



TOPIC: CREDIT SCORE CLASSIFICATION

NÂNG CAO HIỆU QUẢ PHÁT HIỆN RỦI RO TÍN DỤNG VỚI MÔ HÌNH MÁY HỌC

The project made by group K11

TEAM MEMBERS

71131101054

**Nguyễn Việt Dũng
(Nhóm trưởng)**

71131101179

Vũ Tiên Nam

71131101295

Phạm Thu Trang

7123112089

Nguyễn Trung Hiếu

71131101170

Nguyễn Vũ Minh

71131101167

Vũ Quỳnh Mai

71131101156

Lê Quý Long



TỔNG QUAN

- 01 Giới thiệu bài toán**

- 02 Tổng quan nghiên cứu**

- 03 Mô hình đề xuất**

- 04 Thực nghiệm và kết quả**



I. GIỚI THIỆU BÀI TOÁN

- Trong bối cảnh nền kinh tế đang phát triển
- Hoạt động cho vay tín dụng tại các ngân hàng mở rộng hơn tới mọi khách hàng cá nhân và tổ chức
- Sau đại dịch Covid19 nhu cầu về nguồn vốn nhằm phục hồi lại kinh tế tăng cao

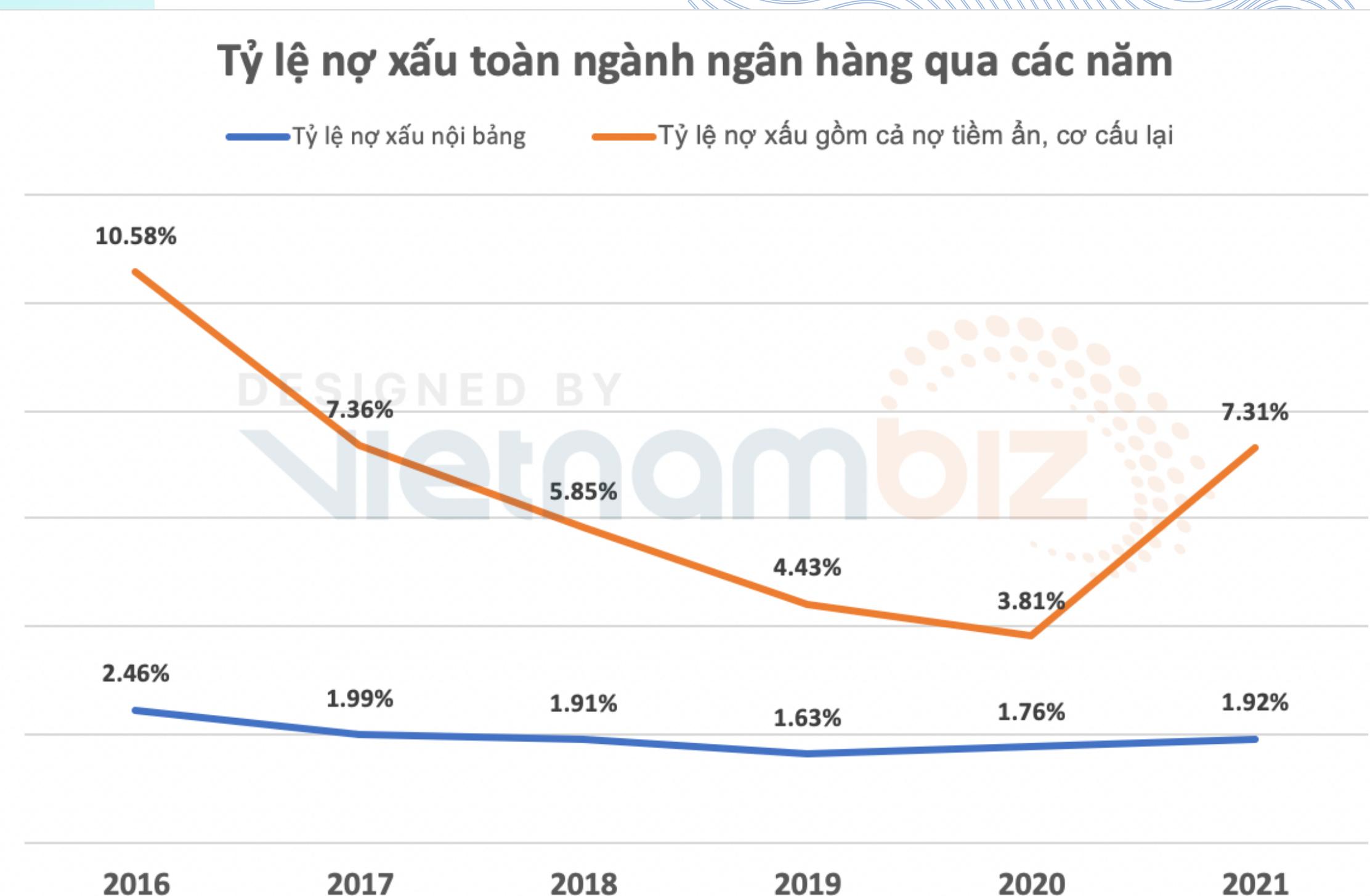


- Số liệu thống kê từ Ngân hàng Nhà Nước Việt Nam (10/03/2022) cho thấy từ giai đoạn bùng phát dịch bệnh thì tỷ lệ nợ xấu tại các ngân hàng đã có xu hướng tăng trở lại

Tỷ lệ nợ xấu toàn ngành ngân hàng qua các năm

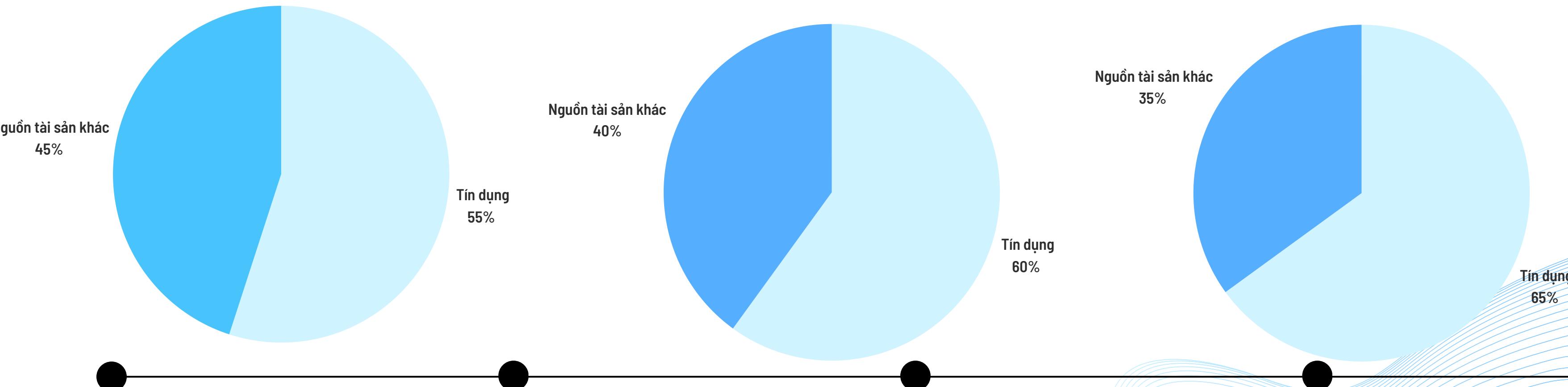
— Tỷ lệ nợ xấu nội bảng — Tỷ lệ nợ xấu gồm cả nợ tiềm ẩn, cơ cấu lại

- Mức tăng tỷ lệ nợ xấu nội bảng và nợ xấu gộp lần lượt 1,6% và 4,4% năm 2019 lên mức 1,9% và 7,3% năm 2021



- **Việt Nam** với tín dụng chiếm tỷ trọng cao nhất trong tổng tài sản và đem lại nguồn thu nhập lớn nhất

=> **Việc hạn chế rủi ro tín dụng là điều cực kỳ quan trọng**



Biểu đồ minh họa tỷ trọng tín dụng chiếm lớn nhất trong tổng tài sản tại Việt Nam từ 2019-2021

