**ĐẠI HỌC UEH**

**TRƯỜNG CÔNG NGHỆ VÀ THIẾT KẾ**

**KHOA CÔNG NGHỆ THÔNG TIN KINH DOANH**



**BÁO CÁO DỰ ÁN**

**MÔN KHOA HỌC DỮ LIỆU**

**ĐỀ TÀI: DỰ BÁO RỦI RO TÍN DỤNG BẰNG PHƯƠNG PHÁP**

**PHÂN LỚP DỮ LIỆU**

|  |  |  |
| --- | --- | --- |
| **Giảng viên** | : | Ths. Võ Thành Đức |
| **Mã HP** | : | 23D1INF50906402 – K47 |
| **Sinh viên**  **Chuyên ngành** | :  : | Võ Minh Nguyên – 31211027655  Khoa học dữ liệu |

**TP Hồ Chí Minh, tháng 3 năm 2023**

**MỤC LỤC**

[**LỜI MỞ ĐẦU 3**](#_Toc129295525)

[**CHƯƠNG 1: GIỚI THIỆU 4**](#_Toc129295526)

[*1. Thực trạng nợ xấu tại các ngân hàng thương mại. 4*](#_Toc129295527)

[*2. Quản lý nợ xấu và vai trò của dự báo rủi ro tín dụng. 4*](#_Toc129295528)

[*3. Mục tiêu nghiên cứu. 5*](#_Toc129295529)

[*4. Câu hỏi dự báo. 5*](#_Toc129295530)

[*5. Phương pháp nghiên cứu. 5*](#_Toc129295531)

[**CHƯƠNG 2: THU THẬP & LÀM SẠCH DỮ LIỆU 6**](#_Toc129295532)

[*1. Giới thiệu sơ lược bộ dữ liệu. 6*](#_Toc129295533)

[*2. Trình bày thuộc tính. 6*](#_Toc129295534)

[*3. Chọn độ lớn mẫu. 7*](#_Toc129295535)

[*4. Tiền xử lý dữ liệu. 8*](#_Toc129295536)

[**CHƯƠNG 3: KIỂM ĐỊNH MÔ HÌNH 13**](#_Toc129295537)

[*1. Phân chia bộ dữ liệu. 13*](#_Toc129295538)

[*2. Phân lớp dữ liệu. 14*](#_Toc129295539)

[*3. Báo cáo kết quả thu được của từng phương pháp phân lớp. 15*](#_Toc129295540)

[**CHƯƠNG 4: ĐÁNH GIÁ VÀ LỰA CHỌN MÔ HÌNH 20**](#_Toc129295541)

[*1. Lựa chọn phương pháp phân lớp. 20*](#_Toc129295542)

[*2. Thực hiện dự báo rủi ro tín dụng. 21*](#_Toc129295543)

[*3. Phân tích số liệu dự báo. 24*](#_Toc129295544)

[*4. Đánh giá mô hình. 27*](#_Toc129295545)

[**CHƯƠNG 5. TRIỂN KHAI MÔ HÌNH 28**](#_Toc129295546)

[*1. Ứng dụng phương pháp phân lớp SVM trong việc dự báo rủi ro tín dụng của khách hàng. 28*](#_Toc129295547)

[*2. Hiệu quả kinh tế. 28*](#_Toc129295548)

[**ĐÁNH GIÁ SỰ ĐÓNG GÓP CỦA CÁC THÀNH VIÊN 29**](#_Toc129295549)

# **LỜI MỞ ĐẦU**

Lời đầu tiên, em xin gửi lời cảm ơn chân thành nhất đến giảng viên bộ môn - Ths. Võ Thành Đức. Trong suốt quá trình học tập và tìm hiểu bộ môn Khoa học dữ liệu, chúng đã nhận được sự quan tâm, giúp đỡ và hướng dẫn vô cùng tận tâm từ thầy. Qua đó chúng đã có thể tiếp thu được những kiến thức mới mẻ và đầy thú vị của bộ môn Khoa học dữ liệu.

Với thái độ học tập nghiêm túc và những kiến thức đã tích lũy được, chúng đã hoàn thành bài tiểu luận của mình dưới đây. Tuy nhiên, trong quá trình thực hiện bài tiểu luận chắc chắn không thể tránh khỏi những thiếu sót. Em rất mong nhận được những sự góp ý từ thầy để em thể sửa chữa những sai sót, rút kinh nghiệm cho bản thân và chắc chắn đó sẽ là hành trang vô cùng quý báu cho bọn về sau.

Em xin chân thành cảm ơn!

# **CHƯƠNG 1: GIỚI THIỆU**

## **Thực trạng nợ xấu tại các ngân hàng thương mại.**

* + Đại dịch Covid-19 bùng phát từ đầu năm 2020 và tiếp tục kéo dài với diễn biến phức tạp, đã ảnh hưởng sâu sắc tới toàn bộ nền kinh tế, gây những tác động tiêu cực tới hoạt động của các doanh nghiệp, ảnh hưởng nghiêm trọng tới khả năng thanh toán của doanh nghiệp.
  + Tỷ lệ nợ xấu toàn ngành ngân hàng tăng mạnh kể từ năm 2020, và dự kiến sẽ còn tăng trong giai đoạn tới. Trên thực tế, để ứng phó với diễn biến tiêu cực của nợ xấu, các thể chế tín dụng đã chủ động phân loại nợ, trích lập dự phòng rủi ro (tỷ lệ bao phủ nợ xấu bình quân của 28 NHTM niêm yết và Agribank (chiếm khoảng 80% thị phần tổng tài sản) đã tăng lên mức 150% cuối năm 2021, là mức cao nhất từ trước tới nay), song không thể phủ nhận nợ xấu vẫn còn tiềm ẩn rất nhiều rủi ro, gánh nặng đối với hệ thống thể chế tín dụng là không nhỏ.
  + Nợ xấu của hệ thống các thể chế tín dụng gia tăng là điều đã được dự báo trước khi mà sự bùng phát của đại dịch Covid-19, và đặc biệt là làn sóng thứ 4 với biến chủng Delta trong năm 2021 đã gây ra các tổn thất nặng nề đối với các hoạt động sản xuất, kinh doanh của doanh nghiệp, sinh kế và đời sống của người dân. Theo báo cáo tài chính năm 2021 mới được các ngân hàng công bố, nợ xấu có xu hướng gia tăng rõ rệt tại một số ngân hàng, thí dụ như VPBank (tăng 60% so với 2020), Vietinbank (49%), VIB (58%), HDB (43%)…v.v.; bình quân số dư nợ xấu 28 NHTM niêm yết và Agribank tăng 17,3% so với năm 2020.

## **Quản lý nợ xấu và vai trò của dự báo rủi ro tín dụng.**

* + Qua việc tiếp nhận thông tin trên, cũng như các dữ kiện từ báo đài và các trang báo xã hội, ta có thể thấy hiện nay việc sử dụng vay tín dụng nhưng không thực hiện trả nợ dẫn đến nợ xấu của nhiều khách hàng đang có xu hướng tăng ở phần lớn các ngân hàng trên cả nước.
  + Nợ xấu là những khoản vay mà người vay không thể thanh toán đúng kỳ hạn trong hợp đồng tín dụng. Điều này có thể có nhiều nguyên nhân khác nhau, nhưng thường bởi các nguyên nhân sau:
* Bên vay không có đủ điều kiện chi trả.
* Không muốn chi trả đúng hạn.
* Nguồn thu nhập của họ gặp rủi ro và chưa thể trả nợ đúng hạn hợp đồng.
  + Vì vậy, việc quản lý và dự đoán những rủi ro của những khoản vay tín dụng là cực kỳ cần thiết để tránh tình trạng ngân hàng gặp phải nhiều trường hợp nợ xấu và dễ dàng phân loại khoản vay và quản lý danh mục đầu tư tín dụng của ngân hàng.
  + Quá trình quản lý nợ xấu được cụ thể bằng việc:
* Thu thập và xử lý thông tin của khách hàng.
* Phân loại danh mục các khoản vay của khách hàng.
* Kiểm tra, giám sát mức độ tuân thủ theo đúng cam kết trong Hợp đồng tín dụng.
* Nắm được tình hình sản xuất kinh doanh hiện tại của khách hàng, sớm nhận biết các dấu hiệu tiềm ẩn để kịp thời có các biện pháp xử lý, tránh để phát sinh hay chuyển thành nợ xấu.
* Thông qua đó, em em quyết định lựa chọn đề tài **DỰ BÁO RỦI RO TÍN DỤNG** để đánh giá và chẩn đoán mức độ rủi ro trong việc vỡ nợ tín dụng của khách hàng, cũng như từ đó có thể đưa ra được các phương pháp nghiên cứu trong thực tiễn.

1. **Mục tiêu nghiên cứu.**
   * Bài toán mà em đưa ra sẽ là tập trung phân tích khả năng vỡ nợ tín dụng. Cụ thể là dự đoán dựa trên những đặc tính đã thu thập được từ bộ dữ liệu của khách hàng để đưa ra kết luận rằng: khách hàng “**có vỡ nợ**” hay “**không vỡ nợ**”. Từ đó, tạo cơ sở để các ngân hàng đưa ra những quyết định chính xác hơn khi cho vay.

## **Câu hỏi dự báo.**

* + Khả năng vỡ nợ của khách hàng: “**Có**” hoặc “**Không**” ?

## **Phương pháp nghiên cứu.**

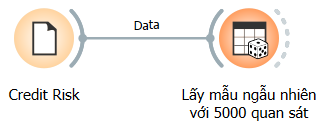
* + Em sẽ thực hiện tiền xử lý dữ liệu trên phần mềm Orange, sau đó trích bộ dữ liệu làm hai phần: 80% dữ liệu sẽ được thực hiện phân lớp và tạo mô hình dự báo. Từ cơ sở đó sẽ sử dụng mô hình để dự báo cho 20% bộ dữ liệu còn lại.

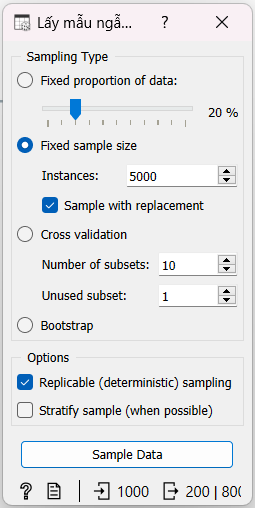
# **CHƯƠNG 2: THU THẬP & LÀM SẠCH DỮ LIỆU**

* + - 1. **Giới thiệu sơ lược bộ dữ liệu.**
  + Bộ dữ liệu Credit Risk được thu thập từ trang thông tin đáng tin cậy là Kaggle.com. Từ dữ liệu khách hàng được thu thập chi tiết ở đây mà em có thể thực hiện được các thao tác phân tích dữ liệu, để từ đó có thể đưa ra được những dự báo tiếp theo cho mình.

