- A bath Pre- processing
ble formal. Real - world dotte is often incomplete, inconsistent
is proven method toresdie such issues. Date pre-processing
Data processing have  Leany Data  Propocessing Data  Data processing Have  Data processing is important  to achive accuracy (whether data entered is  correct or not).
Tim liness  Ontopology - Completness (-11- duta is available or not recorded)  Accuracy - Timeliness (duta should be updated correctly)  - Believlibling (-11 - be trustable)  Totrepretablity (understablity of duta)
- Data Cleaning
it also replaces the incomplete date inaccurate data from incomplete date.
it also smooth out the noise while identifying outliers.  Cornect the inconsistency.
- Handling Missing Values:  Filling manually.  Typore the row (taple) - time consuming
Tynore the row (tuple)  - used when class label or  Primary Key is missing  Lue glubal constant
Typle Contain several attribute constant like I unknown or 100.
- Not available or Na'  Can be used to replaced the  missing values.

	in a column of the column of t
Central tendancy  - use of mean or median  to fill the missing value	mobable
central tendancy	the most programme
- we of mean or and some to make	par in the
to fill the missing value	letermined with regress
dor normal symmetric data distribution mean is used.	Interence boue tooks
distribution mean is used	using Bayesian tomum
and a contract of the contract	or decesion tree.
40x sicensed - An-ORA motivation to the west	•
Moise of the books of the day of the Hall	
Collection Francisco	, 
- 13) value, means random error or contain	ing unecessary duta point
Binning is This mothed to the smooth first data is sorted and Seperated and stored	ar hundle notsy duta
first data is sorted an	d then sorted value are
Seperated and stored	in the form of bird
3 method of Smoothip	elata ·
Smoothing by bin Smoothing by big some	Smoothing by bis
Mean method Smoothing by bin median method	boundary
by mean. Teplaced by medecin.	- more e min
by mean. by medecin.	udlue ove tecken
	by closest bounday
nor parametric inchance.	value.
(2) Kegression is used to consult that	h. Halida kan
D Regression _ used to smooth data 2 unnecessary data is pres	nanque adra When
for analysis pure	re represente telos la
decide the yanable	11 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Clustering  Clustering  Clustering  Clustering  Clustering  Clustering  Clecial the Yanable  Grouping the data  Assed in unsupervised	Stag only by
-> rused for Junding #	re outliers 2 also in
Douping the data.	23 C) \ (+1°)
→ usedin unsupervised	d learning.
જાલમાં લેવિક કર્યા છે.	· · · · · · · · · · · · · · · · · · ·
	19 63 90.
Douter Reduction, helps to reduction of volume make analysis easier and possing result	of the dota which
malce analysis easier and p	roduces almost the
-same result:	- 6 Gola Trongon other
Same result:  thelps to reduce literage de	ace
3 techniques	
of paner arm of girmer that Dimensionality red	900dy
the second constant of the second	ron!
Numerosity red	
Las Data compan.	
Most in the properties of the Action of the Contraction of the Contrac	1840) 1844) 6 Organis 1
property control of the control of	

