**Pandas Profiling :** is a library used to do many things for us. It generates the html file which include max of data analysis.

**Feature Engineering:** is the process of using domain knowledge to extract features from raw data. These features can be used to improve the performance of machine learning algorithm.

Feature engineering can be classified into 4 sub categories:

1. Feature Transformation
2. Feature Construction
3. Feature Selection
4. Feature Extraction

**Feature Transformation:** You transform the feature so it can perform better with your model. It includes *Missing values, Handling Categorical Features, Outlier Detection, Feature Scaling.*

**Feature Construction:** Sometime you construct feature by your own to get better result.

**Feature Selection:** Sometime you didn’t give all the features to algorithm you remove some of them which are not much import is called feature selection.

**Feature Extraction:** Extracting feature from given features pro-grammatically is called feature extraction.

* **Feature Scaling:** is a technique to standardize the independent features present in the data in a fixed range.

1. g when ever we have some data and we are solving for categorical data using KNN algorithm. Let suppose we have two columns age and salary when ever we will try to get distance between two row salary will always dominate because it may in 1000’s and age will be in 10’s. Feature scaling has following technique. *Standardization, Normalization.*

**Standardization:** *In standardization we do two thing mean centering and scaling by the factor of standard deviation. See the notebook for practical example.*

**When to use standardization:** use when you use these algorithms , K-Means, K-Nearest-Neighbours, PCA(Principal Component Analysis), Artificial Neural Networks, Gradient Descents.

When Not to use **standardization: decision tree, Random Forest, etc…**

**Normalization:** *is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting difference in the ranges of the values or losing information.*

E.g if we have weight feature then we can eliminate the unit of this feature that is kg, pound etc

**Types of normalization**: Min-Max scaling, Mean normalization, Max absolute scaling, Robust scaling.

**When to use Min-Max scaling**: let suppose we have weight feature and we are asked to normalize this feature. Formula for this will be = X = Xi - Xmin / Xmax - Xmin. It will always range between [0,1].

**Mean Normalization:** When we have numerical data and we want to normalize it we use mean normalization or mean centering. Formula for this is = X = Xi - Xmean / Xmax - Xmin. It gives us range between [-1 to 1]. There is no class in scikit learn. You have to right its class on your own.

**Max Absolute Scaling :** formula = X = Xi / |Xmax| . It uses when we have large numbers of zero’s in the data. Scikit learn has its class.

**Robust Scaling:** formula = X = Xi - Xmedian / IQR {75 - 25}. It is best when then are large number of outliers in our data.

**Normalization VS Standardization**

1. Question 1 to ask : Is feature scaling required?
2. Max of time standardization is used.
3. Min-Max (normalization) when you have known in future the value of min and max.