TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**DỰ ÁN CUỐI KỲ**

**Một số vấn đề của Machine Learning thông qua xây dựng bài toán phân loại**

*Người hướng dẫn*: **PGS. TS. Lê Anh Cường**

*Người thực hiện*: **Lâm Phúc Nghi – 51403239**

Lớp **: 14050302**

Khoá  **: 18**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**DỰ ÁN CUỐI KỲ**

**Một số vấn đề của Machine Learning thông qua xây dựng bài toán phân loại**

*Người hướng dẫn*: **PGS. TS. Lê Anh Cường**

*Người thực hiện*: **Lâm Phúc Nghi – 51403239**

Lớp **: 14050302**

Khoá  **: 18**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

LỜI CẢM ƠN

Qua nghiên cứu và thực nghiệm, em đã hoàn thành được dự án cuối kỳ môn Nhập Môn Học Máy.

Em xin cảm ơn PGS.TS Lê Anh Cường đã hướng dẫn nhiệt tình những kiến thức cơ bản và hướng thực hiện dự án cuối kỳ này giúp em có được nền tảng vững chắc để hoàn thành đề tài, thầy đã tận tình giải đáp những thắc mắc và hướng dẫn cách phát triển về đề tài của dự án cuối kỳ này trong việc nghiên cứu.

Do kiến thức còn hạn hẹp, nên phần báo cáo của em có thể còn nhiều sai sót. Em mong thầy/cô thông cảm và góp ý, nhận xét để em có thể tiếp tục khắc phục được những sai sót.

Em xin chân thành cảm ơn.

**DỰ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm dự án cuối kỳ của riêng tôi và được sự hướng dẫn của PGS.TS. Lê Anh Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong dựán còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Lâm Phúc Nghi*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Trình bày các vấn đề của machine learning thông qua xây dựng mô hình học máy cho bài toán phân loại.

1. Chuẩn hóa dữ liệu, chọn bài toán với dữ liệu có kiểu numeric và categorical (nominal). (1đ)
2. Xử lý với dữ liệu không cân bằng (unbalanced), tức là trong dữ liệu có loại (nhãn) có ít ví dụ huấn luyện so với dữ liệu thuộc các nhãn khác. (1đ)
3. Vẽ biểu đồ thể hiện thông tin thống kê dữ liệu: số lượng theo từng nhãn (loại). (1đ)
4. Xây dựng mô hình phân loại sử dụng các thuật toán: kNN, NB, SVM, Logistic Regression, MultiLayer Perceptron (i.e FFNN). (2đ)
5. Vẽ biểu đồ thể hiện kết quả training, sử dụng early stopping và validation dataset. (1đ)
6. Huấn luyện mô hình có sử dụng tham số batch size và epoch number. (1đ)
7. Hiển thị kết quả đánh giá mô hình với các độ đo khác nhau. (1đ)
8. Trình bày về vấn đề overfitting và có ví dụ minh hoạ. (1đ)
9. Trình bày một ví dụ minh họa của mạng Convolution Neural Network hoặc Long Short Term Memory (2đ)

Sử dụng các thư viện: numpy, pandas, matplotlib, sklearn, keras.

**Mục Lục**

[TÓM TẮT iv](#_Toc56961817)

[1. Chọn bài toán với dữ liệu kiểu numeric và categorical (nominal) 5](#_Toc56961818)

[2. Chuẩn hóa dữ liệu. 6](#_Toc56961819)

[3. Vẽ biểu đồ theo phân lớp. 10](#_Toc56961820)

[4. Cân bằng dữ liệu 15](#_Toc56961821)

[5. Chia dữ liệu huấn luyện. 15](#_Toc56961822)

[6. Xây dựng các mô hình phân loại. 16](#_Toc56961823)

[6.1. Mô hình Logistic Regression. 16](#_Toc56961824)

[6.2. Mô hình k-Nearest Neighbor. 16](#_Toc56961825)

[6.3. Mô Hình Naïve Bayes. 19](#_Toc56961826)

[6.4. Mô hình SVM. 20](#_Toc56961827)

[6.5. Mô hình multilayer perceptron. 21](#_Toc56961828)

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH ẢNH**

[1‑1 UCI Machine Learning Repository: Bank Marketing Data Set 7](#_Toc56908408)

[1‑2 Bộ dữ liệu bank-additional-full.csv 7](#_Toc56908409)

[2‑1 Bảng mô tả dữ liệu. 9](#_Toc56908410)

[2‑2 Giá trị unknown. 9](#_Toc56908411)

[2‑3 Xử lý giá trị “unknown”. 10](#_Toc56908412)

[2‑4 LabelEncoder cho thuộc tính phân lớp. 11](#_Toc56908413)

[2‑5 One-hot encoder với thuộc tính categorical. 11](#_Toc56908414)

[2‑6 MinMaxScaler dữ liệu thuộc tính numeric. 12](#_Toc56908415)

[3‑1 Biểu đồ hộp các giá trị thuộc tính “age” trái là no, phải là yes kề nhau. 12](#_Toc56908416)

[3‑2 Biểu đồ hộp các giá trị thuộc tính “campaign” trái là no, phải là yes kề nhau. 12](#_Toc56908417)

[3‑3 Biểu đồ hộp các giá trị thuộc tính “cons.price.idx” trái là no, phải là yes kề nhau. 13](#_Toc56908418)

[3‑4 Biểu đồ hộp các giá trị thuộc tính “contact” trái là no, phải là yes kề nhau. 13](#_Toc56908419)

[3‑5 Biểu đồ hộp các giá trị thuộc tính “day\_of\_week” trái là no, phải là yes kề nhau. 13](#_Toc56908420)

[3‑6 Biểu đồ hộp các giá trị thuộc tính “default” trái là no, phải là yes kề nhau. 13](#_Toc56908421)

[3‑7 Biểu đồ hộp các giá trị thuộc tính “education” trái là no, phải là yes kề nhau. 14](#_Toc56908422)

[3‑8 Biểu đồ hộp các giá trị thuộc tính “emp.var.rate” trái là no, phải là yes kề nhau. 14](#_Toc56908423)

[3‑9 Biểu đồ hộp các giá trị thuộc tính “euribor3m” trái là no, phải là yes kề nhau. 14](#_Toc56908424)

[3‑10 Biểu đồ hộp các giá trị thuộc tính “housing” trái là no, phải là yes kề nhau. 14](#_Toc56908425)

[3‑11 Biểu đồ hộp các giá trị thuộc tính “job” trái là no, phải là yes kề nhau. 15](#_Toc56908426)

[3‑12 Biểu đồ hộp các giá trị thuộc tính “loan” trái là no, phải là yes kề nhau. 15](#_Toc56908427)

[3‑13 Biểu đồ hộp các giá trị thuộc tính “marital” trái là no, phải là yes kề nhau. 15](#_Toc56908428)

[3‑14 Biểu đồ hộp các giá trị thuộc tính “month” trái là no, phải là yes kề nhau. 15](#_Toc56908429)

[3‑15 Biểu đồ hộp các giá trị thuộc tính “nr.employed” trái là no, phải là yes kề nhau. 16](#_Toc56908430)

[3‑16 Biểu đồ hộp các giá trị thuộc tính “pdays” trái là no, phải là yes kề nhau. 16](#_Toc56908431)

