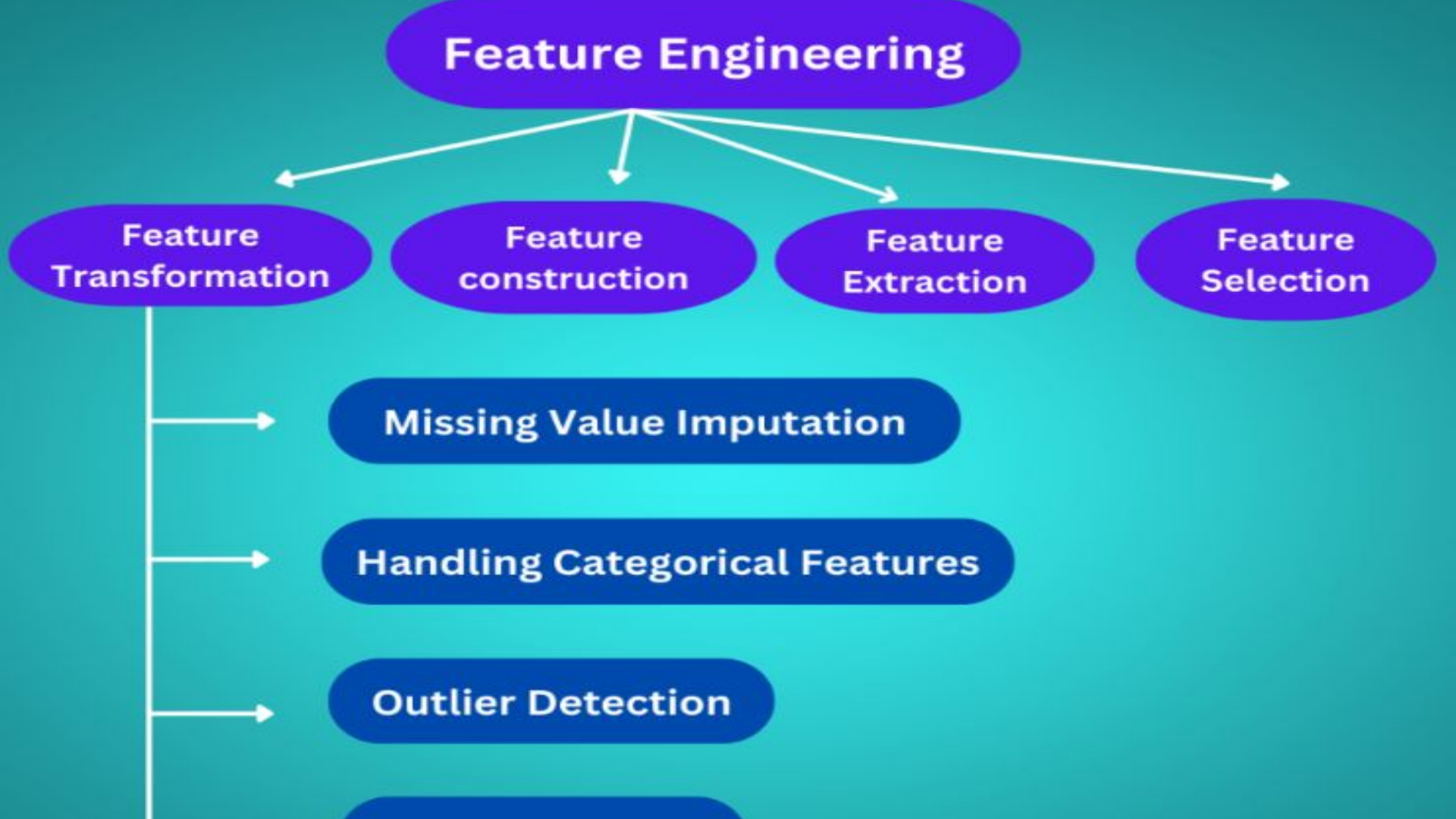
Feature engineering

It is the process of using domain knowledge to extract features (characteristics, properties, or attributes) from raw data that make machine learning algorithms work better


Feature Engineering
Feature Transformation
Feature construction
Feature Extraction
Feature Selection
Missing Value Imputation
Handling Categorical Features
Outlier Detection

1.Feature transformation :-

It involves transforming the raw features of a dataset into a more suitable format for modeling.