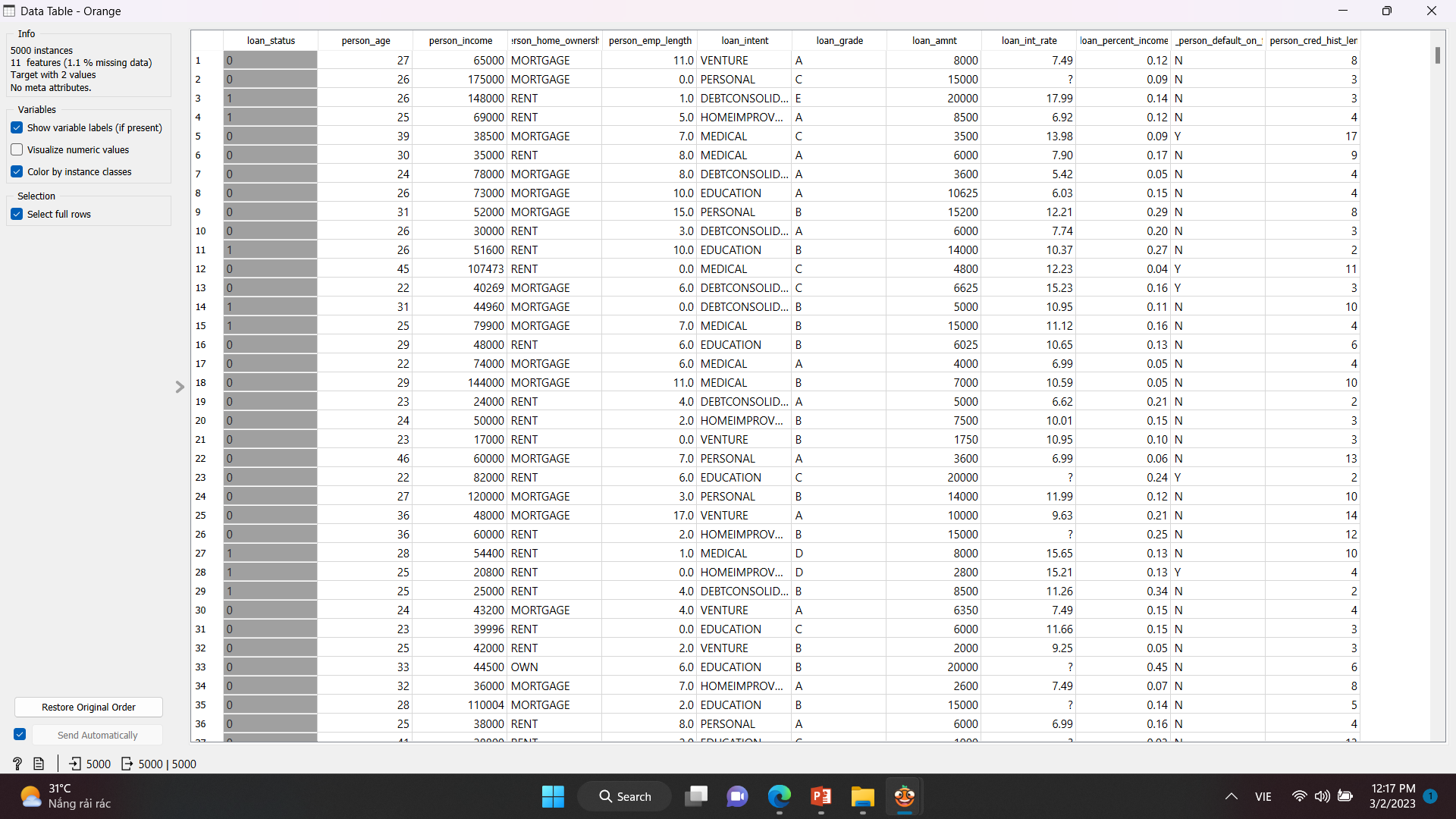
## **Trình bày thuộc tính.**

* + Dữ liệu có tổng cộng **32581** quan sát, vì số quan sát quá lớn khiến cho quá trình phân tích, chia sẻ hay lưu trữ bị chậm đi, thế nên để thuận tiện hơn em sẽ dùng **Date Sampler Widget** có trong phần mềm Orange để giảm số quan sát xuống còn **5000** mỗi khi thực hiện phân tích.
  + Tổng cộng có: **12** biến dữ liệu, trong đó có **1.1%** dữ liệu bị thiếu. Các thuộc tính bao gồm:
* **Biến dữ liệu số (Numeric Data):**
* Person\_age: Số tuổi của khách hàng.
* Person\_income: Thu nhập của khách hàng.
* Person\_p\_length: Thời gian làm việc của khách hàng (tính theo năm).
* Loan\_amnt: Số tiền vay.
* Loan\_int\_rate: Thông tin lãi suất vay.
* Loan\_percent\_income: Phần trăm thu nhập.
* Cb\_person\_cred\_hist\_length: Độ dài lịch sử tín dụng.
* **Các biến thuộc tính (Categorical Data):**
* Person\_home\_ownership: Quyền sở hữu nhà (bao gồm: chủ nhà (Own), nhà cho thuê (Rent), nhà thế chấp (Mortgage)).
* Loan\_intent: Mục đích vay.
* Loan\_grade: Điểm vay (dùng để đánh giá sự uy tín của cá nhân khi sử dụng hình thức cho vay của tổ chức).
* Cb\_person\_default\_on\_file: Tình trạng vỡ nợ.
* **Biến mục tiêu:**
* Loan\_status: Tình trạng nợ (với 1 là đại diện cho vỡ nợ và 0 đại diện cho không vỡ nợ).

1. **Chọn độ lớn mẫu.**
   * Như đã đề cập từ ban đầu, em sẽ thực hiện cắt nhỏ mẫu dữ liệu để thuận tiện hơn trong việc khai thác, cũng như tránh làm mất quá nhiều thời gian. Với **Data Sampler Widget**, ta sẽ lấy được ngẫu nhiên mẫu với **5000** quan sát để tiến hành phân tích.



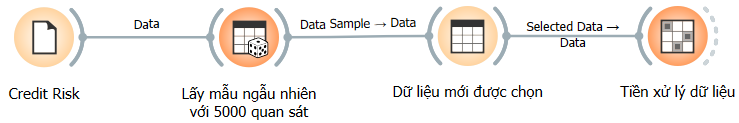
* + Để x được bộ dữ liệu vừa tạo, ta sử dụng **Data Table Widget**:

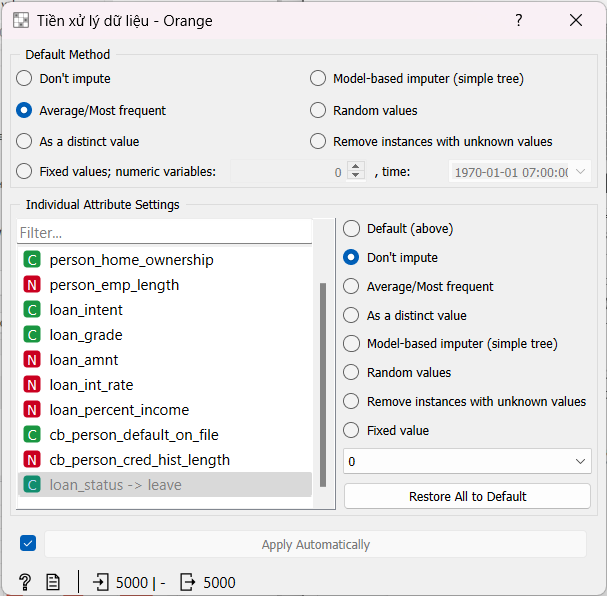


* Sau khi thực hiện lấy mẫu mới, ta vẫn còn thấy rằng: có **1.1%** dữ liệu đang bị thiếu. Vì vậy, em sẽ thực hiện tiền xử lý dữ liệu trong phần tiếp theo.

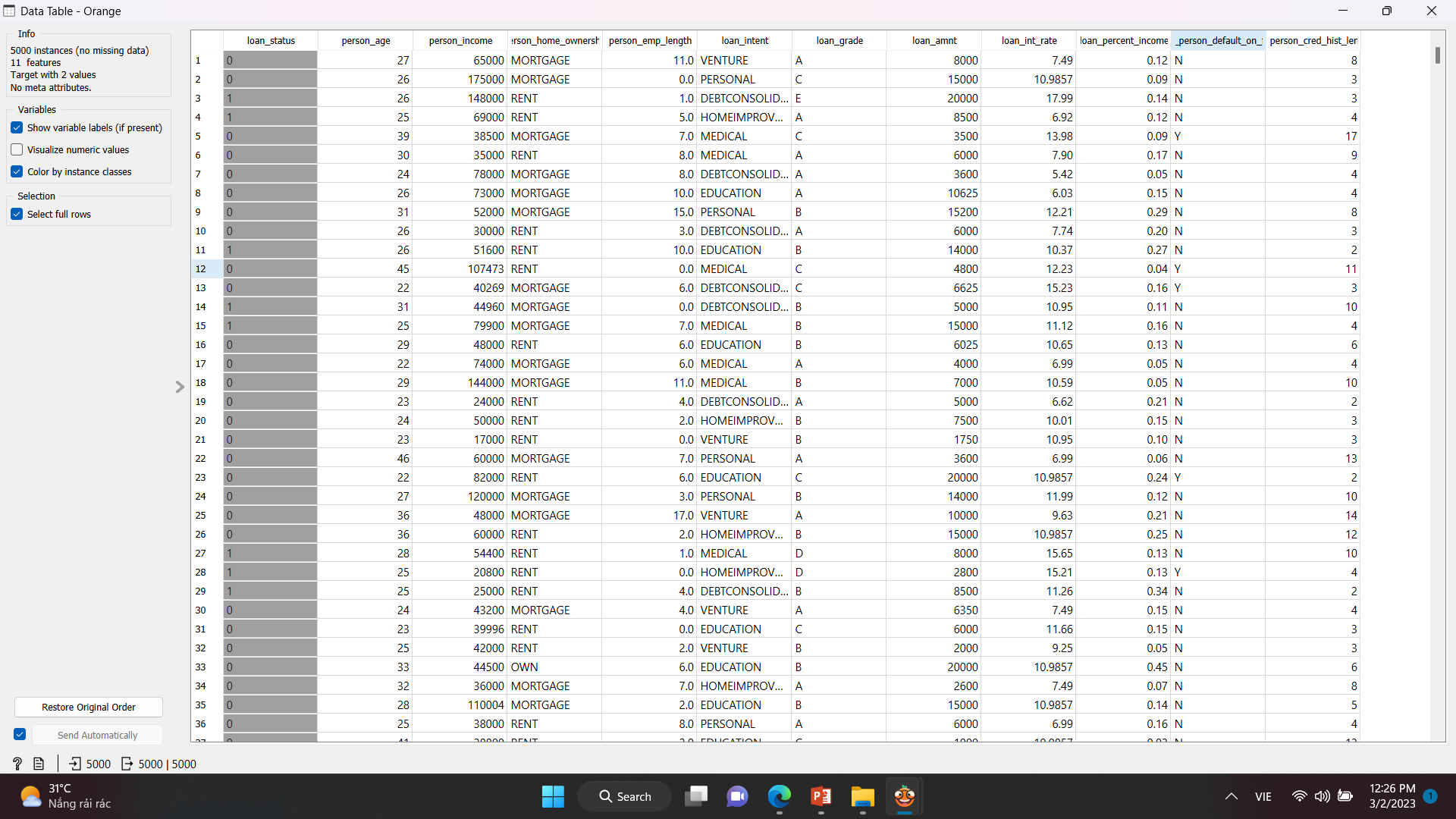
## **Tiền xử lý dữ liệu.**

* + Đối với bài nghiên cứu này, em sẽ tập trung vào bước làm sạch dữ liệu (cleansing). Đây là bước đầu tiên và cũng là bước quan trọng nhất, bởi khi không được xử lý sẽ dẫn đến việc cho ra kết quả khai thác dữ liệu không chính xác.
  + Với bộ dữ liệu bị thiếu, ta chọn **Impute Missing Value Widget** có trong phần mềm Orange, tiếp tục chọn vào **Average/ Most Frequent** để có thể thay thế các

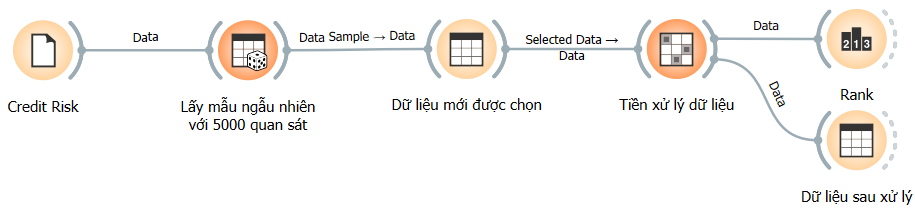
dữ liệu bị thiếu bằng giá trị trung bình của từng biến (hay giá trị có mật độ xuất hiện nhiều nhất trong 5000 quan sát).

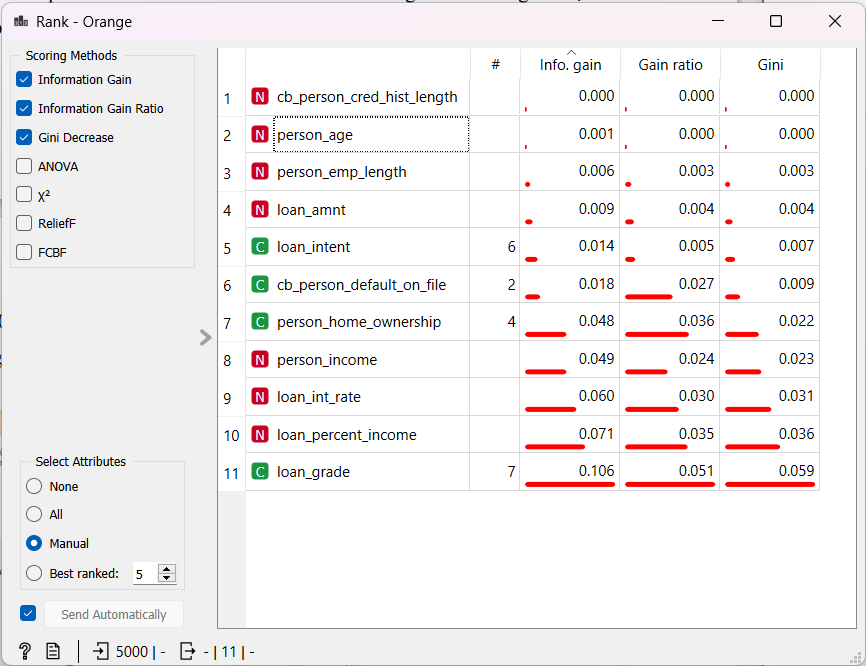


* + Dữ liệu sau khi được xử lý:

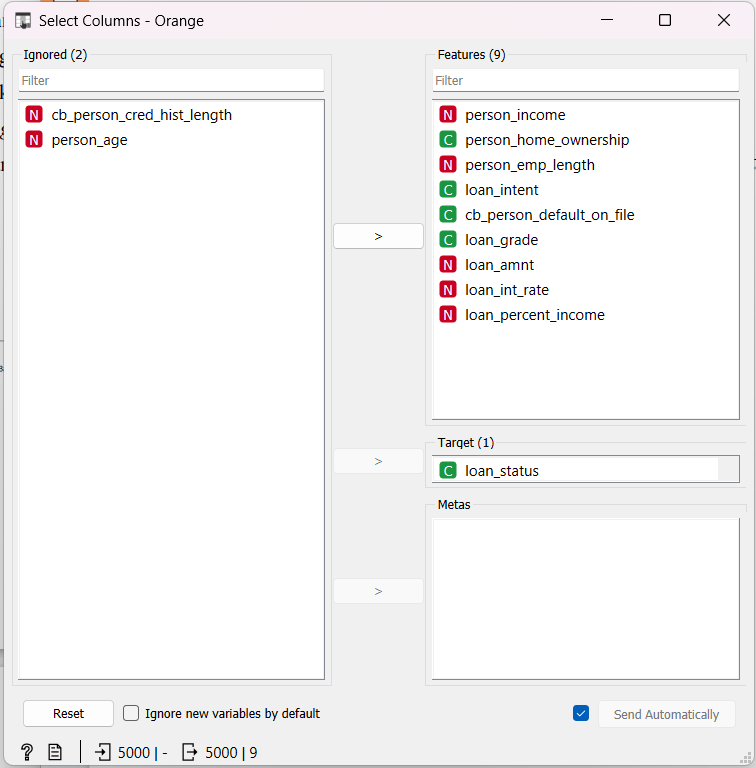


* + Sau khi xử lý dữ liệu bị thiếu, em tiếp tục lược bỏ bớt các biến không cần thiết. Bằng **Rank Widget**, ta có thể quan sát được các chỉ số như: **Information Gain**, **Gain Ratio**, … của từng thuộc tính, từ đó quan sát được sự tương quan của nó đối với biến phụ thuộc là **loan\_status**. Các biến có chỉ số **Gain Ratio** thấp như **cb\_person\_cred\_hist\_length** và **person\_age** (đều là **0.000**) sẽ bị loại bỏ vì không thể hiện sự tương quan cao với biến cần thực hiện phân tích.

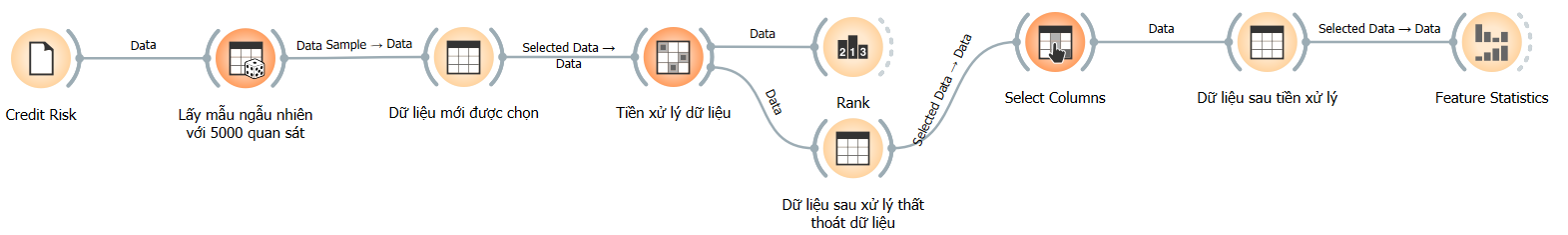


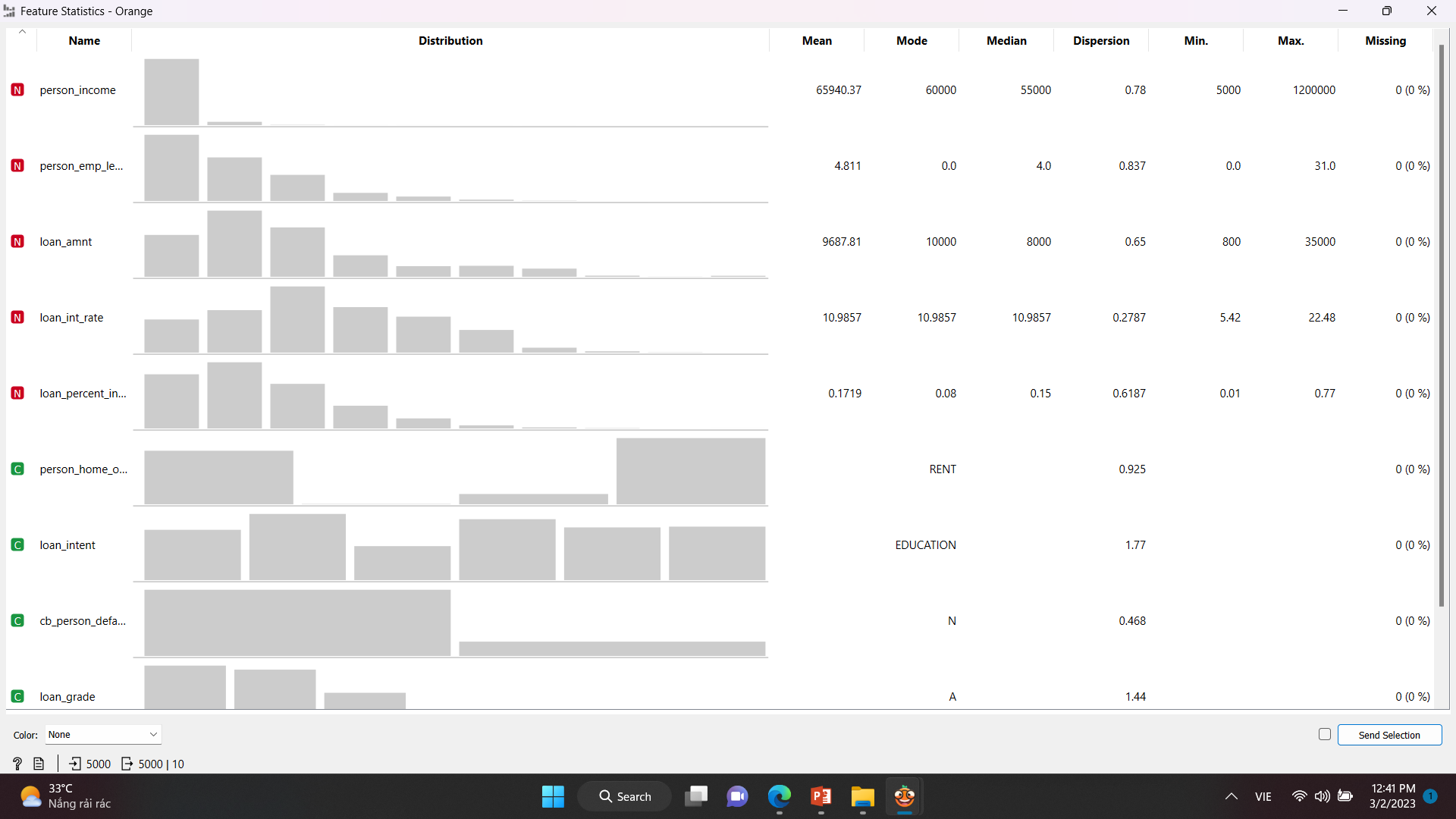


* + Để loại bỏ các biến, ta sử dụng **Select Columns Widget**. Thực hiện kéo thả hai biến **cb\_person\_cred\_hist\_length** và **person\_age** vào phẩn **Ignored**.



* + Để có thể x lại tổng quan bộ dữ liệu sau khi được xử lý, ta sử dụng **Feature Statistics Widget**.

