**Yêu cầu về Logistics Regression:**

1. *Phát biểu bài toán (đề tài) và nêu ý nghĩa đề bài toán:* ***(1 điểm)***

Đề tài: Hiện nay, nhu cầu cần cho vay rất phổ biến, các định chế tài chính trung gian như Ngân hàng cần rất nhiều thời gian và giấy tờ để xứ lí nghiệp vụ cho vay hay không. Để đơn giản hóa và đạt về hiệu suất làm việc, các định chế này muốn tự động hóa quy trình đủ điều kiện cho vay mua nhà trong khoản thời gian gần như là thực tế (real time scenario) khi khách hàng cung cấp các thông tin để đăng ký hình thức vay mua nhà.

Ý nghĩa: Qua mô hình này, chúng ta dự đoán những khách hàng đăng ký khoản vay mua nhà và công ty muốn kiểm tra đồng thời xác thực một cách nhanh chóng rằng khách hàng có đủ điều kiện vay mua nhà hay không.

1. Định nghĩa các Metrics/Variable bằng cách:  ***(1,5 điểm)***
   1. Chia nhỏ các câu hỏi lớn trên thành các câu hỏi nhỏ hơn

* Một khách hàng sẽ cung cấp các thông tin gì cho công ty?
* Đánh giá đủ điều kiện vay mua nhà dựa trên các biến số nào?
* Khách hàng không có gia đình đi kèm có dễ được cho vay hơn không?
  1. Lực chọn Metric hay Variable cần thiết để trả lời từng câu hỏi
* Các biến cần khách hàng cung cấp: giới tính, tình trạng hôn nhân, thu nhập cá nhân, thu nhập của người thân, có đang làm việc tự do không….
* Đánh giá đủ điều kiện vay dựa trên các yếu tố sau: thu nhập cá nhân, thu nhập người thân,khoản vay, thời gian vay vốn….
* Dựa vào biến: số lượng người phụ thuộc (Number of dependents), thu nhập của người thana đi kèm, học thức…

1. Phân tích khám phá dữ liệu (EDA\*\*\*) giữa biến phụ thuộc với các biến độc lập (tối thiểu 03 biến) bằng phương pháp Descriptive analytics (bao gồm: Mean, Standard Error, Median, Mode, Standard Deviation, Sample Variance, Kurtosis, Skewness), Box Plots, Scatter Plots, Histogram. Sau đó, với mỗi giá trị từ kết quả phân tích cần được phân tích, giải thích ý nghĩa và nhận định kết quả. **(1,5 điểm)**

