UNIT - 1

Data - collection of information
Row - horizontal Representation of data
Column - vertical Representation of Data

Q. What is Data Science ?

Data Science
- Knowledge
- Detection
- solving
- Big Data
- classification
- Statistics
- Analyzing

→ understand patterns, trends, correlations with in data
→ make predictions
→ Recomendation based on data
→ make informed decisions
→ solve Real world Problems

(scientific words yhi use honge)

Data science will able to make
→ Better decision
→ what will happen next
→ pattern finding

Data science multidisciplinary field hoti hai jo Data pr scientific methods use karke process karti hai algorithm & system ki help se Extract karti hai, knowledge & insights structured / unstructured data se

Applications of Data Science

(1) Healthcare - (a) pattern dekh ke pta lagate hai ki future mein konsi disease fastly spread hogi.
(b) optimization ke sath treat karti hai patient ko like past data set ki help se deka jata hai yeh syptoms konse disease ke the ya konsi disease us person ko past mein hui thi.

(2) Finance - (a) stock market mein kya badhega kya ghatega
(b) Fraud detection karta hai.

(3) E-commerce - (a) past purchasing/Interest Data dekh ke Recommend Karta hai
(b) product demand batata in future increase hoga ya decrease.

(4) Marketing - (a) social media se customer ka interest dikhata hai.

(5) Education - (a) Decision making sikhata hai.

(6) Human Resources - (a) Forcast Karta hai Business mein aage Kya hoga.

Data Scientist

- SKILLS NEEDED

→ Solid Back-ground in mathematics, statistics, programming & Domain Specific Knowledge.
↳ (python, R) (technical)
→ Data Related task ane chaiee like cleaning, modelling machine learning algo Kaise apply karte hai. (technical)
→ communication. (Non technical)
→ Business understanding (Non technical)
→ critical thinking approach. (Non technical)
→ Machine learning & AI (technical)
→ technical (team work/ collaboration)

Roles & Responsibilities of Data scientist -

Management -

Supporting the planning & Executions

Analytics

Create models for analysics problem.

Strategy

Design strategy by understanding customer's trends & behavior.

Collaboration -

Collaboration with people like stake holders.

Provide Business solution using Data Related techniques.

Create AI tools to utilize data.

Recommend ways to improve Data Reliability Efficiency Quality.

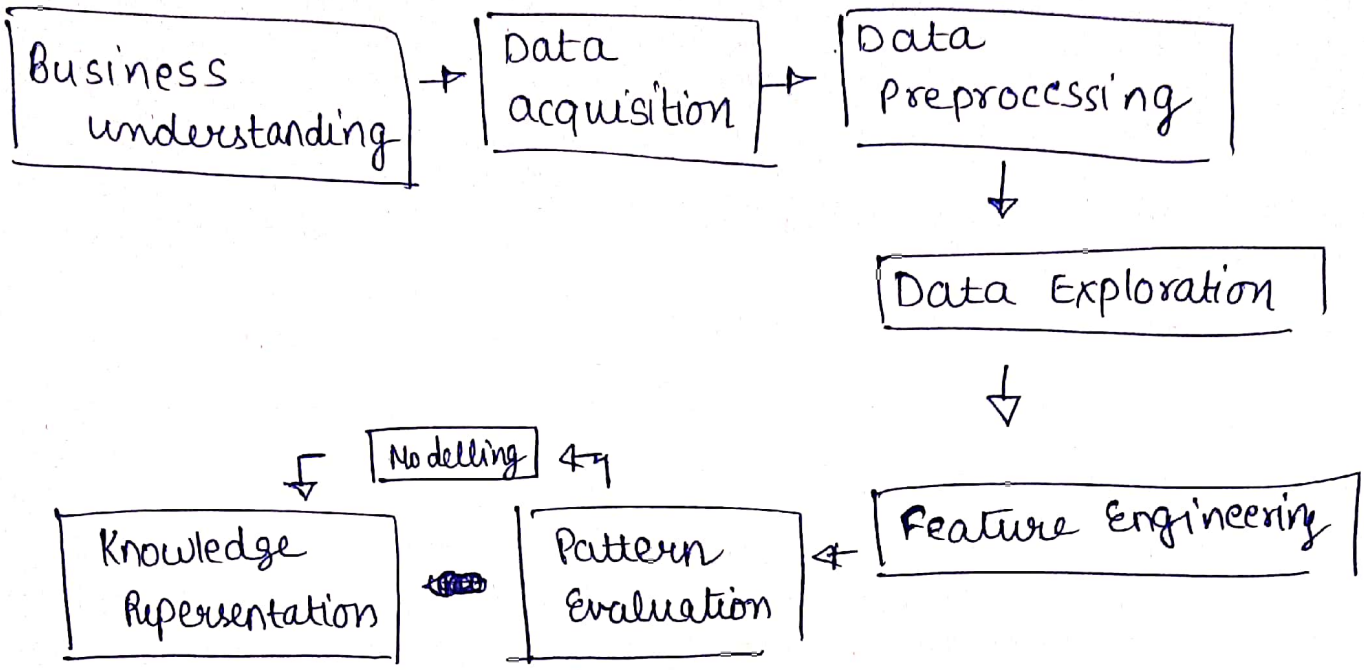
Automate work through the use of predictive analysis

Explore & Examine data find hidden patterns.