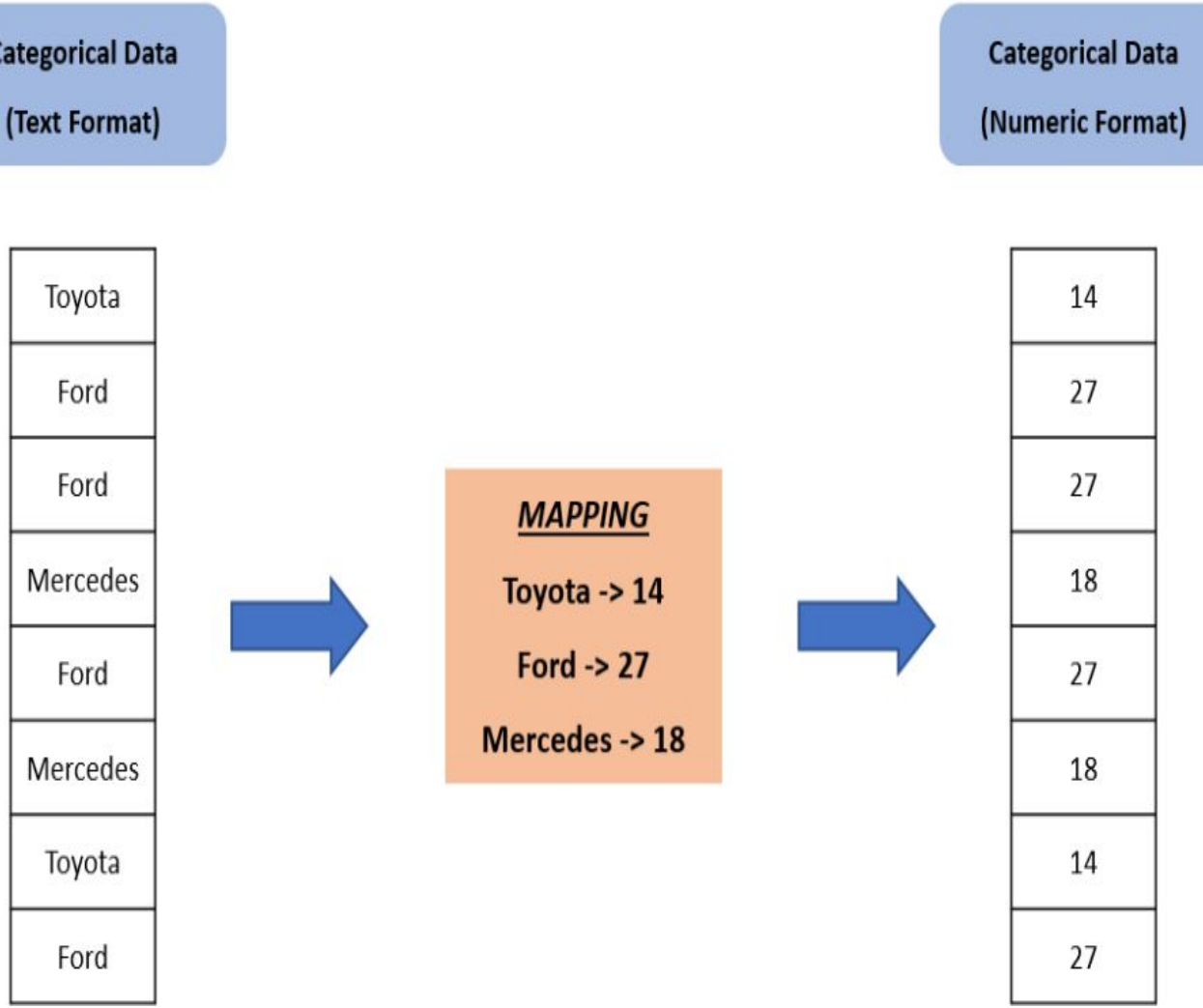
a. Missing value imputation:-

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

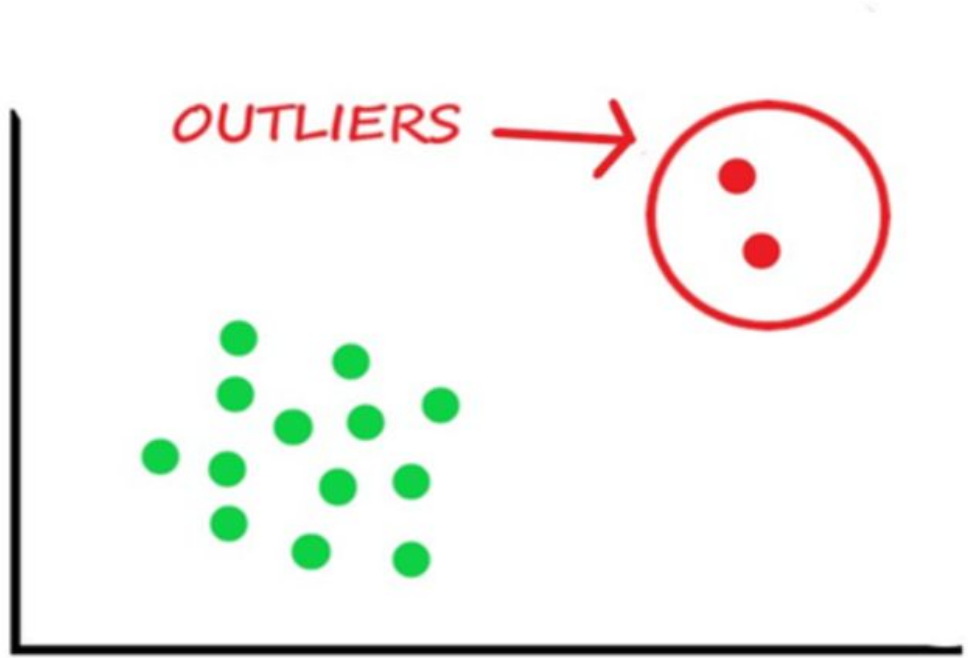
b.categorical features:-