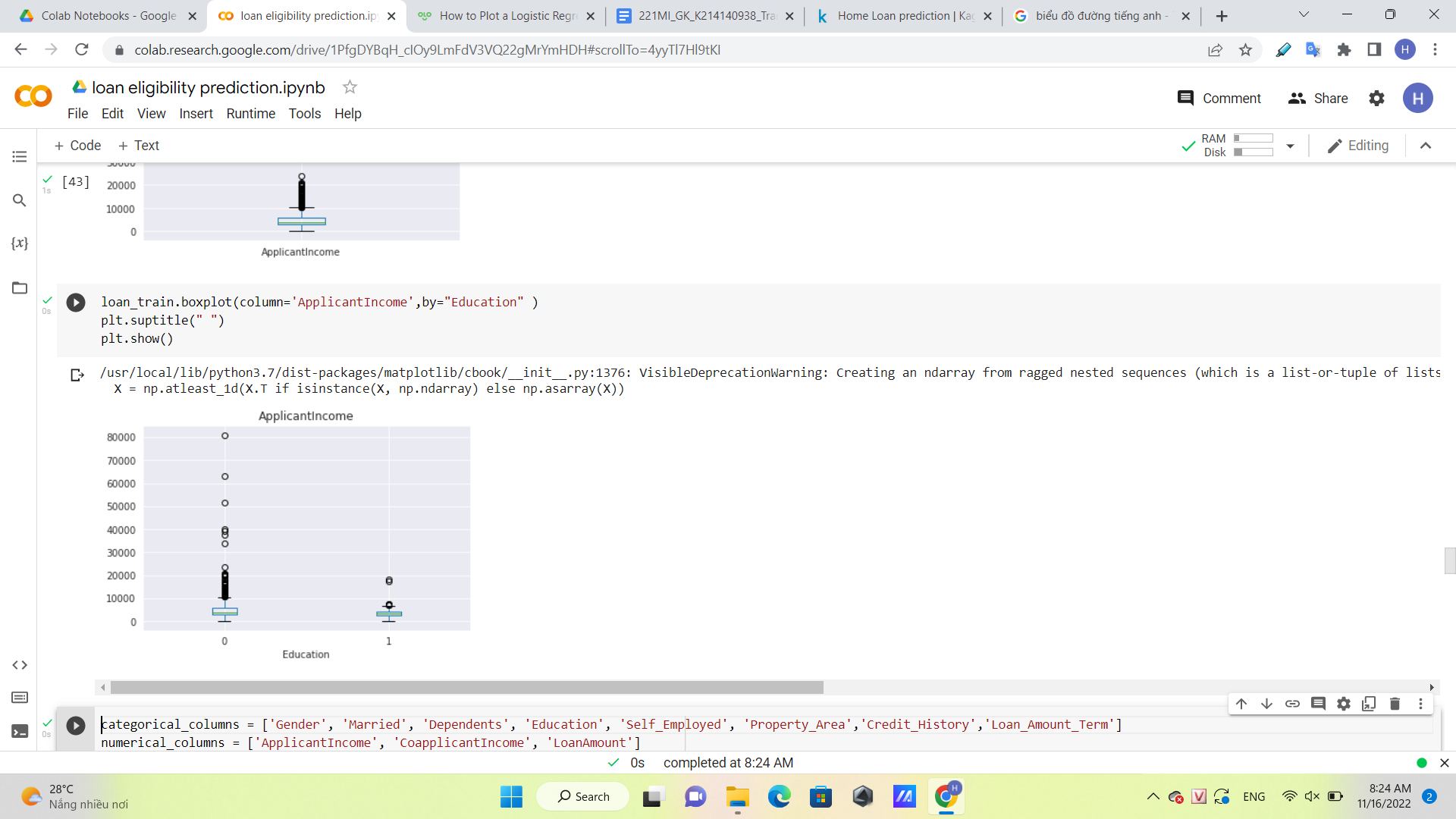
**EDA bằng cách:**

1. Scatter Plots



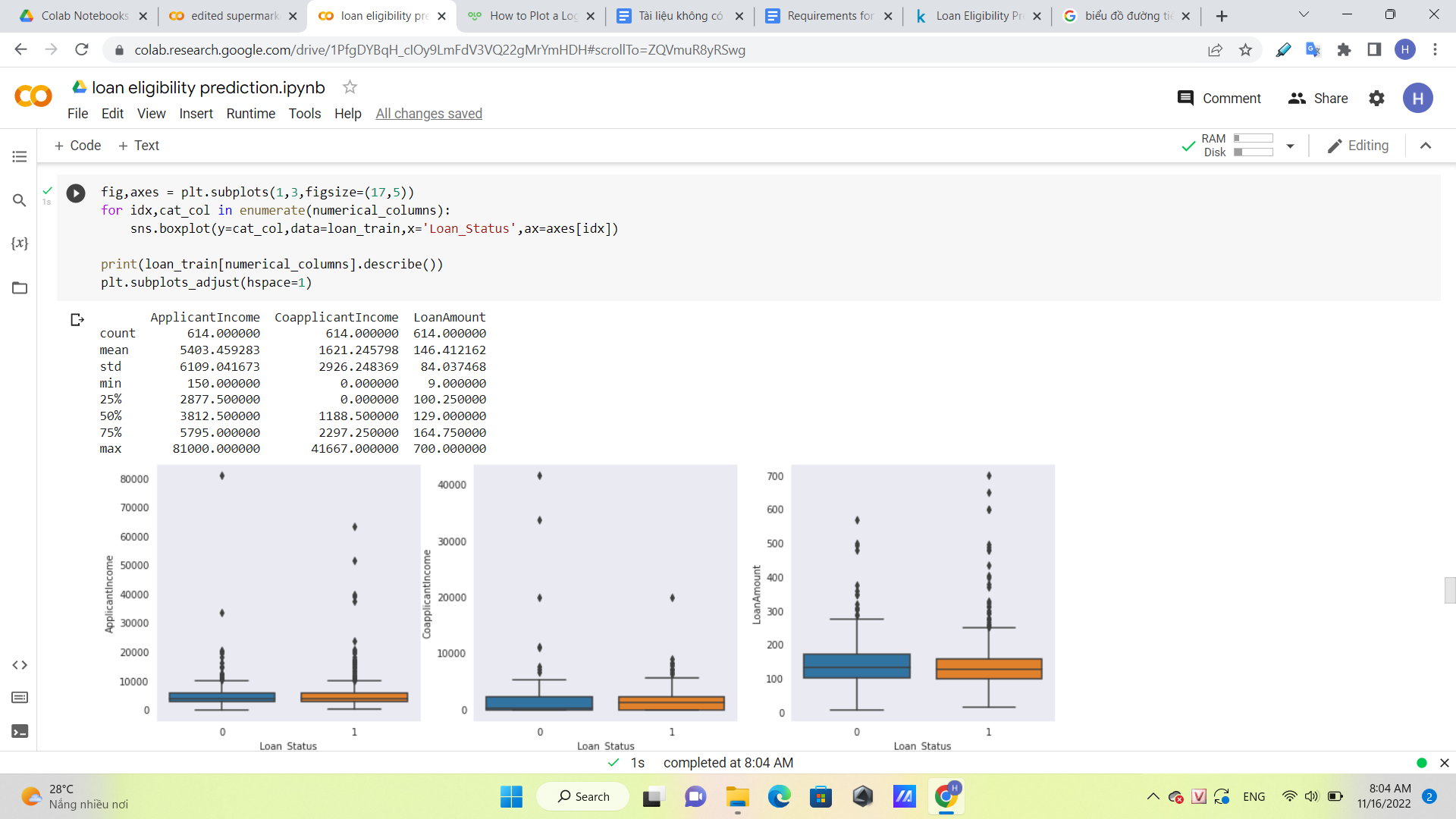
Ý nghĩa: Ta thấy rằng mối quan hệ giữa mức thu nhập và khoản cần vay khá tương đồng và chủ yếu tập trung ở khoản thu nhập từ $0-$10.000 cho mức họ muốn vay là $0 - $250.000. Tuy nhiên, qua biểu đồ này, ta nhận thấy có một vài outliner với mức thu nhập cao (> $40.000) và họ cũng có vay số tiền rất cao (>$40.000). Từ đó, ra có thể nhận định sơ bộ rằng, các khách hàng với thu nhập cao cũng có nhu cầu đi vay để mua nhà.

1. Box Plots



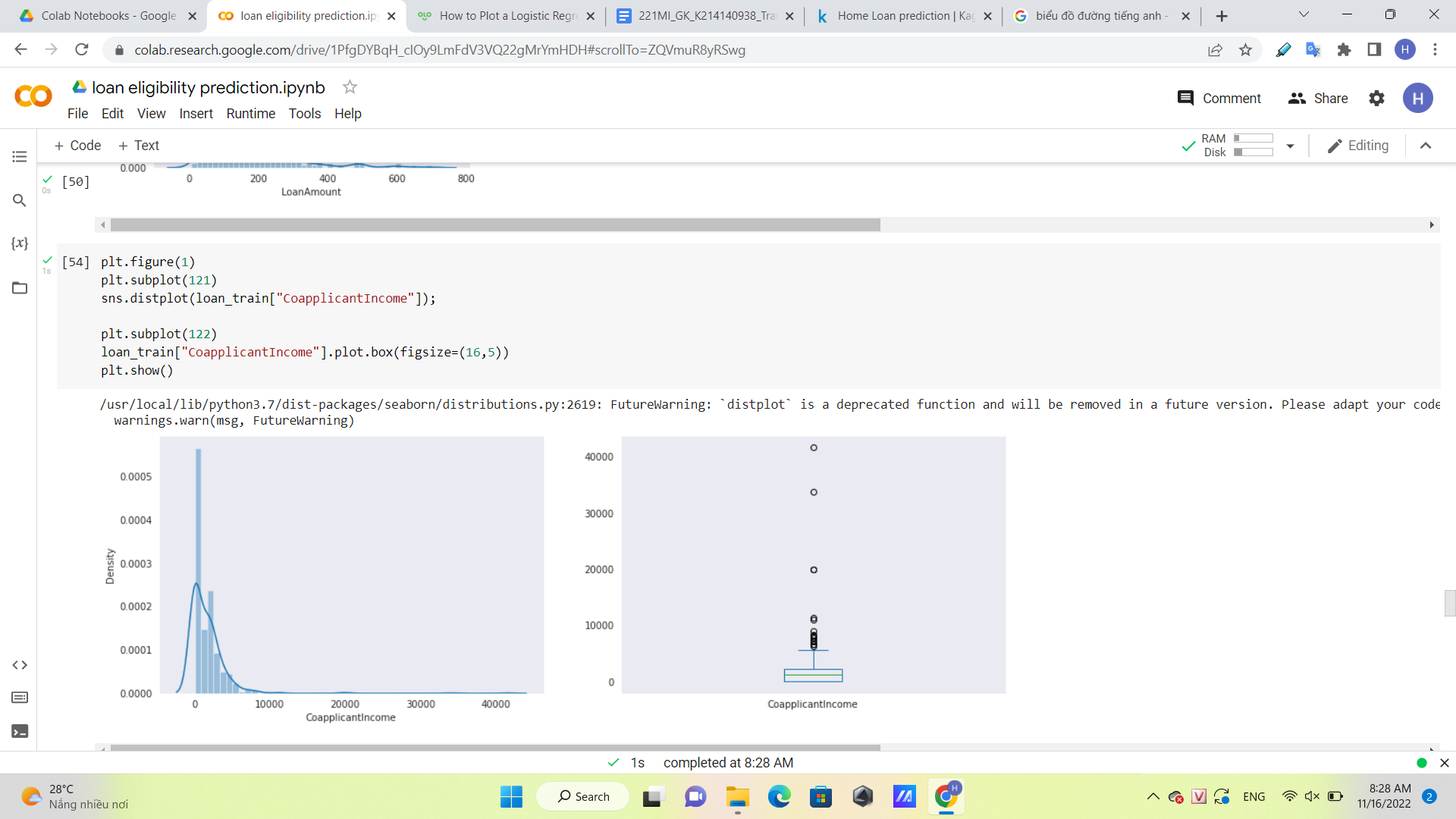
Ý nghĩa: Chúng ta dễ dàng thấy rằng số lượng lớn người đã tốt nghiệp có lượng thu nhập cao và ổn định là những người dạng outliner