Why python used in Data Science –

(1) interpreted language, code is executed immediatly.
(2) flexible programming language.
(3) Easy & free to use.
(4) have large community support.
(5) Extensive supports of libraries.
[Numpy, scipy, pandas, Matplotlib & Seaborn]
(6) Built in Mathematical functions.
(7) versatility - allow Data scientist to perform a wide Range of tasks.
(8) Data handling - cleaning, manipulation & haandling missing values mein help Karti hai.
(9) Integration - Semlesly integrate Karti hai other data Science tools & technologies Ke sath or work Karti hai.
(10) Indentation.
(11) Portable
(12) Straight forward syntax.

Life cycle of Data science



Business understanding - Problem ko define karo, (statement)
Ki problem mein karna kya hai, kis cheez ka solution nikalna hai and uske lie kya Data set needed hai.

Data acquisition - Data ko collect karo jo problem se Related hai, or yeh Data boht sare sources se atta hai like ko .pdf, .image jpeg/ jpg, esv etc. So, in sb ko ek hi format mein convert karna.

Data preprocessing - Data Preprocessing ek process hai, jisme Apne kaam se Related Data ko lete hai or Baaki sb ko Remove kar Dete hai

Data Exploration -
is phase mein Data ko Explore karte hai, Kithe Rows, Kitne columns, Kya values, Mead, median, Variance & all Kitna hai.

Feature Engineering - ~~Python~~ tools Ki help se missing values and Errors ko treat Karte hai jisse jo predections ho vo theek aye.

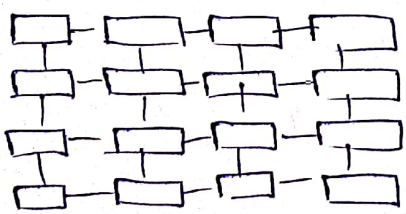
Pattern Evaluations - pattern Evaluation mein hum apne Result Ko Evaluate Karte hai, ki jo predictions ki hai vo Shi hai ki nhi, nhi hota to vapis se Data Ko process Karte hai or Agr Result Shi hota hai to next setep follow Karte hai.

Modelling - Apne predections ko model pe apply Karte hai.

~~Pattern~~ Deployment - Model present Karenge shareholders ko or feedback lenge. jo Kami hongi vo shi Karenge or market mein model Ko ~~Hunch~~ launch Karenge.

Structured Data

- Data follows - well defined, and organized structure, yeh store hota hai databases mein tables & Rows ki form mein.
- organized hota hai Rows & Columns ki form mein.
- tabular formats of Relational databases ki form mein hota hai.
- less flexible
- Efficient for analytical Processing.
- Business transactions, Financial Records.
- Data nature - Quantative
- Storage - Requires less storage
- analysis - Easy
- Graphical Representation.



unstructured Data

Data lack karta hai predefined model or structure yeh kai baar images, videos ki form mein hota hai.

Koi predefined structure nhi hota.

Stored in various formats

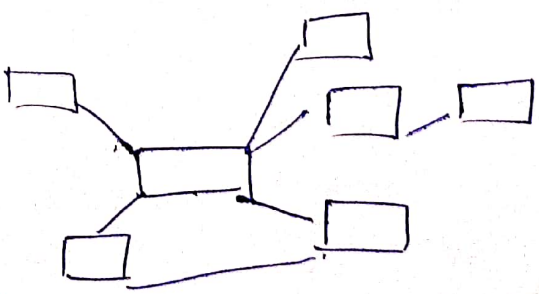
More flexible

processing may Require advanced technique like natural language processing.

Speech & image Recognition

Qualitative.

Storage - Requires More storage different.



techniques to process unstructure Data -

① Tokenization - Breaking down → tex ko break karna small words or phrases main.

② part-of-speech tagging - Assaigning Gramatical categories (noun, verb, adjective) noun, verb, adjective add kardo sentence mein.

③ Sentiment analysis - Determine karna anyli's karna tone ka, tone kaisi hai (Positive, Negetive ya Neutral).