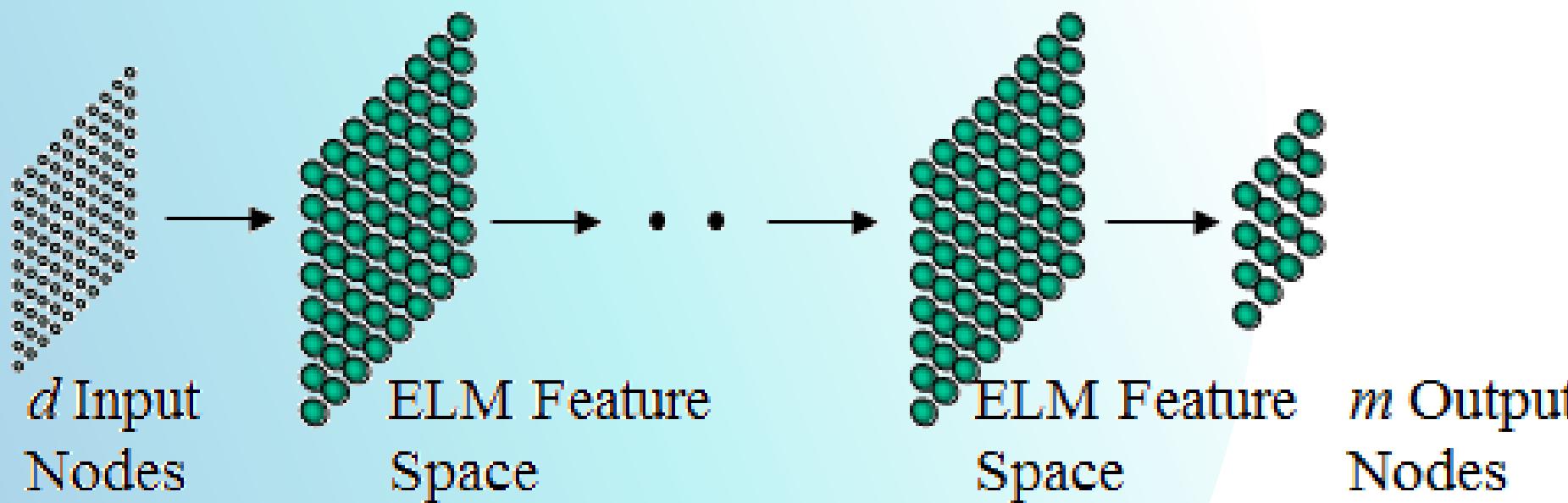
II. TỔNG QUAN NGHIÊN CỨU

- ✓ • Trong quá khứ đã có nhiều bài nghiên cứu đưa ra các đề xuất phương pháp, mô hình nhằm giải quyết bài toán cho rủi ro khi cho vay tín dụng.
- Sau đây là một số kết quả nghiên cứu từ các bài nghiên cứu trong quá khứ



Nghiên cứu "Phân tích rủi ro tín dụng bằng cách sử dụng công cụ phân loại học máy"

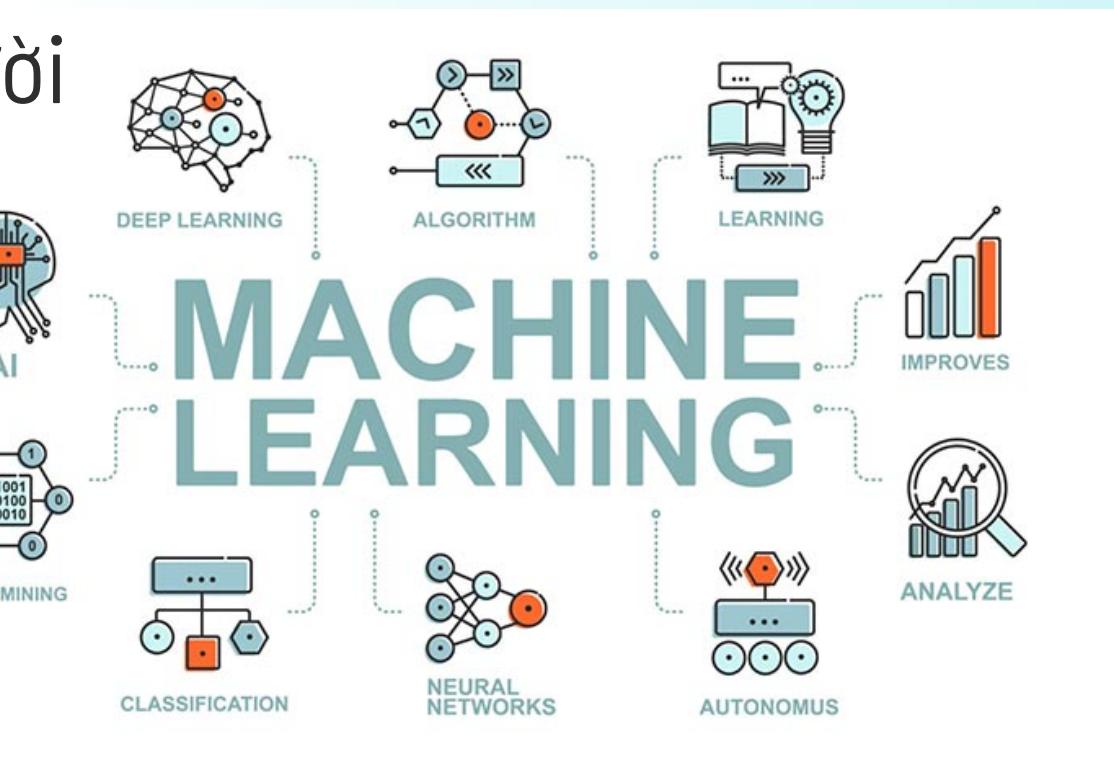
(Trilok Nath Pandey và đồng tác giả (2017)



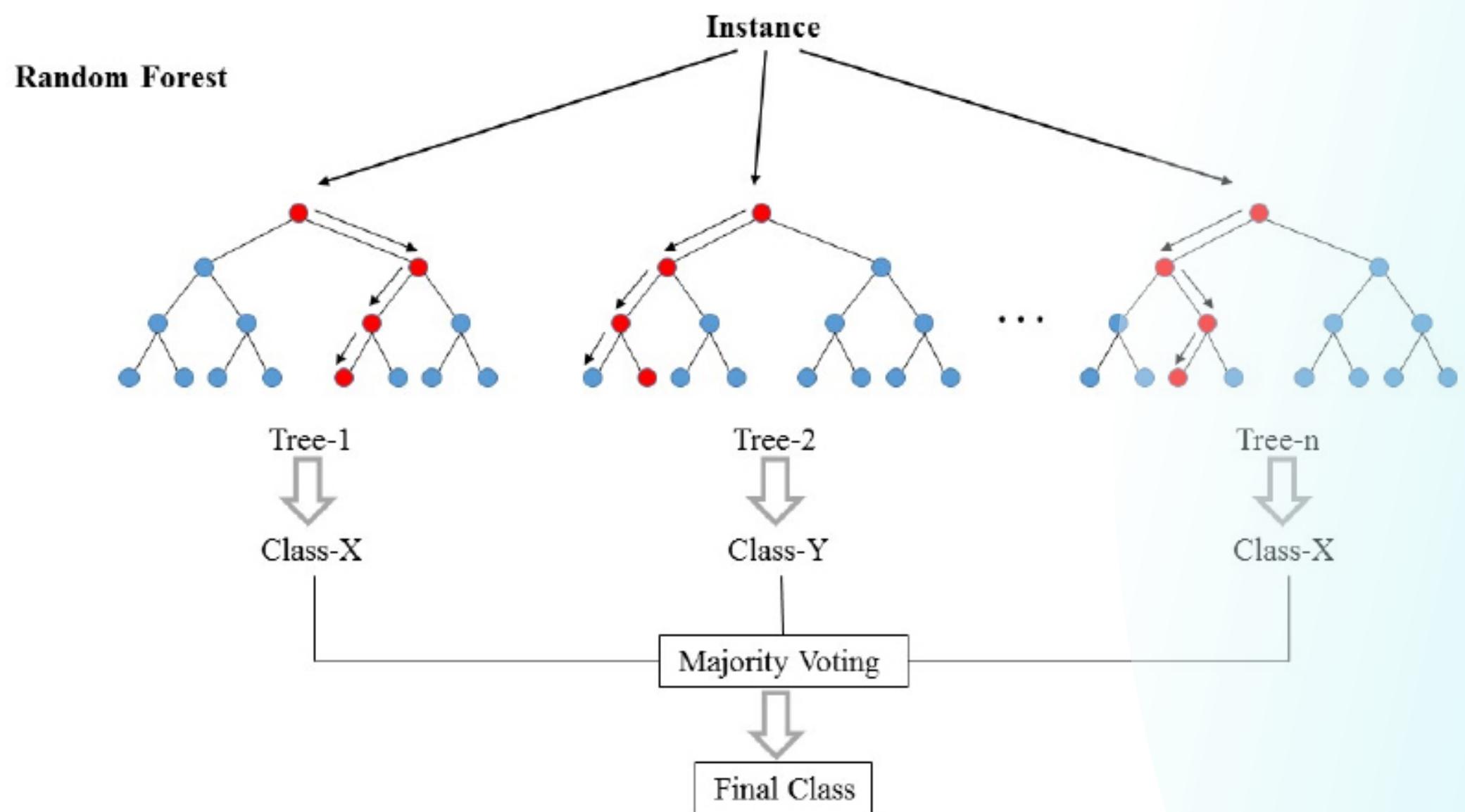
- Cho thấy có hiệu suất cao nhất trong việc dự đoán rủi ro tín dụng.
- Với bộ dữ liệu German, accuracy đạt 0.9633, và với bộ dữ liệu Australian, accuracy đạt 0.9632.
- Điều này làm nổi bật khả năng mạnh mẽ của ELM trong phân loại và dự đoán khả năng trả nợ của khách hàng, hỗ trợ các tổ chức tài chính trong quản lý rủi ro và ra quyết định về tín dụng.

