bố dữ liệu

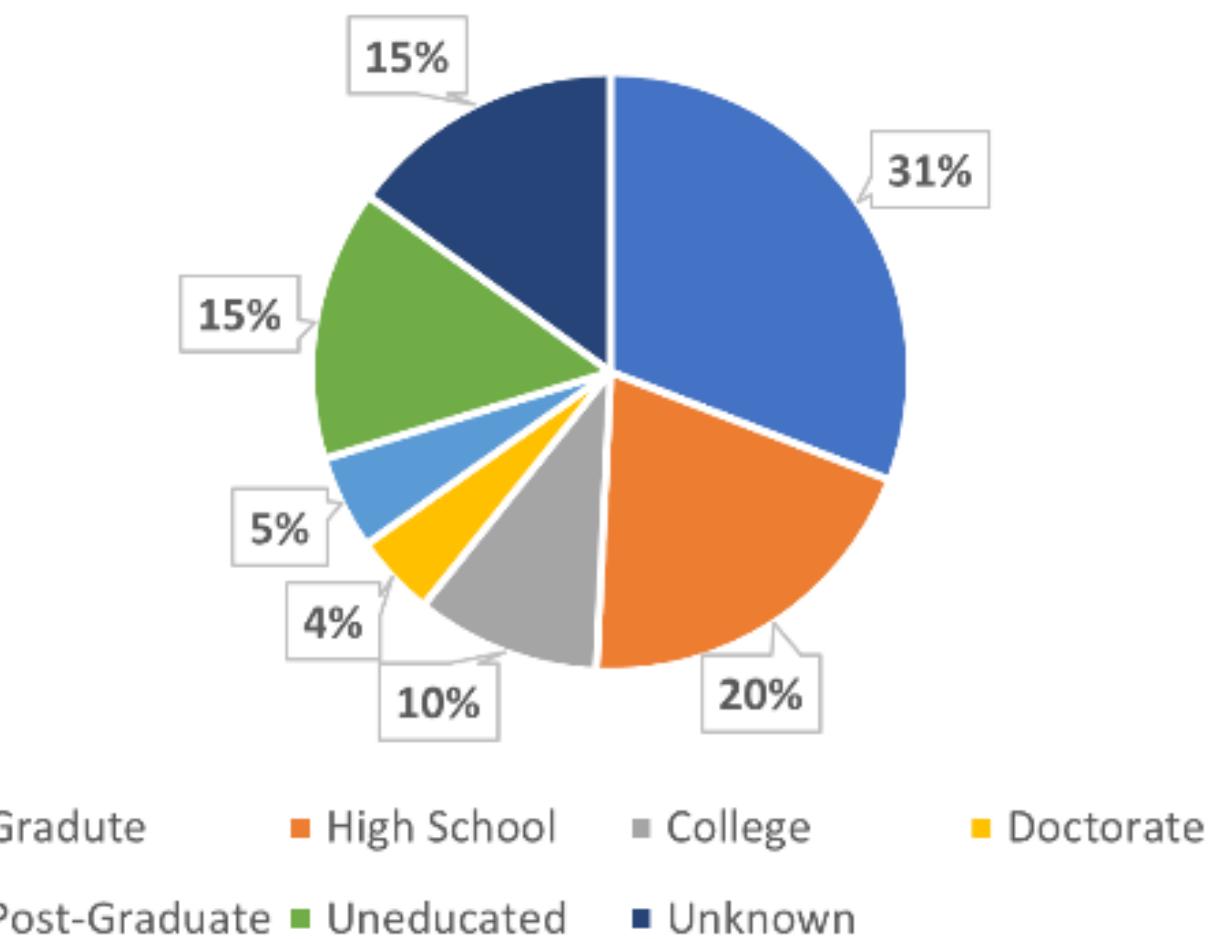


Phân bố dữ liệu

Gender

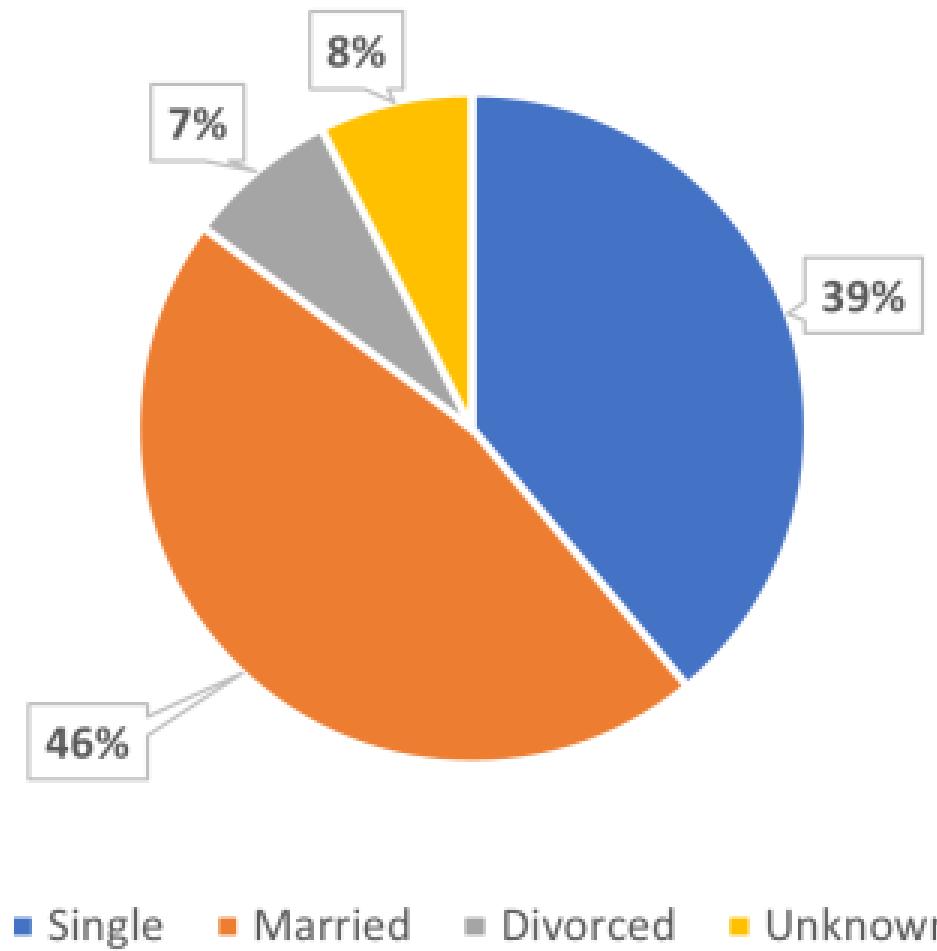