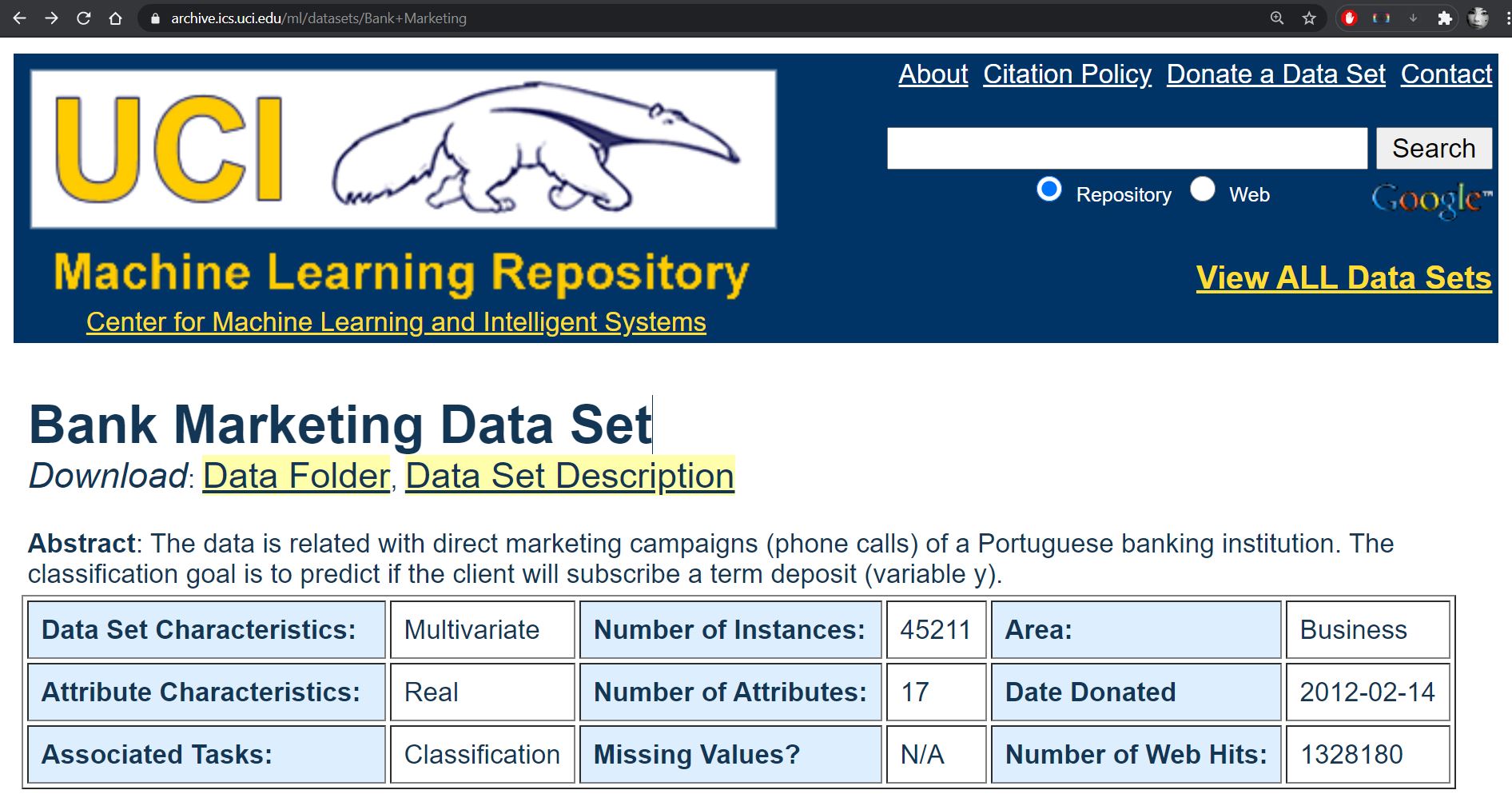
[3‑17 Biểu đồ hộp các giá trị thuộc tính “poutcome” trái là no, phải là yes kề nhau. 16](#_Toc56908432)

[3‑18 Biểu đồ hộp các giá trị thuộc tính “previous” trái là no, phải là yes kề nhau. 16](#_Toc56908433)

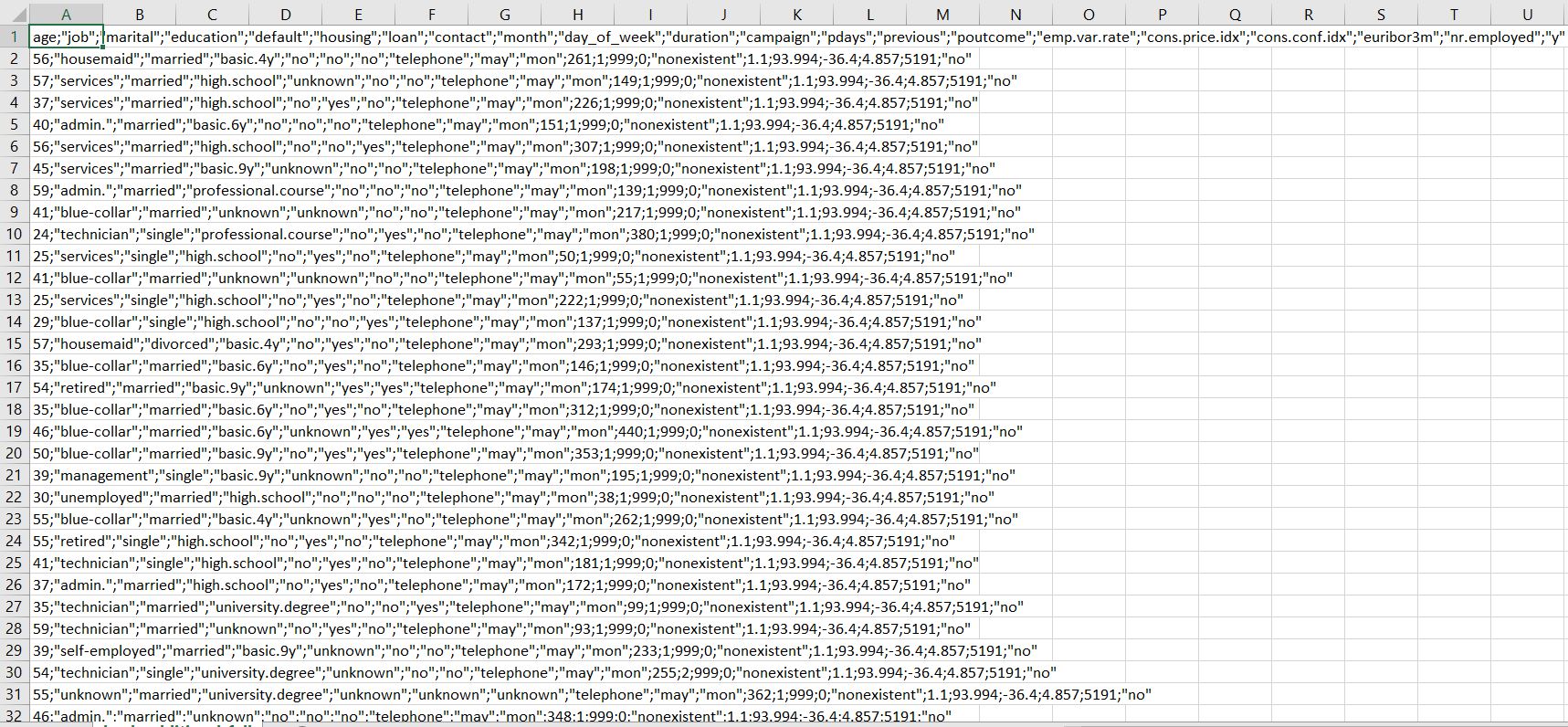
[4‑1 Up – sample: tăng số lượng mẫu nhãn yes. 17](#_Toc56908434)

# Chọn bài toán với dữ liệu kiểu numeric và categorical (nominal)

Bộ dữ liệu “Bank Marketing Data Set” của [Moro et al., 2014] S. Moro, P. Cortez and P. Rita được tải tại UCI Machine Learning Repository.



6.1‑1 UCI Machine Learning Repository: Bank Marketing Data Set



1‑2 Bộ dữ liệu bank-additional-full.csv

Mô tả dữ liệu:

Dữ liệu về chiến dịch tiếp thị của một ngân hàng Bồ Đào Nha dựa trên các cuộc gọi điện thoại. Cần liên hệ nhiều lần với một khách hàng để tiếp thị về mở tài khoản tiết kiệm.

Bộ dữ liệu “bank-additional-full.csv” có 41188 mẫu (dòng), 20 thuộc tính (cột) được xếp theo ngày (từ tháng 5 năm 2008 đến tháng 11 năm 2010).

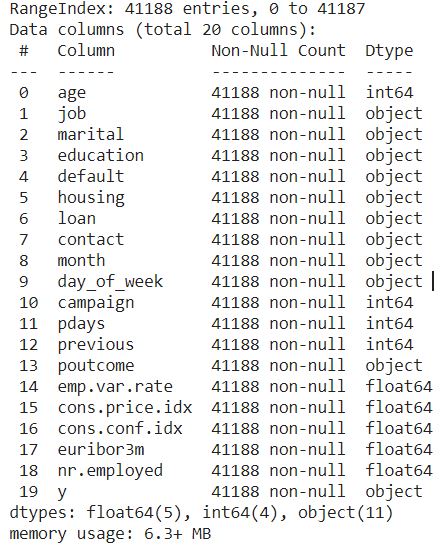
Chi tiết thuộc tính: 10 thuộc tính categorical, 9 thuộc tính numeric và 1 thuộc tính phân lớp (y = yes/no).

1. Age - tuổi: từ 17 đến 98 trừ {90, 93, 96, 97}.
2. Job – Nghề nghiệp: admin. – quản trị; blue-collar – công nhân; entrepreneur – doanh nhân; housemaid – giúp việc; management – quản lý; retired – về hưu; self-employed – tự kinh doanh; services – dịch vụ; student; technician – kỹ thuật viên; unemployed – thất nghiệp; unknown.
3. Marital – hôn nhân: divorced – ly hôn hoặc đơn thân; married – kết hôn; single – độc thân; unknown.
4. Education – học vấn: basic.4y; basic.6y; basic.9y; high.school; professional.course; university.degree; illiterate – thất học; unknown.
5. Default – has credit in default: yes; no; unknown.
6. Housing – có thuê nhà không: yes; no; unknown.
7. Loan – có vay không: yes; no; unknown.
8. Contact – kiểu liên hệ: cellular – di động; telephone – điện thoại bàn.
9. Month – tháng liên hệ gần nhất.
10. Day\_of\_week – ngày liên hệ gần nhất.
11. Campaign – số lần liên hệ trong chiến dịch này.
12. Pdays – số ngày đã trôi qua sau khi liên hệ khách hàng từ chiến dịch trước đó (999: chưa liên hệ).
13. Previous – số lần liên hệ trước chiến dịch này.
14. Poutcome – kết quả của chiến dịch trước đó: failure; nonexistent; success.
15. Emp.var.rate – tỷ lệ thay đổi việc làm – hàng quý.
16. Cons.price.idx – chỉ số giá tiêu dùng – hàng tháng.
17. Cons.conf.idx – chỉ số niềm tin của người tiêu dùng – hàng tháng.
18. Euribor3m – lãi suất 3 tháng Euro interbank offer rate – hàng ngày.
19. Nr.employed – số lượng nhân viên – hàng quý.
20. Y : yes; no.