1. Mean, Standard Error, Median, Mode…



Ý nghĩa: Dựa vào hình này , ta dễ dàng quan sát các giá trị mean,median, tứ phân vị,.... Để từ đó ta có thể đưa ra định mức cho các khách hàng vay trên mức độ tương đối. Đồng thời có được cái nhìn tổng quan về tệp data của mình.

1. Histogram

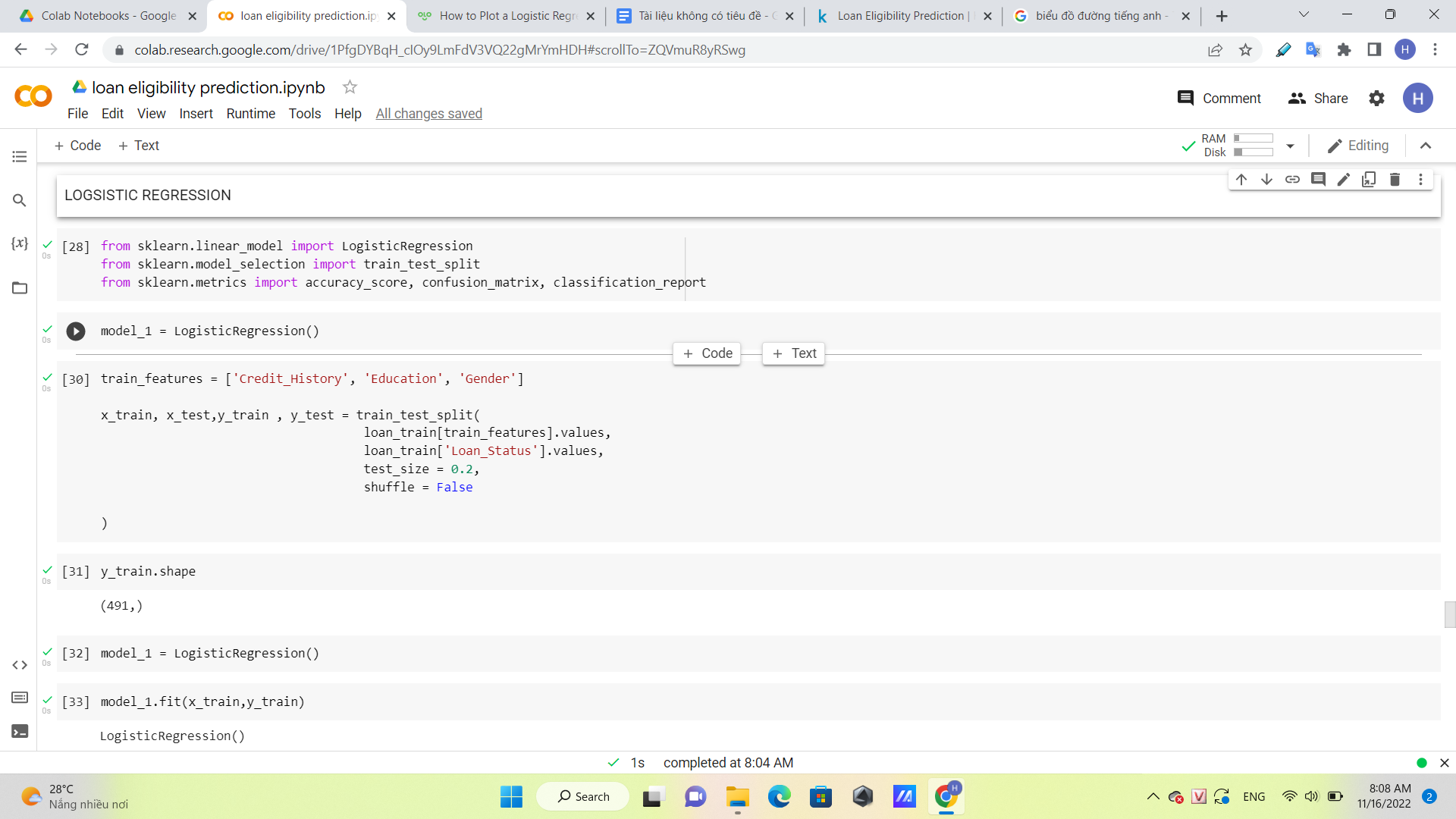


Ý nghĩa: Phần lớn thu nhập của người thân của người đăng ký khoản vay nằm trong khoảng từ $0 đến $5000. Ngoài ra, cũng thấy rất nhiều điểm khác biệt trong thu nhập của người thân của người đăng ký và thu nhập này không được phân phối bình thường.