Trong các nghiên cứu:

- "Phương pháp học máy để dự đoán mặc định khoản vay của người Trung Quốc trên thị trường P2P"
- "Áp dụng mô hình học máy để dự đoán tính đủ điều kiện của khoản vay ngân hàng"
- "Ứng dụng một số mô hình học máy để dự đoán khả năng trả nợ của mỗi người khi nộp đơn đi vay tín dụng"



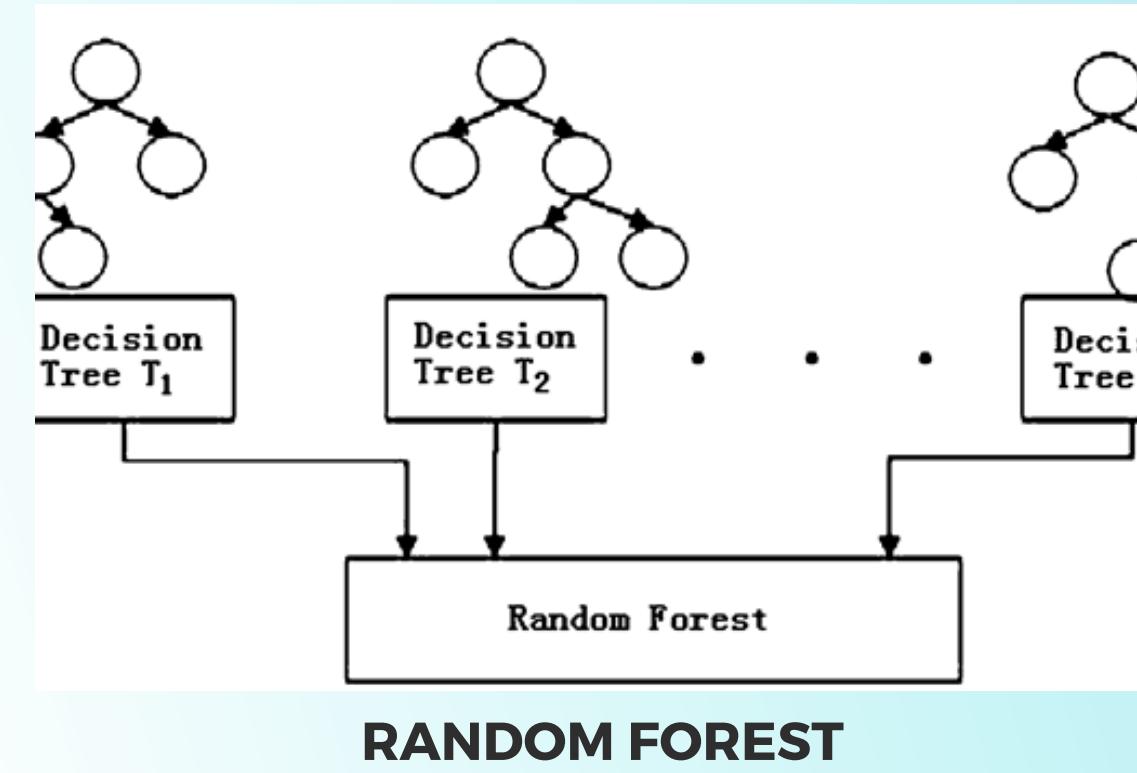
Đều cho thấy các mô hình **Logistic Regression, Random Forest, XGBoost và LightGBM** đã đạt độ chính xác lần lượt là 70.8%, 95.02%, 95.5% và 95.4%.



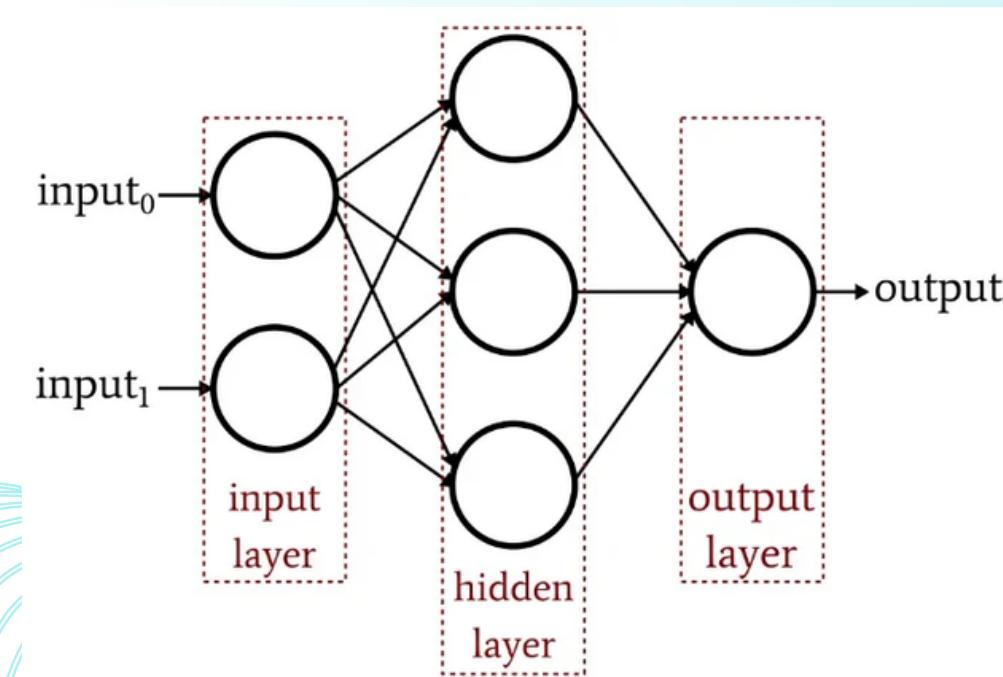
=> Các mô hình học máy có khả năng dự đoán khả năng trả nợ của mỗi cá nhân khi nộp đơn vay tín dụng một cách chính xác và hiệu quả.

III. MÔ HÌNH ĐỀ XUẤT

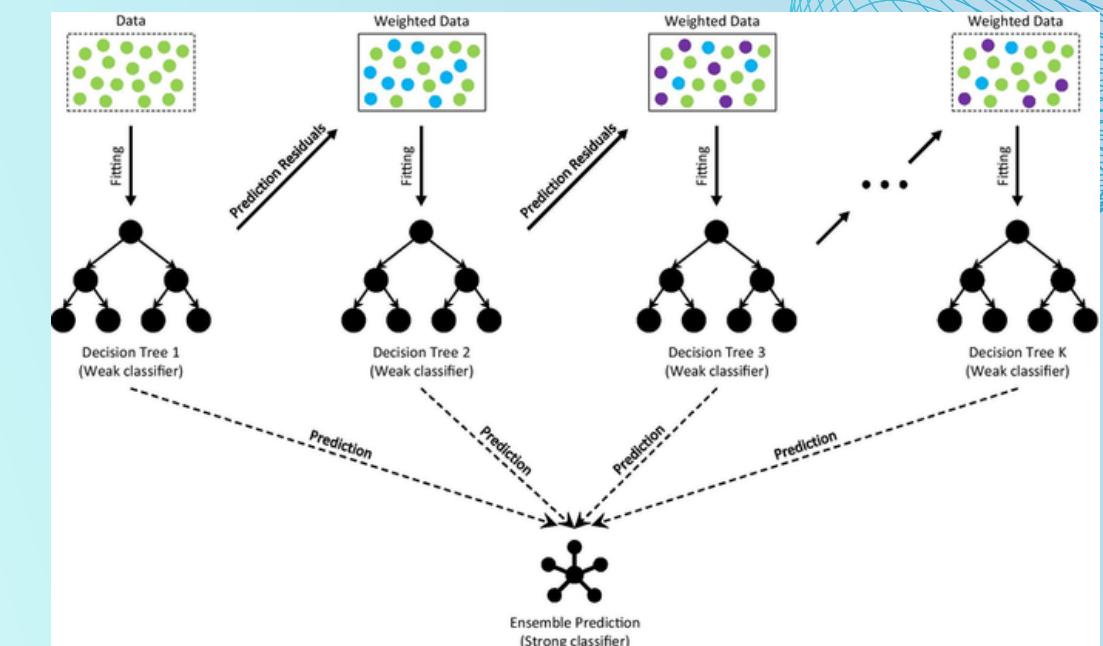
- Rừng ngẫu nhiên
(Random forest)
- GBTC (Gradient
Boosted Trees
Classifier)
- Multilayer Perceptron