Trong 41188 mẫu, có 36548 mẫu nhãn “no” chiếm 88.73% và 4640 mẫu nhãn “yes” chiếm 11.27%.

# Chuẩn hóa dữ liệu.

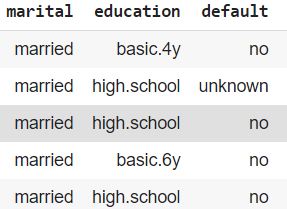
Trong 20 thuộc tính có thuộc tính phân lớp “y”. Có 10 thuộc tính categorical: job; marital; education; default; housing; loan; contact; month; day\_of\_week; poutcome. Có 9 thuộc tính dạng numeric: age; campaign; pdays; previous; emp.var.rate; cons.price.idx; cons.conf.idx; euribor3m; nr.employed.



2‑1 Bảng mô tả dữ liệu.

Các thuộc tính categorical có dtype=object.

Cột Non-Null Count đều là 41188 cho thấy không có dữ liệu null. Tuy nhiên, một số thuộc tính categorical có giá trị là “unknown”. Dùng SimpleImputer của sklearn để xử lý.

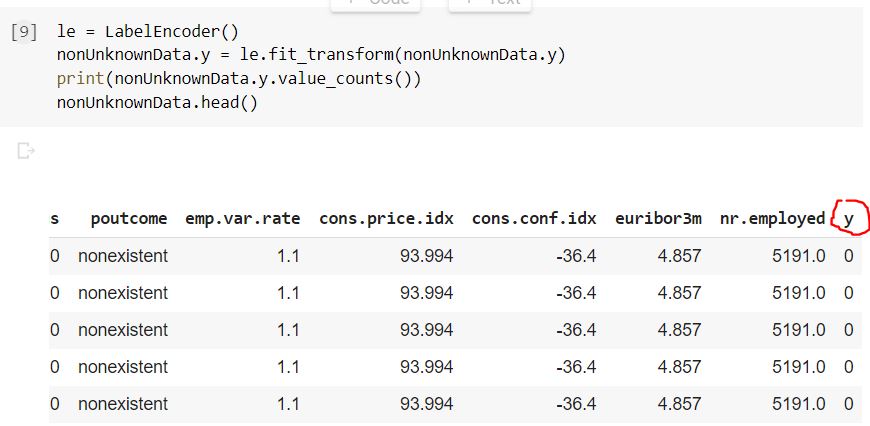


2‑2 Giá trị unknown.



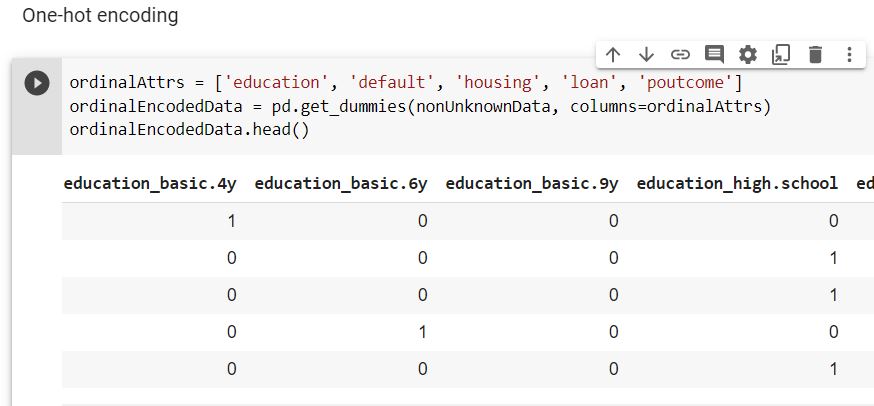
2‑3 Xử lý giá trị “unknown”.

Chuyển dữ liệu của thuộc tính phân lớp “y” yes/no thành 1/0 bằng LabelEncoder.



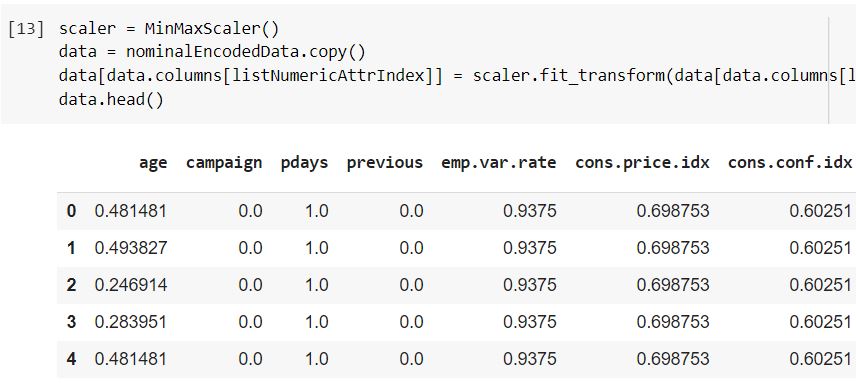
2‑4 LabelEncoder cho thuộc tính phân lớp.

Chuyển dữ liệu của thuộc tính categorical bằng one-hot encoder.



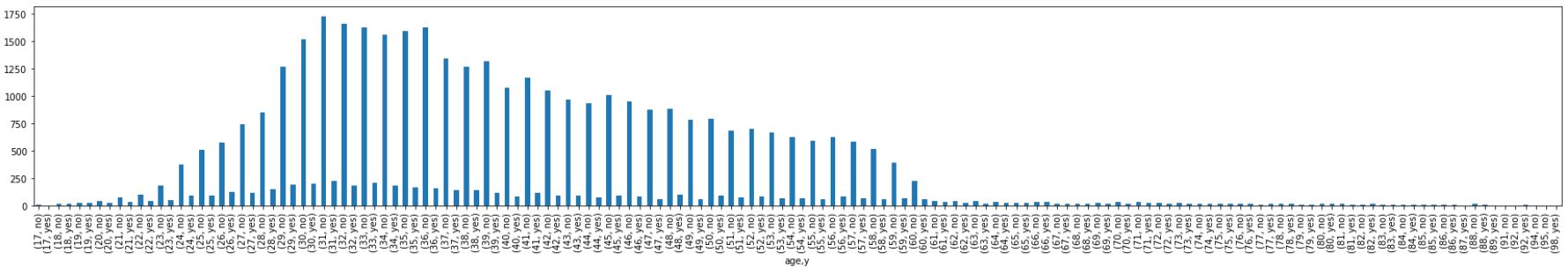
2‑5 One-hot encoder với thuộc tính categorical.

Chuyển dữ liệu của thuộc tính numeric bằng MinMaxScaler.

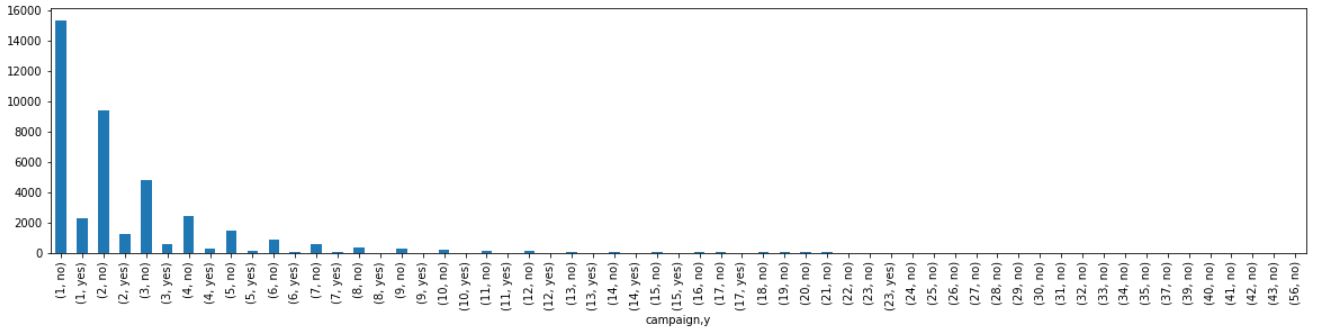


2‑6 MinMaxScaler dữ liệu thuộc tính numeric.