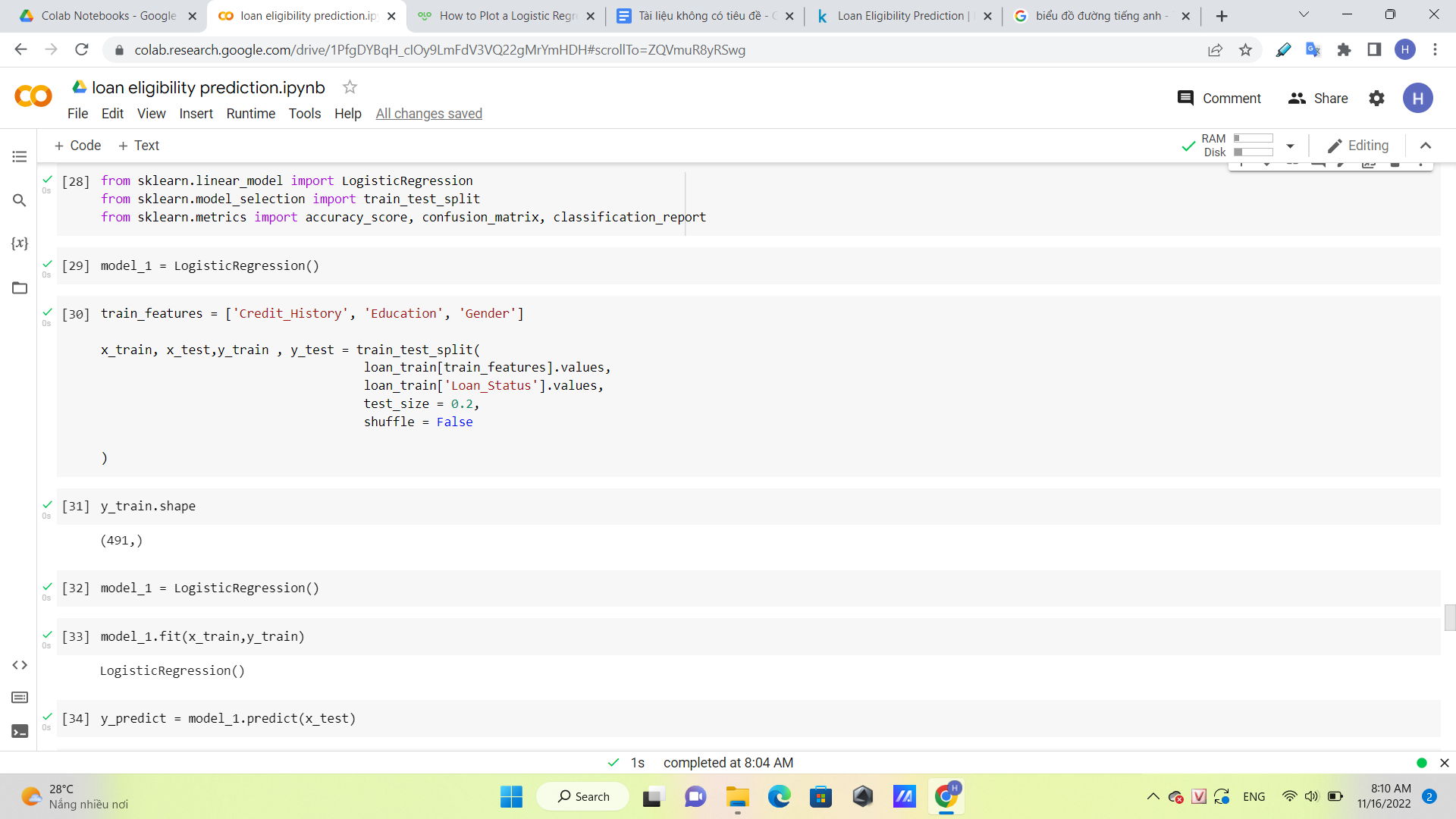
1. Xây dựng mô hình hồi quy logistics để phân loại (thể hiện sự tương quan giữa biến độc lập và phụ thuộc). **(5 điểm)**

**Bài toán đặt ra: Đánh giá khả năng có cho khách hàng vay hay không dựa trên 3 tiêu chí: Lịch sử tín dụng, Học thức và giới tính**

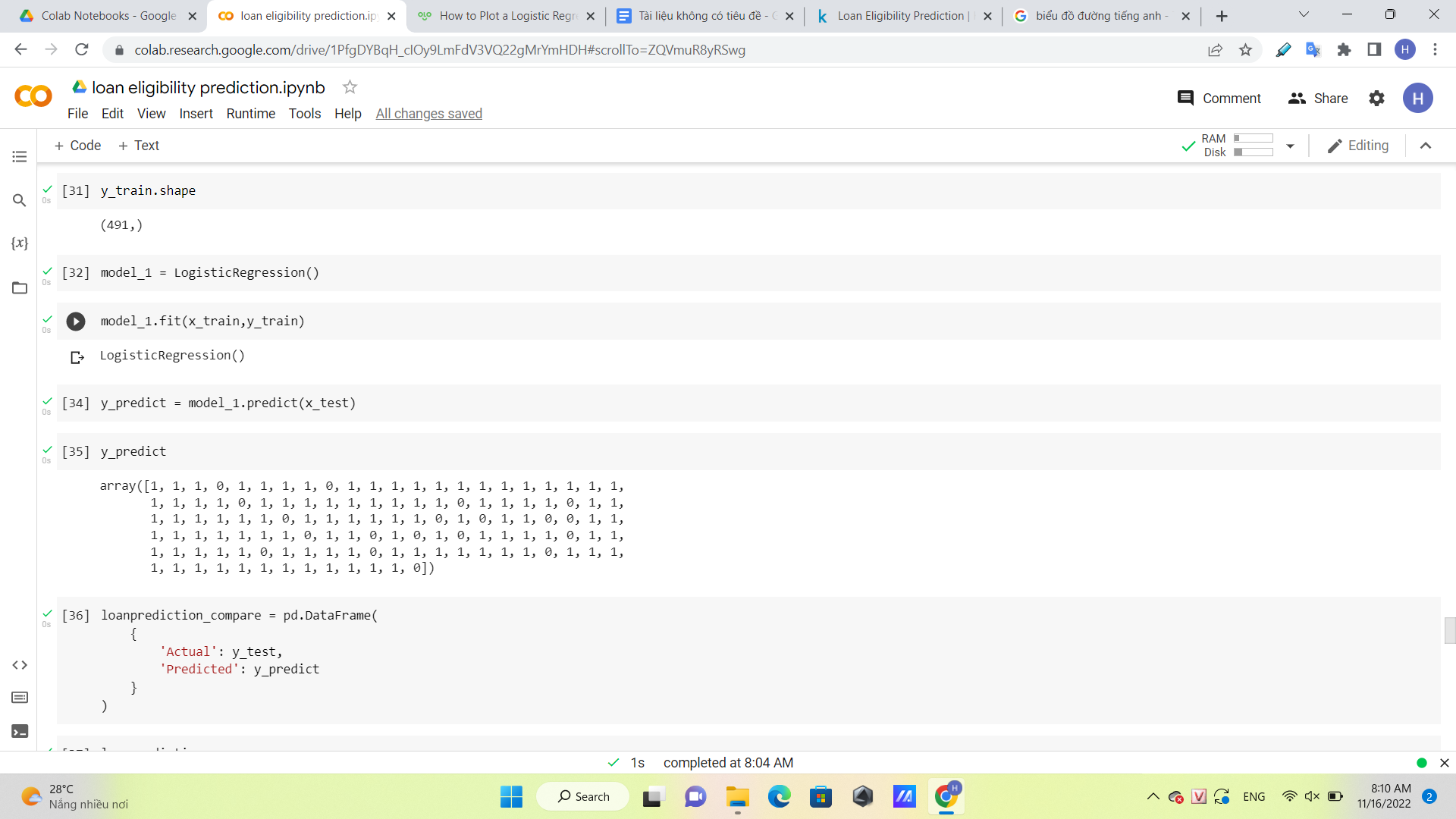
Bước 1: Import các gói và gọi tên mô hình



