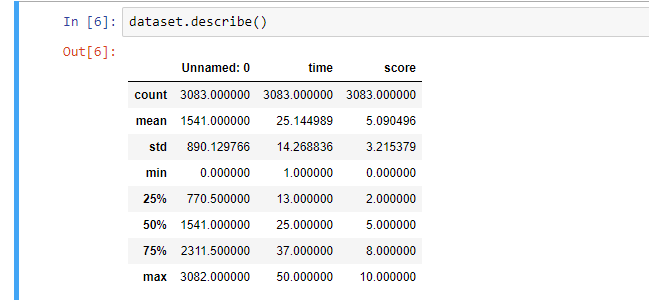
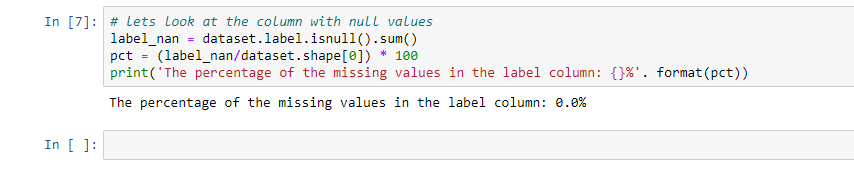
Classifier Training

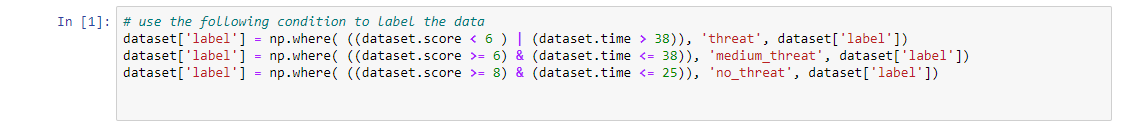
Dataset Preparation



Missing Values

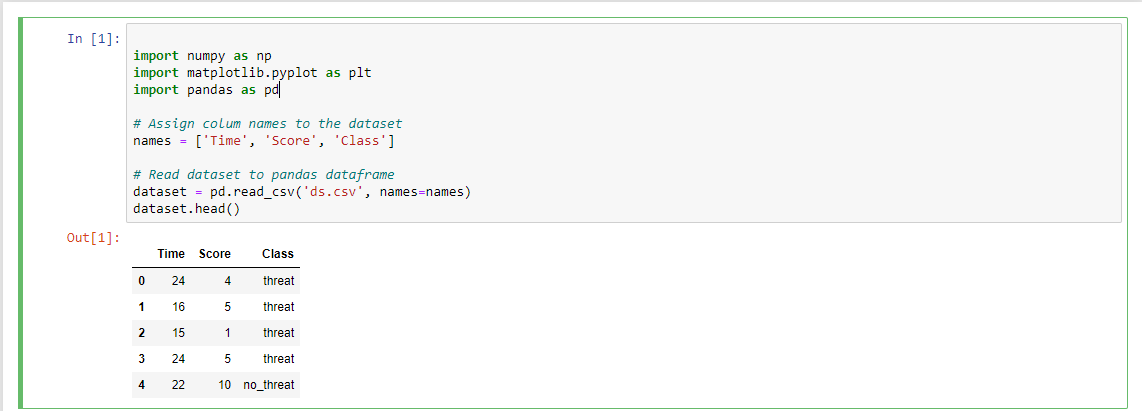


Condition of Classification



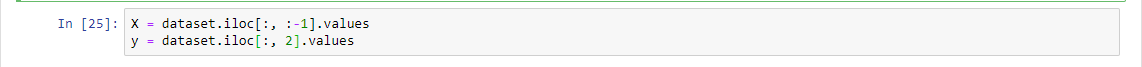
Training to K-Nearest Neighbor

1. Importing the library & Dataset



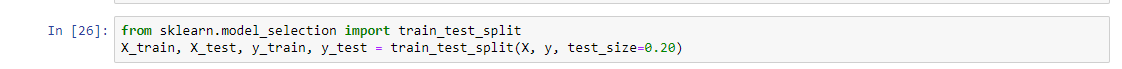
1. Preprocessing

The X variable contains the first four columns of the dataset (i.e. attributes) while Y contains the labels.