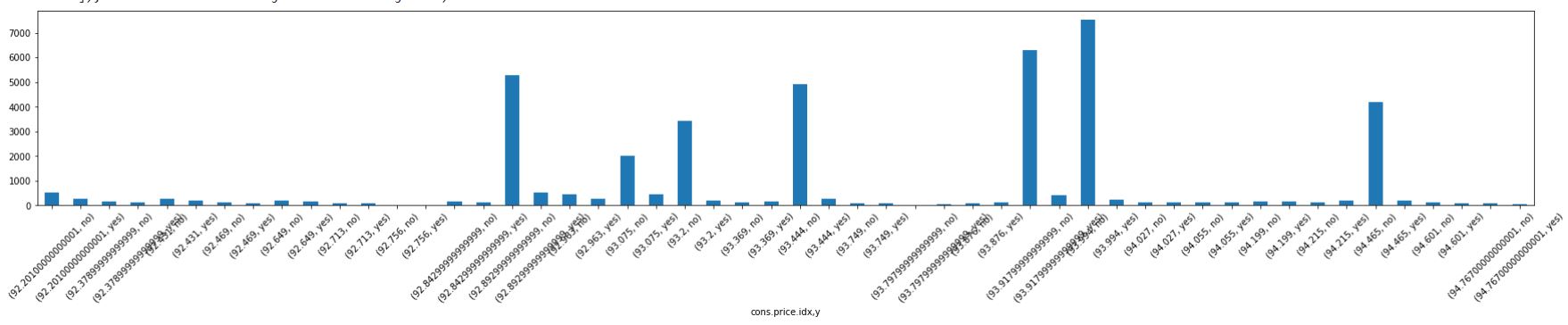
# Vẽ biểu đồ theo phân lớp.



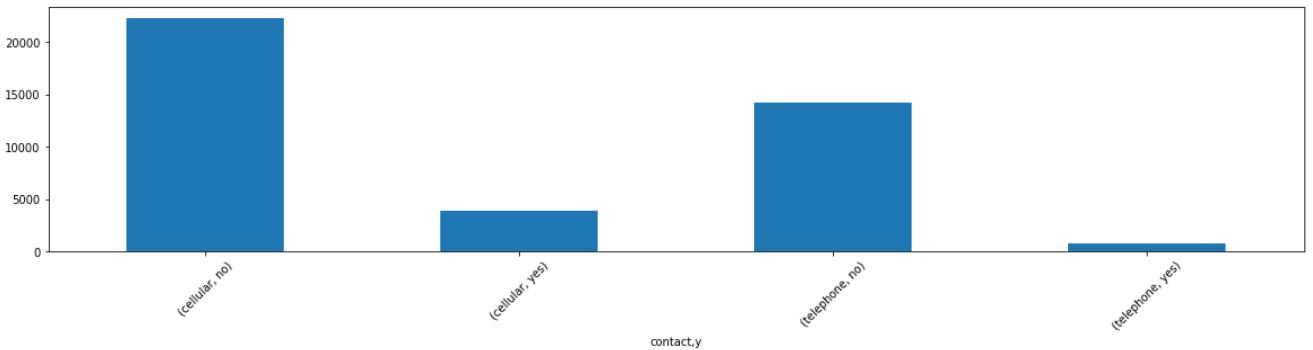
6.1‑1 Biểu đồ hộp các giá trị thuộc tính “age” trái là no, phải là yes kề nhau.



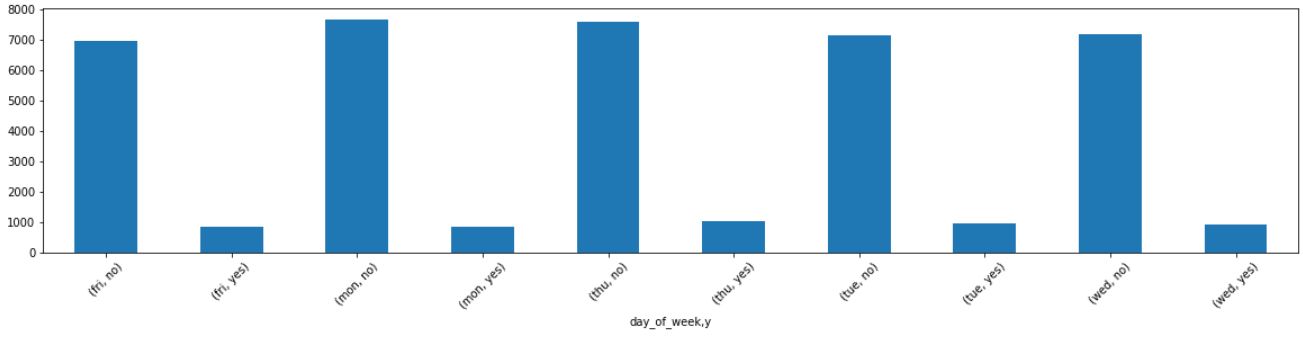
3‑2 Biểu đồ hộp các giá trị thuộc tính “campaign” trái là no, phải là yes kề nhau.



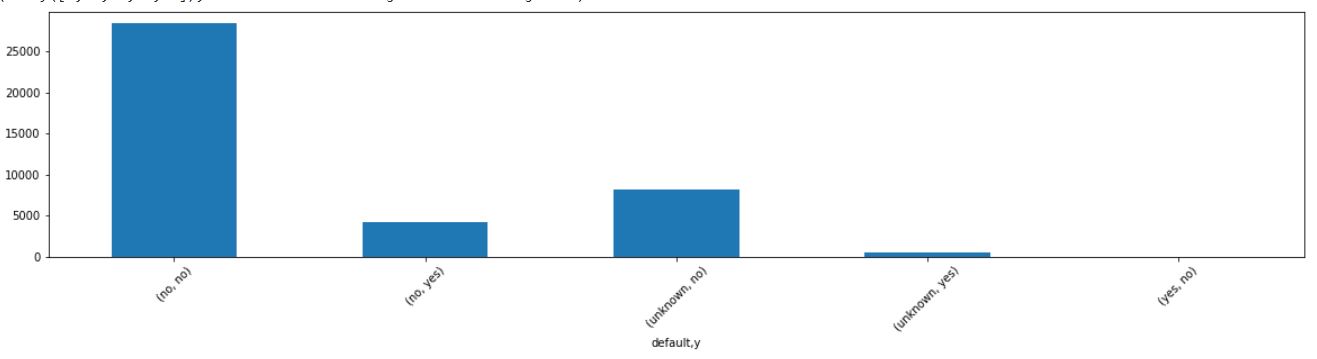
3‑3 Biểu đồ hộp các giá trị thuộc tính “cons.price.idx” trái là no, phải là yes kề nhau.



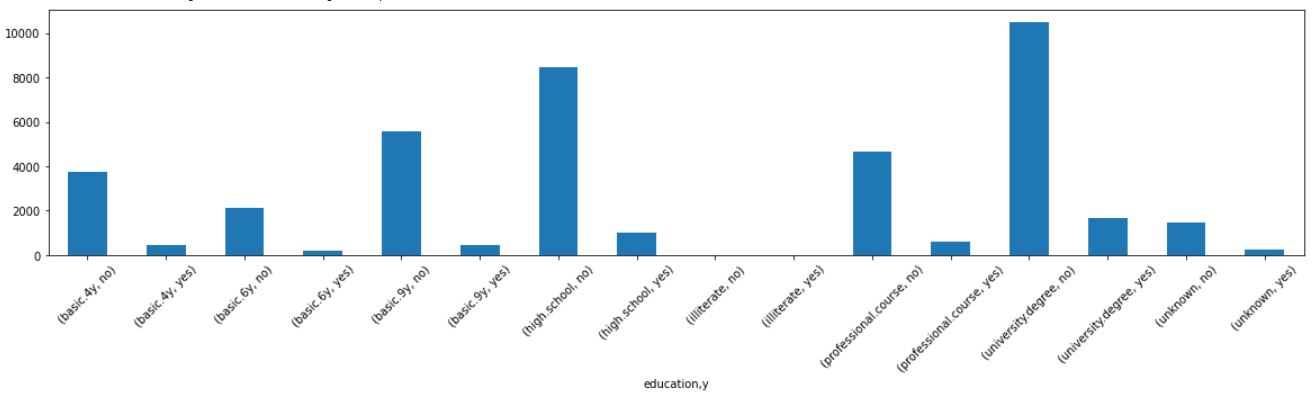
3‑4 Biểu đồ hộp các giá trị thuộc tính “contact” trái là no, phải là yes kề nhau.