The state of the state of	. 0	Landing in the "Mark
Dimensionality redur.		Linday Linday and
	and a death the part	variables (random) of
anogoniza Afrika Iraina maga	well of reducing the 19	- 20
Company of the Compan	Hisburtes lunder consideral	W.A. T.
30	me Melhada allana	Luman Don Str. Tolk
War and the state of the state	weld franctions. (Sine w	ave to ousread full)
· 3214 Chrosp of	Sign Trought M. Chosin	(0)
' Pn	hapal Component ( rect	or red
	Analysis along	ni en Uran .
• A	Hribute Subset (remov	in irreluent or
( E and disk commence to	colleun red	undant attributes)
A.  where the constant following the points.	and tomorms enow or sont	on mod tolog to
Date red Strategies		•
Date red strategies	Numerosity reduction :-	स्तानगढ है।
Rid b harot week in	Tirs days was sense	1 · 1 dula ha olla so
	process of removing or	ginal acuta by ulternatur
	for smallerdam of date	i tepr".
id ed filtres &	O Lune Mindell San 2	and red privilence.
jirkkadad	2+pe Parametri i method	button moon
(NIS) A MOSS	parament inches	when are stored
(NO 3 KDM C.	only a and paran	A. Incomplete Model
Korter & cod to the	- ionly data paran	or of many to
publicate til solate til	non-Parametric method.	
value.	HON- I WICHNOTTE THE TEST	n a dab
	- Stores reduce re	presi gradia
B CALLER PORT & MARCH F	, il aller prossessioner,	clustering & Samply
Data Compression : 1222200	, il obah grossoons	
		0.0000
of the original	Haine a reduced or co	inpressed sepretutation
of the onglice	y coola.	t (2) * -
Veconstructiv	original data from compri	essed data without
any loss of	request sibors as tog	olles.
43.44.5.44.53.44.6	Water and American	
	-n	with some
for possof	o" ·	220%
at the state of the second		The state of the s
the board of the form	helps to reduction of volumence and	** ** ** ** ** ** ** ** ** ** ** ** **
Data Transformation	· House grand :	and data & Ichnim
	hade in the domest of structus	4 of about 17 lovems
as Dut	a Transfermation.	
Data is	transformed so that min	is process may be
more e	Micont & Dalloma en and	more he as the la
	llivent 2 patterns tound	THE CHANGE TO
Chaulania.	"Towns what comp"	
Lemma		Aladkda
- smoothing, Attribute	Aggregation, Norm	Wration, VISUCTION
Constructi	0//	

Smoothing  $(\mathfrak{D})$ All ibute Construction Normalization > used to remove noise , attribute date also known cus from data are scaled so decitive Construction technique isclude cuits smaller where new collab binning, regression, are constructed & range & clustening. such as added from given seta attribute. -1.0,000 1.0 Aggregation and with Discretization - Where Summony raw value of a or aggregation numeric ayribute are oprn is applied replaced by interval late. + Used in Constructing Discretitation by Binning ex!data cube for analysis of deuta histogram Analysis at multiple level of - by cluster, Decesion abstraction. Troe and Correlation Analysis. average of data. (summing up all the no. I then duindy Etidologia Siego (S. Sille = 10+20+30+40+50 = (30) year. move incorest again . Indicated on 160 -10/20130, 40/2001 A. Proceed to a Median e white whether entires e in -no gobs? 1 (even) 17 (odd) Lo, 20, 30,40,50 10,20,30,40,50,60 11 = 51 = 3=> Median = 30 A= 3 → B0,40 -> 30+40 Mcdin=35 1,3,4,6,7,3,3,5,10,3 10 19 fol szála mandiai b Mode = 3 A highest trequency y Primary Provided on 48,52,57,61,64,7,2,76,7781,8588 IQR 93-9, (first (second) 72/36/33/8/185/88 91 =median of = 57+6) Pr= media 1 27781 1st half School half

Marcinum Value - Minimum Value 32,41, 28,54,35, 26, 23,33,38,40 54 - 23 = 31 Roye Standor deviation devication from mean value. n: no g papul?  $\sigma = \int \frac{\sum (x_i - u)^2}{n}$ M= mean = x1+x2+x3+--en Variance : usquared deviated  $Q_{5} = \frac{\sum (x - y_{0})_{5}}{V}$ Popl" voriance .. U= mean  $\frac{\sigma^2}{2} = \frac{\mathcal{E}(\mathbf{x} \cdot \mathbf{x})^2}{2}$ Sample varion X = averge. Box Act - standonized way of displaying the distribution of data based on five no. of summary ("minimum" fixt-quartile (91) -0 91 Can tell you about Your outlier 2 median third quartite (83) where the value. maximum y) -) also tell about your data is symmetrial, sicewed, queuped. outli cas IGR outliers 00 11 Min 4 "MOX " Median Q1-15+ IAR) (25th (South) 7 tell (Q3+1.5 \* IAR) Percentik) perentile) - 2 -1