Data Preprocessing using Scikit-Learn

By
G NagaMounika

Introduction to Scikit-Learn

- Scikit-Learn become one of the most popular open source machine learning libraries for Python
- It Supports Machine Learning models like SVM, KNN, K Means algorithms and so on
- The library is focused on modeling data. It is not focused on loading, manipulating and summarizing data.

Why Data Preprocessing is Required

This technique is used to detect and handle outlier in our data

Scaling

Feature Scaling

Feature scaling is the method to limit the range of variables. It is performed on continuous variables.

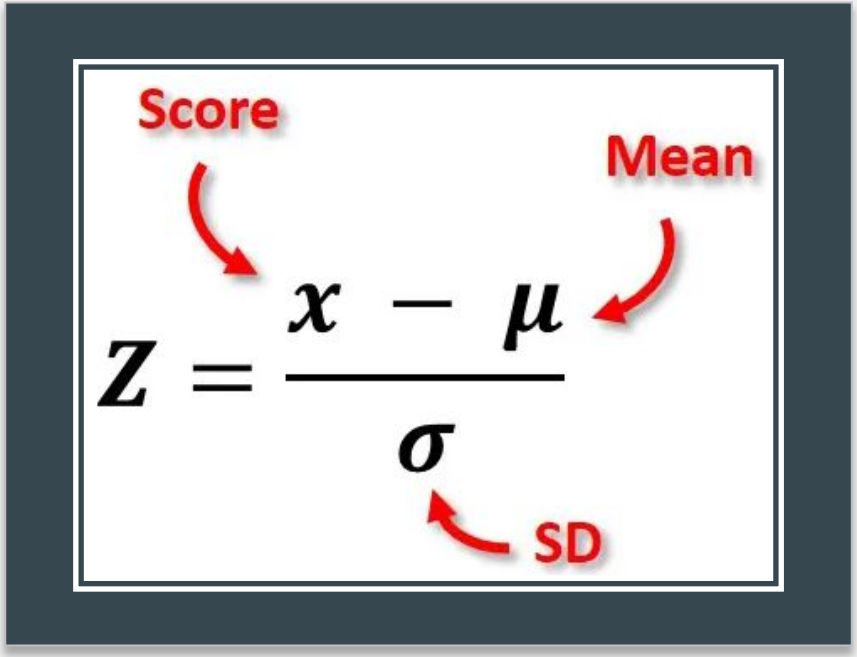
Scaling Techniques

- Standard Scaler
- Robust Scaler
- MinMax Scaler
- Normalization
- etc..

Standard Scaler

To transform data points between range of Mean and Standard Deviation

- `preprocessing.StandardScaler()`



Robust Scaling

Robust Scaler transforms the data points between Inter Quartile Range (IQR)

- `preprocessing.RobustScaler()`

$$\frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)}$$

Min Max Scaler

It transforms the data points, range between min value to max value in given data set

- `preprocessing.MinMaxScaler()`

$$\frac{x_i - min(x)}{max(x) - min(x)}$$

Normalizer

It scales the data in a range between 0 to 1

- `preprocessing.Normalizer()`

$$Z = \frac{Xi}{\text{Sqrt}(X1^2 + X2^2 + X3^2)}$$


THANK YOU