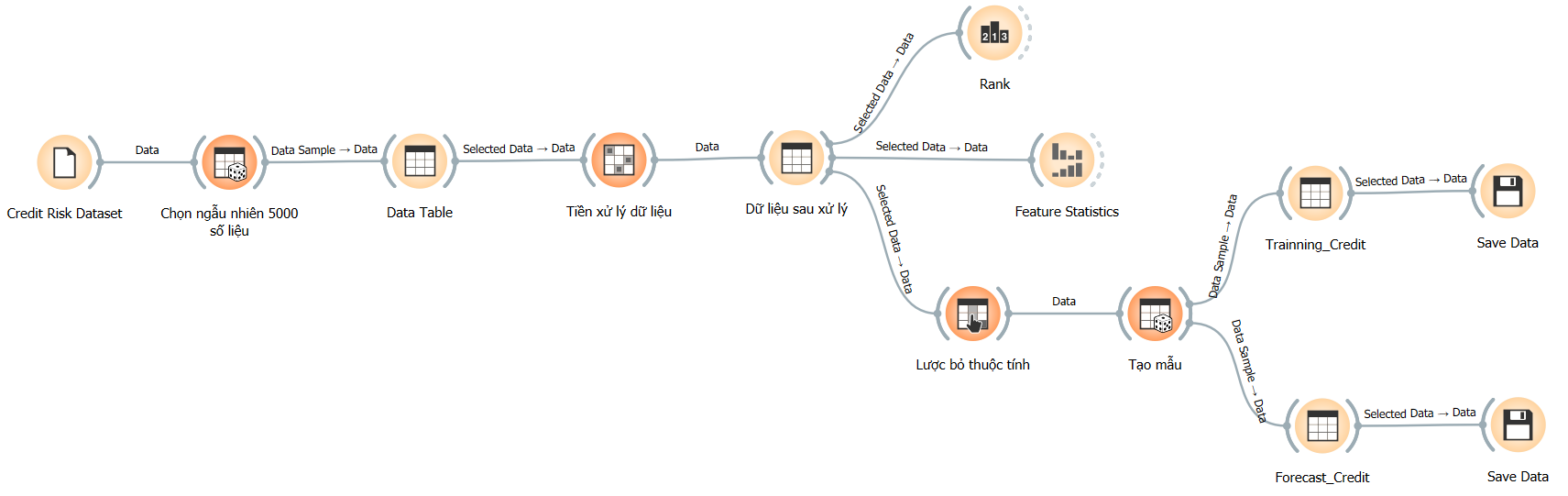




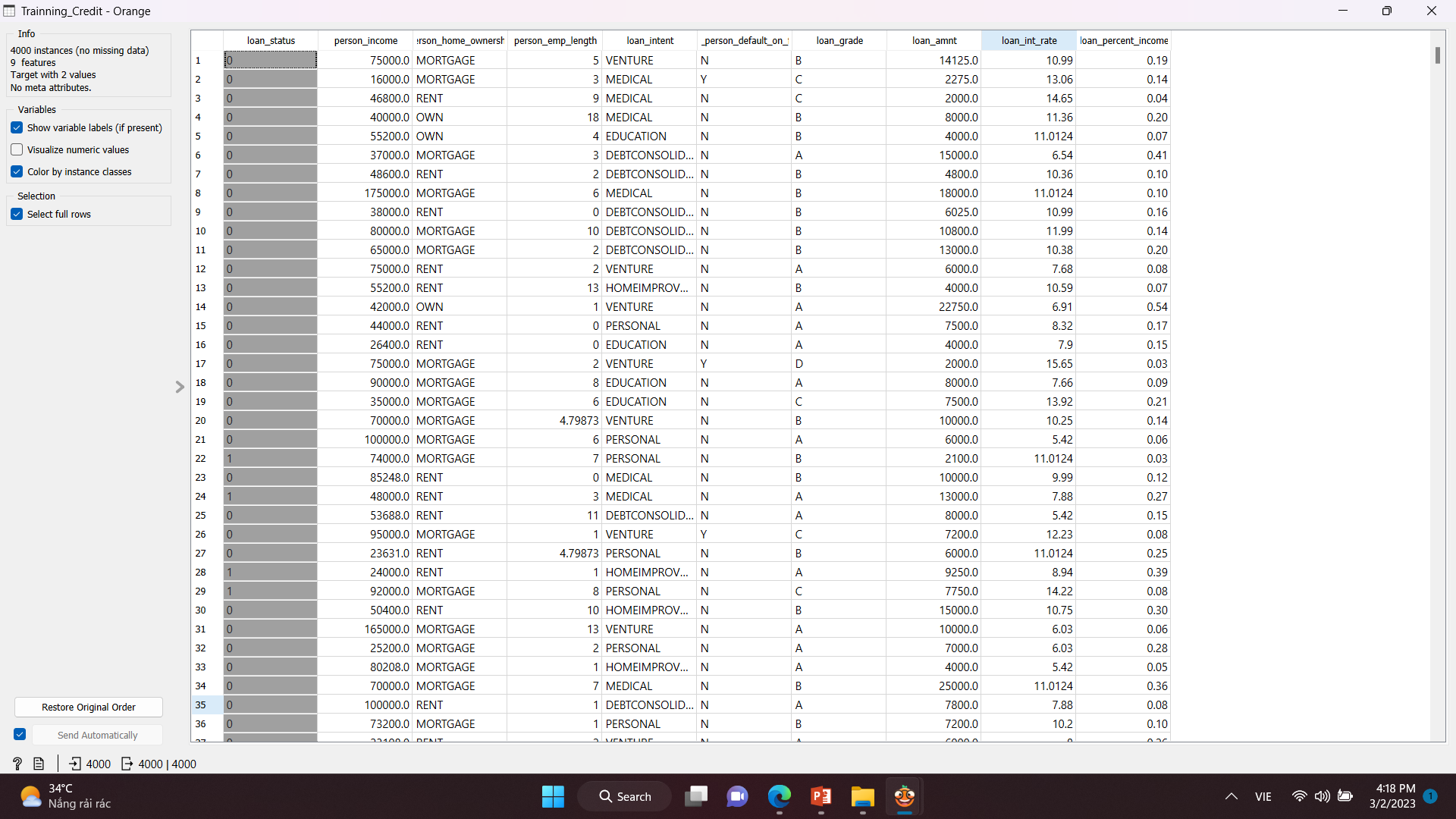
# **CHƯƠNG 3: KIỂM ĐỊNH MÔ HÌNH**

## **Phân chia bộ dữ liệu.**

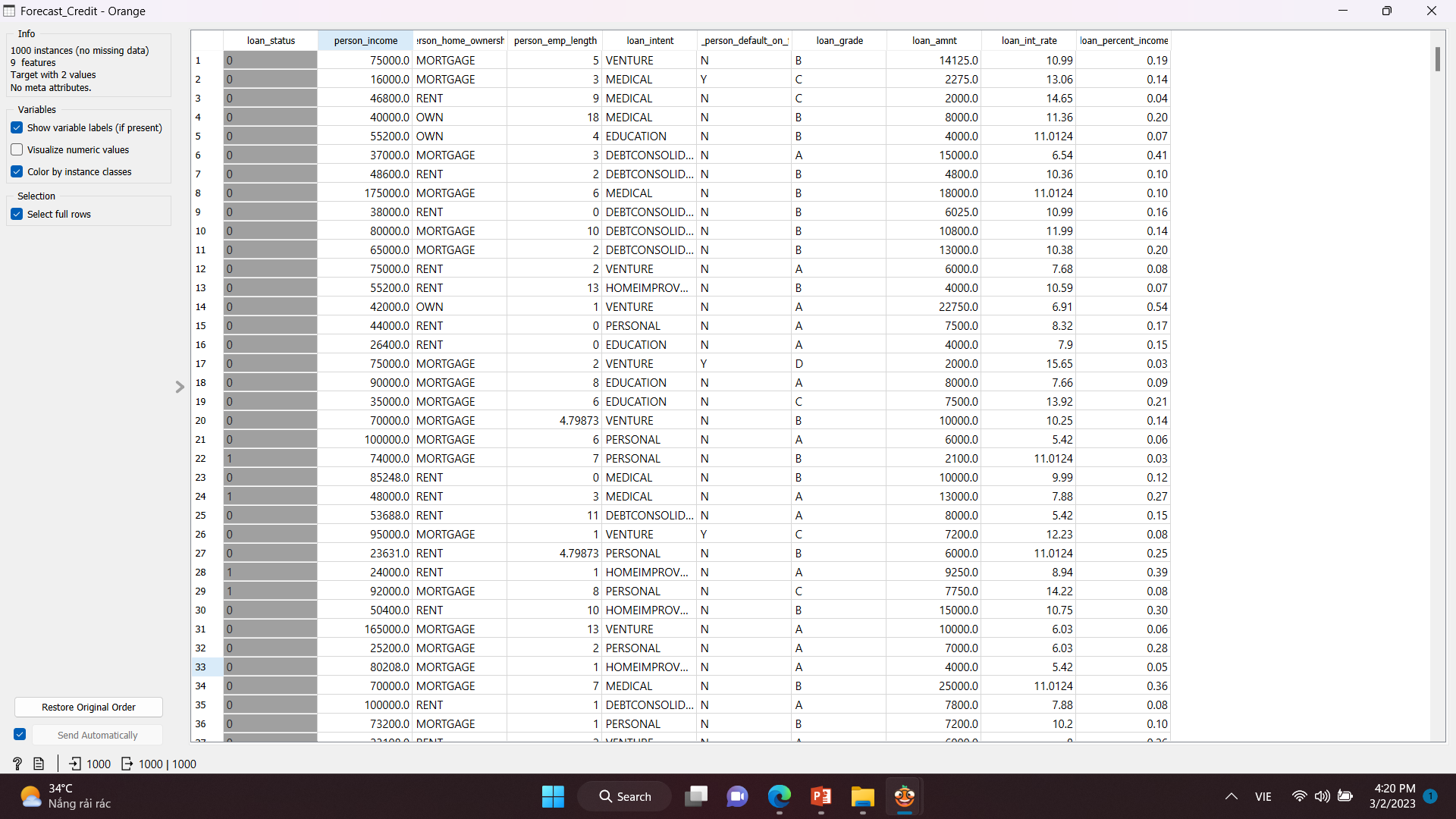
* + Như thông tin đã nêu ở phần phương pháp nghiên cứu, em thực hiện tách bộ dữ liệu đang sử dụng với 5000 quan sát thành hai phần. Phương pháp tách như sau: dùng **Data Sampler Widget** lấy ngẫu nhiên **20%** dữ liệu để làm mẫu dự báo được lưu dưới tên **Forecast\_Credit**. Phần còn lại của bộ dữ liệu sẽ lưu dưới tên **Trainning\_Credit**, cũng như chọn ra kỹ thuật phân lớp tốt nhất để huấn luyện bộ dữ liệu vừa tạo.



* + Với bộ dữ liệu có tên **Trainning\_Credit**, ta thu được mẫu có **4000** quan sát và không có bất kỳ dữ liệu nào bị thiếu.

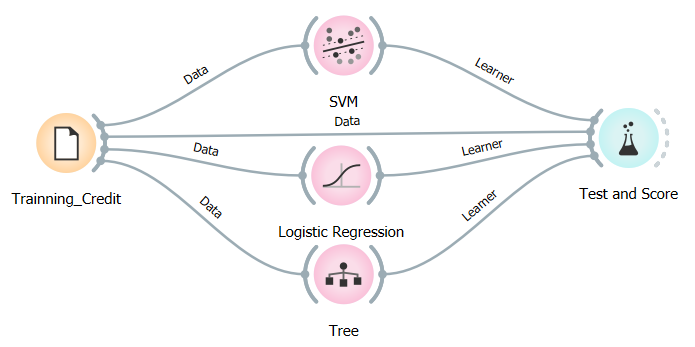
****

* + Bộ dữ liệu được trích với tỷ lệ **20%** có tổng cộng **1000** quan sát và cũng không có bất kỳ dữ liệu nào bị thất thoát.

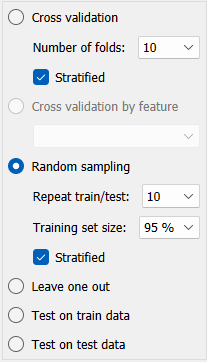
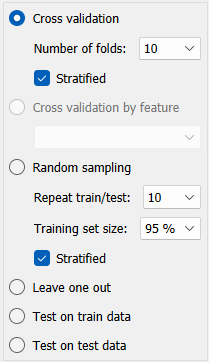


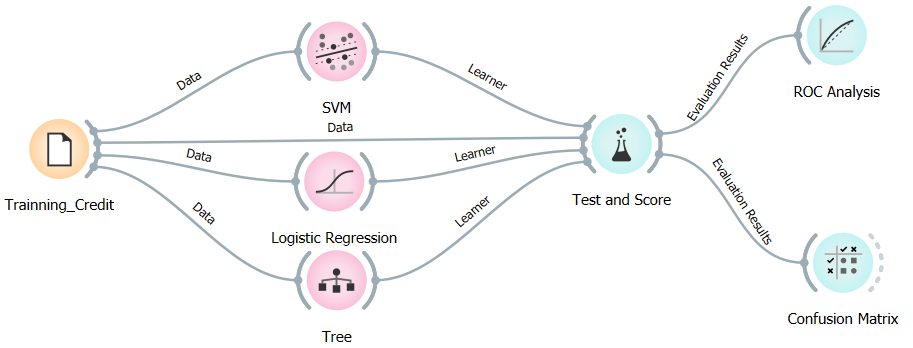
## **Phân lớp dữ liệu.**

* + Sử dụng các **Widget** sau để thực hiện phân lớp bộ dữ liệu: **SVM**, **Tree** và **Logistic Regression** (lần lượt là các phương pháp phân lớp: Support Vector Machine, cây quyết định và hồi quy Logistic). Bên cạnh đó, kết hợp sử dụng công cụ **Test and Score** để x được các chỉ số đánh giá giữa 3 phương pháp phân lớp đó.



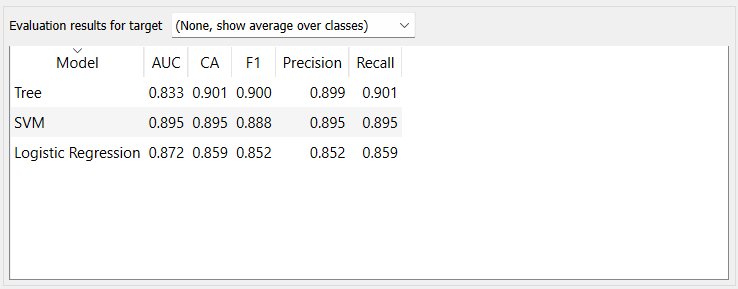
* + Tiếp tục tại **Test and Score Widget**, ta thực hiện phân chia dữ liệu theo từng phương pháp là: **Cross Validation** hoặc **Random Sampling** nhằm tìm ra phương pháp có chỉ số phù hợp nhất để dự báo. Bên cạnh đó, để có thể đánh giá được sâu và chính xác hơn ta sẽ tham khảo chỉ số của từng công cụ sau, là: **Confusion Matrix** và **ROC Analysis**.

****

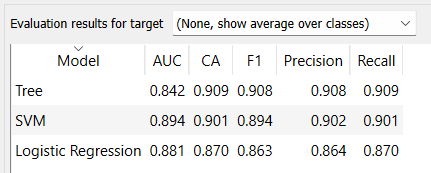
****

## **Báo cáo kết quả thu được của từng phương pháp phân lớp.**

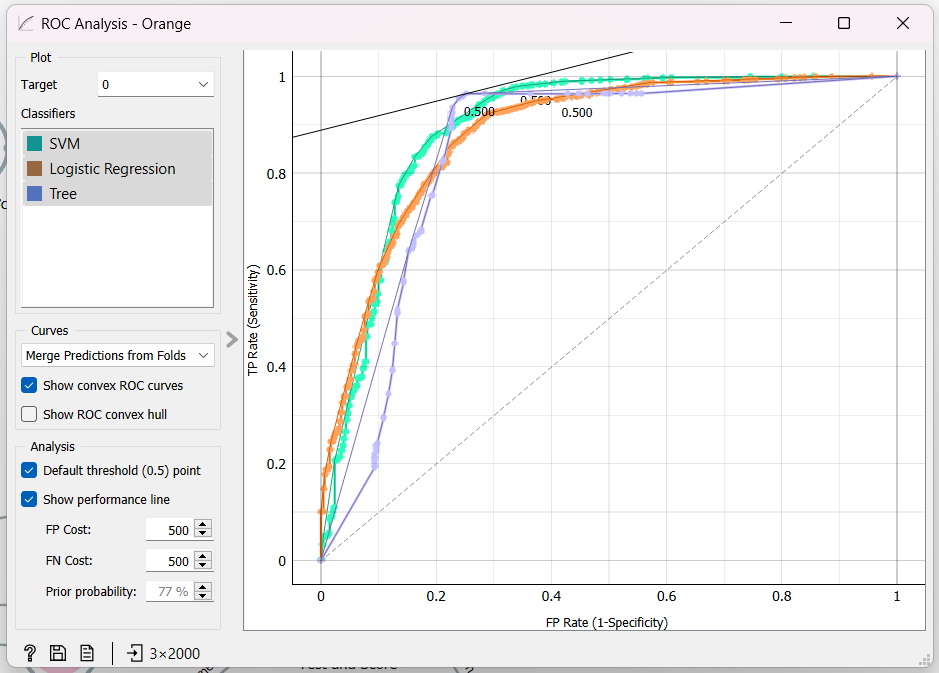
* + Với việc sử dụng **Cross Validation** và **Number of Folds** để chia mẫu dữ liệu thành **10** phần, ta thu được các chỉ số sau:



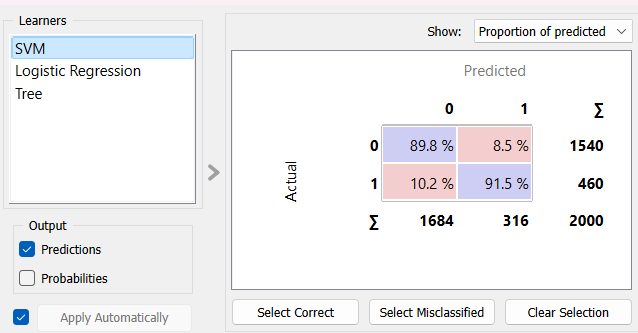
* + Kết quả trả về của phương pháp **Random Sampling** cũng gần như là tương tự:



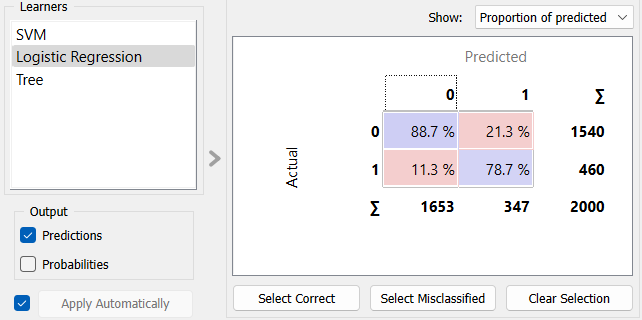
* + Ý nghĩa của các chỉ số như sau:
* AUC: diện tích nằm dưới đường cong, với giá trị càng lớn thì mô hình càng tốt.
* CA: là tỷ lệ quan sát được phân lớp đúng trong toàn bộ tập dữ liệu.
* F1: giá trị trung bình điều hòa của hai độ đo *Precision* và *Recall*.
* Precision: độ chính xác (cho biết trong số m mẫu được phân vào lớp i thì có tỷ lệ bao nhiêu mẫu có đúng).
* Recall: là độ phủ hay độ nhạy.
  + Trực quan hóa số liệu thu được bằng cách sử dụng **đồ thị ROC**:



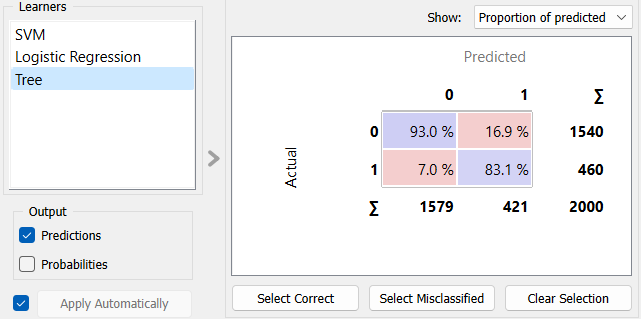
* + Các chỉ số thu được từ ma trận nhầm lẫn (**Confusion Matrix**):



* Kết quả phân lớp của phương pháp **SVM**, em thu được các thông tin sau:
* 89.8% người được dự đoán đúng với thực tế là không vỡ nợ.
* 8.5% người được dự đoán là vỡ nợ nhưng thực tế là không vỡ nợ.
* 91.5% người được dự đoán đúng với thực tế là vỡ nợ.
* 10.2% người được dự đoán là không vỡ nợ và trên thực tế là vỡ nợ.



* Kết quả phân lớp của phương pháp **hồi quy Logistic**, em thu được các thông tin sau:
* 88.7% người được dự đoán đúng với thực tế là không vỡ nợ.
* 21.3% người được dự đoán là vỡ nợ nhưng thực tế là không vỡ nợ.
* 78.7% người được dự đoán đúng với thực tế là vỡ nợ.
* 11.3% người được dự đoán là không vỡ nợ và trên thực tế là vỡ nợ.

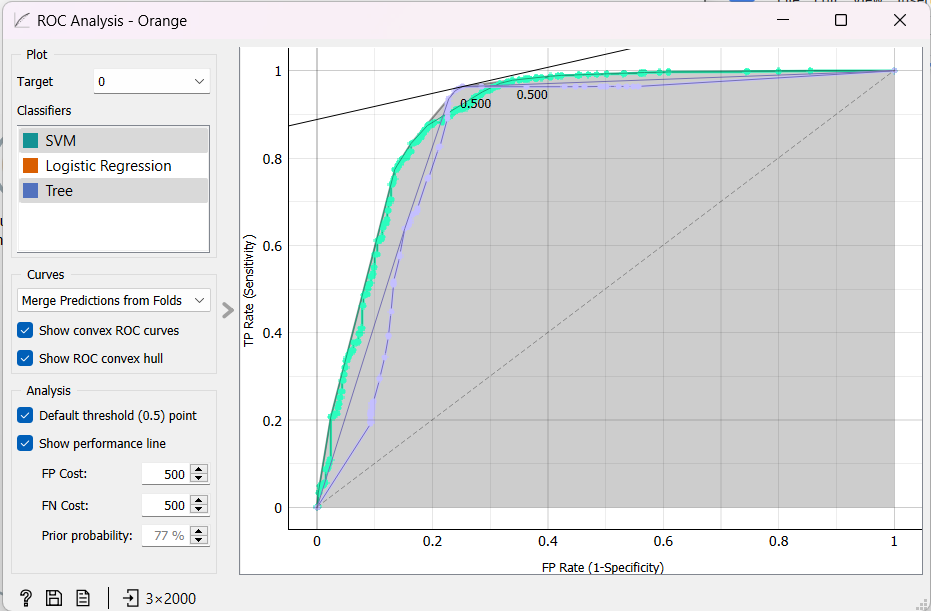


* Kết quả phân lớp của phương pháp **cây quyết định**, em thu được các thông tin sau:
* 93.0% người được dự đoán đúng với thực tế là không vỡ nợ.
* 16.9% người được dự đoán là vỡ nợ nhưng thực tế là không vỡ nợ.
* 83.1% người được dự đoán đúng với thực tế là vỡ nợ.
* 7.0% người được dự đoán là không vỡ nợ và trên thực tế là vỡ nợ.

# **CHƯƠNG 4: ĐÁNH GIÁ VÀ LỰA CHỌN MÔ HÌNH**

## **Lựa chọn phương pháp phân lớp.**

* + Từ số liệu thu được ở bảng kết quả phân lớp bằng **Test and Score Widget**, ta thấy được rằng: hai phương pháp **cây quyết định** và **SVM** đều có chỉ số gần như là tương đồng nhau (sử dụng kết hợp **biểu đồ ROC** để theo dõi các chỉ số được trực quan hơn):



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **AUC** | **CA** | **F1** | **Precision** | **Recall** |
| **SVM** | 0.894 | 0.901 | 0.894 | 0.902 | 0.901 |
| **Tree** | 0.842 | 0.909 | 0.908 | 0.908 | 0.909 |

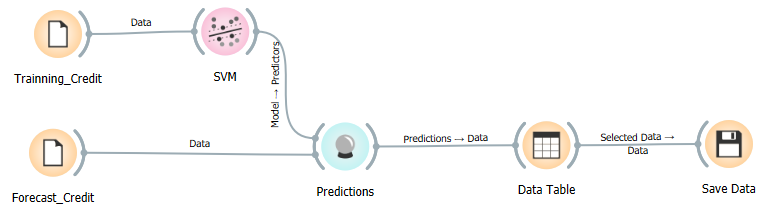
* Đối với **SVM**, chỉ số **AUC** dường như là nổi bật nhất. Nhưng với **Tree** thì cả 4 chỉ số còn lại đều vượt một khoảng nhỏ so với **SVM**.
* Từ đây ta thấy được đường cong ROC màu xanh là (đại diện cho phương pháp SVM) là đường tiệm cận với điểm (0;1) nhất hay diện tích nằm dưới đường cong ROC là lớn nhất, cho thấy hiệu quả khi thực hiện mô hình phân lớp này là tốt hơn.
  + Để có thể đưa ra được cái nhìn bao quát hơn, cũng như lựa chọn được chính xác phương pháp phân lớp thích hợp nhất, ta quan sát các kết quả còn lại ở **ma trận nhầm lẫn**:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Sai lầm loại 1** | **Sai lầm loại 2** | **Tổng** |
| **Hồi quy Logistic** | 21.3% | 11.3% | 32.6% |
| **Tree** | 16.9% | 7.0% | 23.9% |
| **SVM** | 8.5% | 10.2% | 18.7% |

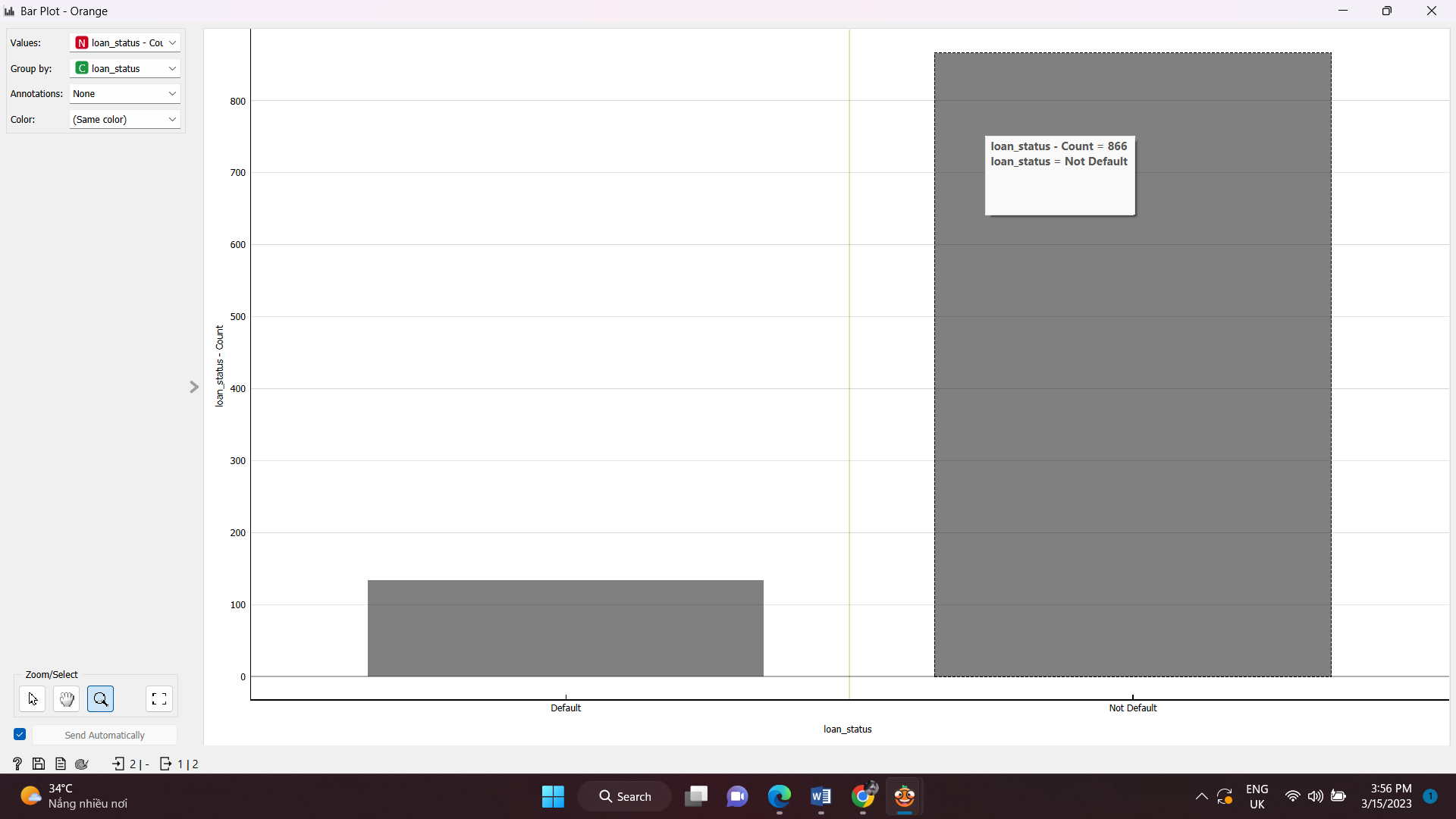
* Với sai lầm loại 1 (FP): Dự báo có mà thực tế là không.
* Và sai lầm loại 2 (FN): Dự báo không mà thực tế có.
* Với số liệu trên ta tiếp tục thấy được: **Tree** có tổng chỉ số sai lầm là **23.9%** nhưng với sai lầm loại 2 (sai lầm được x là nguy hiểm nhất, hậu quả mang lại cũng rất lớn và khó giải quyết đối với các ngân hàng) lại thấp nhất (chỉ với **7.0%**), tuy nhiên **SVM** là phương pháp có tổng chỉ số sai lầm thấp nhất trong 3 phương pháp (**18.7%**).
* Chính vì thế ta sẽ lựa chọn phương pháp **SVM** để thực hiện huấn luận bộ dữ liệu.