④ topic modelling - identify karna themes or topics ka jo document ke hote hai { like Related ki's cheez se hai }

⑤ Text classification - text ko categories mein divide karna
For EXP- Email < spam / not spam

Challenges in processing unstructured data -

1. Lack of structure - organized nhi hota isliye analyze karne main problem hoti hai.

2. volume & variety - Kai baar data kaafi Badi Quantity mein hota hai, isliye or kaafi different formats mein bhi hota hai isliye problems aati hai analyze karne mein.

3. Data noice - irrelevent information/mistakes (in the form of typos, abbreviations or irrelvent details)

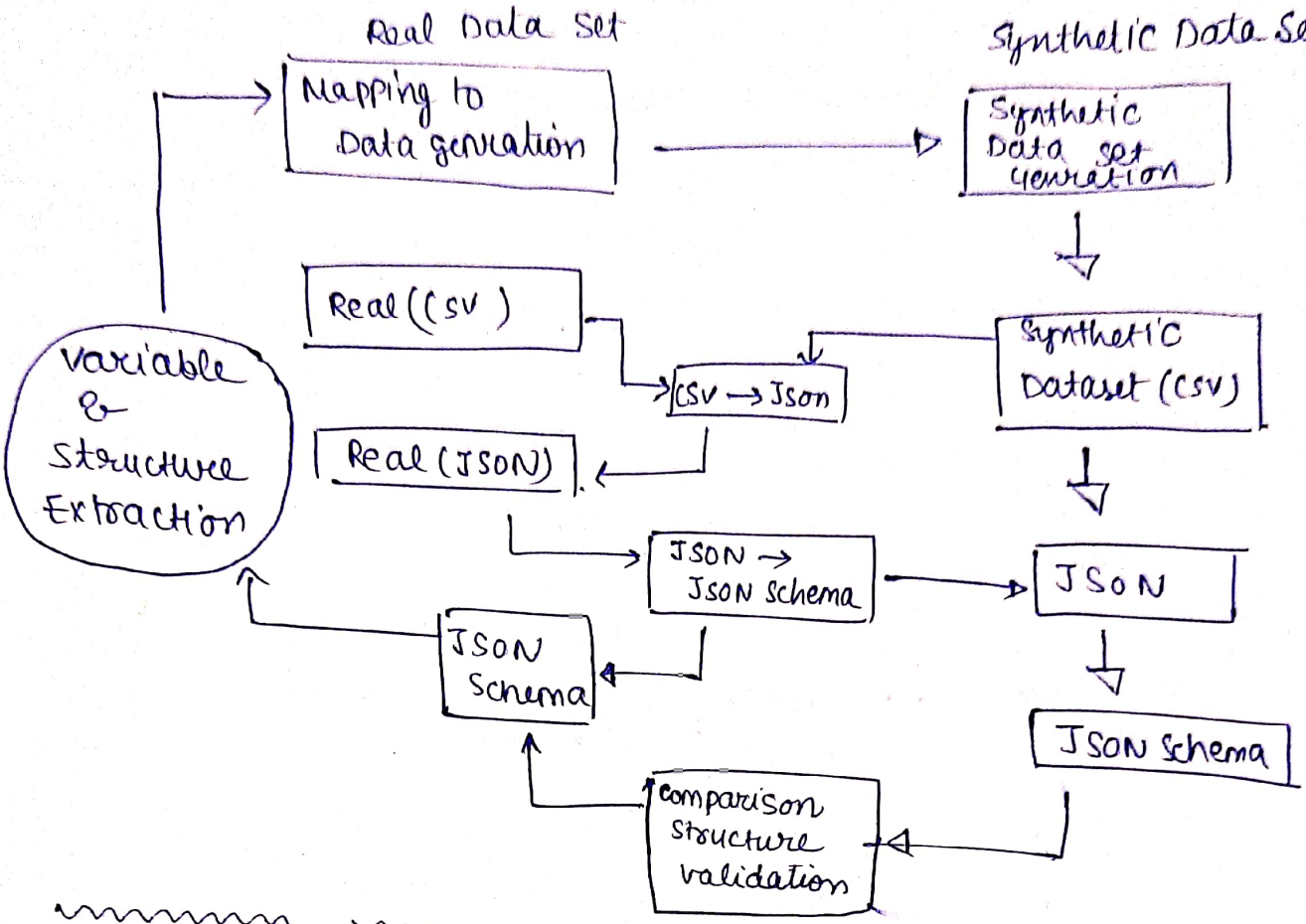
Data integration - storage methods alag hote hai structured or unstructured ke.