RANDOM FOREST

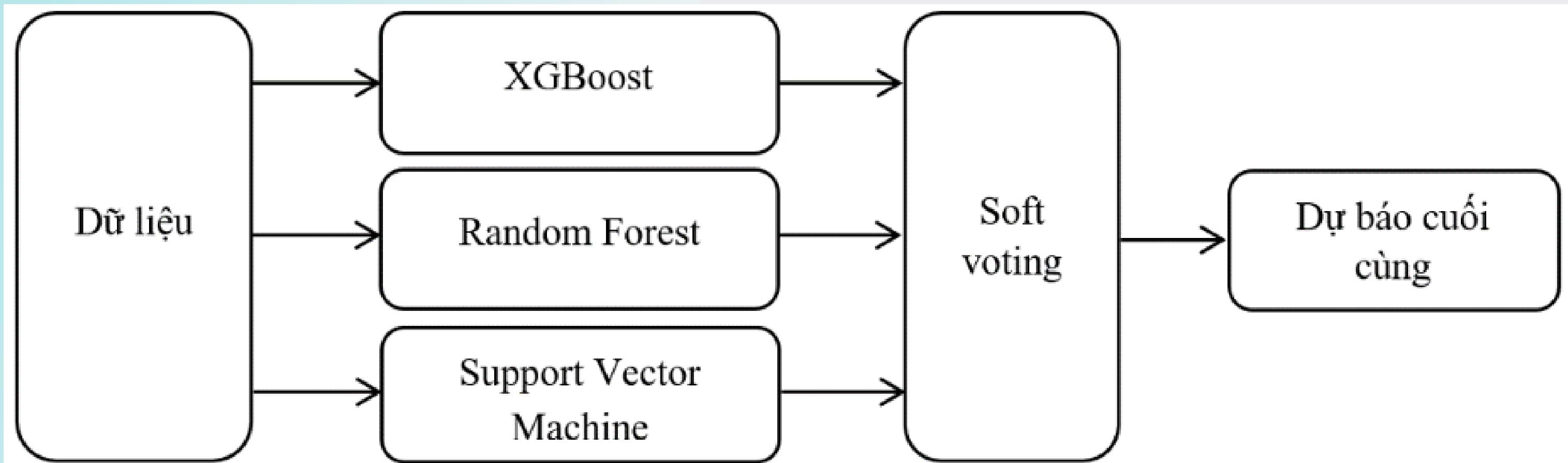


MULTILAYER PERCEPTRON



GRADIENT BOOSTED TREES CLASSIFIER

CẤU TRÚC CỦA MÔ HÌNH ĐỀ SUẤT



3.1 PHƯƠNG PHÁP PHÂN LOẠI

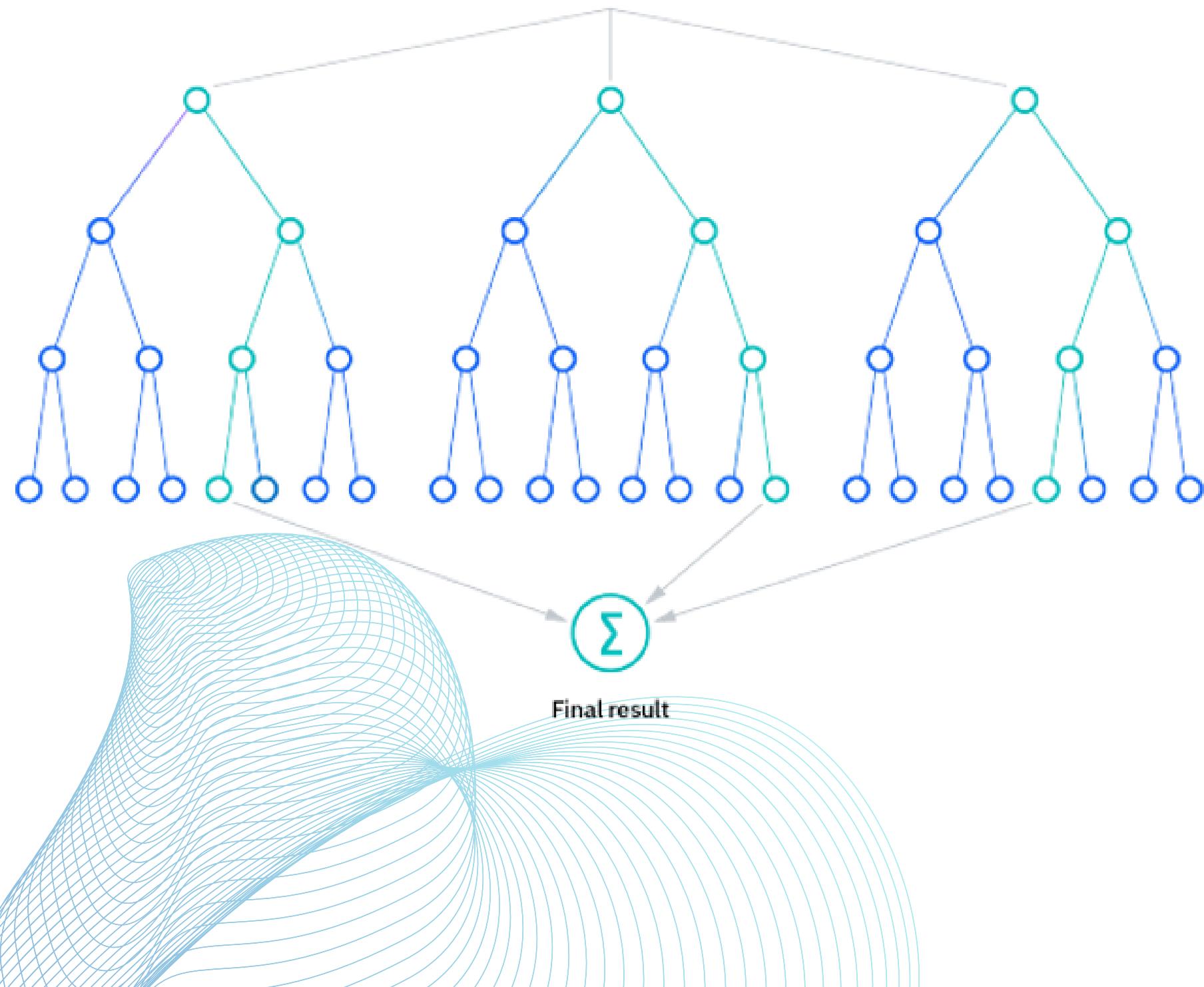
Random Forest (RDF)

- Là một ensemble của các cây quyết định
 - Tổng hợp thông tin từ nhiều cây để tạo ra một mô hình dự đoán
 - Nó giải quyết vấn đề quá khớp của cây quyết định bằng cách sử dụng nhiều cây nhỏ với một phần nhỏ dữ liệu và các thuộc tính ngẫu nhiên.
- => Random Forest là sự kết hợp thông tin từ các cây để tạo ra một mô hình tổng hợp có hiệu suất cao hơn và ít bị chêch hơn

Tương tự như "The Wisdom of Crowds" tổng hợp thông tin từ nhiều người sẽ cho kết quả tốt hơn so với một cá nhân.



