Before Data processing we need to know about feature scaling

Feature Scaling:

Always applied to columns. We cannot apply across rows

Types:

Normalization

Standardization

Imagine a dataset with average salary of a person and age

A 70000$ 45 years

B 60000$ 44years

C 52000$ 40 years

Note the salary difference and age difference

The values will vary a lot in a dataset

We group person c with person b because they seem closer when compared to person A

We do normalization to both salary and years column

A,B,C

Salary: 1, 0.444, 0

Age: 1, 0.8, 0

Data Preprocessing:

Every machine learning model has to be pre processed

Step 1: Import the libraries (import the data preprocessing toolkit)

Step 2: import the dataset

Step 3: taking care of missing data

Step 4: Encoding Categorical data

Step 5: Splitting the dataset into training and Test set

Step 6 : Feature Scaling (this will be determined in machine learning models to use or not depending on dataset)

Important

Any data set we train in machine learning model, common factors are

Features (this will mostly be in the first column of dataset)

Dependent variable (this vector is always in last column

Features are independent

Attribute functions we use:iloc

: in pyrthon means range

How to take all columns excpt last one (dependent variable vector)

[ :, :-1 ]

-1 in pthon means index of large column

Range in python includes lower bound but excludes upper bound

Therefore, they take all columns excluding the last one

Never forget to import the data in google colab