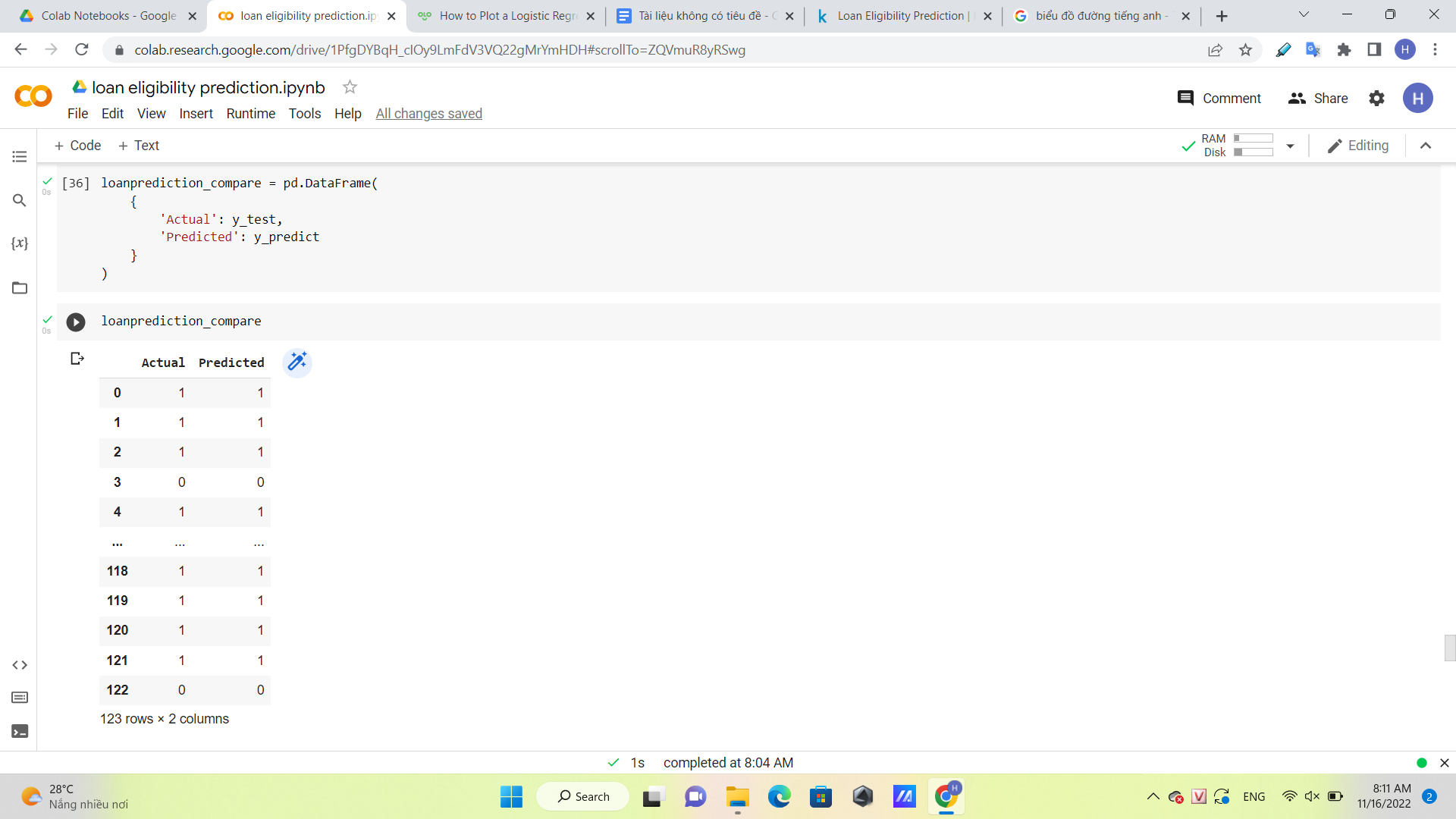
Bước 2: Định nghĩa các biến x\_train,y\_train,x\_test,y\_test và chia tập dữ liệu ra 80:20



Bước 3: Xuất ra kết quả y\_train



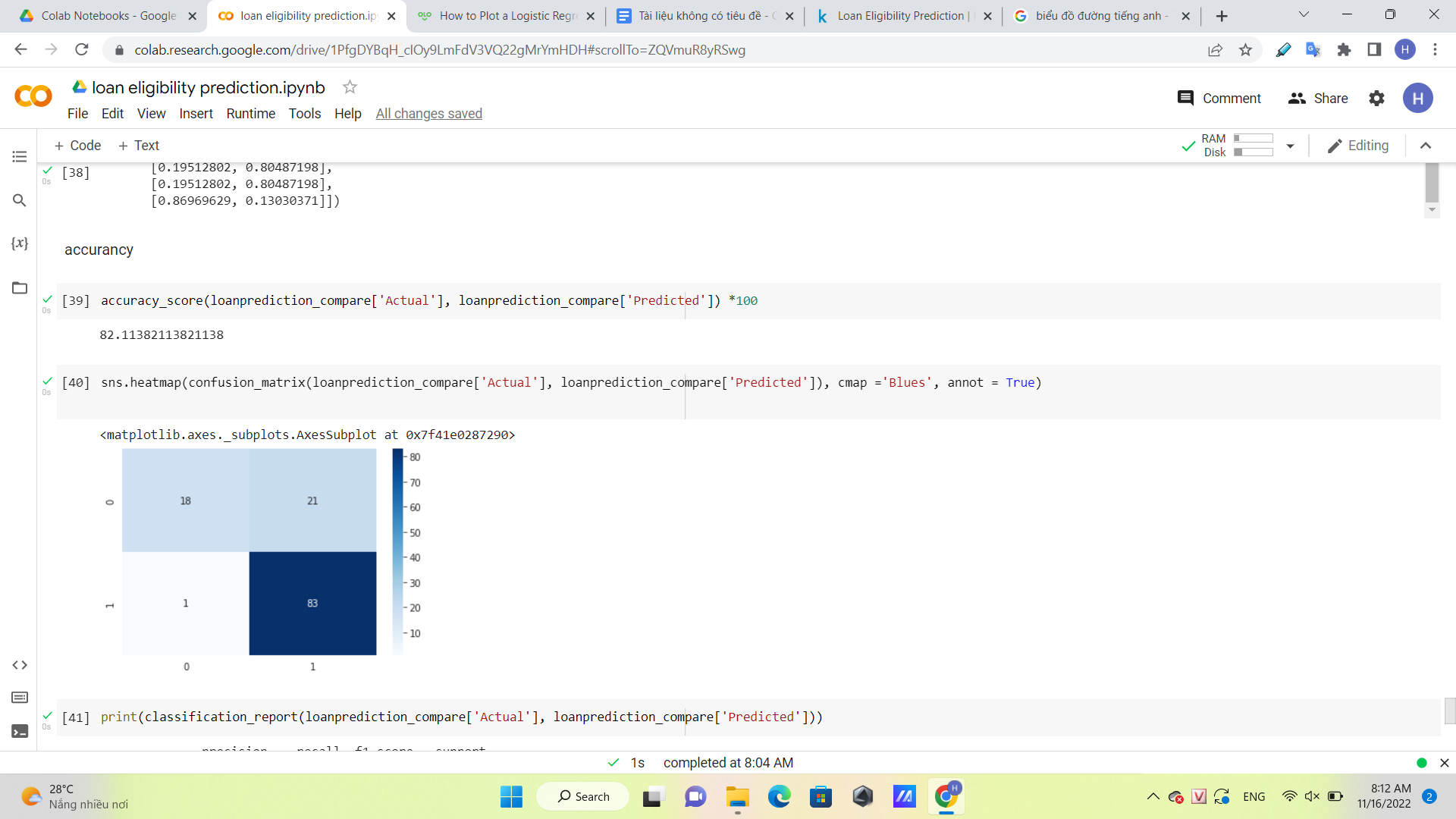
Bước 4: So sánh kết quả y thực tế (y\_test) và y predict (y\_train)



