1) Data Preprocessing:-

=> The process of transforming raw data into an Understandable format. eg:- Marks of students.
60 dm stand (a, b, c ---- z) } The data is different
(90, 91, 92 --- 100) } here.

=> Databases have noisy, missing and inconsistent data due to their huge size

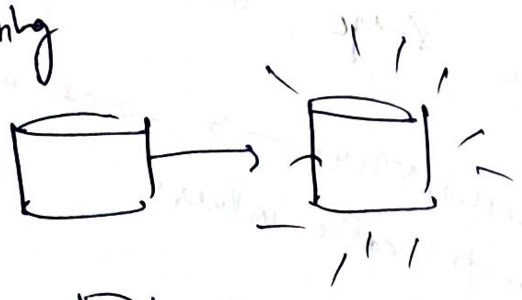
=> Low Quality data leads to low quality data-mining

=> Data pre-processing is used to improve the quality of data and Mining results.

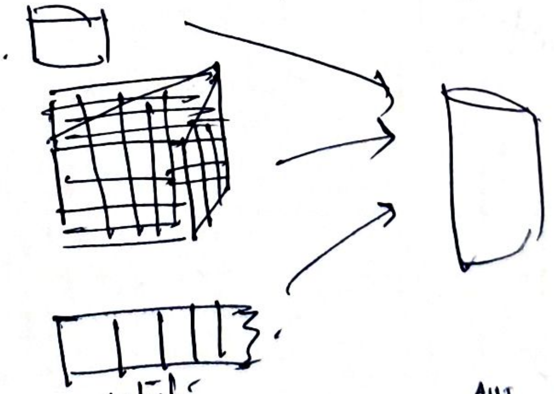
=> Various techniques like, data cleaning, data integration, data reduction, and data transformation are used in data pre-processing.

=> steps in data pre-processing

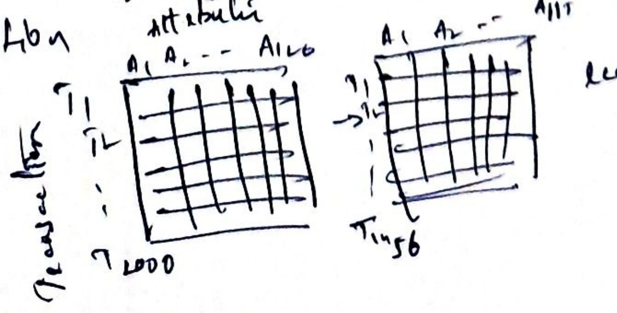
Data cleaning



Data Integration



Data reduction



reduce the data size

1) **Data cleaning** :- is applied to remove noise and Correct in data fills missing values, smooth out noise while identifying outliers

Outliers :- "class A" if the data is not fitti in to the memory other then type of data is called Outlier.

2) **Data Integretion** :- Merges data from multiple sorces into a single data source such as data ware house which helps to reduce the redundent data. [Here reduce the redundancy data]

3) **Data Reduction** :- The size of data by using aggregation clustering methods or by eliminating redyndent data.

4) **Data Transformation** :- Data is scalled to fall within a smaller range like 1.0 → 0.0

Eg: -2, 32, 100, 59, 48 ⟶ -0.02, 0.32, 1.00, 0.59, 0.48

Convert in to smaller Mattres.

ta cleaning :

Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

=> Approaches in data cleaning :

1. Missing values.
2. Noisy data.

1. Missing values :-

i) Ignore the tuple.

ii) Fill in the missing values manually

iii) Use a global constant to fill in the missing values eg (NA)

iv) Use a measure of central tendency for the attribute (e.g. the mean (or) median) to fill in the missing value.

v) Use the most probable value to fill in the missing value (eg using a decision tree)

2) Noisy data :

Noise is a random error or variance in a measured variable.

Approaches in Noisy data:

i) Binning

ii) Regression

iii) Outlier analysis.

Eg :- 6, 10, 17, 22, 22, 25, 27, 30, 36 [9 values numeric]

Partition into equal frequency bins :

Bin 1: 6, 10, 17

Bin 2: 22, 22, 25

Bin 3: 27, 30, 36.