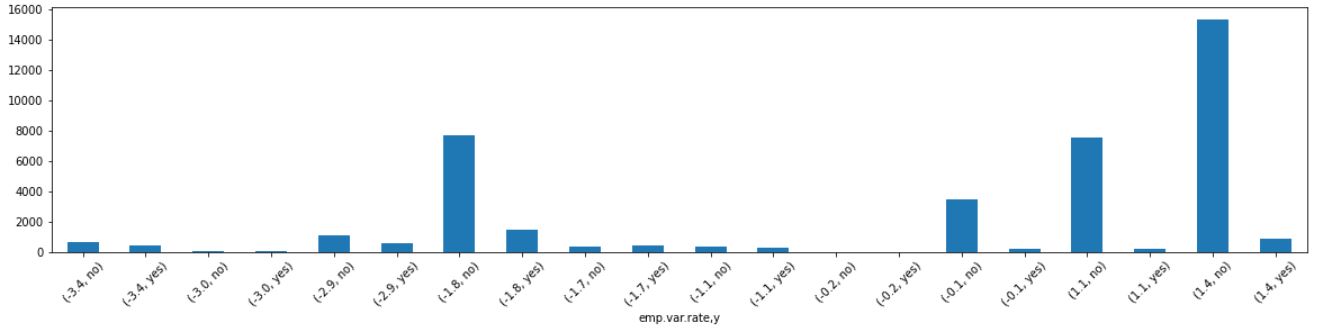
3‑5 Biểu đồ hộp các giá trị thuộc tính “day\_of\_week” trái là no, phải là yes kề nhau.



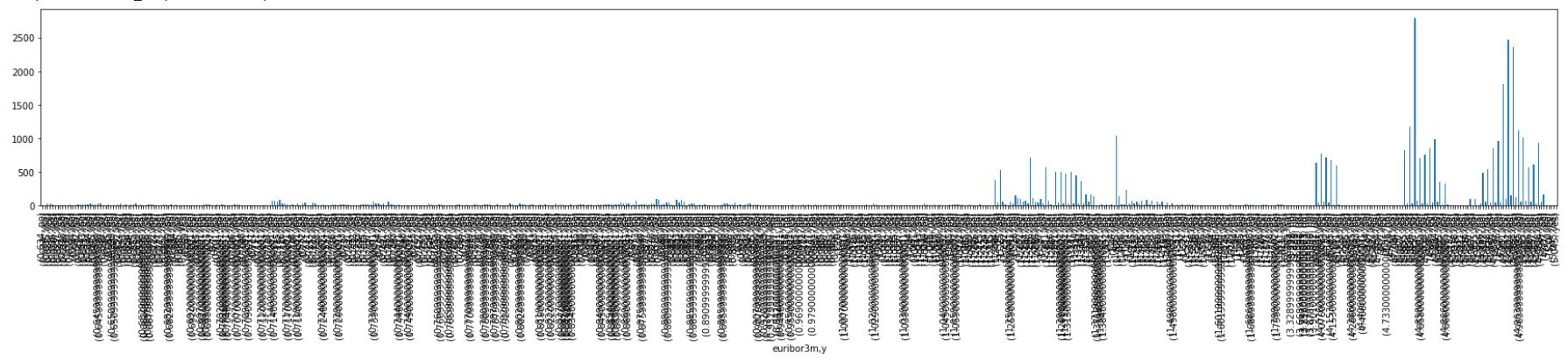
3‑6 Biểu đồ hộp các giá trị thuộc tính “default” trái là no, phải là yes kề nhau.



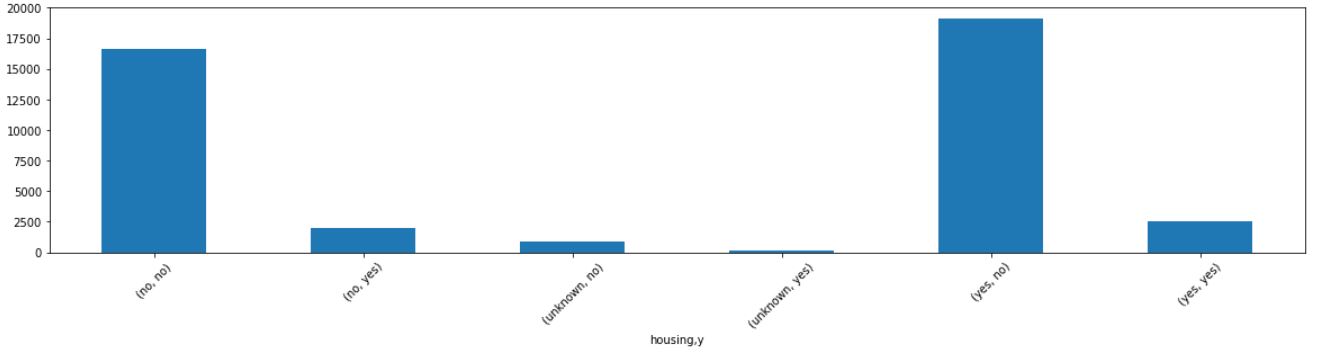
3‑7 Biểu đồ hộp các giá trị thuộc tính “education” trái là no, phải là yes kề nhau.



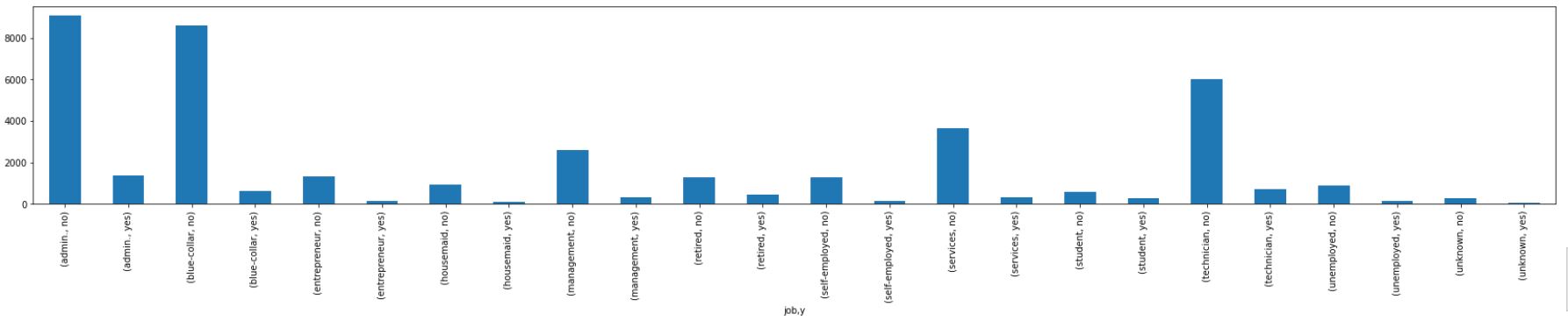
3‑8 Biểu đồ hộp các giá trị thuộc tính “emp.var.rate” trái là no, phải là yes kề nhau.



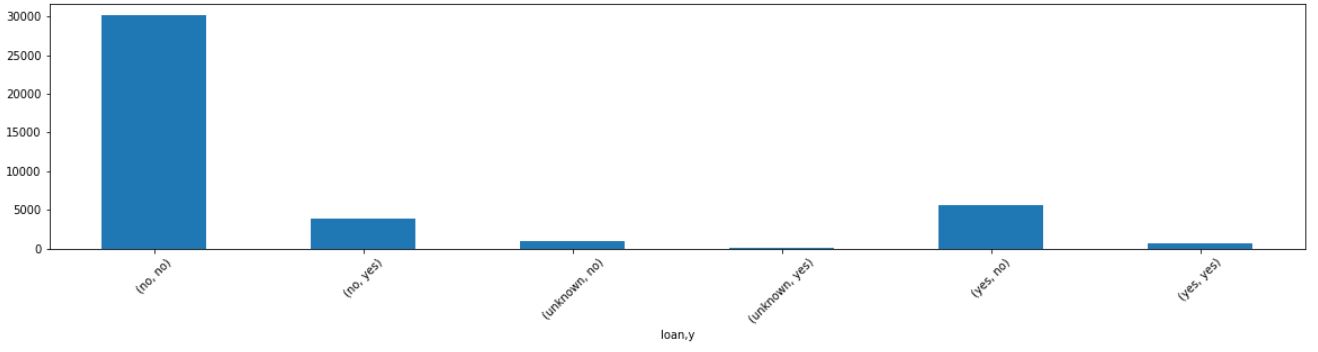
3‑9 Biểu đồ hộp các giá trị thuộc tính “euribor3m” trái là no, phải là yes kề nhau.



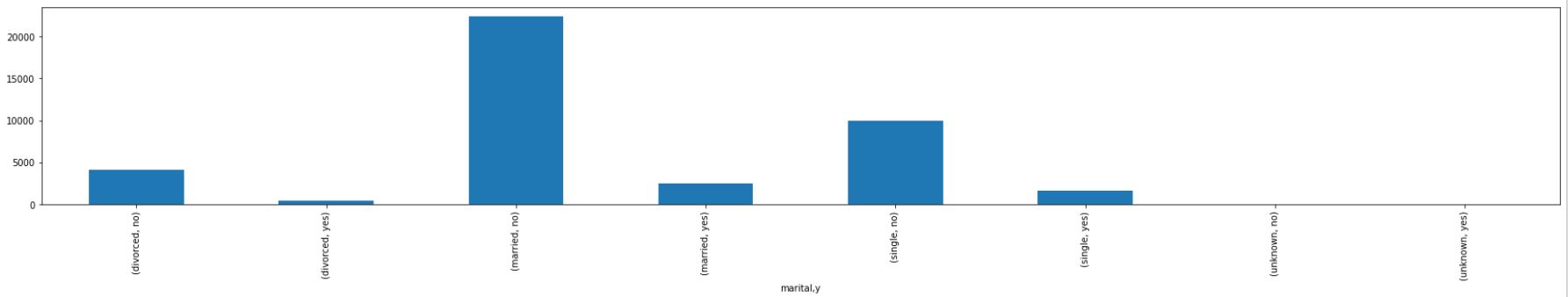
3‑10 Biểu đồ hộp các giá trị thuộc tính “housing” trái là no, phải là yes kề nhau.