Education_Level

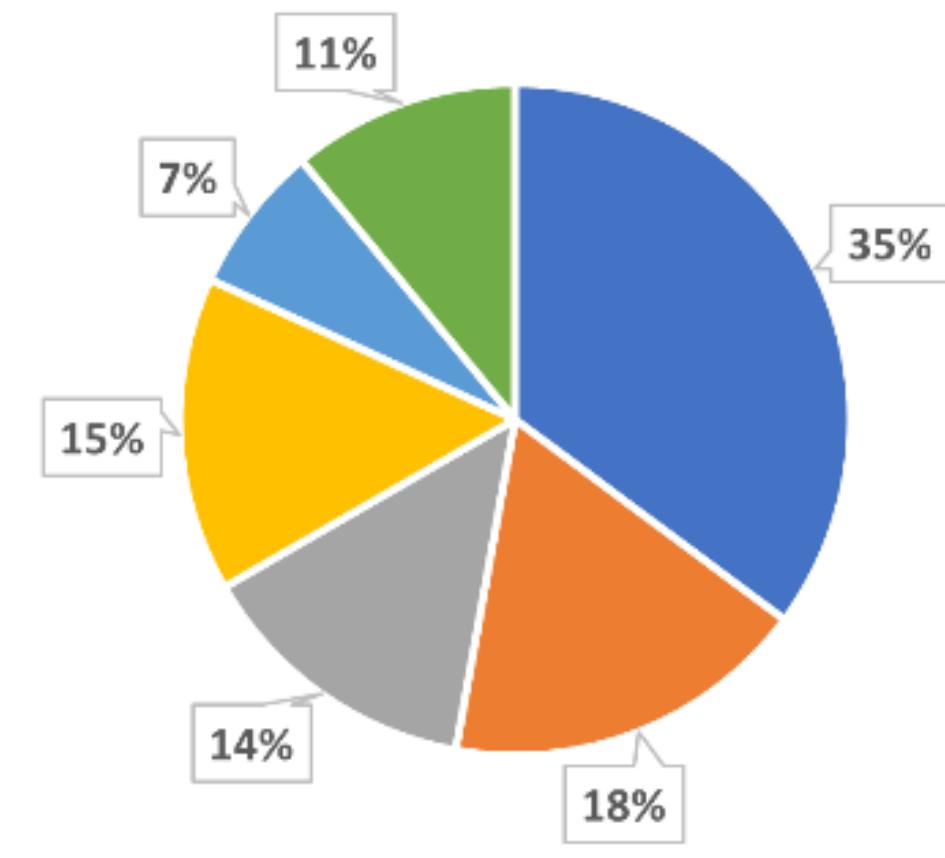


Phân bố dữ liệu

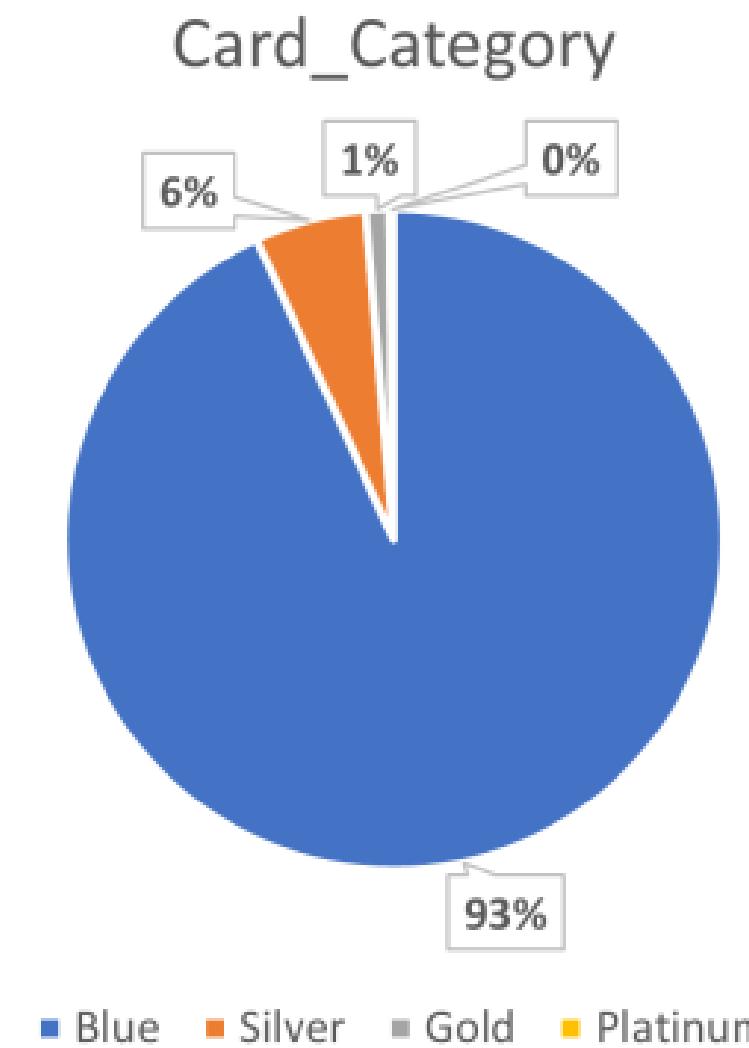