Scalability - Scalable solutions chaiee hote hai data ko process karne ke lie

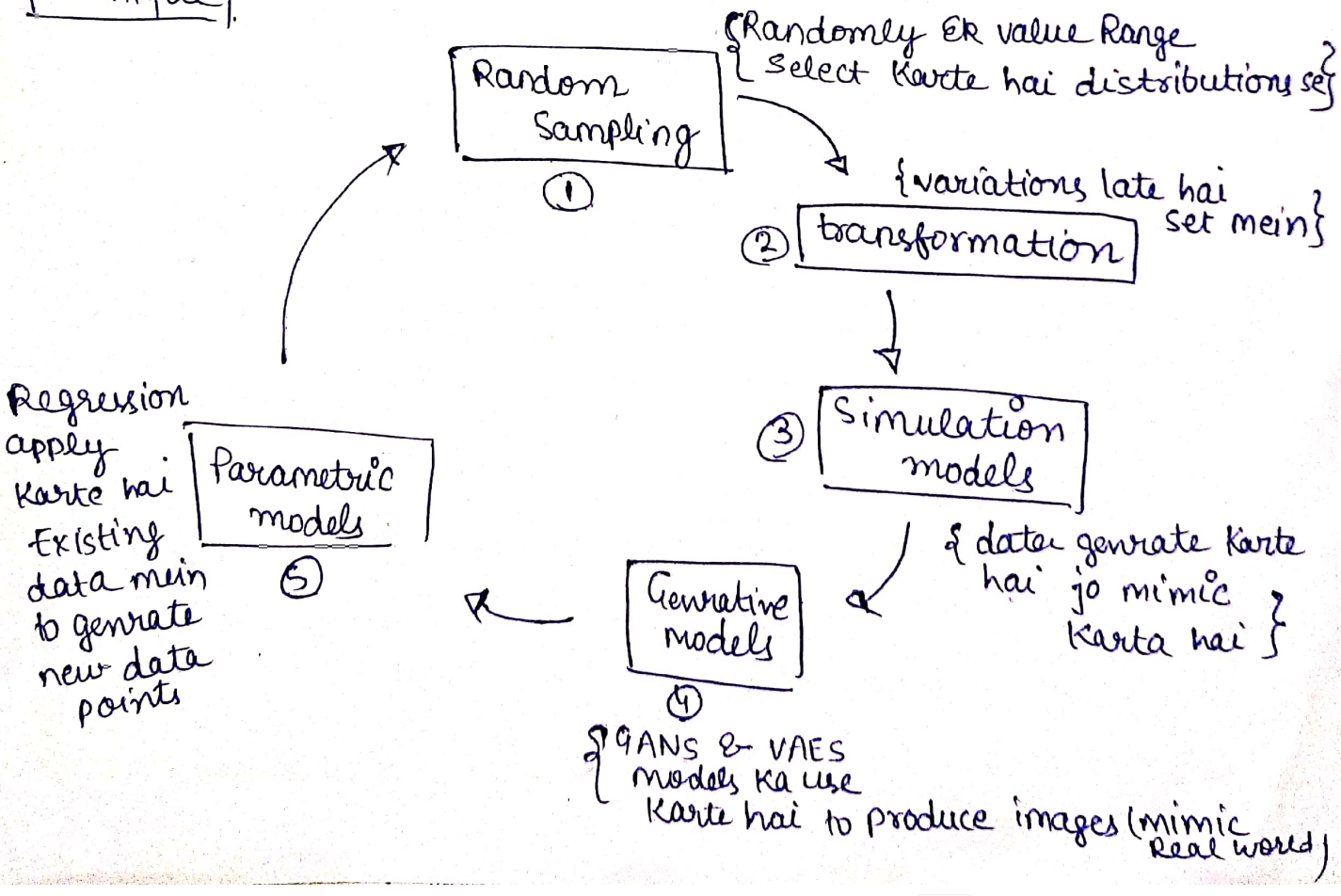
Privacy concerns - data sensitive information carry karta hai, so, Privacy ka bhi dhyan Rakhna padta hai.

continous Evaluation - unstructured data kaafi Quanity mein genrate hota hai, so, tools chaiee hote hai large unstructured ko process karne ke lie.

CSV → JSON Synthetic Data Set



Technique



Privacy in Synthetic Data generation –

⇒ Synthetic Data generation involve karta hai creating artificial Data jo mimic karta hai Real Data set ko.

① **Ethical considerations** – Respecting individual privacy is an Ethical Responsibility. It should not compromise the privacy or confidentiality of individuals represented in orignal Data set.

② **Building trust** – maintaining privacy in synthetic Data generation fosterst among stake holders, Encourage karta hai unko participate karne ke lie. data share karne ke lie bina preshan (concern) hue.

③ **Mitigating Risks** – Privacy volitation hoga, Synthetic Data main sensitive information provide karana jiske lie jail bhi ho skti hai.

