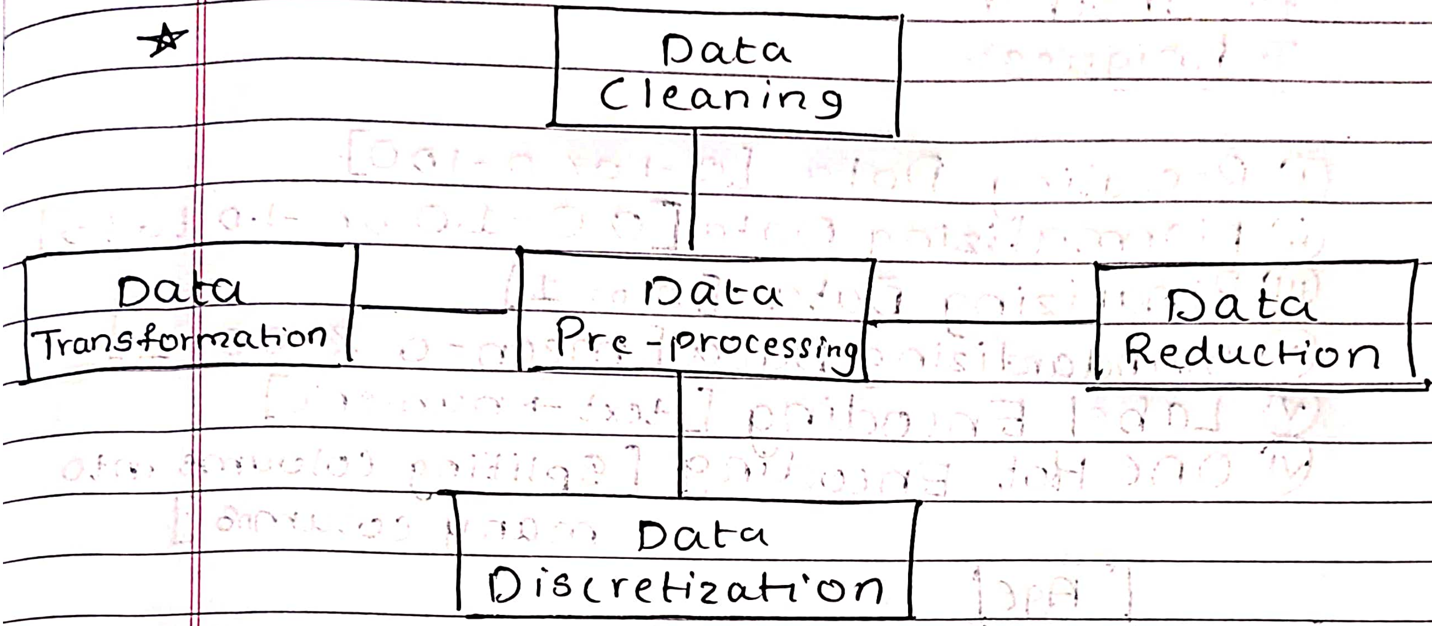
Data Preprocessing

☆ Process of transforming raw data into an understandable format.

☆



1.] Data Cleaning

(i) Handling Missing Values

(a) Replace missing values manually
(b) Replace missing values by zero
(c) Dropping or ignoring rows with missing values.
(d) Use Global constants to fill missing values.

(ii) Noisy Data

(a) Duplicate entry
(b) Multiple entries for a single Entity
(c) Nulls
(d) Huge Outliers.

2] Data Transformation

⇒ Process of converting data from one for
to another.

Techniques :-

(i) Rescaling Data [0-1 or 0-100].
(ii) Normalizing Data. [0.0-1.0 or -1.0 to
(iii) Binarizing Data [0 or 1]
(iv) Standardizing Data [mean=0 & S.D=
(v) Label Encoding [text → numeric]
(vi) One Hot Encoding [Spliting coloums in many columns]

	Age
India	34
Japan	15

⇓

	Age
0	34
1	15

Label Encoding

⇓

0	1	Age
1	0	34
0	1	15

One hot Encoding

3] Data Discretization :

⇒ method of attribute of continuous data into finite set of intervals

Age	10	20	21	50	51	60	70

⇓

Data Discretization

Age	Age	Age
10,20,21	50,51	60,70
Young	Mature	old

Page No.
Date

4] Data Reduction

⇒ reducing a large capacity of the data into small datasets.

Techniques

(i) Data cube Aggregation.
(ii) Numerosity reduction.
(iii) Data compression

★ Types of Attributes

(I) Nominal :[names or symbols]

(II) Binary [0 or 1]

(III) Ordinal [ranking among them]

(IV) Numeric [measurable quantity, represented in integer or real values]