categorical features can be used as inputs to machine learning algorithms, they must be encoded as numerical values



c. outlier :

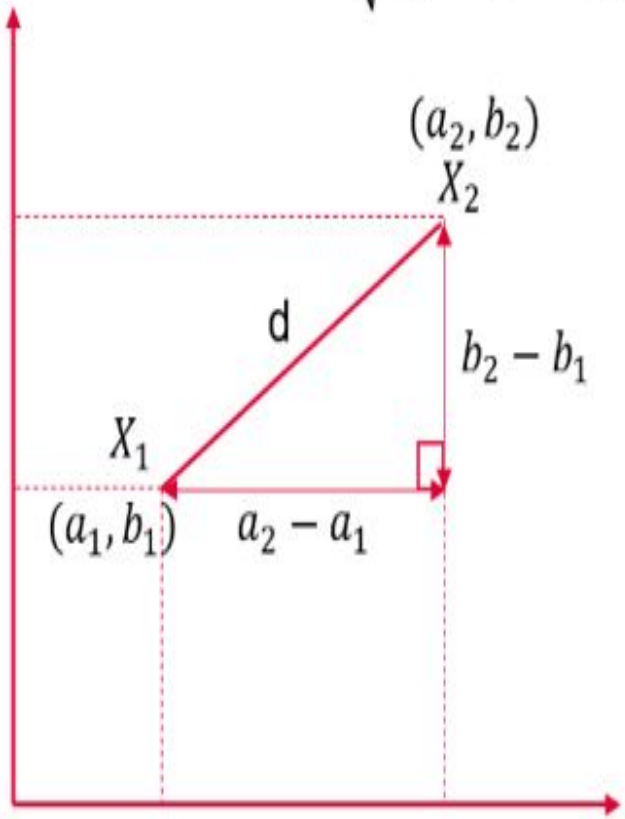
outlier is a data point that stands out a lot from the other data points in a set.