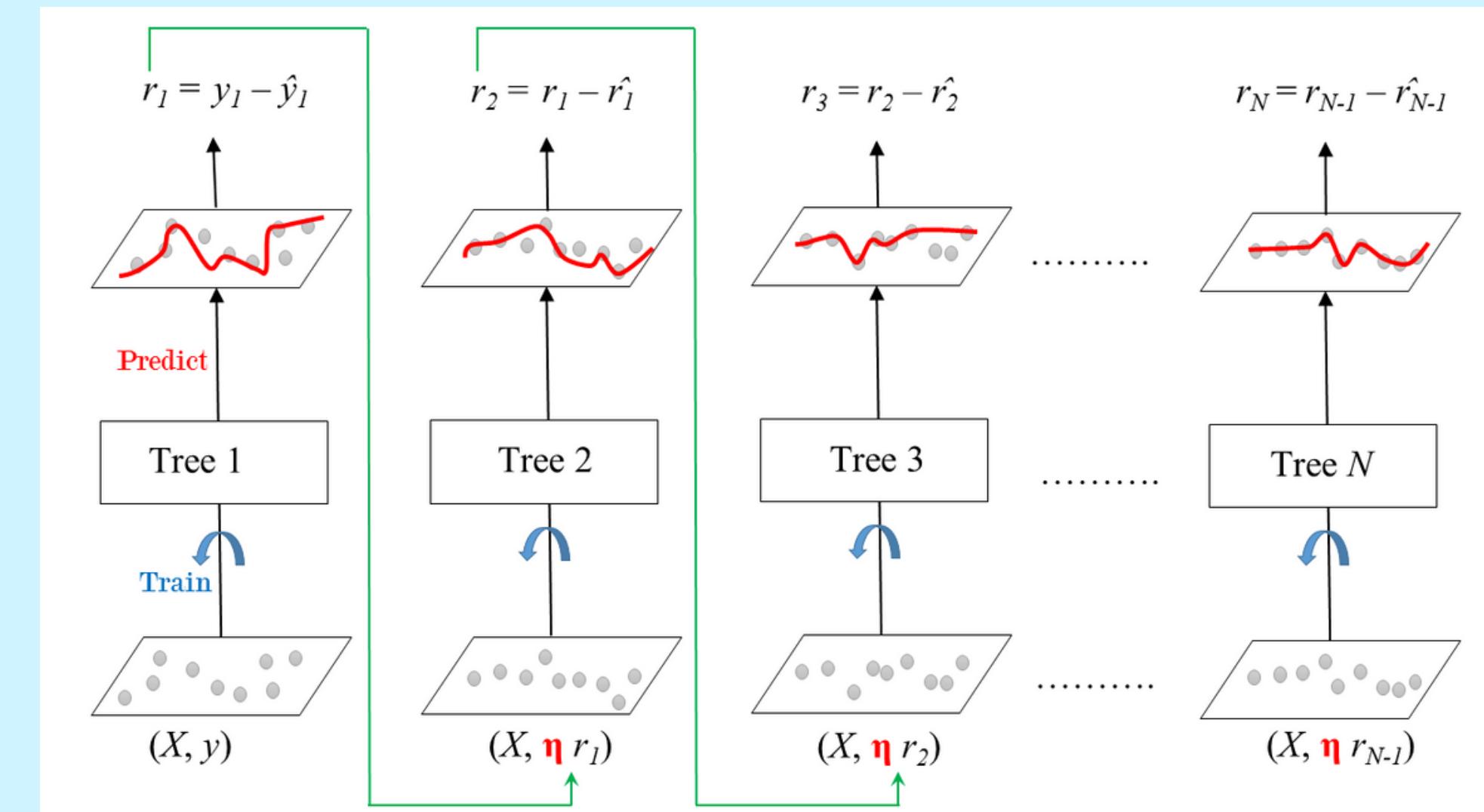
RỪNG NGẪU NHIÊN KHÁC BIỆT VỚI CÂY PHÂN LOẠI CHỦ YẾU THEO HAI CÁCH



1. Dữ liệu cho mỗi cây được ước tính từ các mẫu ngẫu nhiên, thường được lấy ra từ tập dữ liệu huấn luyện với sự thay thế.
2. Cách mà mỗi cây được ước tính là thông qua việc sử dụng các mẫu ngẫu nhiên và một mẫu có thể được sử dụng nhiều lần.

GRADIENT BOOSTED TREES (GBTC)

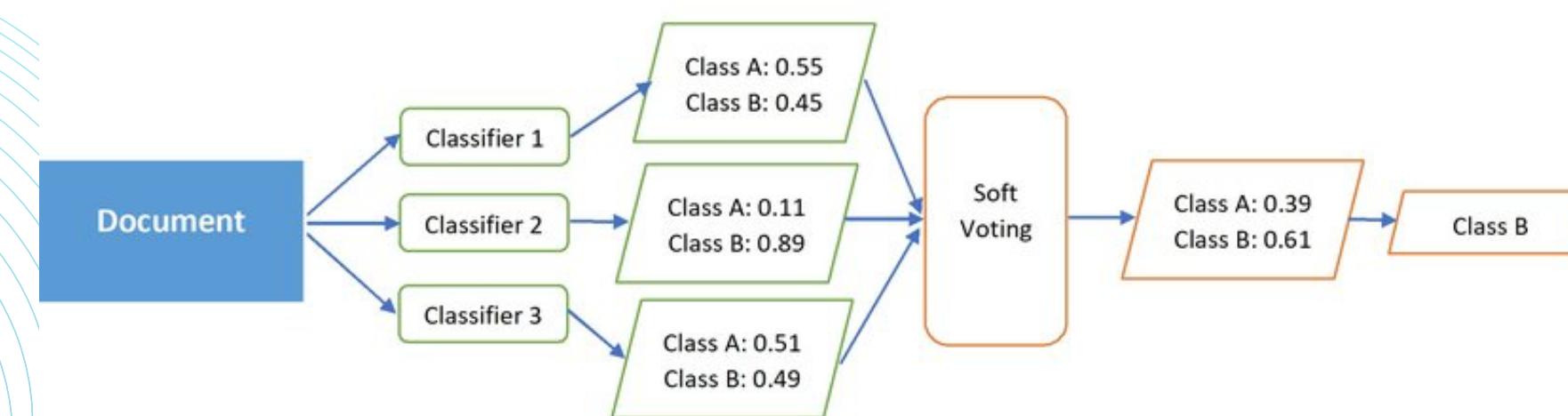
- GBTC là phương pháp Ensemble Model kết hợp các cây quyết định để cải thiện hiệu suất dự đoán thông qua boosting.
- GBTC thích hợp cho phân loại và hồi quy.
- Tuy nhiên, GBTC có thể tốn nhiều tài nguyên tính toán.
- GBTC dễ bị overfitting nếu không được cấu hình đúng.
- MLP là một kiến trúc mạng nơ-ron feedforward linh hoạt.
- MLP có khả năng học được các hàm phức tạp.
- Tuy nhiên, MLP đòi hỏi nhiều dữ liệu huấn luyện và thời gian lâu.
- MLP dễ bị overfitting khi có nhiều lớp và nơ-ron.



3.2 PHƯƠNG PHÁP VOTING

- Kết hợp các thuật toán khác nhau để ước tính xác suất lớp, tăng cường độ chính xác bằng cách đánh trọng số cao nhất cho bộ phân loại có xác suất dự đoán cao nhất.
- Đây là cách tiếp cận tập trung vào tổng hợp xác suất từ các mô hình để đưa ra dự đoán cuối cùng.
- Soft voting được định nghĩa với công thức (2) như sau

$$\hat{y} = \arg \max_i \sum_{j=1}^m \omega_j l$$



1

Sử dụng hàm **argmax** là hàm cho ra giá trị lớn nhất

2

w_j là trọng số biểu thị mức độ tin cậy của mỗi bộ phân loại

3

p_{ij} là xác suất của bộ phân loại trong việc dự đoán một lớp nhất định

-> Các bộ phân loại có độ chính xác và độ tin cậy cao sẽ được đánh trọng số cao hơn trong quá trình ước tính (Classifier 2 đưa ra dự báo gần và cao nhất nên ta sẽ chọn Class B)

IV. THỰC NGHIỆM VÀ KẾT QUẢ

- Bài nghiên cứu này sẽ thực nghiệm trên bộ dữ liệu credit_risk_dataset.csv được tải xuống từ trang Kaggle.com [9] vào ngày 01/3/2024. Bộ dữ liệu mang mục đích nghiên cứu, thực nghiệm, trong bộ dữ liệu có 25 trường dữ liệu và 100.000 quan sát.