Marital_Status



Income_Category



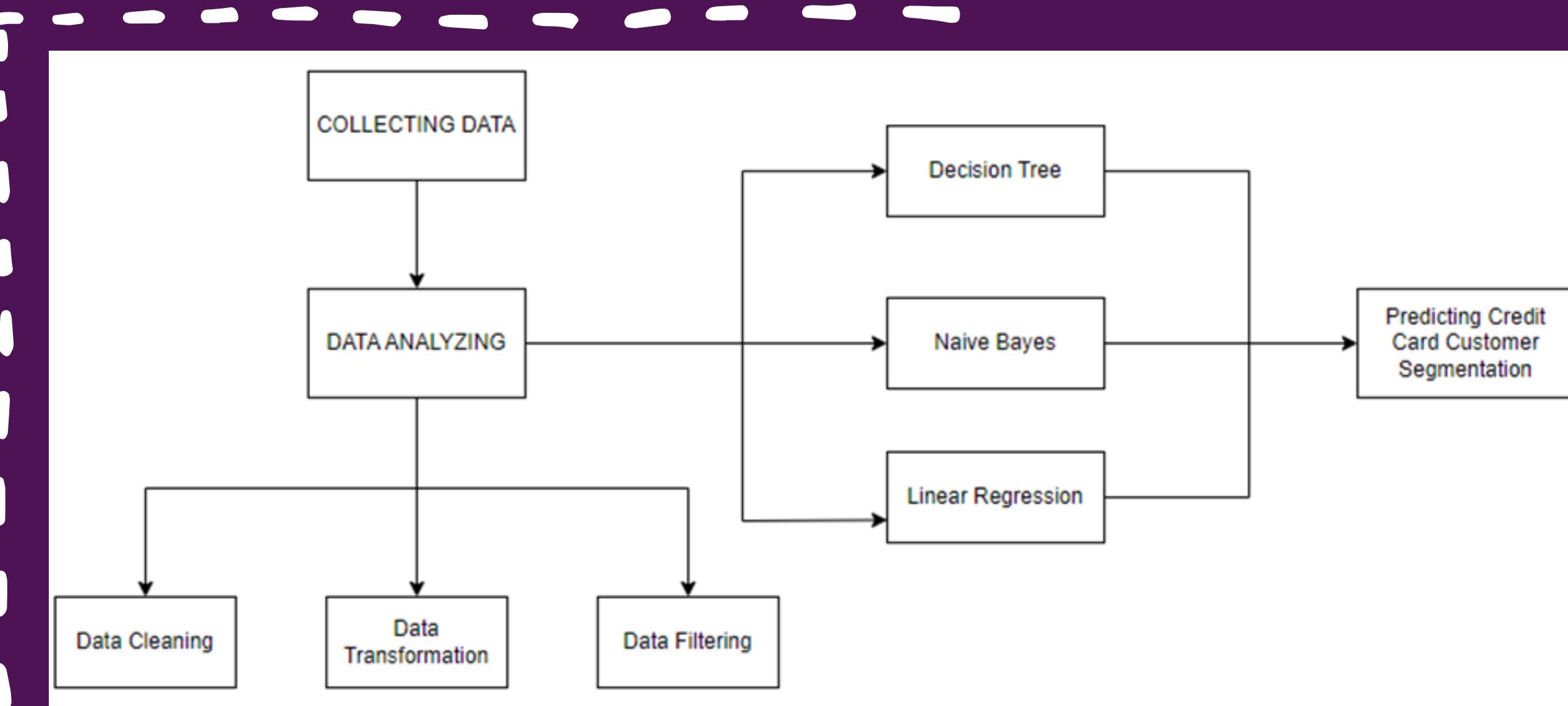
Phân bố dữ liệu



04

THỰC NGHIỆM

Mô hình thực nghiệm



Tiền xử lý dữ liệu

Loại bỏ những thuộc tính không phù hợp

Xử lý các giá trị "Unknown"

