**Data Pre-processing**

1. Handling missing data
2. If amount of missing data is way less than complete data, deleting those rows is a go-to idea.
3. If not, then replace the missing data with average of all data in that column.(Recommended)
4. Replace with Median salary
5. Replace with most frequent

Make use of SimpleImputer function in sklearn library

1. Encoding Categorical Data

One hot encoding :- creating binary vector of distinguished category (not recommended for just true/false or yes/no value… use 0,1 instead)

1. Encoding Dependent Variable

LabelEncoder : - for 0 and 1 type of encoding

* Feature Scaling should be done after splitting dataset into training and test set, because test set is going to be a brand new dataset which we are not going to work with, if feature scaling is applied before splitting, it’ll provide information leakage before training is done on test set by giving mean, median ,standard deviation etc info before training is done.

1. Feature Scaling

To avoid some features to be dominated by some other features in such a way that dominated features are not considered by ML model.

No need to apply for all ML model

2 techniques :-

1. Standardisation

Xstand = (X-mean(X))/standard deviation(X)

1. Normalisation

Xnorm = (X – min(X))/(max(X) – min(X))