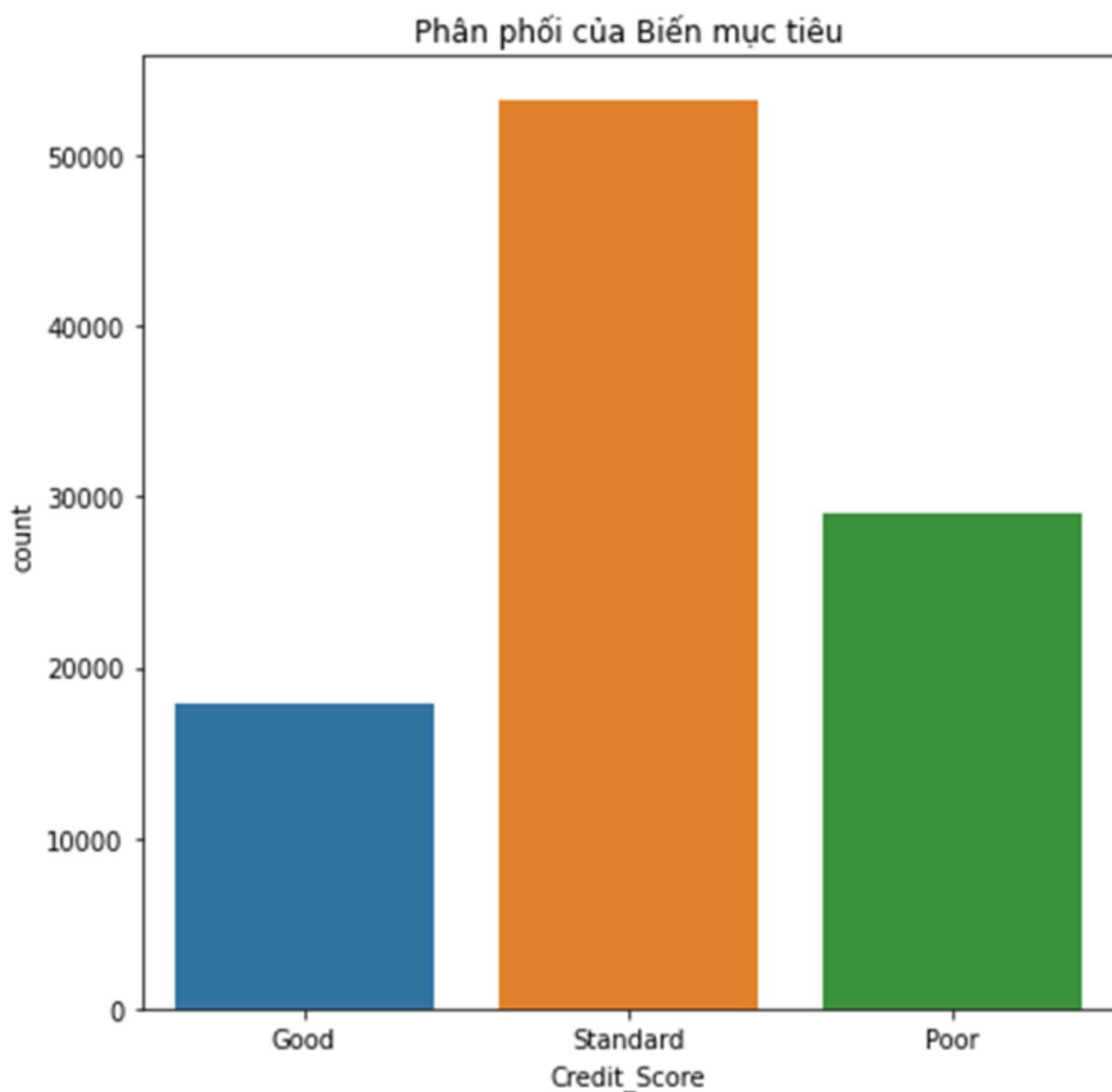


4.1 MIÊU TẢ DỮ LIỆU

Trường dữ liệu “Credit_Score” thể hiện khách hàng có khả năng rủi ro, trong đó có 3 nhãn dữ liệu bao gồm Poor, Standard, Good. Tới số lượng lần lượt bằng:

- Poor với 28998 quan sát (28,99%)
- Standard với 53174 quan sát (53,17%)
- Good với 17828 quan sát (17,82%)

Dữ liệu bao gồm thông tin như tuổi, thu nhập hàng năm, số tài khoản ngân hàng, và nhiều trường dữ liệu khác.

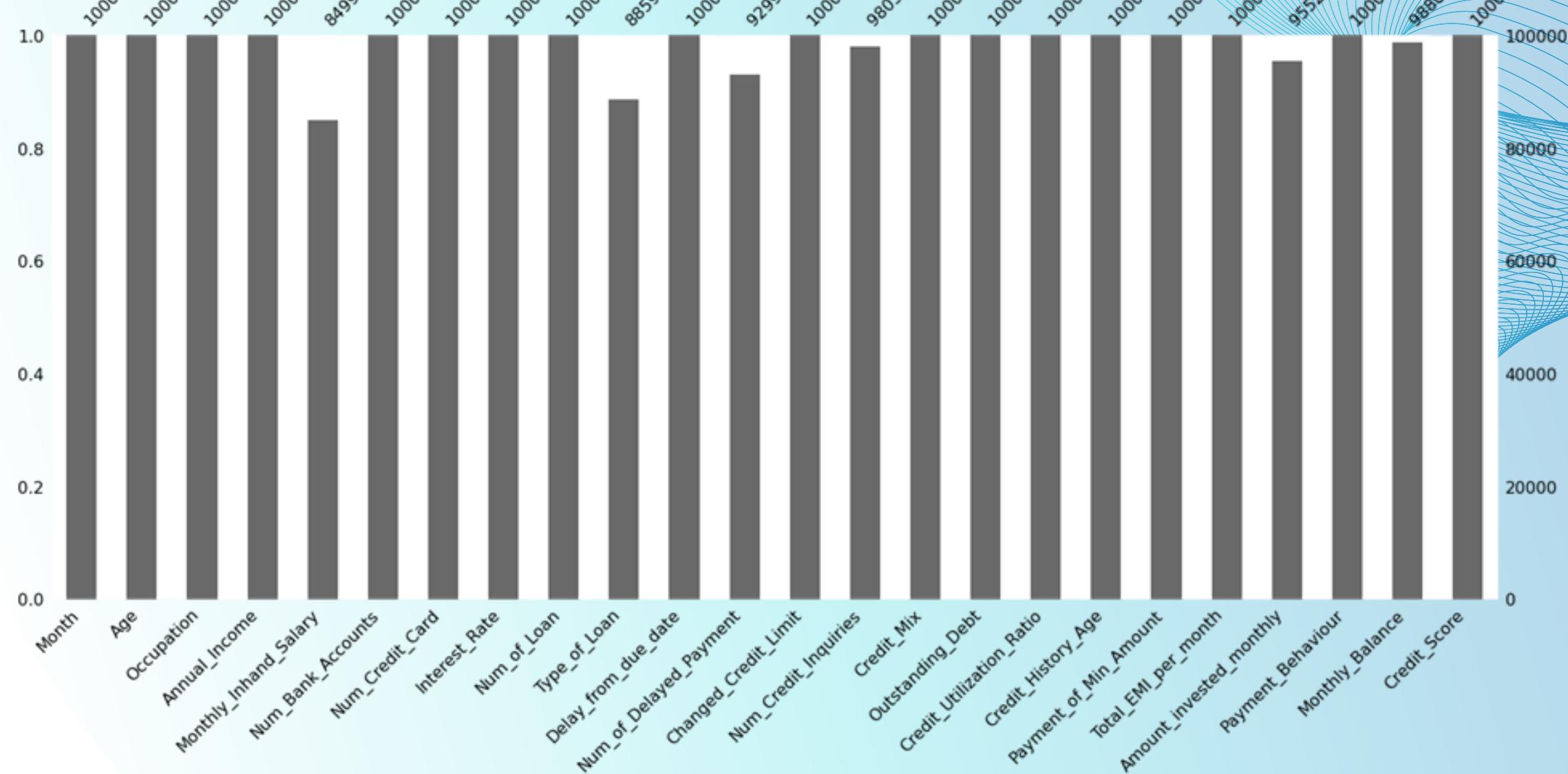


4.2 TIỀN XỬ LÝ DỮ LIỆU

Xử Lý Thiếu Dữ Liệu

Trước khi xử lý, 35.18% dữ liệu bị thiếu, phân bố tương đối đều trong các trường sau