*Xem xét độ tương quan
của các thuộc tính
với Correlation Matrix*

Chuyển đổi kiểu dữ liệu

Tiền xử lý dữ liệu

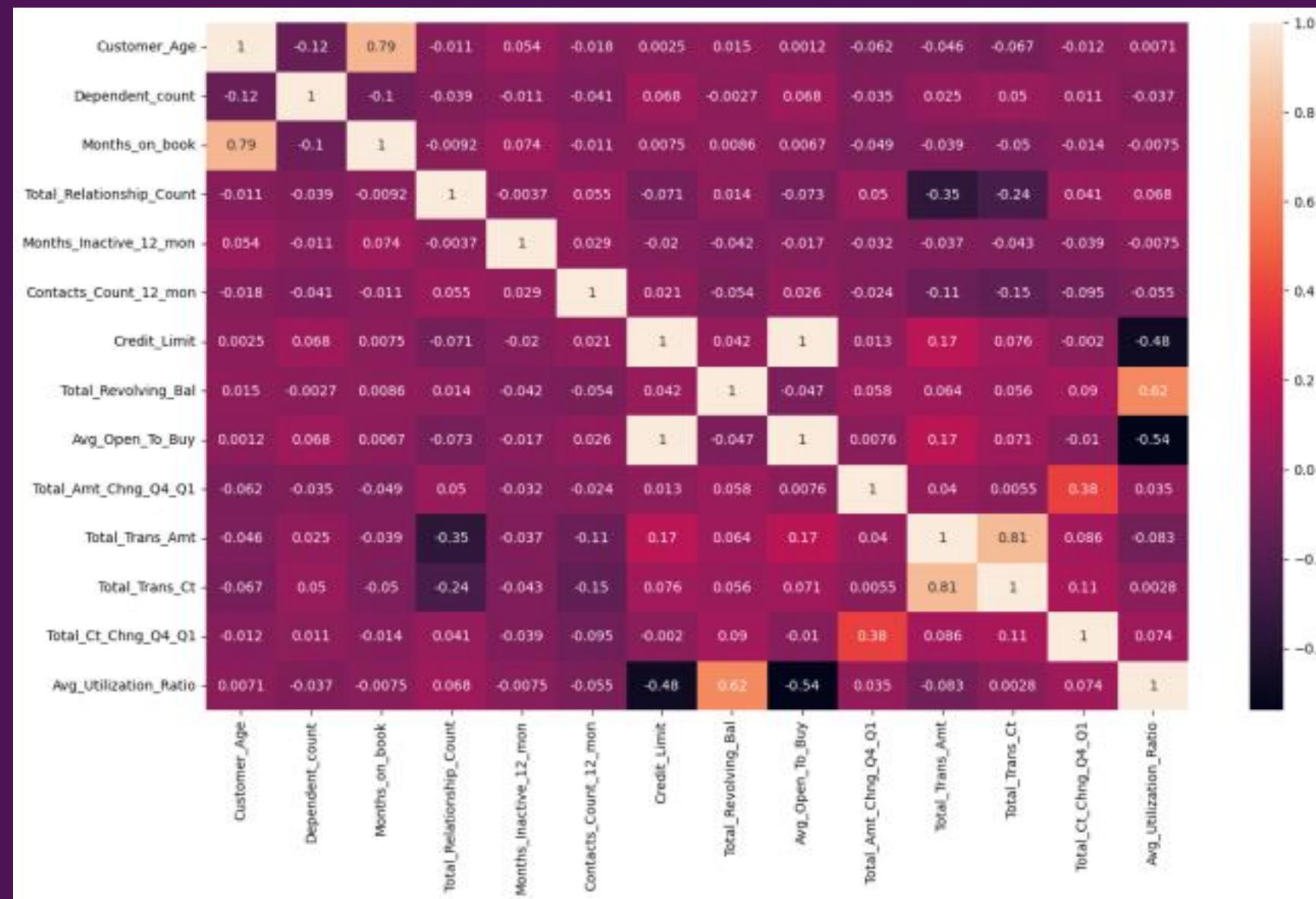
Loại bỏ những thuộc tính không phù hợp



Bỏ thuộc tính CLIENTNUM
và 2 thuộc tính cuối

Tiền xử lý dữ liệu

Xem xét độ tương quan của các thuộc tính với Correlation Matrix



Tiền xử lý dữ liệu

Xử lý các giá trị "Unknown"

- *Chia dataset thành 2 nhóm (dựa vào 2 giá trị của thuộc tính quyết định).*
- *Thay thế giá trị “Unkhown” thành giá trị xuất hiện nhiều nhất trong mỗi nhóm.*

Tiền xử lý dữ liệu

Chuyển đổi kiểu dữ liệu

Chuyển những thuộc tính thuộc kiểu category

về dạng one-hot vector:

- *Gender*
- *Card_Category*
- *Income_Category*
- *Education_Level*
- *Marital_Status*

Chạy các mô hình và dự đoán

Trong bài toán này, chúng em chia tập dữ liệu thành 2 phần: tập Train – tập Test với tỉ lệ là 8:2 để phân tích, dự đoán và đánh giá hiệu suất.