④ **Protection Against threats** – koi bhi kisi bhi information ka misuse na kare islie synthetic Data avoid karta. actual sensitive information ko share karna

⑤ **Preserving long term Data utility** – Orignal Data Set outdate ho jate hai Regulation, technology or Data Quality issues ki wjh se. jabki privacy controlled synthetic Data kaafi time tk sustainable Rehte hai.

It can be done by – {mask like jaise fase mask se face chupate hai vo wale}

① **Anonymization** – mask kar dete hai personal information ko

② **Perturbation** – introduction of small/controlled changes or disturbances in data set jinse Data set ki Prediction affect nhi hoti.

③ **Data Generalization** – Aggregate karna / generalize karna Data ko hide karne ke lie fine grained details ko.

④ **Differential privacy** – noise add karna records mein re-identification ko Prevent karne ke lie.

UNIT-2

Python Basics

* **constants** - Fixed values like numbers, strings

```
Print(82)
   ↳ constant
Print('I love python')
```

* **Reserved words** - Predefined words jo as a variable use nhi kie ja skte, Ex - true, false, finally, for, try, except, import etc.

* **variables** - define karta hai information ko memory mein jo as a fetch Baad me use ki ja skti hai

```
a = 23
b = 'Python is Great!!'
```

* **Assignment** - assigning value to valiable

```
a = at 5.
```

* **Numerical Expression** -

+	→ Addition
-	→ subtraction
*	→ multiplication
/	→ Division
**	→ Exponentiation
%	→ Remainder.

for solving precedence

- Parenthesis
- Exponentation.
- Remainder, multiply, divide.
- Addition, subtraction
- Left to Right

* **Boolean & comparison** - have two possible values (T/F)

comparison operator

Less than	<
less than Equal to	<=
Greater than	>
Greater than or Equal to	>=
Equal to	==
Not Equal to	!=

* **Strings** - Created by using single or double Quotes.
string_1 = "I am valid!".

concatenate using + operation.
length can be find by len(string_1)
.title make the Every 1st word capital.

* **Type** - Integer
- Float
- String
- Boolean

* **Functions** - can write code in blocks, It makes code more Readable. & Easy to Reuse.

```
amount = hours * Pay_per_hour
```

Hum isko multiple locations pr use karna chante hai to isko copy paste kar skte hai
but Ek din amount mein 500 add karna tha.
to formula

```
amount = hours * Pay_per_hour + 500
```

ab code mein yeh line dhund ke Edit karni padgi
islie yeh sb baar na karna pade,
function introduce kia.

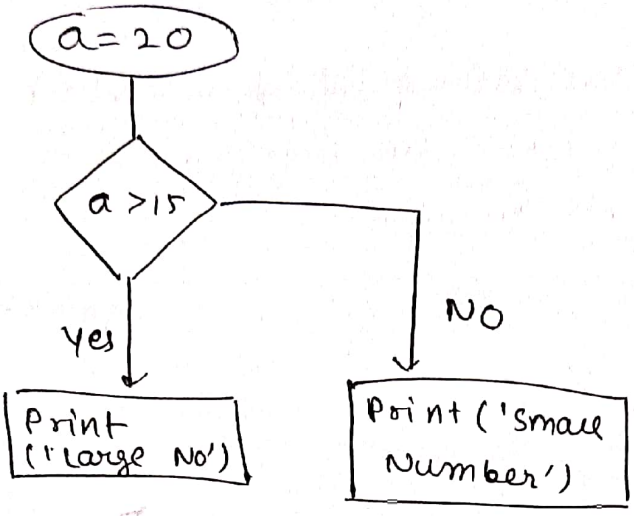
```
def calculate_pay(hours, Pay_per_hour):
    amount = hours * Pay_per_hour
    return amount
```

→ used to define a function
→ values passed to function
→ to return a value
indentation

now jab bhi hume is formula ki zarrot hogi to call karenge function ko.

```
calculate_pay(20, 120)
```

Conditional statement -



Syntax

```
if <condition>:
    do - something
```

conditional Expression → True hai to aage badho execute karo varna skip karo

4 tab spaces

if Else

```
Price = 200
if Price<150:
    Print ('Buy !!')
Else
    Print ('sell !!')
```

if Elif - Else

```
Price=200
if price < 150 :
    Print ("Buy")
Elif price >=150 and price < 250 .
    Print ("Hold")
Else :
    Print ("sell").
```

and → dono statement ko check karta hai "dono true hone chaiee"

or → 1 bhi true hoti hai to Statement Execute hoti hai

not → If koi statement true hogi to False bolenge or False hogi to true.