1. Đánh giá và lựa chọn kết quả mô hình (dựa trên Confusion matrix…) **(1 điểm)**

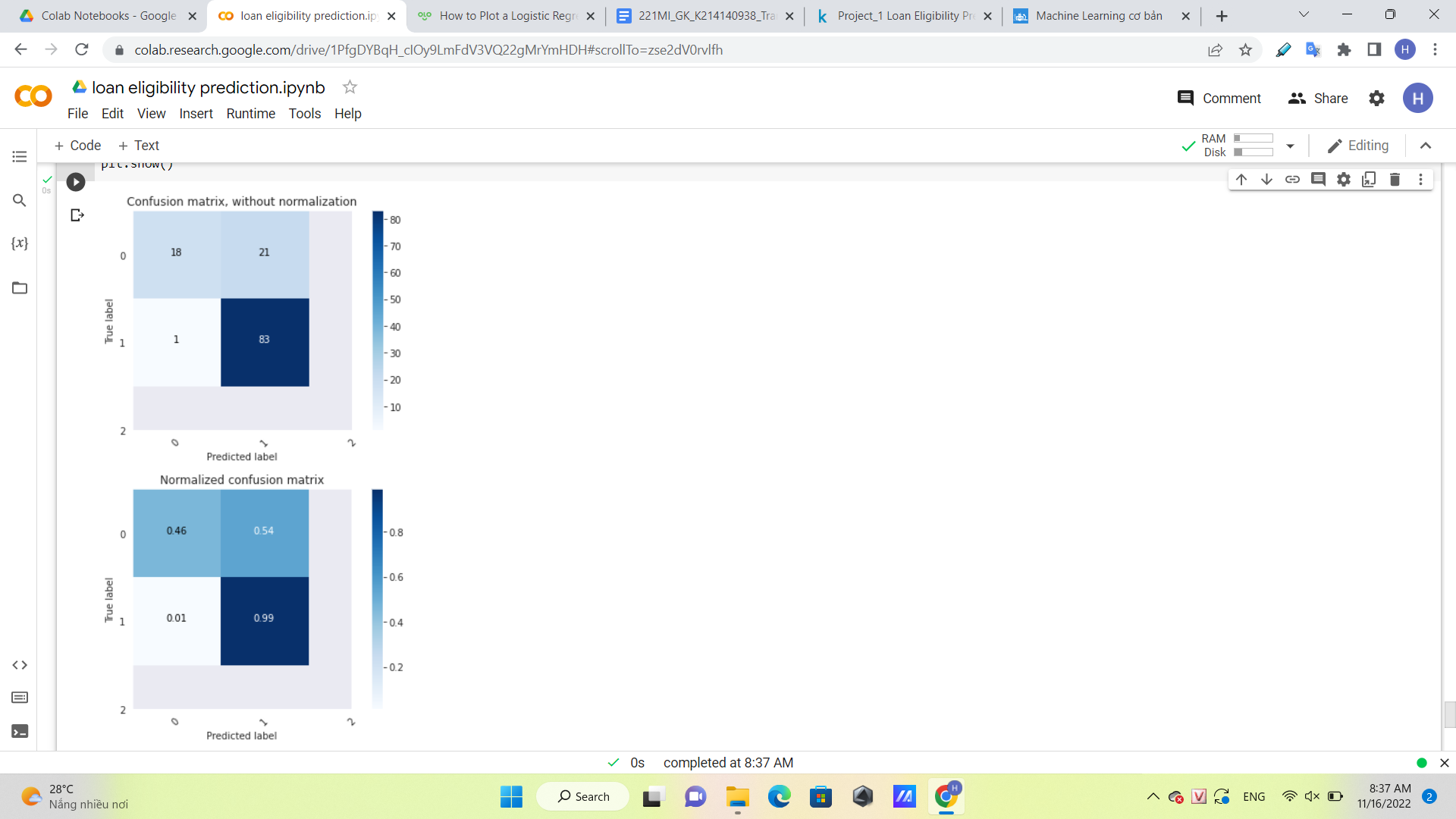
**Đánh giá kết quả qua các chỉ số sau:**