05

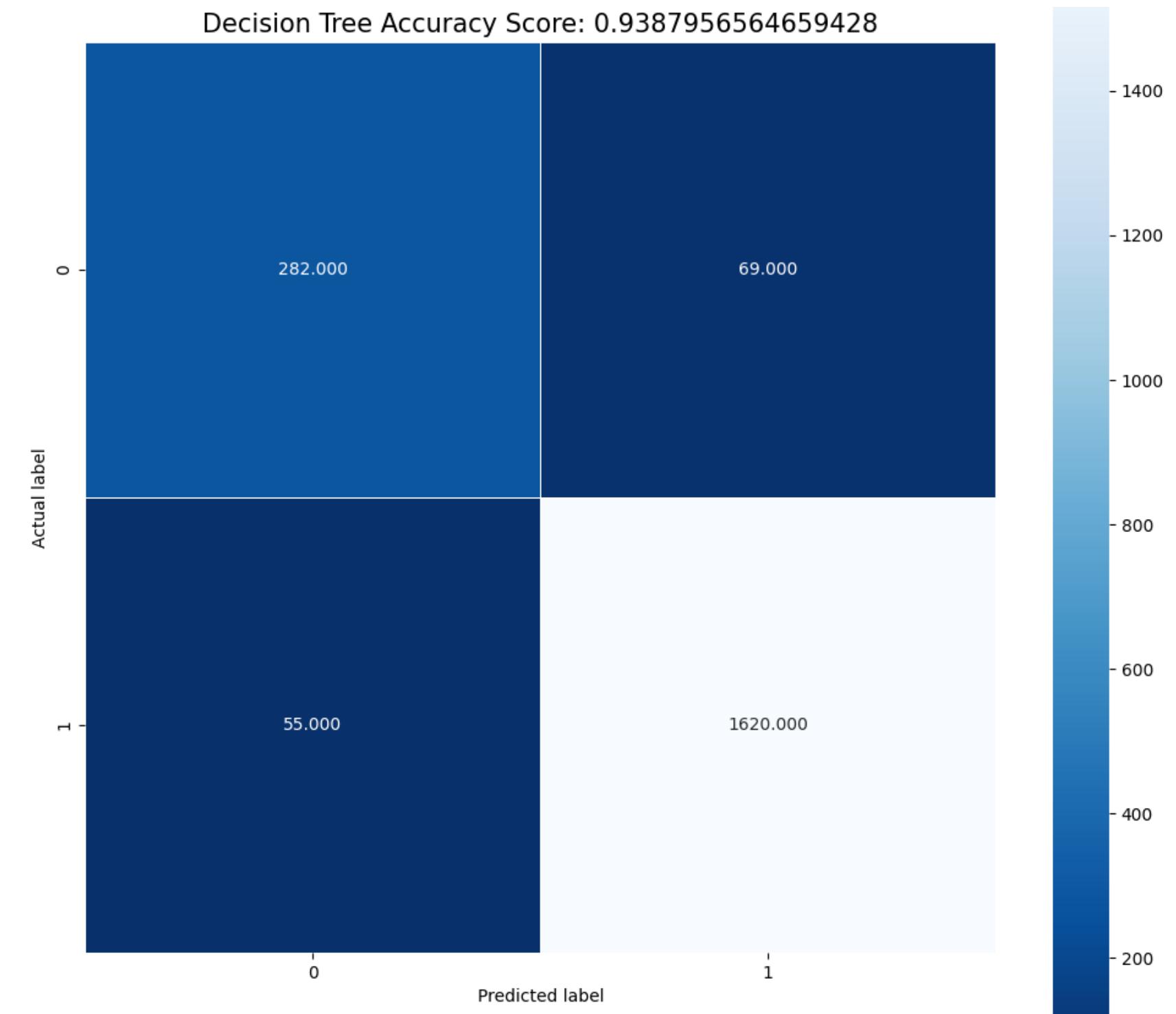
KẾT QUẢ & SO SÁNH

So sánh Naive Bayes và Decision Tree

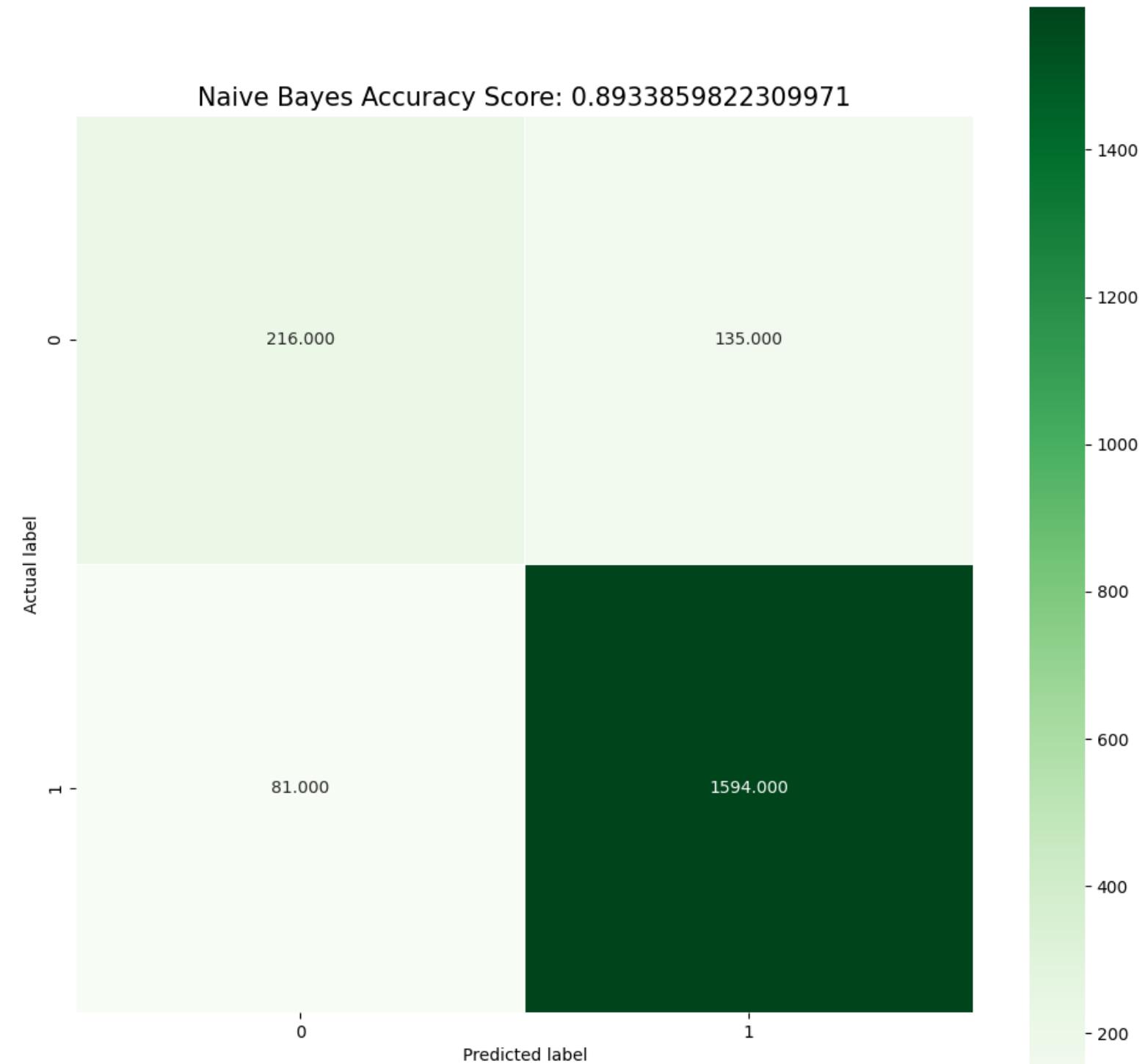
Với 2 mô hình phân lớp là Tree và Naive Bayes, ta đều có thể đánh giá từ ma trận nhầm lẫn (Confusion matrix) và các chỉ số hiệu suất liên quan như sau:

- Accuracy
- Precision
- Recall
- F1_score

Ma trận nhầm lẫn của Decision Tree



Ma trận nhầm lẫn của Naive Bayes



Các chỉ số đánh giá

Decision Tree			
Accuracy	Report		
	Precision	Recall	F1-score
Attrited Customer	0.84	0.8	0.82
Existing Customer	0.96	0.97	0.96

Naïve Bayes			
Accuracy	Report		
	Precision	Recall	F1-score
Attrited Customer	0.73	0.62	0.67
Existing Customer	0.92	0.95	0.94

Đánh giá Linear Regression

Linear Regression là một mô hình hồi quy. Do đó, Linear được đánh giá bằng các chỉ số metrics như sau:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- R-Squared (R²)

Các chỉ số đánh giá

MSE	0.125133246
RMSE	0.404171859
MAE	0.250266492
R2	0.023092464

06

KẾT LUẬN

Xác định được hiệu suất của từng thuật toán trong bài toán ứng dụng

- Naive Bayes: 88%
- Decision Tree: 93%
- Linear Regression: hiệu suất tương đối thấp.

Mô hình Naive Bayes và Decision Tree đạt được độ chính xác cao và Linear Regression không đạt được kết quả như mong muốn.

"Lựa chọn mô hình phù hợp với bài toán và tập dữ liệu là vô cùng quan trọng."

- Kết luận -

Báo cáo Bài tập lớn

**Hiệu suất và ứng dụng của Decision Tree,
Naive Bayes và Linear Regression trong
phân loại khách hàng về thẻ tín dụng**

THE END!