$\frac{6+10+17}{3} = \frac{33}{3} = 11$

$= \frac{69}{3} = 23$

$= \frac{93}{3} = 3$

Smoothing by bin Mean's

Bin1 : 11, 11, 11

Bin 2 : 23, 23, 23

Bin 3: 31, 31, 31.

Smoothing by bin boundaries

Bin1 : 6, $\circled{6}$, 17

Bin2 : 22, $\circled{22}$, 25

Bin3 : 27, $\circled{27}$, 36.

ii) Regression :

=> i) Linear Regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other.

=> ii) Multiple linear Regression :- is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

iii) outlier analysis :

Outliers may be detected by clustering, for eg:- where similar values are organized into groups, or "clusters". Intuitively, values that fall outside of the set of clusters may be considered outliers.

Data Integration in data pre-processing

⇒ Merging of data collected from multiple sources. Careful integration can help reduce redundancies and inconstancies in the resulting dataset.

Approaches in Data Integration

1. Entity Identification problem.
2. Redundancy and correlation analysis.
3. Tuple Duplication.
4. Data value conflict Detection and Resolution.

⇒ 1. Entity Identification problem

1. method

Correlation coefficient :- for the Numaric data.

How two attributes are strongly relatated each other with the availability of attributes

$$r_{A,B} = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B}$$

$$= \frac{\sum_{i=1}^{n} (a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

1) $r_{A,B} > 0$, A, B are positively " correlated.

2) $r_{A,B} = 0$ independent

4 $r_{A,B} < 0$; both are negetively correlated.

A ↑ increases B ↓, like wise versa

→ Avoiding the eliminating the redundancy.

2- Correlation Analysis (Nominal Data)

* For the nominal data, a relationship between the two attributes "A, B" can be discovered by a "χ^2"

A → c distinct values, $a_1, a_2, \ldots a_c$.

B → r distinct values $b_1, b_2, \ldots b_r$.

⇒ A & B are values are shown as "Contingency table".

⇒ with the c values of A making up the Columns and the 'r' values of B making up "the rows."

⇒ Let (A_i, B_j) denote the joint event that attributes 'A' takes on value a_i and attribute B takes on value b_j.

Where $(A = a_i, B = b_j)$

$$\chi^2 \text{(chi-square)} = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\boxed{\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}}$$

⇒

Contingency table

	Male	Female	Sum (row)
Like Science fiction	250(90)	200(360)	450
Not like it	50(210)	1000(840)	1050
Sum(col)	300	1200	1500

y using Eq (3.2) we can verify the expected frequency for each cell

Eg:- expected frequency for the cell (male, fiction) is

$$e_{11} = \frac{\text{Count (male)} \times \text{Count (fiction)}}{n}$$

$$= \frac{300 \times 450}{1500} = 90.$$ 4 cells values

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93.$$

degree of freedom are $(2-1)(2-1) = 1$

⇒ Covariance of Numeric data :-

* Measures how much two numeric attributes A & B change together

* for n observations (a_1,b_1), (a_2,b_2) --- (a_n,b_n) the covariance between A & B is (−1 to +)

Where $\bar{A}$ & $\bar{B}$ are the mean values of A & B.

$$Cov(A,B) = E(A-\bar{A})(B-\bar{B})$$

$$= \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n} = \sigma_A \sigma_B.$$

⇒ The co-relation coefficient $r_{A,B}$ is related to covariance by

$$\boxed{r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}}$$

→ positive covariance :- if A & B tend to change together in the same direction (both increase or decrease), their covariance is positive.

→ Negative Covariance :- if A&B tends to change in opposite direction (one ↑ whilee the other decreases). their covariance is negative.

→ Zero Covariance :- if A & B are independent, their covariance is zero.

All electronics $\rightarrow$ 6, 5, 4, 3, 2

High Tech $\rightarrow$ (20, 10, 14, 5, 5)

Mean value = $\bar{A} = \frac{6+5+4+3+2}{5} = 4$

Mean value $\bar{B} = \frac{20+10+14+5+5}{5} = 10.8$.

