**Train\_Test\_split**

1.Tổng quan train\_test\_split:

-Train-Test-split là một kỹ thuật để đánh giá hiệu suất của một thuật toán học máy.

-Nó có thể được sử dụng cho các vấn đề phân loại hoặc hồi quy và có thể được sử dụng cho bất kỳ thuật toán học có giám sát nào.

-Thủ tục bao gồm việc lấy một tập dữ liệu và chia nó thành hai tập con. Tập hợp con đầu tiên được sử dụng để phù hợp với mô hình và được gọi là tập dữ liệu huấn luyện. Tập hợp con thứ hai không được sử dụng để huấn luyện mô hình; thay vào đó, phần tử đầu vào của tập dữ liệu được cung cấp cho mô hình, sau đó các dự đoán được thực hiện và so sánh với các giá trị mong đợi. Tập dữ liệu thứ hai này được gọi là tập dữ liệu thử nghiệm.

+Tập dữ liệu huấn luyện : Được sử dụng để phù hợp với mô hình học máy.

+Tập dữ liệu thử nghiệm : Được sử dụng để đánh giá mô hình học máy phù hợp.

-Khi nào cần sử dụng Train\_Test\_split:khi dữ liệu data là lớn có rất nhiều ví dụ học

-Định cấu hình Train\_test\_split:khi định cấu hình cần cân nhắc:

+Chi phí tính toán trong đào tạo mô hình.

+Chi phí tính toán trong đánh giá mô hình.

+Đào tạo thiết lập tính đại diện.

+Kiểm tra tính đại diện.

+Thường định là train 80% và test là 20%

2.Tại sao cần phải phân tách dữ liệu(train\_test\_split):

-Học máy có giảm sát là việc tạo ra một mô hình ánh xạ chính xác các đầu vào đã cho(các biến độc lập hoặc các yếu tố dự đoán) với các đầu ra nhất định(các biến phụ thuộc hoặc các phản hồi).

🡪Các cách để đo độ chính xác của mô hình:phụ thuộc vào vấn đề của bài toán đó.

Ví dụ:với bài toán phân tích hồi quy: dùng hệ số xác định,root-mean-square-error,…với các bài toán phân loại:dùng độ chính xác(accuracy),điểm F1(F1-score),…

Do đó vấn đề đặt ra là cần đánh giá khách quan về hiệu suất dự đoán của mô hình và xác nhận mô hình.

🡪Điều này có nghĩa:ta không thể đánh giá độ chính xác của một mô hình dựa trên tập dữ liệu mà ta dùng để đào tạo,nên ta cần đánh giá nó với bộ dữ liệu mới chưa được qua đào tạo.-->Để làm được ta sẽ chia nhỏ bộ dữ liệu ban đầu thành bộ để đào tạo,bộ để kiểm tra,để đánh giá khách quan thuật toán.

-Vấn đề Underfitting và Overfitting:

+Train\_test\_split có thể phát hiện xem mô hình của ta có gặp phải 2 vấn đề underfitting hay overfitting không.

* Underfitting:là hậu quả của việc mô hình không thể đóng gói các mối quan hệ giữa các dữ liệu.
* Overfitting: là khi mô hình quá phức tạp và học được cả quan hệ giữa dữ liệu và nhiễu🡪nó sẽ cho ta đánh giá tốt với tập dữ liệu đào tạo nhưng mang đến hiệu suất kém với các dữ liệu kiểm tra.

3.Cách sử dụng train\_test\_split:

#khai báo thư viện:

>>> from sklearn.model\_selection import train\_test\_split

#cú pháp

>>>sklearn.model\_selection.train\_test\_split(\*arrays, \*\*options) -> list

Trong đó:

-Arrays:là chuỗi danh sách,mảng numpy,hoặc DataFrame,…tạo lên dữ liệu đào tạo

-options:là các đối số từ khoá tuỳ chọn mà mình thêm vào để có được kết quả mong muốn:

* Train\_size:xác định kích thước của tập đào tạo
* Test\_size:xác định kích thước của tập kiểm tra
* Random\_state:đối tượng điều khiển sự ngẫu nhiên trong quá trình tách.
* Shuffle:kiểu boolean :xác định xem tập dữ liệu có xáo trộn trước khi tách hay không.
* Stratify:một đối tượng giống mảng,nó xác định cách sử dụng phân chia phân tầng.

Ví dụ 1:các thông số mặc định

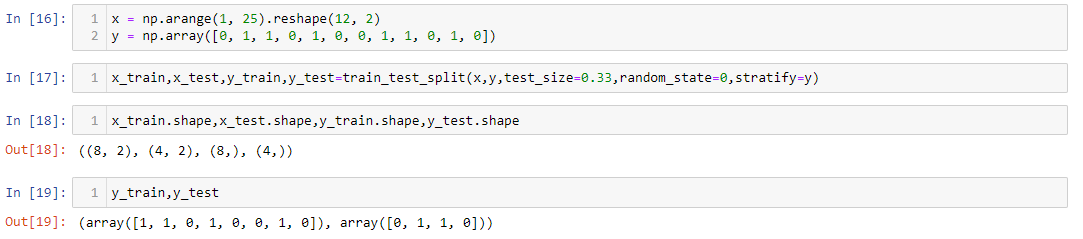
Graphical user interface, text, application, email

Description automatically generated

Ví dụ 2:các thông số:random\_state=0,shuffle=False,test\_size=0.2

Graphical user interface, text, application

Description automatically generatedVí dụ 3: random\_state=0, test\_size=0.33,stratify=y



Ví dụ 4:trong bài toán hồi quy

Graphical user interface, text, application, email

Description automatically generatedví dụ 5:trong bài toán phân loại

Graphical user interface, text, application

Description automatically generated