1. Train Test Split

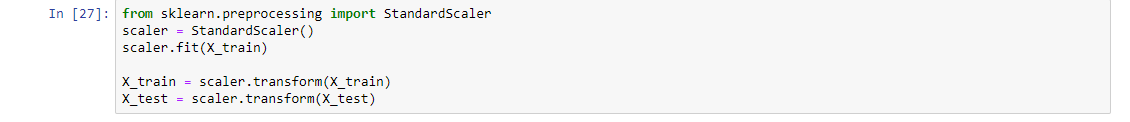
To avoid over-fitting, we will divide our dataset into training and test splits, which gives us a better idea as to how our algorithm performed during the testing phase. This way our algorithm is tested on un-seen data, as it would be in a production application. To create training and test splits, execute the following script:



The above script splits the dataset into 80% train data and 20% test data. records and the test set contains 30 of those records.

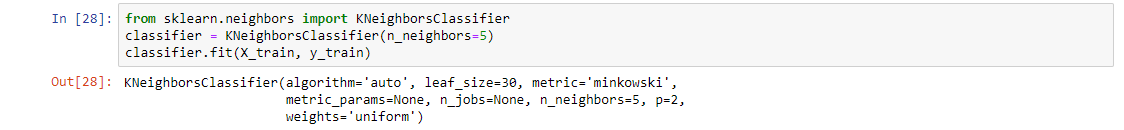
1. Feature Scaling

Before making any actual predictions, it is always a good practice to scale the features so that all of them can be uniformly evaluated.



1. Training and Prediction:

The first step is to import the KNeighborsClassifier class from the sklearn. neighbors library. In the second line, this class is initialized with one parameter, i.e. n\_neigbours. This is basically the value for the K. There is no ideal value for K and it is selected after testing and evaluation, however to start out, 5 seems to be the most commonly used value for KNN algorithm.

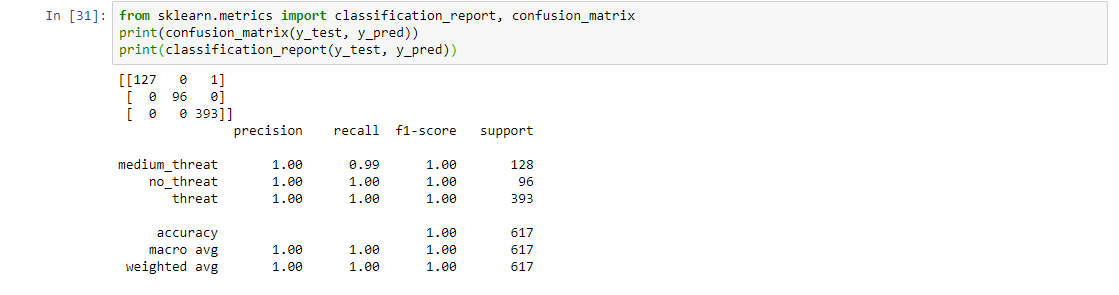


The final step is to make predictions on our test data



1. Evaluating The Algorithm

For evaluating an algorithm, confusion matrix, precision, recall and f1 score are the most commonly used metrics. The confusion matrix and classification report methods of the sklearn. Metrics can be used to calculate these metrics.



The results show that our KNN algorithm was able to classify all the 617 records in the test set with 100% accuracy, which is excellent.

1. Comparing Error Rate with the K-Value

. We randomly chose 5 as the K value and it just happen to result in 100% accuracy. One way to help you find the best value of K is to plot the graph of K value and the corresponding error rate for the dataset. In this section, we will plot the mean error for the predicted values of test set for all the K values between 1 and 40. To do so, let's first calculate the mean of error for all the predicted values where K ranges from 1 and 40.