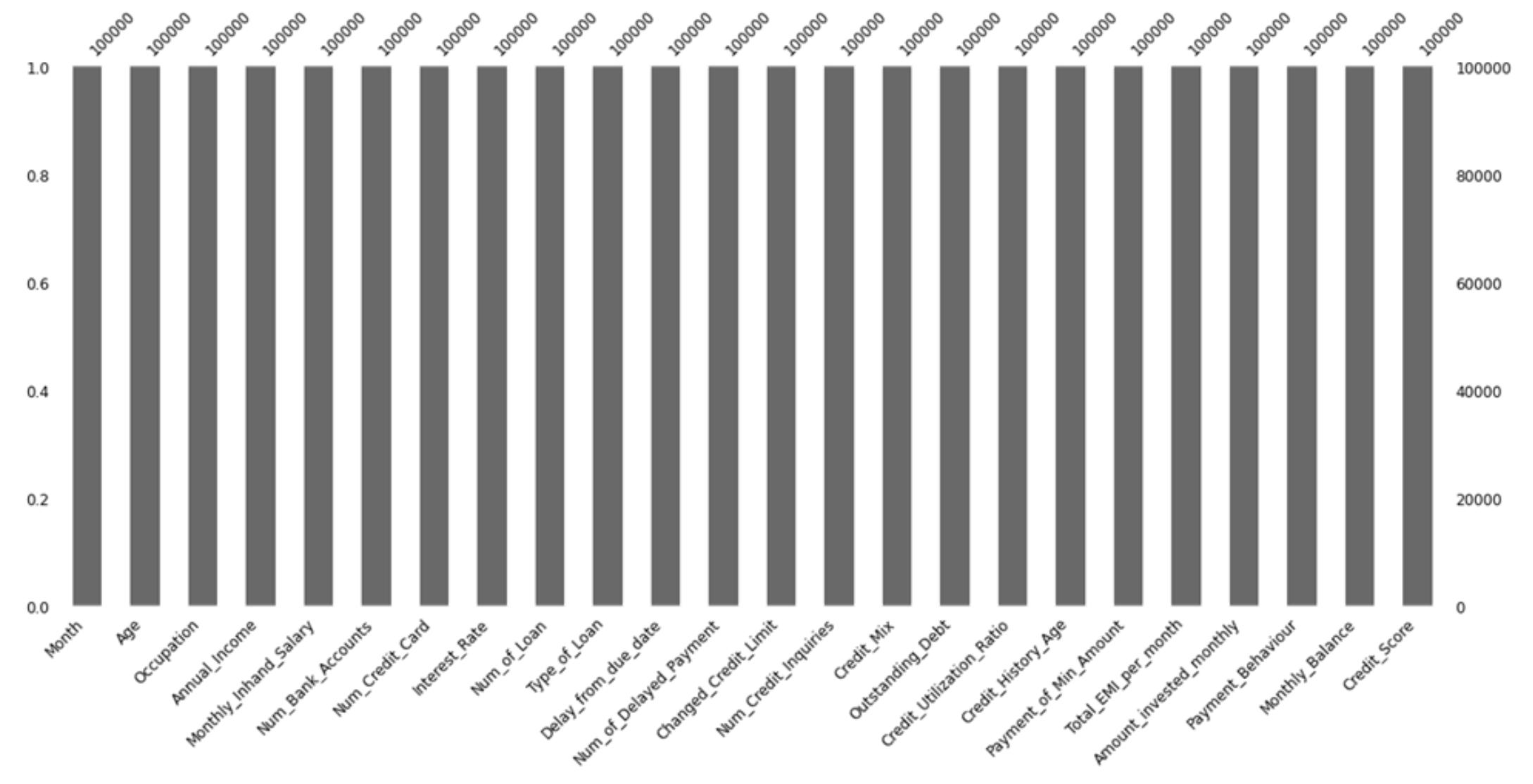
- Monthly_Inhand_Salary: 15.0%
- Type_of_Loan: 11.41%
- Num_of_Delayed_Payment: 7.0%
- Amount_invested_monthly: 4.48%
- Num_Credit_Inquiries: 1.97%
- Monthly_Balance: 1.2%



Số lượng cột bị thiếu dữ liệu

Để giải quyết vấn đề dữ liệu thiếu, chúng tôi áp dụng hai phương pháp khác nhau. Chúng tôi điền giá trị trung bình cho dữ liệu số và 'NAH' cho dữ liệu phân loại. Lựa chọn 'NAH' không chỉ là để điền vào các ô trống, mà còn mang ý nghĩa rằng trong quá trình thu thập dữ liệu, có thể có khách hàng chủ định để không cung cấp thông tin này.

Ngoài ra chúng tôi xử lý dữ liệu ghi sai bằng cách thay thế các giá trị không hợp lệ bằng "None" để đảm bảo tính nhất quán trong dữ liệu.

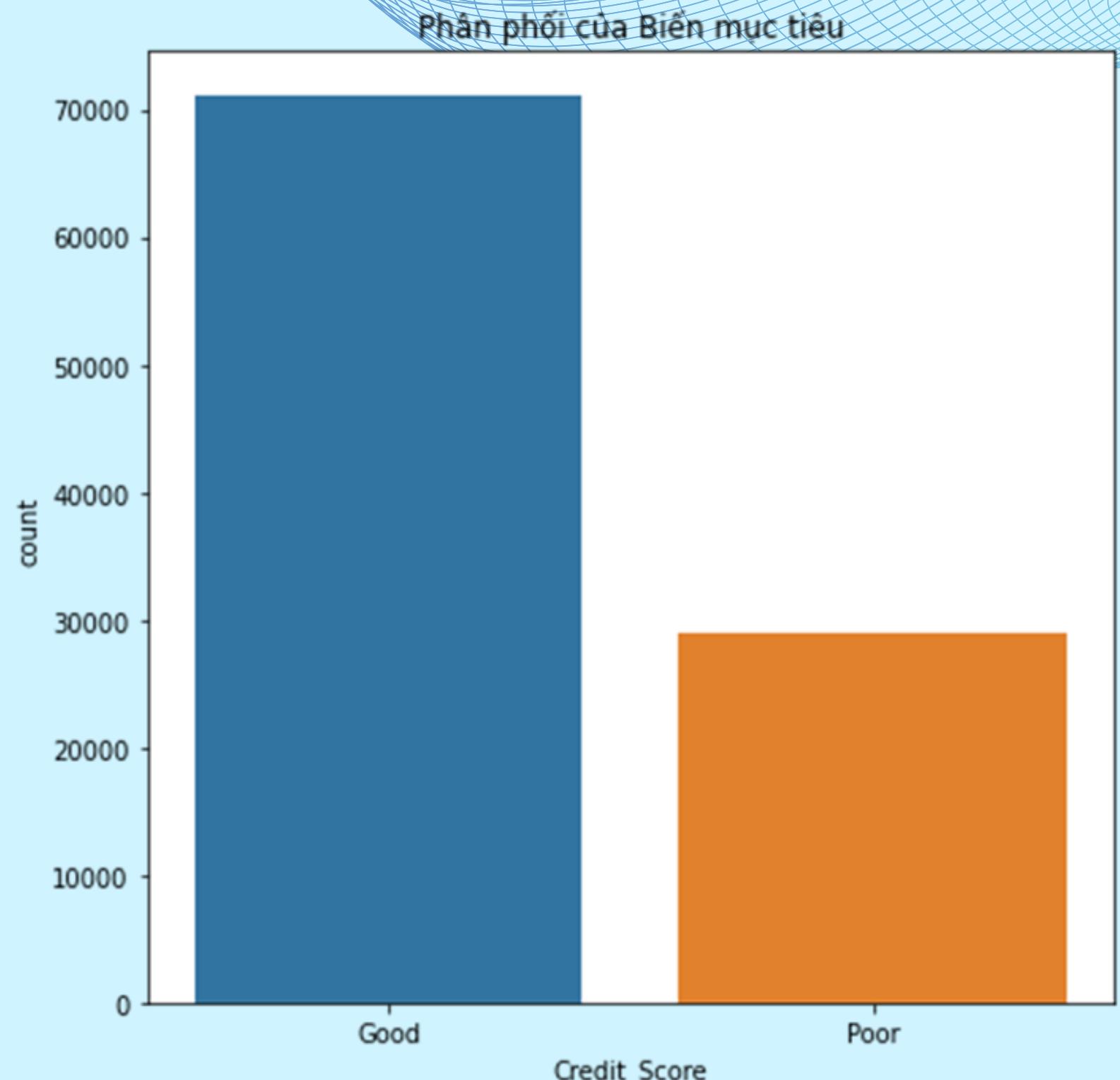


XỬ LÝ MẤT CÂN BẰNG

Nghiên cứu về biến mục tiêu "Credit_Score" nhận thấy sự mất cân bằng trong dữ liệu, khi nhãn "Good" chiếm 71.002 quan sát (71% tổng số) và "Poor" chiếm 28.998 quan sát (28,99% tổng số).

Để giải quyết vấn đề này, phương pháp under sampling được sử dụng bằng cách loại bỏ một phần của các mẫu gán nhãn 'Good'.

Quyết định này giúp cân bằng dữ liệu giữa các nhãn, tăng hiệu suất và đảm bảo tính công bằng và chính xác của mô hình trong các tình huống thực tế.