Data structures & loops

List - Data type.
Values same type ki honi chaiee
[ ] → mein elements likhte hai

function in list

max(list) → maximum element deta hai [Print karta hai]
min(list) → minimum element
sorted(list) → sorted copy deta hai
list.append(element) → Element ko last mein insert karta hai
list.clear() - Delete karta hai sare elements ko.
list.count(element) → Kitne elements hai list mein.
↓
baar
list.Remove(Element) → Remove karta hai element ko
list.Reverse() → Reverse karta hai list ko.
list.pop → last element delete karta hai.

Sets - collection of unique element in no order.
(iska order nhi hota)

Dictonaries - Key value pairs hoti hai python mein

Tuples - Odered collection of items.
immutable

Loops

FOR loop –

```
for i in Range(1,100): ← colon
    Print i
    ↑ indented
```

for → Keyword declare the loop

i in Range(1,100) → condition for iteration/Repetition.

while loop – yeh to tab stop hota hai jab condition false hoti hai

```
i = 1
while i<5:
    Print(i*i)
    i = i+1
    ↑ indented
```

while → keyword

Break – used to come out of loop

```
i = 1
while i<5:
    Print(i*i)
    i = i+1
    if (i==3):
        break
```

sirf 3 pr break hogi statement baaki time chalegi.

continue – used to skip rest of the loop & transfer the controls to the top of the loop.

```
i = 0
while i<5:
    i = i+1
    if (i==3):
        continue
    Print(i*i)
```

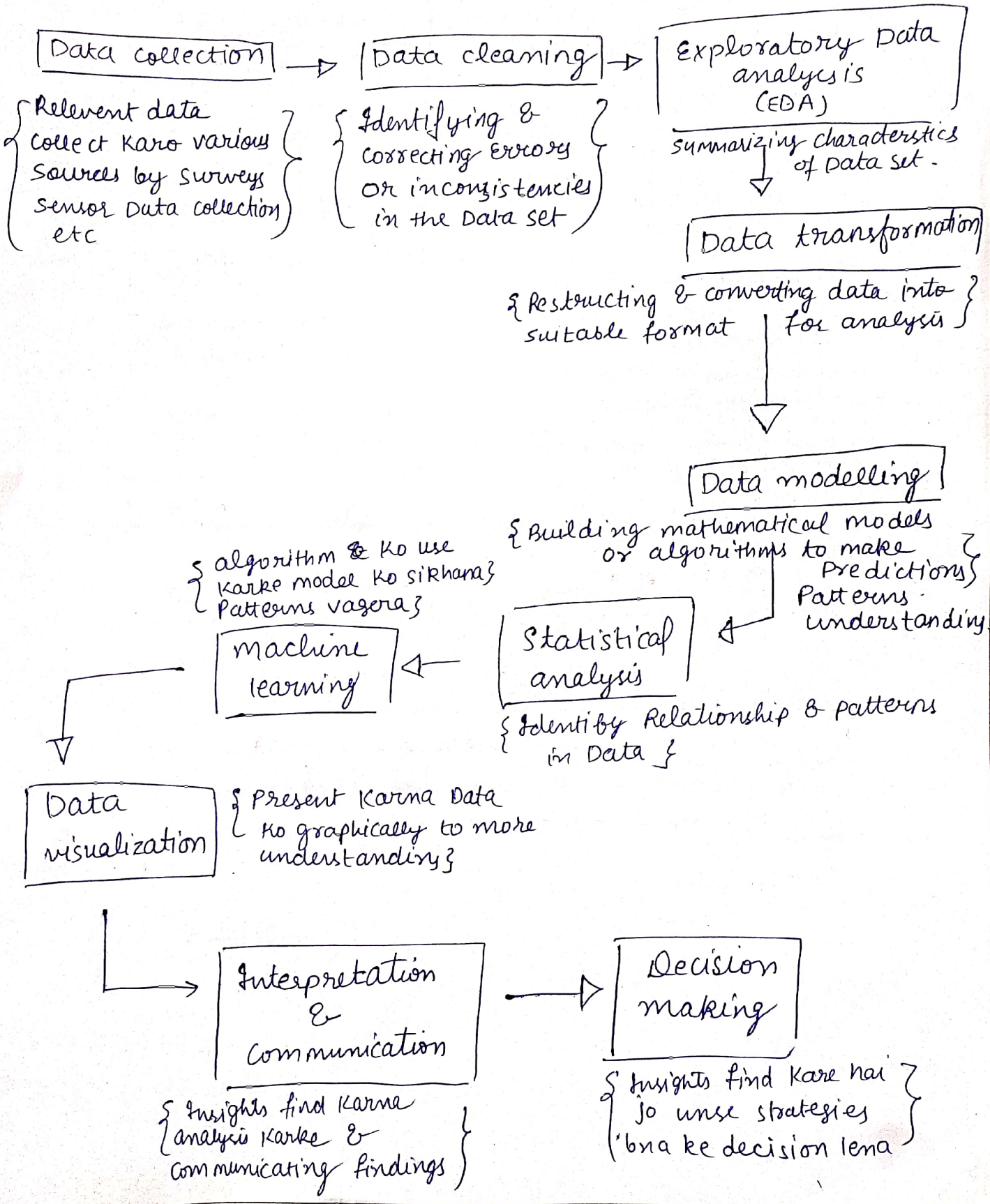
wapis yaha ajayega

i==3 ke lie Print(i*i) wali statement skip hogi

Data wrangling - process of Removing Errors & combing data sets to make them more accessible & Easier to analyze