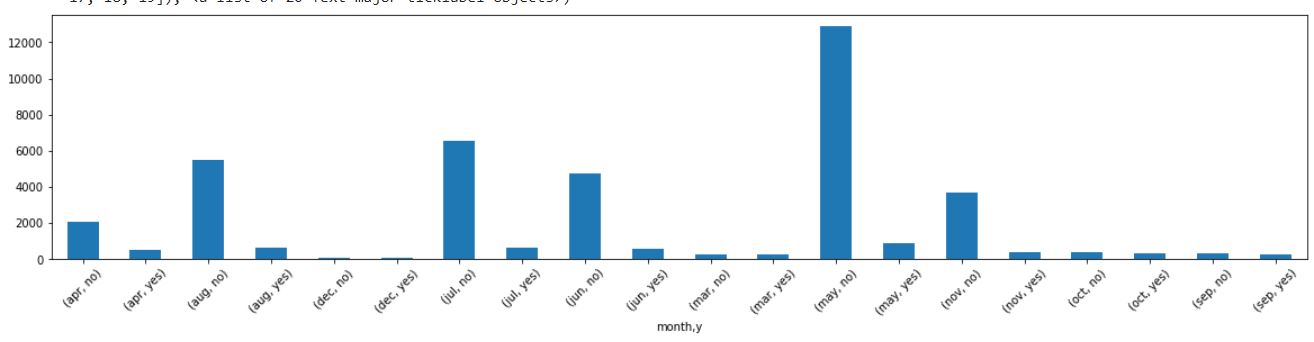
3‑11 Biểu đồ hộp các giá trị thuộc tính “job” trái là no, phải là yes kề nhau.



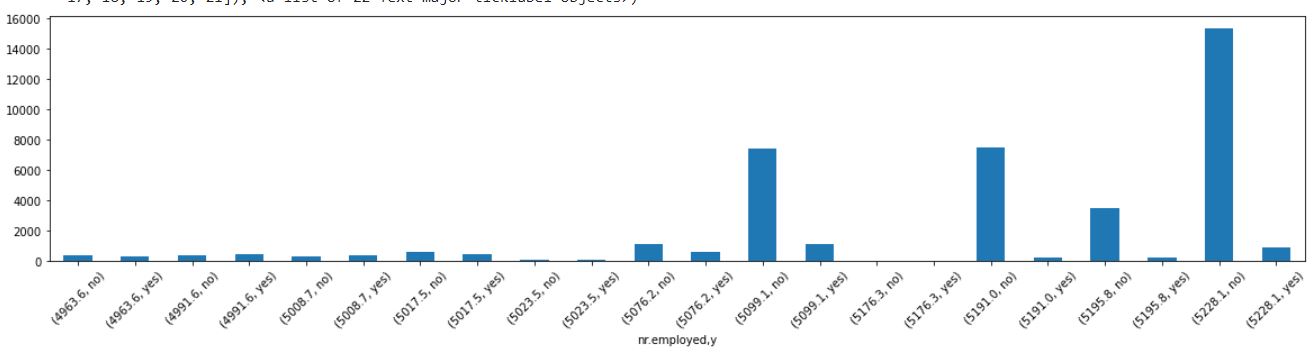
3‑12 Biểu đồ hộp các giá trị thuộc tính “loan” trái là no, phải là yes kề nhau.



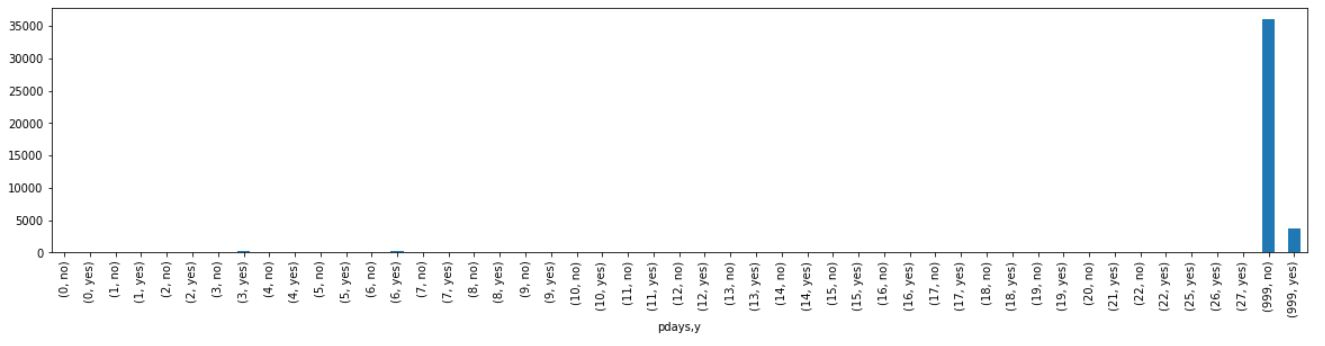
3‑13 Biểu đồ hộp các giá trị thuộc tính “marital” trái là no, phải là yes kề nhau.



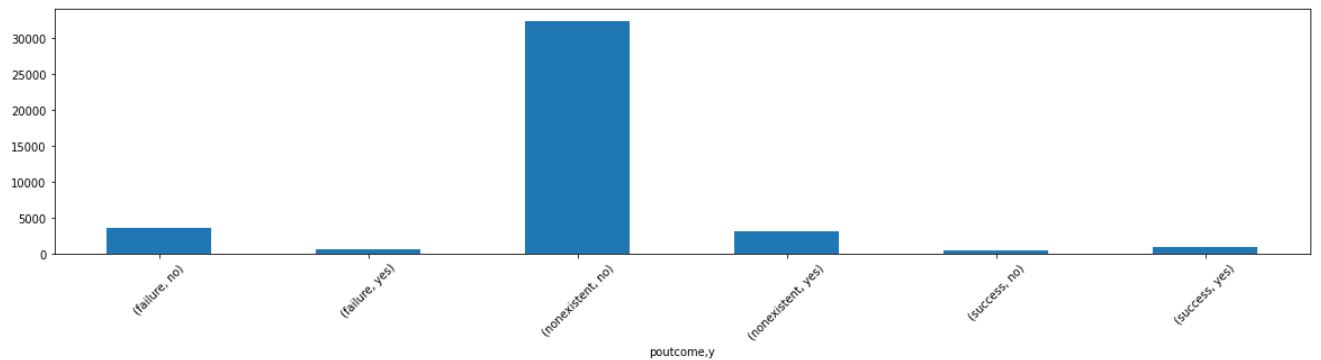
3‑14 Biểu đồ hộp các giá trị thuộc tính “month” trái là no, phải là yes kề nhau.



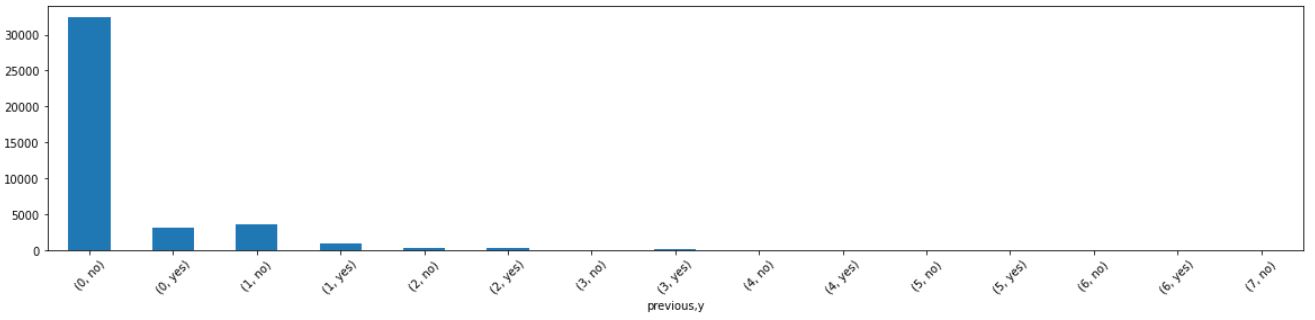
3‑15 Biểu đồ hộp các giá trị thuộc tính “nr.employed” trái là no, phải là yes kề nhau.



3‑16 Biểu đồ hộp các giá trị thuộc tính “pdays” trái là no, phải là yes kề nhau.