d. feature scaling:-

- In general Data set contains different types of variables having different magnitude and units (kilograms, grams, Age in years, salary in thousands etc).
- The significant issue with variables is that they might differ in terms of range of values.
- So the feature with large range of values will start dominating against other variables.
- So to overcome this problem, we do feature scaling.
- The goal of applying Feature Scaling is to make sure features are on almost the same scale so that each feature is equally important and make it easier to process by most ML algorithms.

$$\textbf{Euclidean Distance } d = \sqrt{(a_2-a_1)^2 + (b_2-b_1)^2}$$



- X_1 and X_2 are two data points means two different rows in data set. And Data set is two dimensional feature space
- For three dimensional feature space suppose two data points are $X_1(a_1, b_1, c_1)$ *and* $X_2(a_2, b_2, c_2)$ then distance can be calculated as below:

$$d = \sqrt{(a_2-a_1)^2 + (b_2-b_1)^2 + (c_2-c_1)^2}$$

- In case of multi-dimensional vector space, More generic formula will be

$$\textbf{Euclidean Distance } d = \sum_{i=1}^{k} \sqrt{(X1_i-X2_i)^2}$$

#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

Distance between Emp2 and Emp1 = $\sqrt{(27-44)^2+(47000-73000)^2}$ = 31.06

Distance between Emp2 and Emp3 = $\sqrt{(30-27)^2+(53000-47000)^2}$ = 6.70

Now see if the task is to identify the neighbors of emp2 then by above distance we can say emp3 is more close to emp2 as emp2 and emp3 share less distance between them. but think now if we simply increase the salary number then this distance will increase and then it will imply that emp2 and emp3 are not similar. So idea behind feature scaling is that value range should not impact the feature behavior. When we want to do the comparison between two entities we should first bring them in same page or same level or same scale then only we can do the comparison.

Feature Scaling Techniques:-

1. Normalization
2. Standardization

Normalization

- Normalization is also known as min-max normalization or min-max scaling.
- Normalization re-scales values in the range of 0-1
- If your data does not follow a normal distribution, normalization can be a better choice.
- Use with algorithms like k-NN, neural networks, or distance-based algorithms

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

Normalization

Age	Normalized Age	Salary	Normalized Salary
44	0.80952381	73000	0.838709677
27	0	47000	0
30	0.142857143	53000	0.193548387
38	0.523809524	62000	0.483870968
40	0.619047619	57000	0.322580645
35	0.380952381	53000	0.193548387
48	1	78000	1
	Range 0-1		**Range 0-1**

How to calculate Normalized value?
X = 35, min = 27, max = 48 for column Age.
Xnorm(for 35) = $\frac{35-27}{48-27} = 0.3809$

After Normalization

Distance between Emp2 and Emp1 = $\sqrt{(0-.80)^2+(0-.83)^2} = 1.15$

Distance between Emp2 and Emp3 = $\sqrt{(.14-0)^2+(.19-0)^2} = 0.23$

Comparison will be more significant

Standardization

- Standardization is also known as z-score Normalization.
- In standardization, features are scaled to have zero-mean and one-standard-deviation.
- It means after standardization features will have mean = 0 and standard deviation = 1
- Use standardization when data follows a normal distribution.
- Use with algorithms like linear regression, logistic regression, or linear discriminant analysis