Data analysis -

Process of inspecting, clecning, transforming & Modelling data with the Goal discovering useful information, drawing conclusions & supporting decision making.



Data importing –

↳ Process jisme External files / data ko system mein dalte hai or analysis, manipulation, interpretation karte hai

↳ Fundamental process hota hai

↳ Done by python libraries like "Pandas" ← data frame (work for 2D info) / (work for 1D info series)

```
import pandas
csv_file = Pandas.read_csv (file name)
type(csv_file).
```

If file is Excel file

```
Exl_file = Pandas.read_Excel (filename)
```

If file is JSON File

```
Json_file = Pandas.read_Json(filename
```

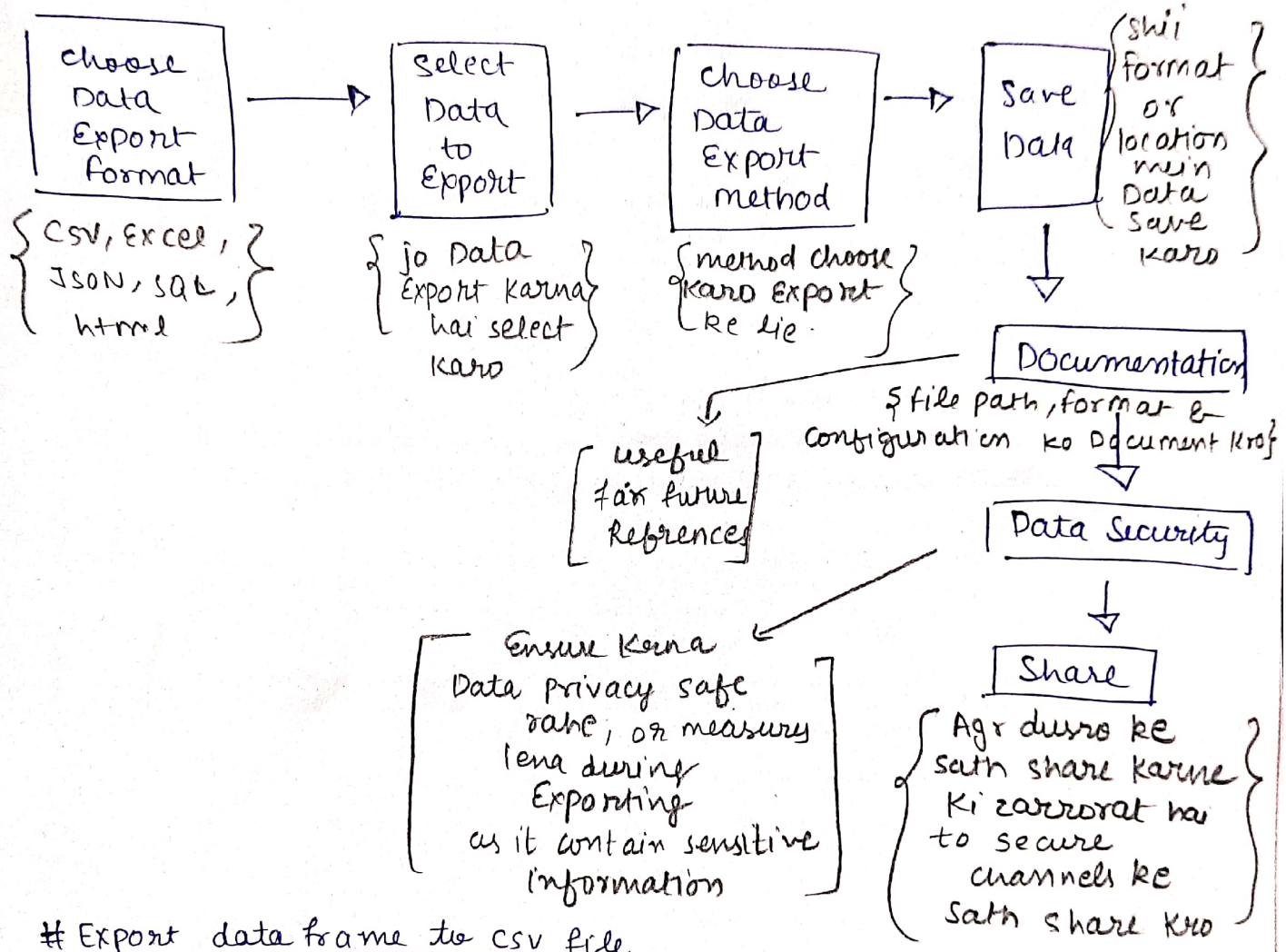
If file is html file

```
html_table = Pandas.read_html(url)
```