Covariance:

$\text{Cov}(A,B) = \frac{\sum_{i=1}^{n} (a_i-\bar{A})(b_i-\bar{B})}{n} \rightarrow \sigma_A \sigma_B$

$= \frac{(6-4)(20-10.8)+(5-4)(10-10.8)+(4-4)(14-10.8)+(3-4)(5-10.8)+(2-4)(5+10.8)}{5}$ No. of values (-1 to 1)

$= \underline{\underline{7}}$

The positive co-variance indicates that the stock prices for both compencies tend to increase together.

=> Data Reduction in data preprocessing :-

* Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume.

=> Mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

=> Methods of data Reduction :-

1. Dimensionality reduction
2. Numerosity reduction
3. Data Compression.

1. "Dimensionality reduction":-

=> D.R is the process of reducing the number of random variables or attributes under consideration.

=> "It eliminates the redundant attributes" which are weakly important across the data.

① att ② attribute

Eg: DOB age

→ present is fixed to stay.

→ redundant attribute here age

↓

we are eliminating the age here so it is called as redundant attribute.

① Stepwise forward selection.

② stepwise backward elimination.

③ Decision tree induction.

stepwise forward selection.

i) Eg:- Initial attribute set : {A, B, C, D, E, F}

initial reduce set { } ✓ emptyset.

{A, d, E} relevant → not irrelevant attributes here

considering redundant attribute and Eliminating here.

{C} most relevant attribute in the dataset.

{C, E} next relevant attribute set in the list.

final reduced set :→ {C, F, B}

2) Here we have to calculate the redundent element in the list and ~~[illegible]~~ elimenate the attribute in the entire data set.

ii) stepwise backward Elimination :

Eg: Initial attribute set {A, B, C, D, E, F) ⟶

Here we are deleting the attributes.

Initial reduced set {A, B, C, D, E, F}

↳ reverse procedure

↳ {F is the irrelevant attributes to eliminating the attribute here}

→ {A, B, C, D, E}

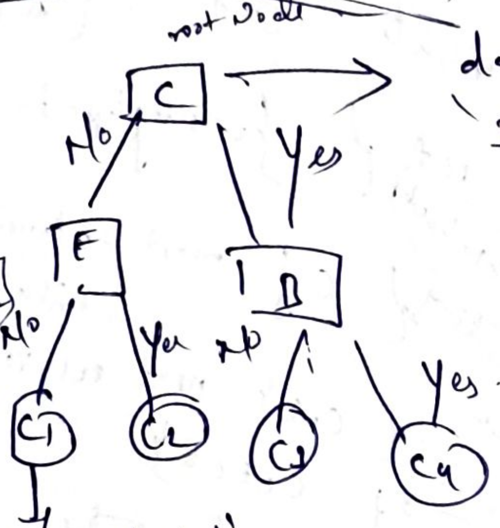
→ {A, B, C, D, E} ⟹ Elimate here

{A, B, C, E}

| final reduced set {C, F, B} |

Here we are Eliminating the {F, D, A} these are the redundent attribute and finally Eliminating here.

iii) <u>Decision tree Induction</u> :- This is classification concept.



→ dataset: '3' attributes are present here {C, E, B}

{A, B, C, D, E, F} → set }
{C, E, B} reduced set }
→ We are eliminating the remaining things.

class label's → final Outcome's.

2. "Numerosity Reduction" :

⇒ Replaces the original data with small form of data representation. There are two methods Parametric and Non-Parametric reduction.

1. parametric method:- Used to estimate the data, so that only parameters of data are required to be stored, instead of actual data.

a) Regression :- Simple linear regression (to fix in a stright line) ($y = ax+b$)

multiple linear regression [with 2 or more predictor variables]

b) ~~Long~~

Long — Linear model :

Used to estimate the probability of each data point in a multidimensional space for a set of discretized attributed, based on a smaller subset of dimensional combinations.

$\boxed{\log(y) = ax+b}$ → This allows a higher-dimensional data space to be constructed from lower-dimensional attribute.

2. ~~Non-parametric~~ Method
2. Non-parametric :- Used to store reduced representation of the data. It includes.

a) Histograms.

b) Clustering.

c) Sampling

d) Data cube aggregation.