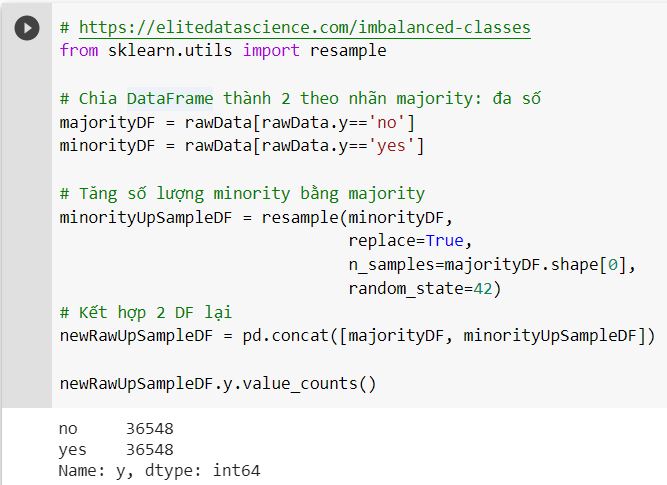
3‑17 Biểu đồ hộp các giá trị thuộc tính “poutcome” trái là no, phải là yes kề nhau.



3‑18 Biểu đồ hộp các giá trị thuộc tính “previous” trái là no, phải là yes kề nhau.

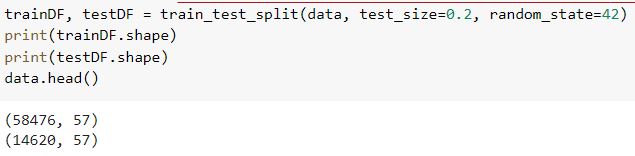
# Cân bằng dữ liệu

Tăng số lượng lớp ít hơn bằng số lượng lớp nhiều hơn, dùng thư viện sklearn resample.

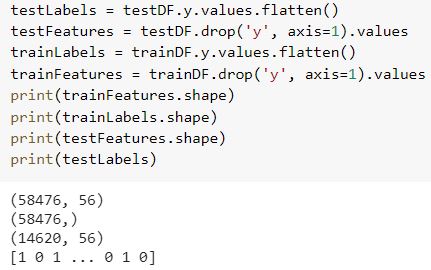


4‑1 Up – sample: tăng số lượng mẫu nhãn yes.

# Chia dữ liệu huấn luyện.



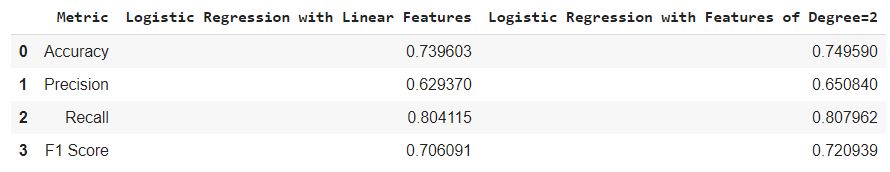
6.1‑1 Chia dataframe theo tỉ lệ train/test = 8/2.



6.1‑2 Chia Features và Labels.

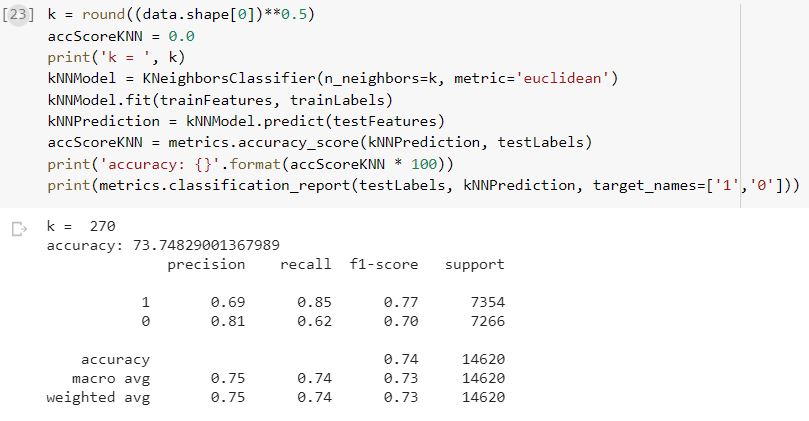
# Xây dựng các mô hình phân loại.

## Mô hình Logistic Regression.

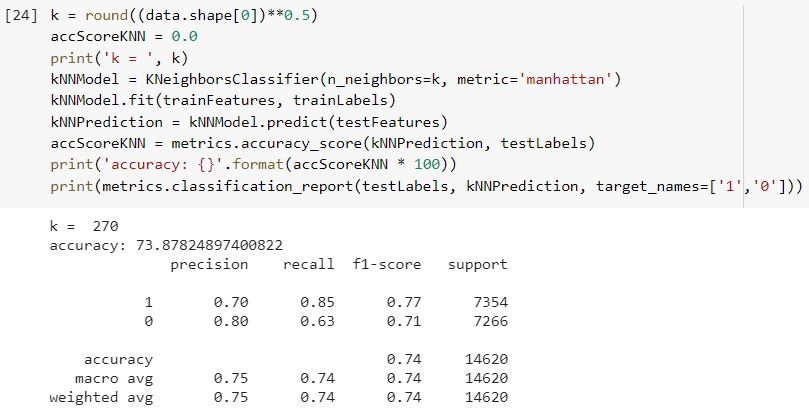


6.1‑1 Kết quả mô hình Logistic Regression với 2 tham số Linear Feature và Polynomial degree 2 Feature.

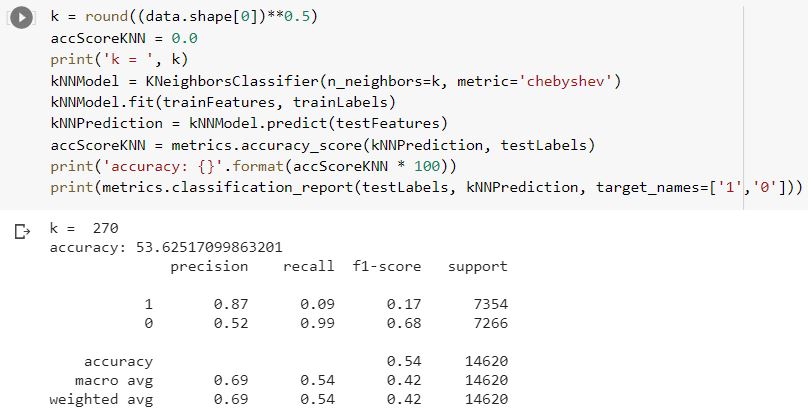
## Mô hình k-Nearest Neighbor.

Kết quả mô hình với các thuộc tính khoảng cách: ‘euclidean’, ‘manhattan’, 

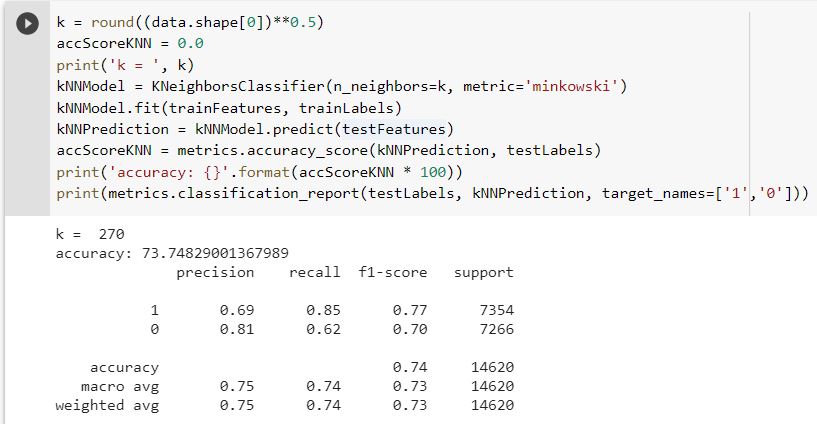
6.2‑1 kNN với độ đo khoảng cách euclidean.



6.2‑2 kNN với độ đo khoảng cách manhattan.

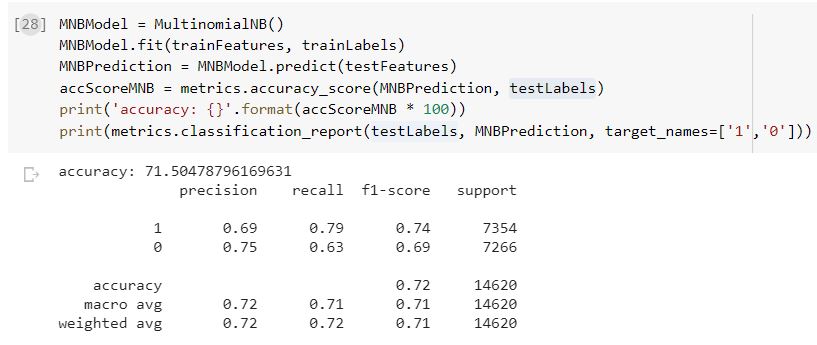
**

6.2‑3 kNN với độ đo khoảng cách chebyshev.