1. Accuracy score

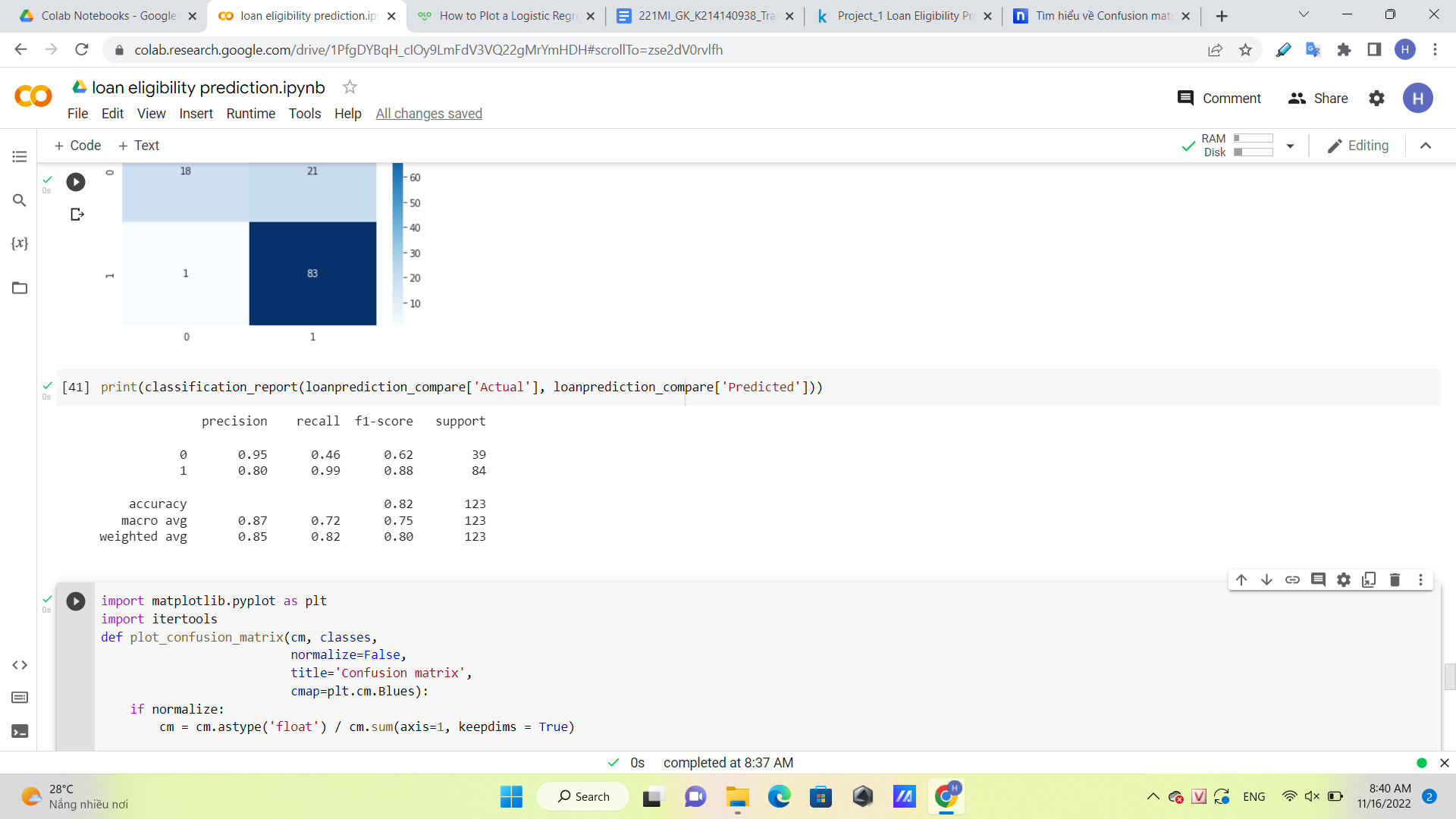


Kết quả cho thấy mô hình đạt độ chính xác xấp xỉ 82.1%. Đây là con số khá lý tưởng khi xây dựng một mô hình Logistic Regression.

1. Confusion matrix



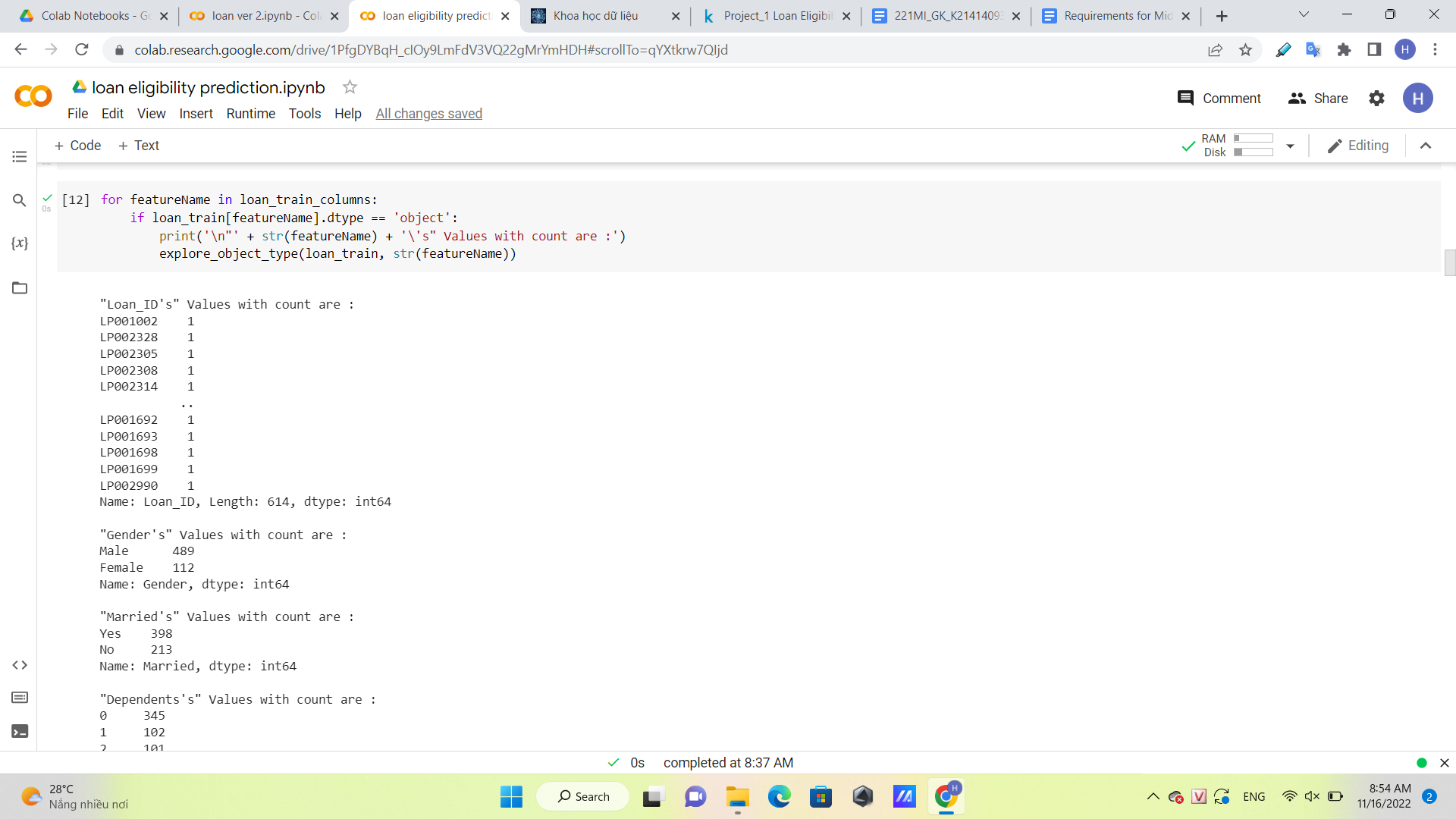
* TP (True Positive): Có 18 người là số lượng dự đoán chính xác được đồng ý cho vay.
* TN (True Negative): Có 21 người được dự đoán là không được cho vay.
* FP (False Positive - Type 1 Error): Chỉ có duy nhất 1 trường hợp dự đoán là người này không được cho vay nhưng thực tế là người này được cho vay.
* FN (False Negative - Type 2 Error):Có tới 83 người được dự đoán là không được vay nhưng người này trên thực tế là được cho vay.



**Điểm cộng khi thực hiện thêm các yêu cầu sau:** Trình bày nội dung và kết quả khi thực hiện một trong hai yêu cầu bên dưới

* Feature selection (1 điểm)

Lựa chọn các đặc điểm nổi bật:

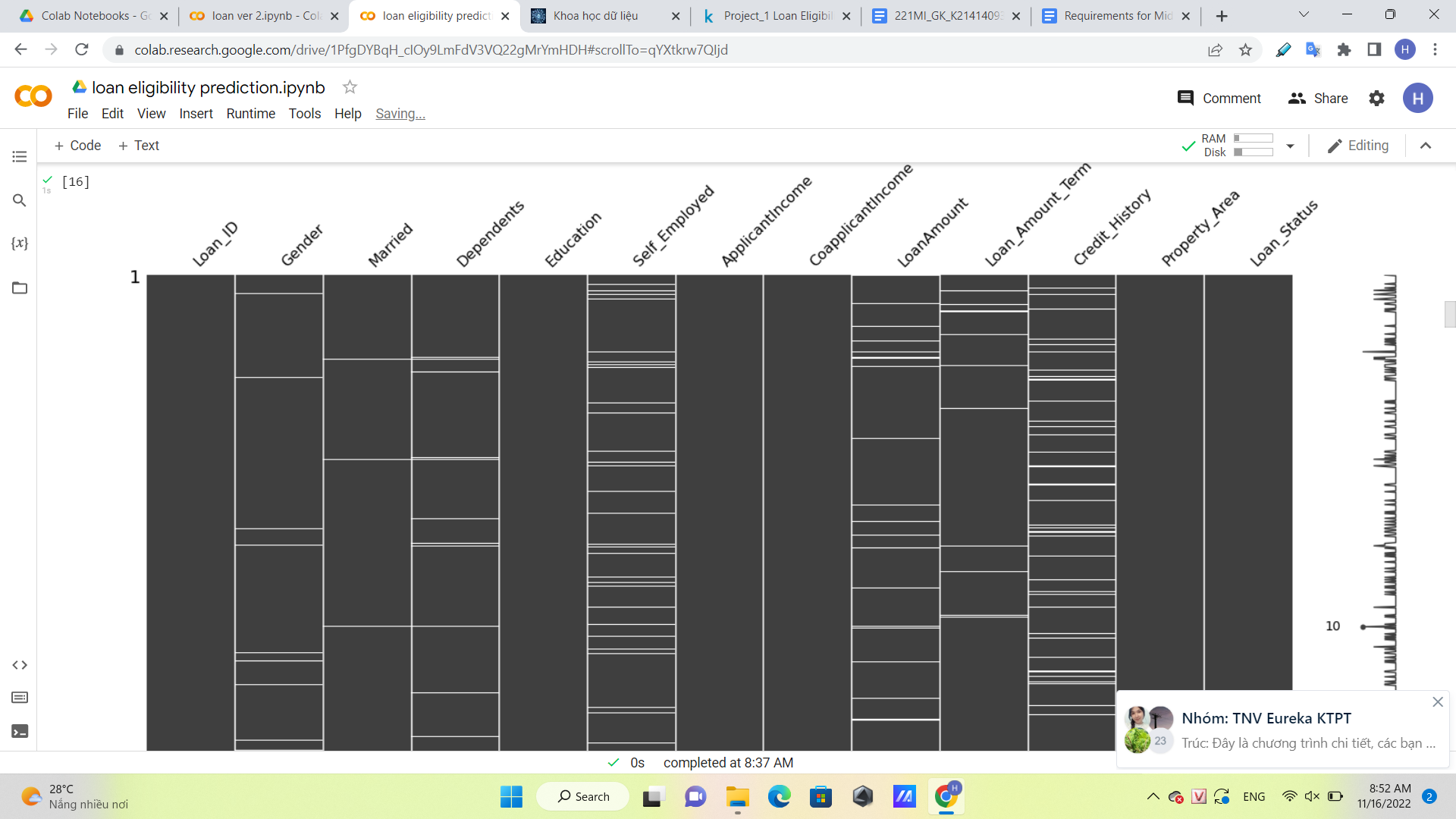


* Normalization data (1 điểm)

Sửa và bổ sung các giá trị bị thiếu:



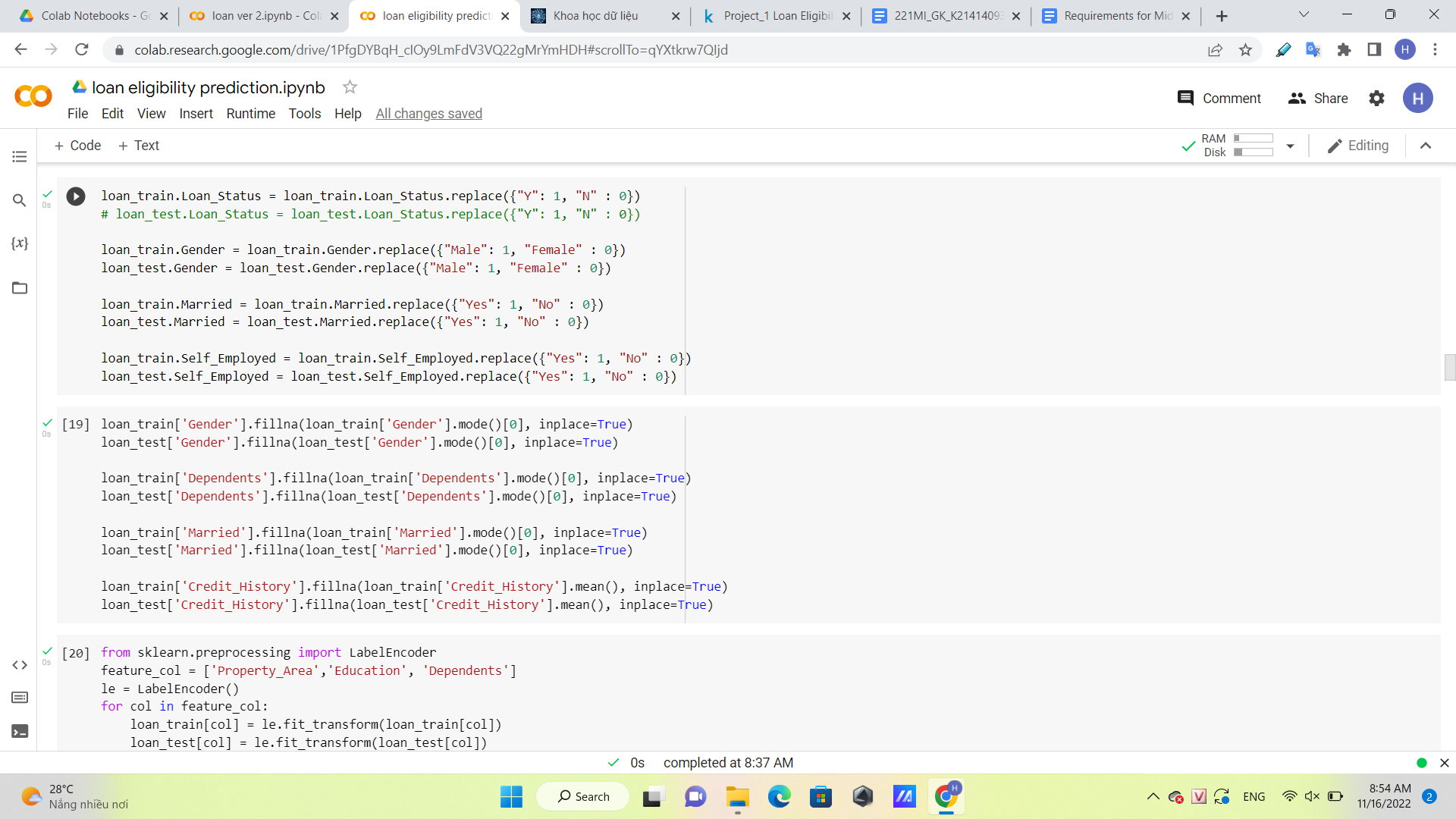
Dưới đây là biểu đồ thể hiện các giá trị còn thiếu trong các biến



Sau đó ta thêm các giá trị còn thiếu vào bằng mean,median…



Để bài toán trở nên khách quan hơn, ta chuyển các giá trị từ chữ về dạng số như sau:



Ghi chú: tổng điểm không quá 10 điểm