Its necessary to deal with null values. Supervised learning when target and input both given.

Classification is like different categories of data you apply classification algos on it can be numeric too.

LabelEncoder: takes in target variable

Data Preprocessing

* Null values handle
* Categorical features conversion

Model Selection

1. Use KNN(Train select) out of everyone select one and train it
2. Test the one you selected
3. Evaluate it’s performance compared to original ans.

Separate input and Output

Separate X and Y. use loc, iloc or anything. One more method to do that is pop. Pops the column you want(Y) and stores it in a variable.

2 splits:

X test will result in y predict.

Train->Train: y train, xtrain

Predict->Test: y test, x test.

SKL->Model selection class train test split in it

Test size, train size= per data (0.3…)if you write numeric then it will count 30 data not the per. To train 0.3% data.

Random state= that sample will be used which was used first, second… if we do not use it then it will take different data when each time run and accuracy will be diff.

Num of sample should be same.

Col can be same since supervised learning.

Maintain sequence.

KneighbourClassifier //performs classification and regression too. But here using as classifier.

* Num of neighbours?
* Training knn(use .fit method(x train, y train)->pass your data->model trained)
* If unsupervised learning then only pass x train since no target variable.
* Y predict= Knn.predict(x test)
* Now we will check performance of y(accuracy)->compares and to original ans.
* Num of outcomes wrong ad right? Curr right ans/total ans(accuracy score).
* Accuracy score()Higher acc place above.
* Y predict->model ans.
* Y test-> actual ans.
* We compare y test and y predict.
* Model evaluation->metrics class.
* Model peromance evaluated on unseen data.
* Model overfit and Model unfit
* 2 accuracy
* Y train, Y predict->Training accuracy = how accurate model is trained
* Y Test, Y predict->testing accuracy= how??
* .score (does 2 steps work in 1 line)(.predict and .train on y test and y predict) method or .accuracy\_score methods used for both.
* Model unfit: no conceptual vision, no training acc good and no accuracy good
* Model over Fit= Trainng acc good but testing not good (a student who does ratta)
* Genrealized model: where both are good.
* Confusion matrix (y test, y predict)
* Classification types:

Multi classification, macro= (tells type) by default binary.

* Lazy learner?
* Will compare one sample to all sample using distance formula, and record the distances (Euclidean distance) nearest neigbour value jitni sleect ki huwi thi who itne distance lega. N = 3. That means will take 3 neighbours. Will check their properties and check majority target and then set that target to the sample data we took in the beginning.
* If 2 different output are there(when seeing neihbours target value ) then how we elect majority? Neighbours should always be odd nums that’s why no even it can be same. Odd soo that decision is easy.

Training working on partial derivative. See the math behind .fit function.

KNN different evaluation. Using distance matrix. Kispe zyda ache se kaam kr raha hai amd all?

* Confusion Matrix:
* TP, FP, TN, FN(True, False, Positive, Negative). On which model is working better and whose classfiication s better positive or neg?. this means what is the input and wether its output is true or not, if not true then misclassification

|  |  |  |  |
| --- | --- | --- | --- |
|  | A | B |  |
| TP | +ve | +ve |  |
| TN | -ve | -ve |  |
| FP | +ve | -ve |  |
| FN | -ve | +ve |  |
|  |  |  |  |

* Precision, recall, f1 scores…. Are made from TP, FP, TN, FN.

APPLY STD SCALAR IF THE DATA VALUES ARE NOT OF ONE STD. MEANS SOMEVALUES IN POINTS, SOME IN 3-4 DIGITS.

Q) HOW TO INCREASE ACCURACY OF DATA?

Ask miss how to know wether something is of good accuracy or not.