Standardization:

$$z = \frac{x - \mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} (x_i)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000
		Mean = 37.42857	Mean = 60428.5714
		Std. Dev. = 6.883876	Std. Dev = 10499.7570

Standardization

How to calculate Standardized value?
X = 35, mean = 37.42, Std. Dev. = 6.88
for column Age.
Xstd(for 35) = $\frac{35-37.42}{6.88}$ = -0.3527

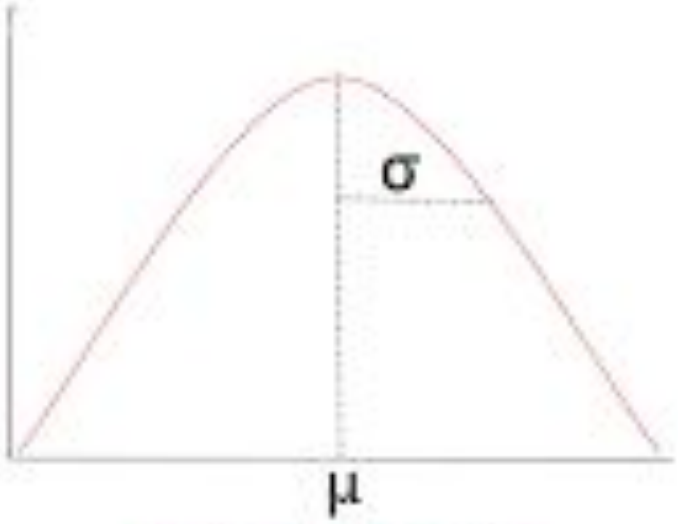
Age	Standardized Age	Salary	Standardized Salary
44	0.954611636	73000	1.197306616
27	-1.514927162	47000	-1.278941158
30	-1.079126198	53000	-0.707499364
38	0.083009708	62000	0.149663327
40	0.373543684	57000	-0.326538168
35	-0.352791257	53000	-0.707499364
48	1.535679589	78000	1.673508111
	Mean = 0 **Std. dev. = 1**		**Mean = 0** **Std. dev. = 1**

After Standardization

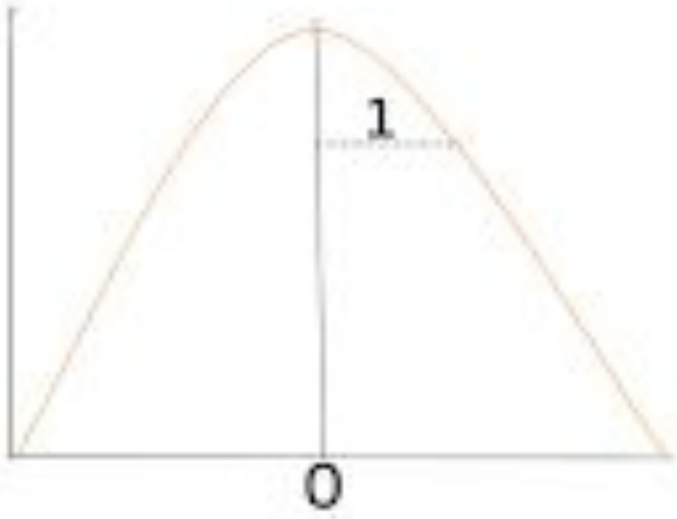
Distance between Emp2 and Emp1 = $\sqrt{(-1.51 - 0.95)^2+(-1.27 - 1.19)^2} = 3.47$

Distance between Emp2 and Emp3 = $\sqrt{(-1.07 + 1.51)^2+(-0.70 + 1.27)^2} = 0.71$