3. Data Compression :-

Reduce the size of the files using different "Encoding mechanisms". There are ② types.

i) Lossless compression :- without loss after compression.

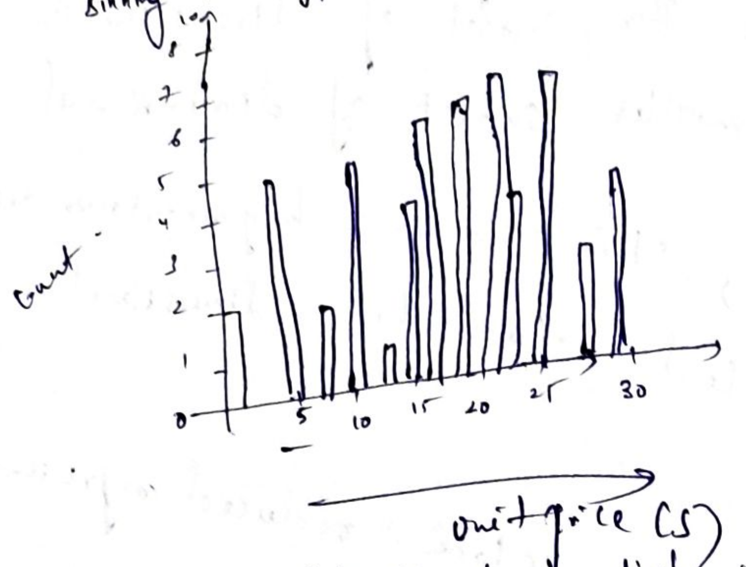
ii) Lossy compression :- The decompressed data may differ to the original data but are useful enough to retrieve information from them. They are.

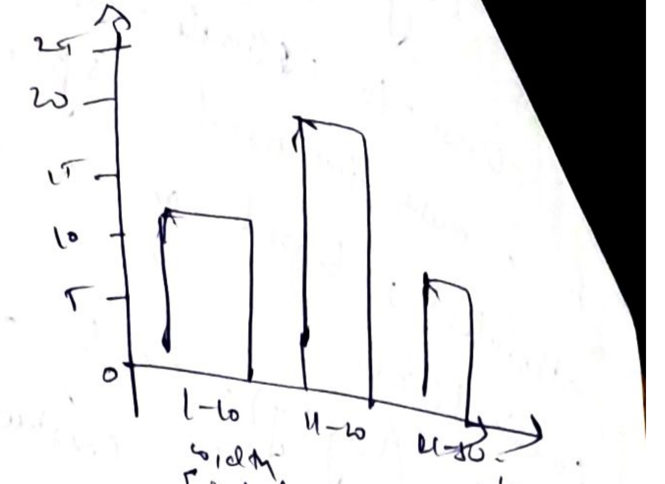
ⓐ Discrete Wavelet Transforms.

ⓑ Principal Component Analysis.

⇒ a) <u>Histograms</u> :-

Binning to approximate data distributions.





[include 2R, 10R, TR → calc Mean Value and represent in data]

→ Based on the data we have display in the graph manner.

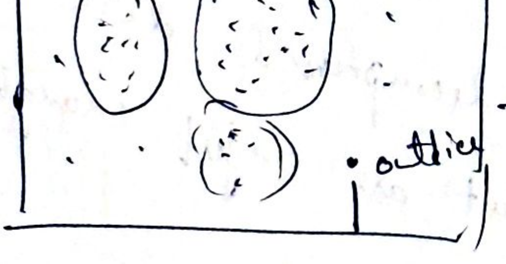
→ Before we construct the Equal width in the above histogram.

→ {0, 5, 10, 15, 20, 25, 30 → we have equal unit prices] [0-30]
unit price → Based on Equal width we have to dividing the data distributions

→ Equal frequency potal ~~ro~~ three, 1st price items, however!
2nd price items
3rd price items } frequency based.

b) <u>Clustering</u> :-

Partitions the whole data into different clusters, Centroid distance is an alternative measure of cluster quality and is defined as the average distance of each cluster object from the cluster centroid.



→ when we clustering the Object some data is missing here at that time [ignoring the unimportant data.

→ [group] → outlier elimination / clustering value data reduce arithmetic