1. Tiền xử lí ( ngồi nhìn viết nhá )

* Thì bảng dữ liệu nó có 11 cột
* 10 cột là dữ liệu đầu vào . cột cuối là cột mục tiêu sẽ có 4 mục tiêu : **rain cloudy sunny snowny**
* Df.info kiểm tra bảng dữ liệu của mình xem có đủ không 13200 dữ liệu
* Tiếp thì chia 2 biến X – bảng 10 cột , Y – 1 cột cuối ra để train
* Xem những cột mà nó chưa mã hóa về các số ấy thì chuyển nó về số dùng : mã hóa nhị phân hoặc là LabelEncoder()
* Như bài mình là đang có 3 cột này : 'Cloud Cover\_cat', 'Season\_cat', 'Location\_cat

1. Chia Mô hình

Dùng sklearn.model chia dữ liệu ra thành 70 /30 or 80 / 20 đều được . kiểu cái mình thử độ chính xác cao hơn thì mình lấy :

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

**np.random.seed(42)**

**X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X,Y, test\_size=0.3)**

Đấy thì sau khi chia xong thì dùng mô hình có sẵn để train thôi. Như DecisionTree,randomfore, SMV ,logicticclassìier,..

MÔ HÌNH CỦA MÌNH NÓ LÀ PHÂN LOẠI NÊN LÀ CHỌN CÁI CHỮ ĐOẠN SAU LÀ Classifier

Nhóm mình sẽ đánh giá các mô hình mà đc học , kiểu như cùng là 1 bài của mình nhưng mình dùng nhiều mô hình để train - > xong đánh giá độ chính xác của nó bằng mấy cái hàm có sẵn thôi **from sklearn.metrics import accuracy\_score**, ko tự tính cũng đc

+ hoặc có thể dùng K – đéo gì ấy quên r để mở lại K – FOLD để so sánh

Nói chung là 3 bạn huy thảo đạt chưa biết cái gì thì xem 4 cái video đi cái đã .