Comparison will be more significant



Normalization



Standardization

2.Feature Construction

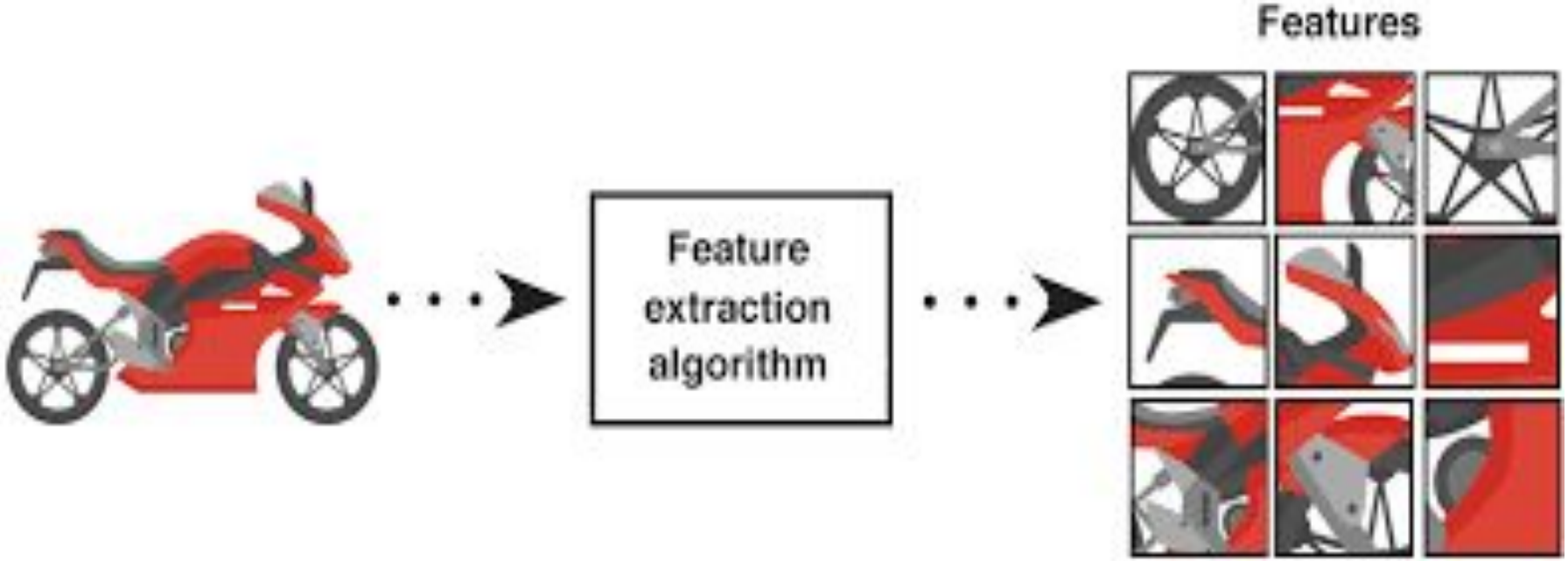
It is the process of creating new features or variables from existing data that can be used to improve the performance of machine learning model

apartment_ length	apartment_ breadth	apartment_ price
80	59	23,60,000
54	45	12,15,000
78	56	21,84,000
63	63	19,84,000
83	74	30,71,000
92	86	39,56,000

apartment_ length	apartment_ breadth	apartment_ area	apartment_ price
80	59	4,720	23,60,000
54	45	2,430	12,15,000
78	56	4,368	21,84,000
63	63	3,969	19,84,500
83	74	6,142	30,71,000
92	86	7912	39,56,000

3.Feature extraction

process in machine learning and data analysis that involves identifying and extracting relevant features from raw data



4. Feature Selection

It is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data

Name	Employee ID	No. of year experience	Previous salary	Salary
Rahul	1	2	20000	40000
Aman	34	3	30000	50000
Ritika	31	5	50000	70000

Not useful Not useful

No. of year experience	Previous salary	Salary
2	20000	40000
3	30000	50000
5	50000	70000