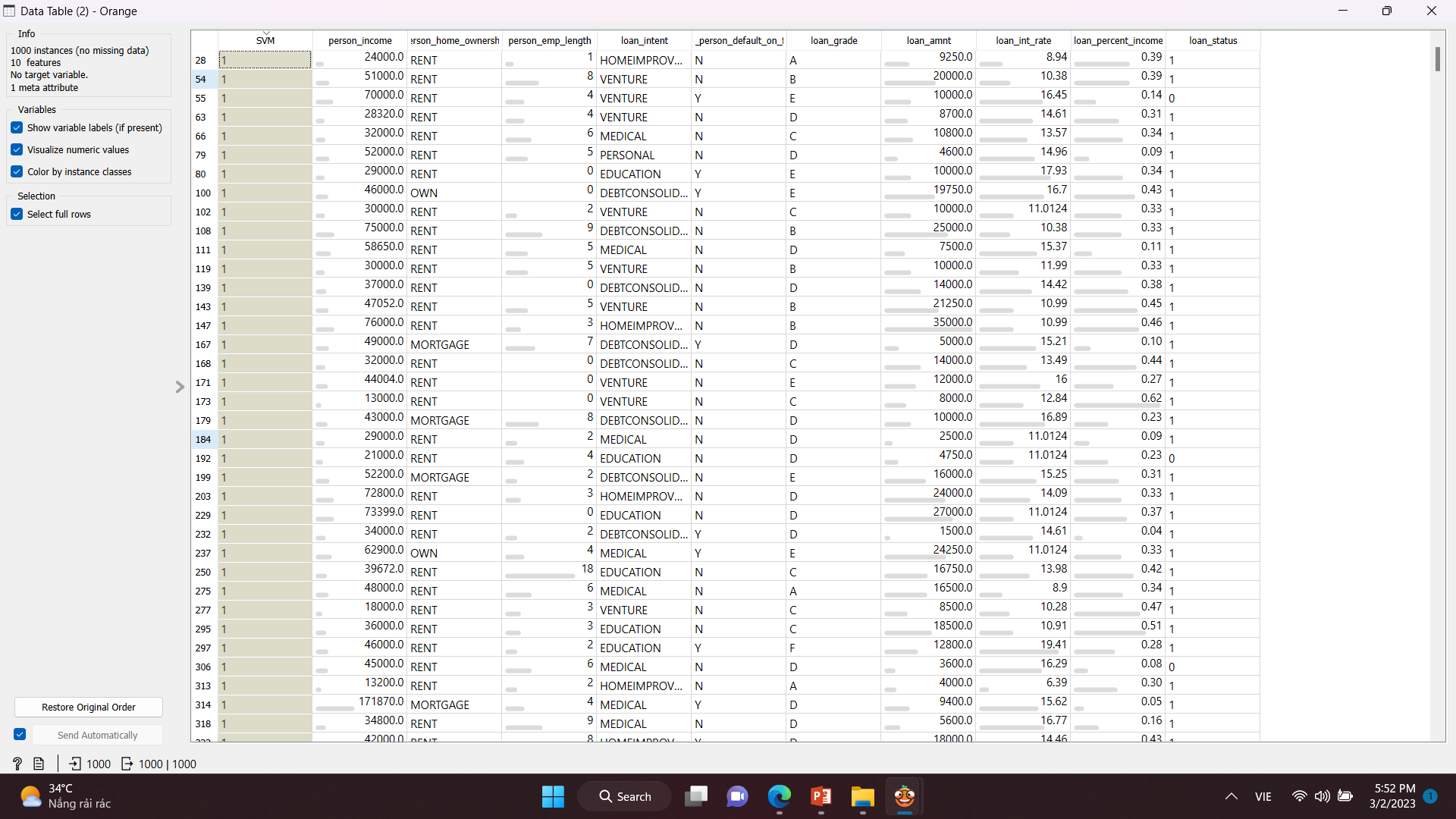
## **Thực hiện dự báo rủi ro tín dụng.**

* + Sau khi lựa chọn được phương pháp phân lớp thích hợp nhất là **SVM**, ta tiến hành dự báo cho bộ dữ liệu **Forecast\_Credit** đã chia trước đó bằng **Prediction Widget** có trên Orange:

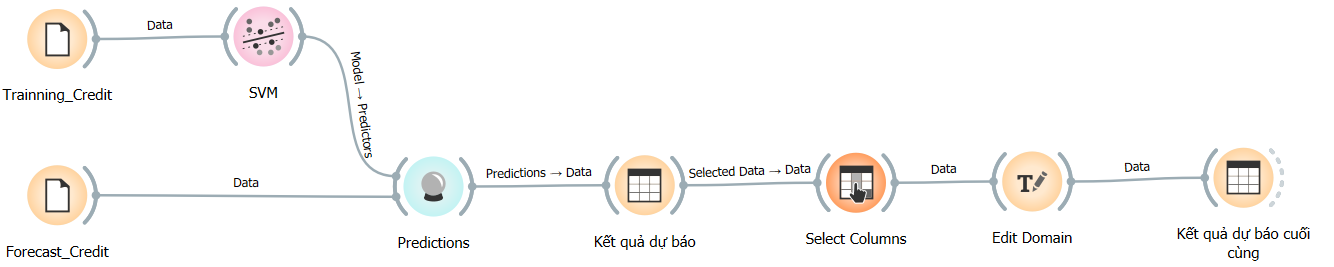


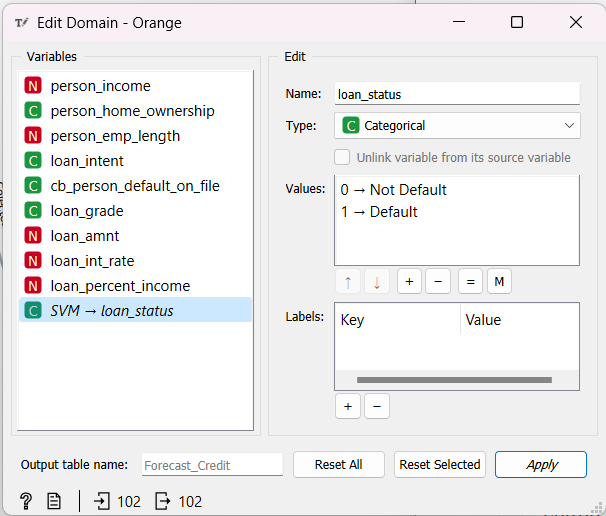
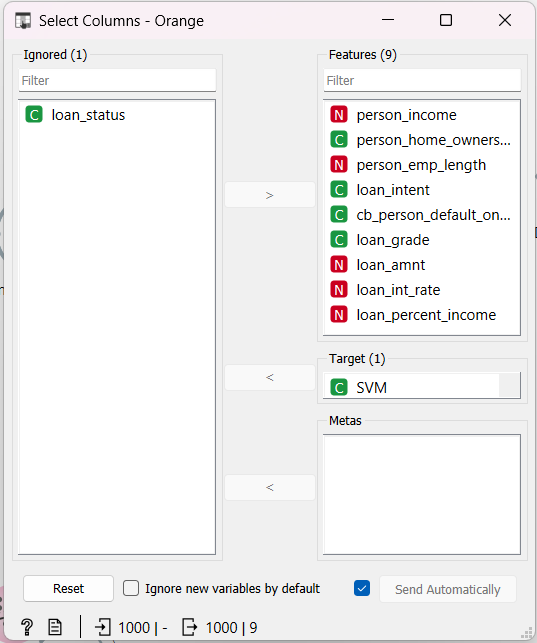
* + Kết quả dự báo thu được như sau:
* Tổng cộng có **134** quan sát cho kết quả là **1 - vỡ nợ** (**13.4%**) và **866** quan sát có kết quả **0 - không vỡ nợ** (**86.6%**).
* Thay vì **102** quan sát **vỡ nợ** và **898** quan sát **không vỡ nợ** như số liệu cũ.

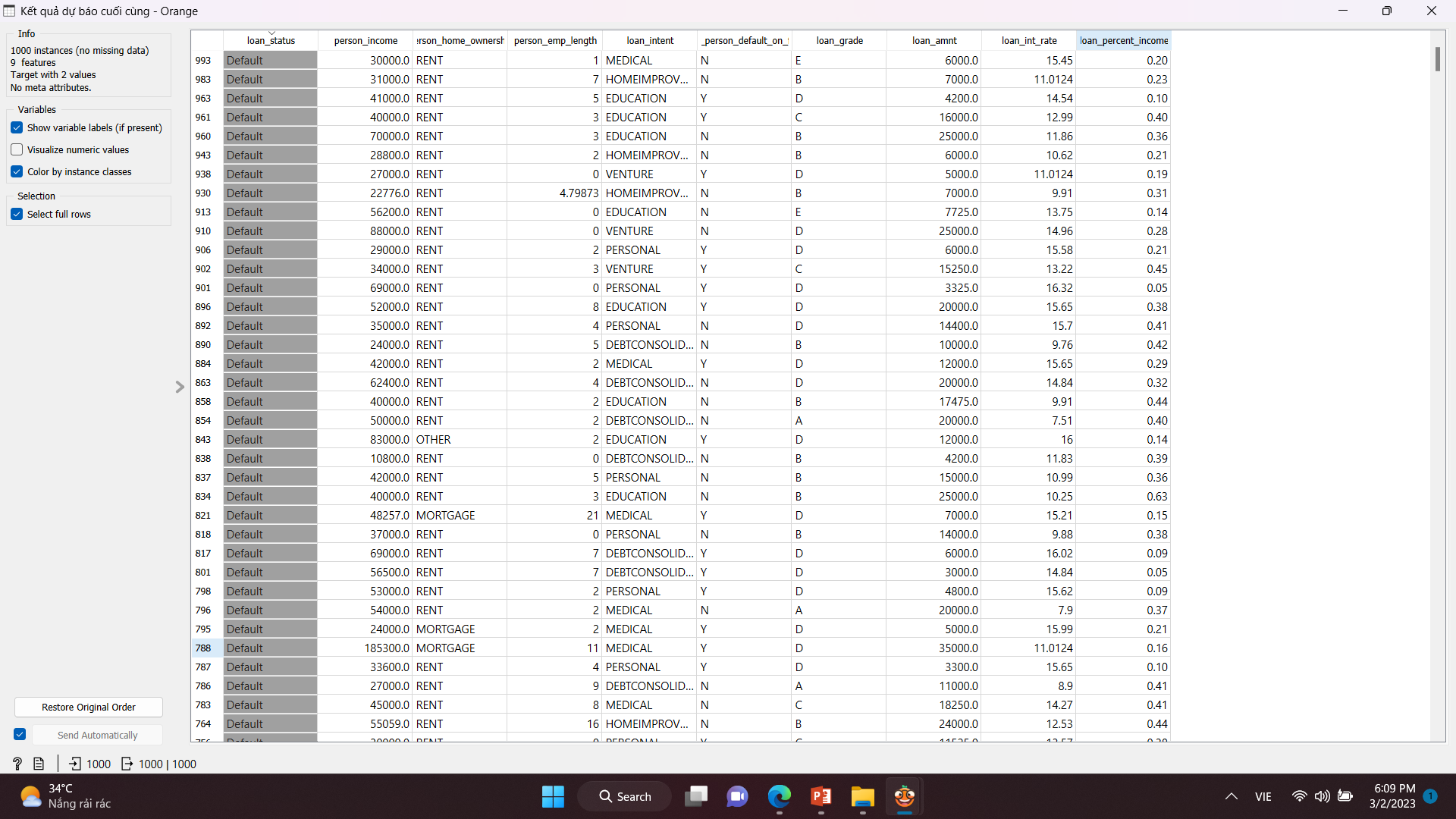




* Thực hiện đổi kết quả của cột **loan\_status** cũvà thay giá trị mới dự báo vào để cho ra kết quả chính xác hơn. Mặc khác, để dữ liệu mang trực quan hay dễ hiểu hơn, ta thực hiện đổi tên giá trị **0, 1** thành **default** và **not default** (tương ứng với **vỡ nợ** và **không vỡ nợ**). Kết quả dự báo cuối cùng nhận được là:

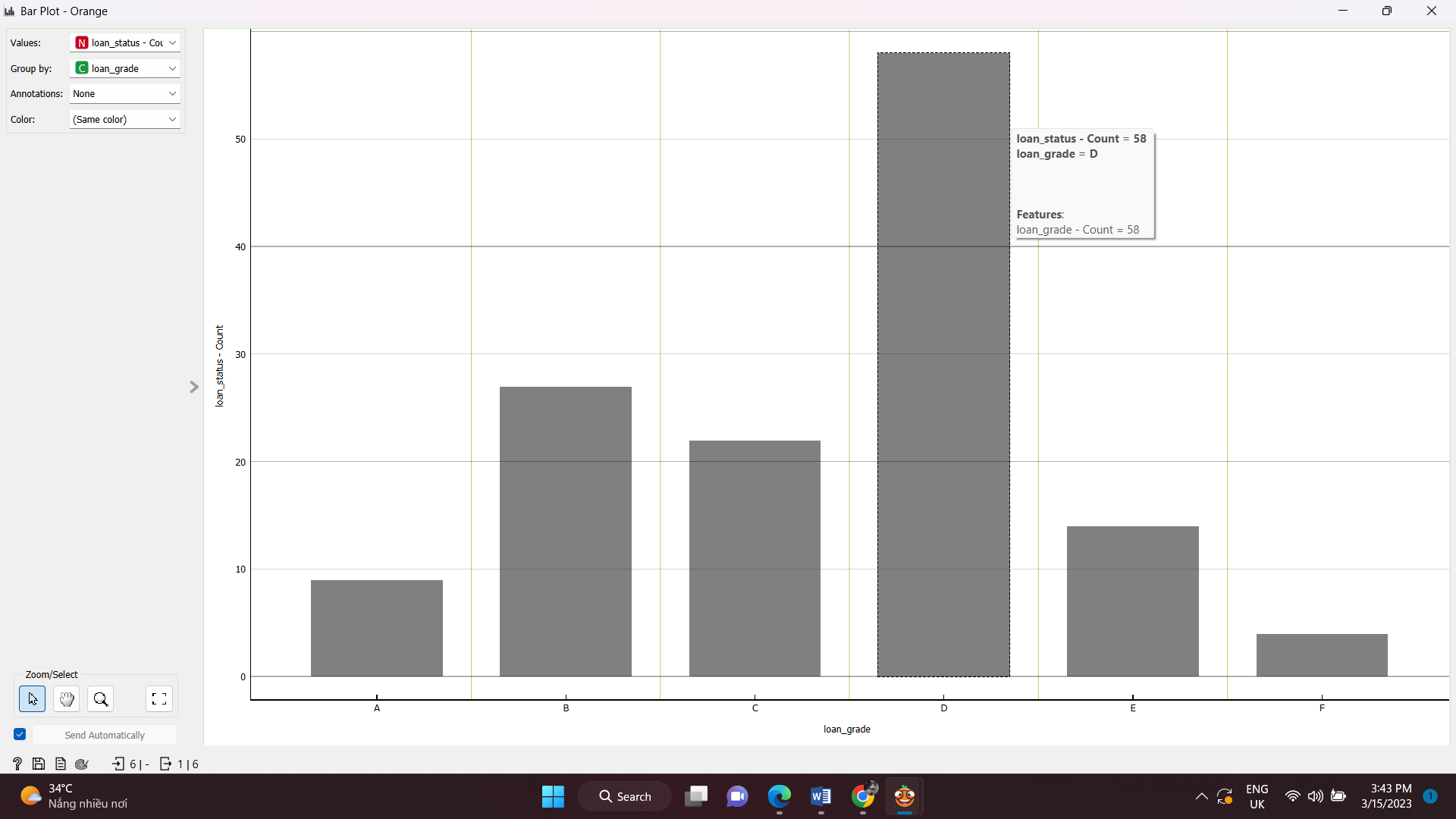


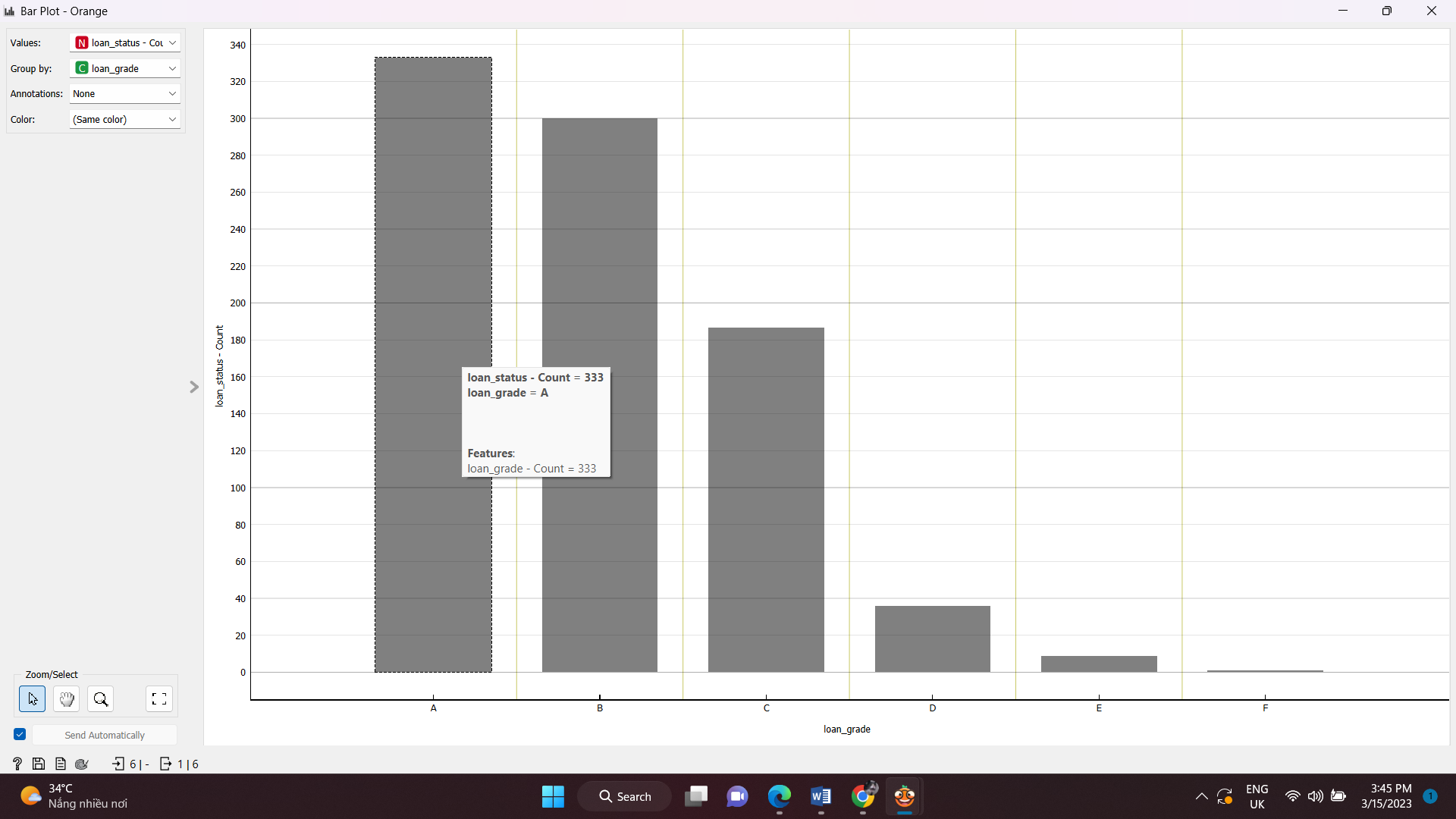




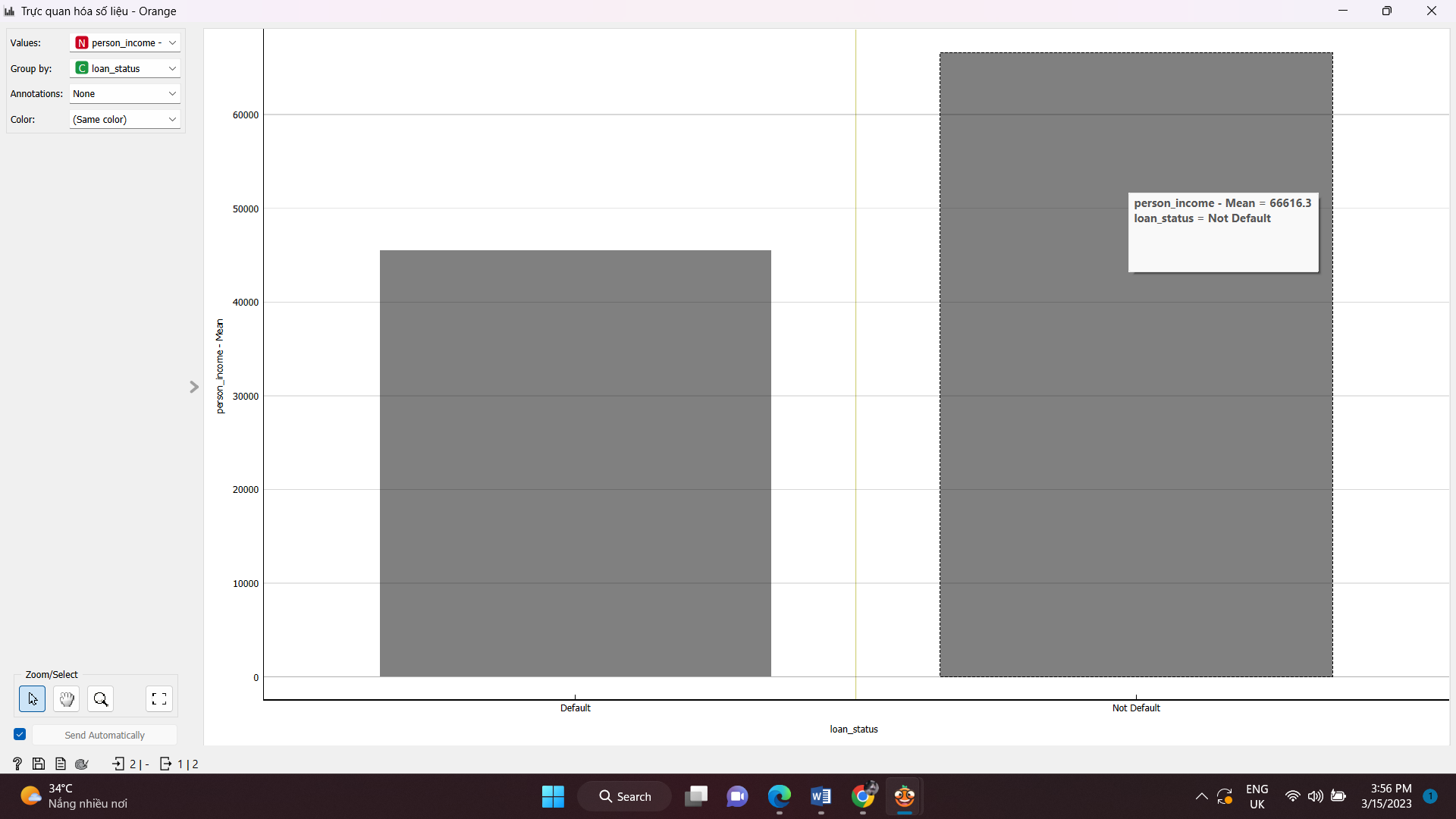
## **Phân tích số liệu dự báo.**

* + Từ kết quả trên, ta có thể đưa ra một số đánh giá sơ bộ cho thuộc tính **tình trạng nợ** (**loan\_status**) của khách hàng như sau:
* Loan\_grade: điểm vay của em khách hàng vỡ nợ xuất hiện nhiều nhất là **D** (~**43.3%**), trong khi đối với em khách hàng không vỡ nợ là **A** (~**38.5%**). Cho thấy khi cá nhân có số điểm vay càng thấp thì khả năng vỡ nợ của cá nhân đó càng cao.