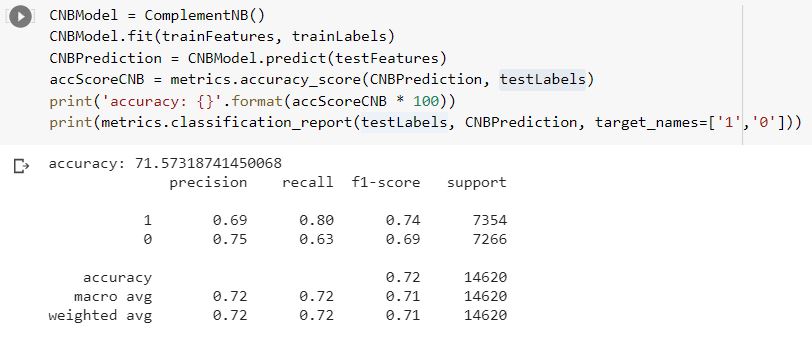


6.2‑4 kNN với độ đo khoảng cách minkowski.

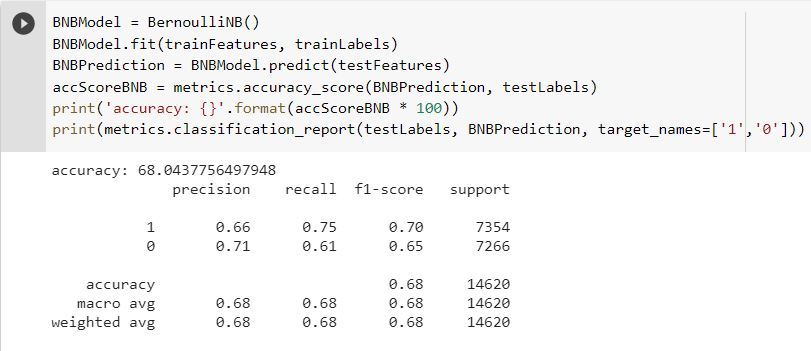
## Mô Hình Naïve Bayes.



6.3‑1 Kết quả mô hình Multinomial Naïve Bayes.

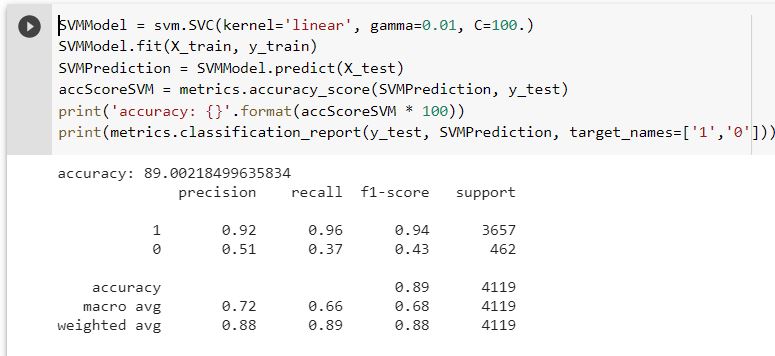


6.3‑2 Kết quả mô hình Complement Naïve Bayes.



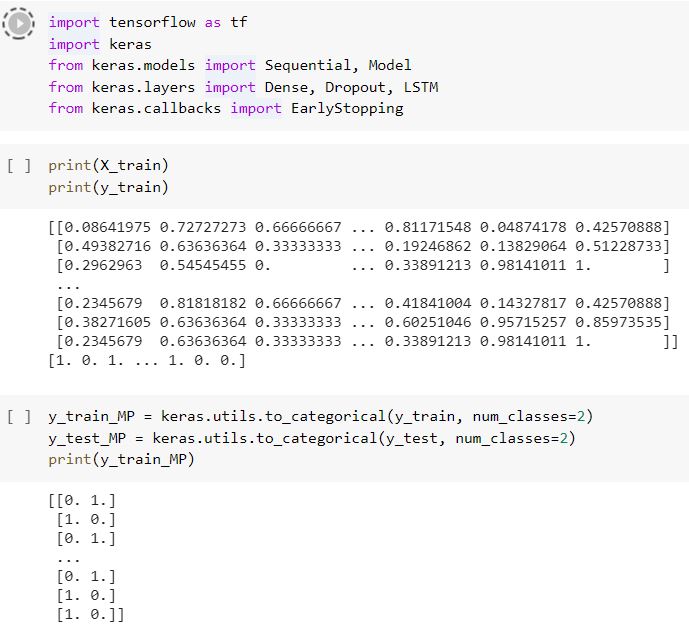
6.3‑3 Kết quả mô hình Bernoulli Naïve Bayes.

## Mô hình SVM.

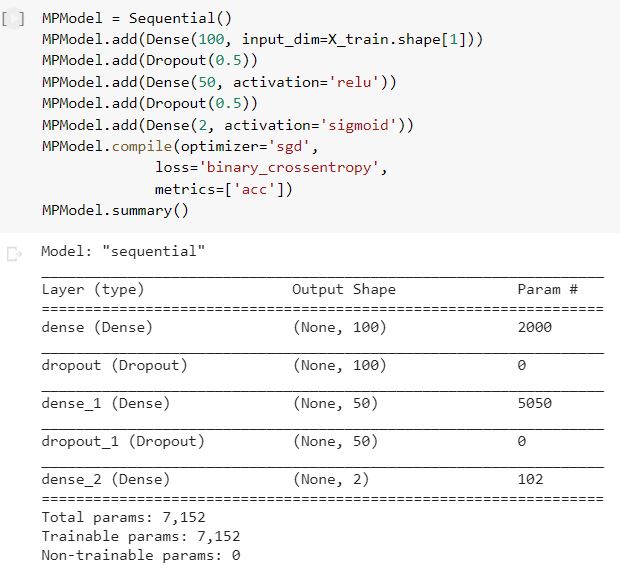
**

6.4‑1 Mô hình SVM

## Mô hình multilayer perceptron.



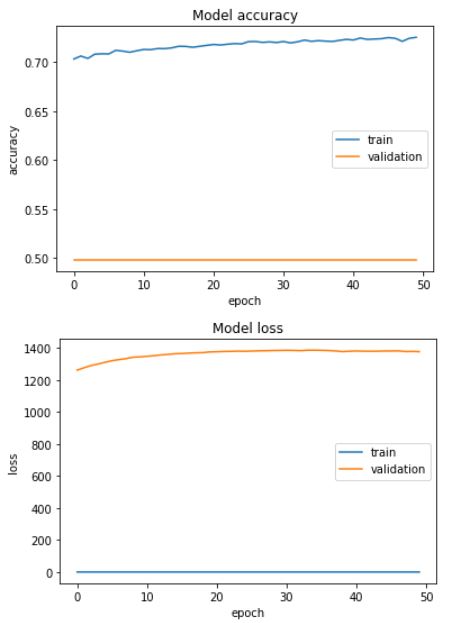
6.5‑1 Sử dụng thư viện keras tạo mô hình multilayer perceptron.



6.5‑2 Tầng input 100 neurons, tầng ẩn 1 50 neurons dùng hàm relu, tầng output dùng hàm sigmoid.



6.5‑3 epochs=50, batch\_size=1000.



6.5‑4 Kết quả multilayer perceptron.

**TÀI LIỆU THAM KHẢO**

1. [Cao Minh Ngọc, Fearture Engineering (Phần 2): Feature engineering với dữ liệu dạng số liên tục (continous numeric data).](https://viblo.asia/p/feature-engineering-phan-2-feature-engineering-voi-du-lieu-dang-so-lien-tuc-continuous-numeric-data-1VgZvXr25Aw)
2. [Cao Minh Ngọc, Feature Engineering (Phần 3): Feature engineering với dữ liệu dạng phân loại (Categorical Data)](https://viblo.asia/p/feature-engineering-phan-3-feature-engineering-voi-du-lieu-dang-phan-loai-categorical-data-GrLZDQx2lk0)
3. [Sklearn Imputation of missing values.](https://scikit-learn.org/stable/modules/impute.html)
4. [Education in Portugal.](https://en.wikipedia.org/wiki/Education_in_Portugal)
5. [Machinelearningcoban.com. Bài 33: Các phương pháp đánh giá một hệ thống phân lớp Classification.](https://machinelearningcoban.com/2017/08/31/evaluation/)
6. [Machinelearningcoban.com. Bài 36: Giới thiệu về Keras.](https://machinelearningcoban.com/2018/07/06/deeplearning/)
7. [Machinelearningcoban.com. Bài 32: Naïve Bayes Classifier](https://machinelearningcoban.com/2017/08/08/nbc/)
8. [Machinelearningcoban.com. Bài 19: Support Vector Machine](https://machinelearningcoban.com/2017/04/09/smv/)
9. [Machinelearningcoban.com. Bài 6: K-nearest neighbors](https://machinelearningcoban.com/2017/01/08/knn/)
10. [Machinelearningcoban.com. Bài 15: Overfitting](https://machinelearningcoban.com/2017/03/04/overfitting/)

**Dữ Liệu Huấn Luyện**

UCIDataset: https://archive.ics.uci.edu/ml/machine-learning-databases/00222/