ENCODER

Sau khi cân bằng dữ liệu, chúng tôi thực hiện việc encoder cho các trường dữ liệu 'Month', 'Type_of_Loan', 'Credit_Mix', 'Credit_History_Age', 'Payment_of_Min_Amount', và 'Payment_Behaviour'. Điều này giúp chuyển đổi các biến phân loại thành dạng số, làm cho dữ liệu dễ hiểu và xử lý hơn đối với các thuật toán máy học.

VECTOR HÓA

Tiếp theo, chúng tôi sử dụng VectorAssembler để tổ chức dữ liệu. Đây là một công cụ quan trọng trong việc chuẩn bị dữ liệu cho mô hình học máy trong Apache Spark. VectorAssembler giúp chúng tôi kết hợp các cột dữ liệu thành một vectơ đặc trưng lớn, là đầu vào cho các thuật toán học máy trong Spark. Việc này tăng hiệu quả của quá trình chuẩn bị dữ liệu và đảm bảo rằng mô hình học máy có thể sử dụng dữ liệu một cách dễ dàng và hiệu quả. Đồng thời, việc tự động hóa quá trình này giúp giảm thời gian phát triển mô hình và giảm thiểu các lỗi có thể xảy ra.

CHUẨN HÓA DỮ LIỆU

Cuối cùng, chúng tôi thực hiện chuẩn hóa dữ liệu bằng phương pháp z-score.

Đây là một trong những phương pháp chuẩn hóa dữ liệu phổ biến nhất trong học máy

Phương pháp z-score normalization được thực hiện bằng cách áp dụng công thức sau cho mỗi giá trị của biến đặc trưng

$$z = \frac{(x - \mu)}{\sigma}$$

- z là giá trị chuẩn hóa của biến đặc trưng.
- x là giá trị ban đầu của biến đặc trưng.
- μ là giá trị trung bình của biến đặc trưng trong tập dữ liệu.
- σ là độ lệch chuẩn của biến đặc trưng trong tập dữ liệu.

Phương pháp này chuyển đổi mỗi giá trị của biến đặc trưng sao cho giá trị trung bình là 0 và độ lệch chuẩn là 1.

Quá trình này giúp đồng bộ hóa và định lượng biến đặc trưng theo cùng một tỷ lệ, từ đó cải thiện hiệu suất và độ chính xác của mô hình học máy trên các biến đa dạng.

Ngoài ra, việc này cũng giúp tăng tốc quá trình học và giảm thiểu ảnh hưởng của các giá trị ngoại lai, đồng thời tạo điều kiện thuận lợi cho việc so sánh giữa các biến đặc trưng khác nhau.

CHIA DỮ LIỆU HUẤN LUYỆN VÀ KIỂM ĐỊNH

Chúng tôi chia tập dữ liệu thành hai phần: 80% dành cho huấn luyện (43868 quan sát) và 20% dành cho kiểm định mô hình (10786 quan sát).

Tỷ lệ này được chọn để đảm bảo mô hình có đủ dữ liệu để học và hiểu các mẫu, đồng thời đảm bảo rằng có một tập dữ liệu đủ lớn để kiểm định hiệu suất mô hình một cách đáng tin cậy.

Điều này giúp cân bằng giữa việc huấn luyện mô hình một cách hiệu quả và đánh giá hiệu suất của nó trên dữ liệu mới.

4.3 PHƯƠNG PHÁP ĐÁNH GIÁ

Trong bài nghiên cứu, phương pháp đánh giá hiệu suất của mô hình phân loại dựa trên ma trận nhầm lẫn (Confusion Matrix), sử dụng các chỉ số như accuracy, precision, recall và F1 score. Ma trận nhầm lẫn được biểu diễn như sau:

	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Trong đó:

- TP (True Positive) là số trường hợp dự báo đúng mẫu tích cực.
- TN (True Negative) là số trường hợp dự báo đúng mẫu tiêu cực.
- FP (False Positive) là số trường hợp dự báo sai mẫu tích cực.
- FN (False Negative) là số trường hợp dự báo sai mẫu tiêu cực.

Các chỉ số đánh giá mô hình bao gồm:

- Accuracy: tỷ lệ giữa số lượng dự đoán đúng trên tổng số mẫu.
- Precision: tỷ lệ giữa số lượng dự đoán tích cực đúng trên tổng số dự đoán tích cực.
- Recall: tỷ lệ giữa số lượng dự đoán tích cực đúng trên tổng số mẫu tích cực.
- F1 score: trung bình trọng số của Precision và Recall, thể hiện sự cân bằng giữa chúng.

4.4 CÁC THAM SỐ VÀ MÔI TRƯỜNG CÀI ĐẶT

Dưới đây là giới thiệu về 4 mô hình học máy cùng các tham số quan trọng:

Random Forest (RDF) với Entropy:

- numTrees = 101 (Số cây)
- maxDepth = 25 (Số tầng tối đa)
- seed = 42 (Mã hạt giống cho quá trình ngẫu nhiên)
- impurity = 'entropy' (Phương pháp đo lường độ)

Random Forest (RDF) với Gini:

- numTrees = 101 (Số cây)
- maxDepth = 25 (Số tầng tối đa)
- seed = 42 (Mã hạt giống cho quá trình ngẫu nhiên)
- impurity = 'gini' (Phương pháp đo lường độ)

Gradient Boosted Trees (GBT):

- maxDepth = 5 (Số tầng tối đa của cây)
- maxBins = 32 (Số lượng bins tối đa cho các tính năng)
- minInstancesPerNode = 1 (Số lượng mẫu tối thiểu cần phải có trong mỗi nút của cây)
- lossType = 'logistic' (Loại hàm mất mát được sử dụng trong quá trình tối ưu hóa)

Multilayer Perceptron (MLP):

- solver = 'l-bfgs' (Thuật toán tối ưu hóa)
- maxIter = 100 (Số lượng vòng lặp tối đa)
- stepSize = 0.03 (Kích thước bước cho quá trình tối ưu hóa)
- layers = [22, 128, 64, 32, 16, 8, 2] (Cấu trúc của các lớp trong mạng nơ-ron)

4.5 KẾT QUẢ THỰC NGHIỆM

Mô hình đề xuất cùng các mô hình đơn lẻ đã được huấn luyện và kiểm định trên tập dữ liệu thực nghiệm, cho ra các chỉ số đánh giá mô hình phân loại sau:

Mô hình	Accuracy	Recall	Precision	F1-score
RDF (1)	0.7904	0.7544	0.7696	0.7810
RDF (2)	0.7878	0.7453	0.7612	0.7768
GBT	0.8090	0.8055	0.8082	0.8069
MLP	0.7760	0.7612	0.7811	0.7710
Voting	0.8136	0.8042	0.8087	0.7858

Dựa trên bảng đánh giá hiệu suất của các mô hình học máy, có các nhận xét sau:

Random Forest (RDF):

Các mô hình Random Forest (RDF) (Entropy và Gini) có độ chính xác và F1-score ổn định, khoảng 0.79 và 0.78 tương ứng. RDF (Entropy) được ưu chuộng cho mô hình Voting vì F1-score cao hơn.

Gradient Boosted Trees (GBT) đạt accuracy cao nhất (0.8090) và F1-score gần nhất ở mức 0.8069, cho thấy khả năng dự đoán mạnh mẽ trên tập dữ liệu.

Multilayer Perceptron (MLP) có hiệu suất thấp nhất với accuracy là 0.7760 và F1-score là 0.7710.

Mô hình Voting (RDF (Entropy), GBT, MLP) kết hợp kết quả từ các mô hình khác, cải thiện về accuracy và F1-score.

Mặc dù mô hình đề xuất chưa đạt hiệu suất tốt nhất, điều này có thể do cần hiệu chỉnh tham số tốt hơn, tính phức tạp của bài toán, dữ liệu thực tế chứa nhiều hoặc giá trị thiếu, và sự lựa chọn mô hình không phù hợp.

4.6 KẾT LUẬN

Nghiên cứu này phát triển một công cụ dự đoán khả năng trả nợ của khách hàng ngân hàng.

Sử dụng kết hợp của Random Forest, GBT, MLP và phương pháp Voting, mô hình GBT đạt kết quả cao nhất với Accuracy: 0.8136 và F1-score: 0.8069. Kết quả Voting có Precision cao nhất là 0.8087.

Trong tương lai, chúng tôi sẽ cải thiện mô hình bằng cách thử nghiệm các phương pháp mới và cải thiện quá trình chuẩn bị dữ liệu.

**Thank You
For
Watching!**