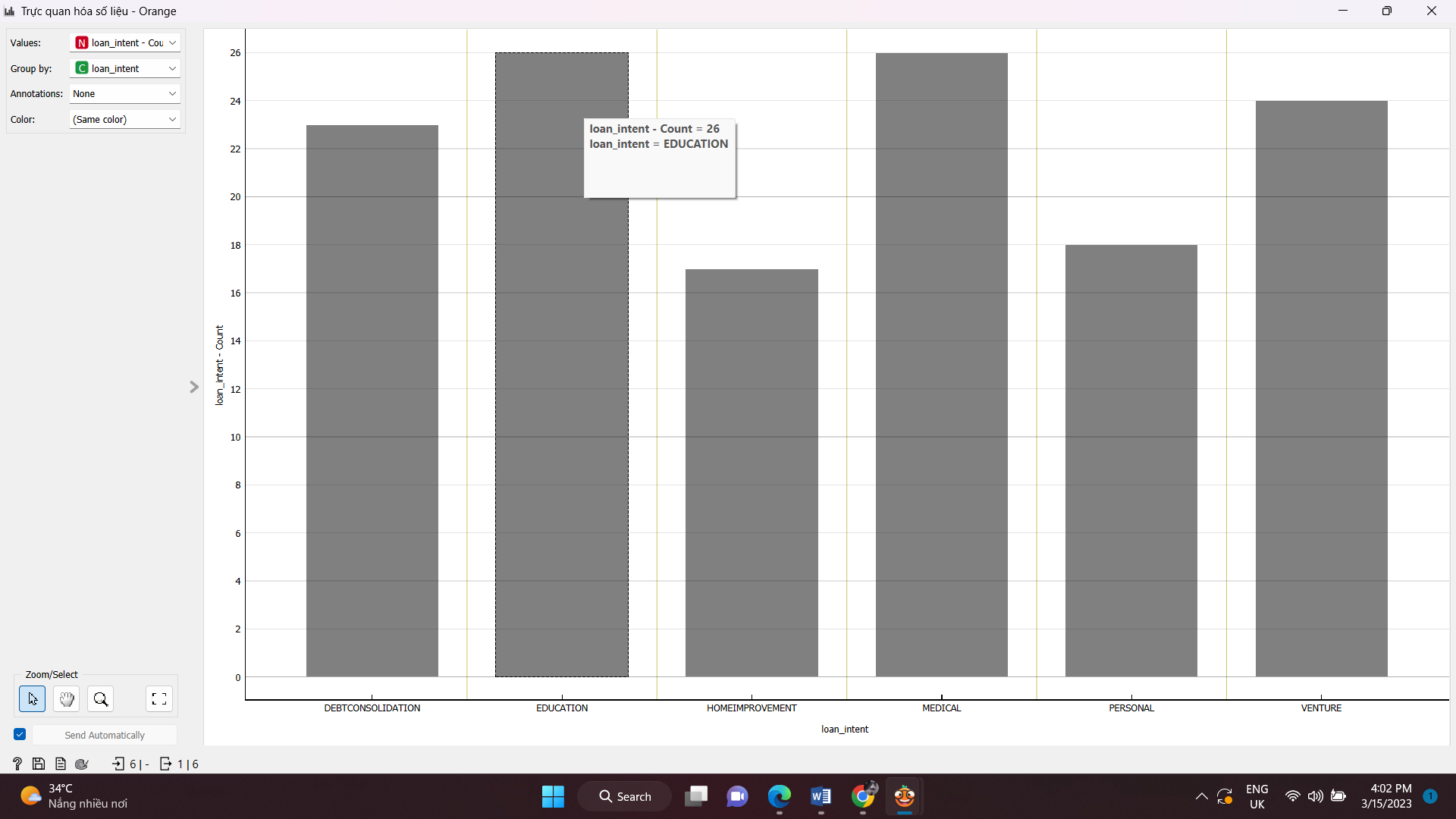


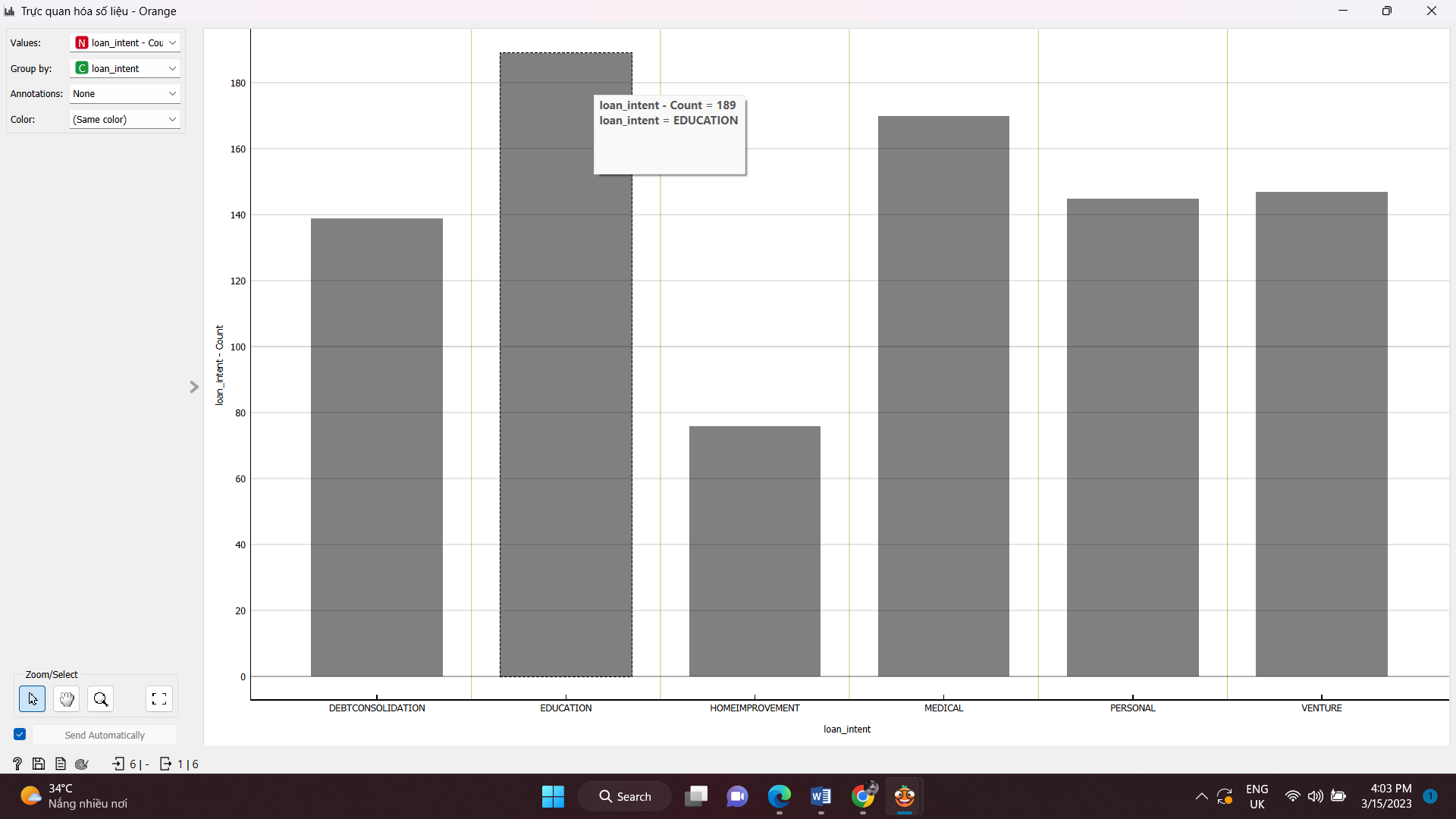


* Person\_income: Từ dữ liệu dự báo, thấy được thu nhập bình quân của em người vỡ nợ thấp hơn em người không có khả năng vỡ nợ (trung bình: **45557.052 < 66616.314**). Cho thấy: khi cá nhân không có khả năng kiếm được đủ tiền hoặc cân bằng được số tiền vay từ thu nhập bình quân, thì khả năng vỡ nợ của cá nhân đó cũng rất cao.



* Cho dù là em người vỡ nợ hay không vỡ nợ thì mục đích được đề cập đến nhiều nhất của cả hai em này đều là cho giáo dục (**EDUCATION**). Với em vỡ nợ là **19.4%** và **21.82%** là em còn lại. Tuy nhiên, đối với em vỡ nợ thì song song với lý do đó y tế cũng chính là mục đích vay của họ (cũng với **19.4%**).





## **Đánh giá mô hình.**

* + Ưu điểm:
* Đây là thuật toán hoạt động hiệu quả với không gian cao chiều (High Dimensional Dpaces).
* Thuật toán tiêu tốn ít bộ nhớ vì chỉ sử dụng các điểm trong tập hỗ trợ để dự báo trong hàm quyết định.
* Chúng ta có thể tạo ra nhiều hàm quyết định từ những hàm Kernel khác nhau. Thậm chí sử dụng đúng Kernel có thể giúp cải thiện thuật toán lên đáng kể.
  + Nhược điểm:
* Trong trường hợp số lượng thuộc tính (p) của tập dữ liệu lớn hơn rất nhiều so với số lượng dữ liệu (n) thì SVM cho kết quả không tốt.
* SVM không trực tiếp cung cấp các ước tính xác suất, chúng được tính toán bằng cách sử dụng Five-Fold Cross-Validation.

# **CHƯƠNG 5. TRIỂN KHAI MÔ HÌNH**

* + - 1. **Ứng dụng phương pháp phân lớp SVM trong việc dự báo rủi ro tín dụng của khách hàng.**

Thông qua nghiên cứu, em đã mô phỏng việc sử dụng mô hình SVM trong việc **Dự báo khả năng vỡ nợ tín dụng của khách hàng**. Việc áp dụng mô hình này vào các công tác quản trị và dự báo rủi ro tín dụng tại các ngân hàng thương mại ở Việt Nam sẽ mang lại tính khả thi và chính xác cao, bởi mô hình sẽ là cơ sở giúp bộ phận thẩm định đánh giá và tái thẩm định các khoản vay nhằm đảm bảo các khoản vay lớn sẽ được x xét khách quan hơn, tạo tính độc lập trong thẩm định cho vay. Từ đó giúp người phê duyệt tín dụng nhận dạng rõ hơn các rủi ro tiềm ẩn và ra quyết định cấp vốn hoặc không cấp vốn vay cho khách hàng. Hơn nữa, việc xây dựng cũng như áp dụng các mô hình vào hệ thống đánh giá tín dụng của các ngân hàng trở nên tự động hóa được quá trình cho vay, tiết kiệm được một lượng chi phí đáng kể và tăng hiệu quả sử dụng vốn của doanh nghiệp. Việc tự động hóa quy trình giúp cả hai bên, bên đi vay và cho vay tiết kiệm thời gian từ việc bỏ đi một số bước thẩm định tín dụng truyền thống, người đi vay sẽ nhanh chóng nhận được khoản tiền để kịp thời đáp ứng nhu cầu tiêu dùng, sản xuất và kinh doanh, ngân hàng cũng sử dụng đồng vốn một cách nhanh chóng, hiệu quả hơn và tăng tốc độ quay vòng vốn.

Mô hình dự báo khả năng vỡ nợ của khách hàng được xây dựng dựa trên những phương pháp thống kê khoa học và sự phát triển của công nghệ thông tin, thông qua đó tiến hành kiểm định để lựa chọn ra mô hình phù cho ra kết quả dự báo tốt nhất, phù hợp trong thực tế. Tuy nhiên, trong quá trình áp dụng và triển khai mô hình vào thực tế, phải thường xuyên có sự kiểm định và điều chỉnh kịp thời cho phù hợp. Vì vậy, em đề xuất một vài khuyến nghị để mô hình được hoàn thiện và đi vào hoạt động ổn định trong tương lai như sau:

* Đầu tư xây dựng một hệ thống lưu trữ dữ liệu lớn có thể đáp ứng được việc gia tăng mạnh mẽ của dữ liệu đầu vào và tốc độ xử lý dữ liệu cao.
* Xây dựng một mô hình kết hợp nhiều mô hình cơ sở thành một siêu mô hình nhằm đạt được độ chính xác dự báo cao hơn cho việc ra quyết định.
* Thường xuyên bảo trì, cập nhật, thực hiện các phân tích, nghiên cứu các ngành và quy mô,... giúp cho các doanh nghiệp mở rộng mô hình nếu có nhu cầu.

1. **Hiệu quả kinh tế.**

Các ngân hàng thương mại trong làn sóng chuyển đổi số phải đối mặt với sự thay đổi rất lớn và phức tạp, đồng thời cũng là nhiều thách thức mới. Cuộc chạy đua để giữ chân khách hàng giữa các ngân hàng thương mại cũng ngày càng căng thẳng, mỗi ngân hàng đều đưa ra những chính sách thu hút khách hàng sử dụng dịch vụ của mình nói chung và sử dụng dịch vụ tín dụng nói riêng. Sau khi triển khai mô hình SVM trong việc **Dự báo khả năng vỡ nợ của khách hàng**, em nhận thấy rằng các thông tin thu được đóng vai trò quan trọng trong việc quản trị tín dụng của các ngân hàng thương mại ở Việt Nam, từ đó khiến cho hiệu quả kinh tế ngày càng chuyển biến tích cực hơn.

Từ những dữ liệu và đặc điểm của khách hàng vay tín dụng, các ngân hàng sẽ dự báo được đặc điểm của những khách hàng có xác suất lớn hay không khả năng vỡ nợ, phân loại khách hàng nhằm đưa đến quyết định cấp tín dụng phù hợp cho từng đối tượng khác nhau và giám sát việc hoàn trả vốn gốc và lãi. Chính vì thế, việc áp dụng các mô hình phù hợp, và sửa đổi mô hình kịp thời sẽ giúp quá trình chấm điểm tín dụng diễn ra suôn sẻ và nhanh chóng hơn. Hiệu ứng tích cực mô hình này sẽ tạo động lực cho nhiều khách hàng tìm đến ngân hàng để vay tín dụng hơn, nhu cầu của khách hàng vay tín dụng tăng sẽ là lực đẩy cho ngân hàng ngày càng nâng cao chất lượng, phát triển quy mô để có thể đáp ứng được lượng cầu ngày càng tăng, từ đó mang về nhiều lợi nhuận hơn cho ngân hàng nhưng vẫn đảm bảo một mức độ rủi ro thấp và đảm bảo một nền kinh tế ổn định.

Và cuối cùng, các ngân hàng thương mại có thể tận dụng sự liên kết giữa các mô hình dự báo để điều tiết các chính sách thắt chặt tiền tệ hoặc nới lỏng chính sách cho vay, tạo niềm tin với khách hàng, tăng danh tiếng cho ngân hàng.

**ĐÁNH GIÁ SỰ ĐÓNG GÓP CỦA CÁC THÀNH VIÊN**

|  |  |  |
| --- | --- | --- |
| **Họ Tên Sinh Viên** | **Chi tiết thực hiện** | **Đóng góp** |
| Võ Minh Nguyên  **(Trưởng em)** | * Xây dựng, phát triển ý tưởng đề tài. * Soạn nội dung phần II, III, IV, V. * Thiết lập các Widget trên Orange. | 100% |
| Nguyễn Thanh Vy | * Soạn nội dung phần I, V. * Viết báo cáo tổng hợp đề tài (hỗ trợ). * Tổng hợp, sửa chữa nội dung. | 100% |
| Hoàng Đức Dân | * Thiết kết PPT, lựa chọn bộ dữ liệu. * Viết báo cáo tổng hợp đề tài (thực hiện chính). | 100% |
| Nguyễn Thiện Nhân | * Soạn nội dung phần III, VI. | 90% |