Data Exporting

Process jisme save or write karte hai Data system mein or use External file or storage medium mein store karte hai

Process



Export data frame to csv file

```
CSV_file.to_csv(file name)
```

Export dataframe to Excel file

```
Exl_file.to_excel(file name)
```

Export dataframe to json file object

```
json_file.to_json(file name)
```

Export dataframe to pickle object

```
pickle_frame.to_pickle(file name)
```

#1 Export_data frame to html table

```
html_table.to_html(filename)
```

Basic Insights from data set -

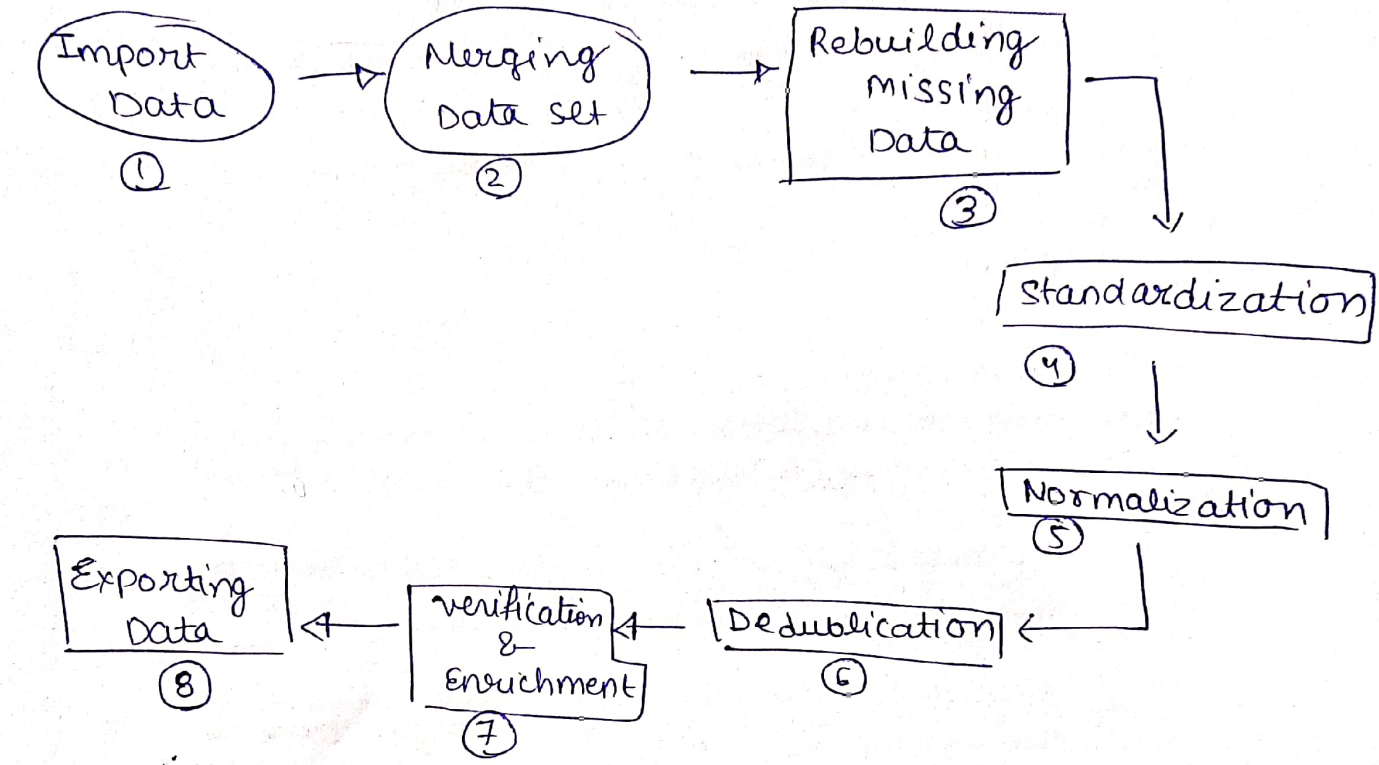
Extract karna Basic insights ka data se involve karta hai Performing Exploratory data analysis jo understand karta hai characteristics, patterns or trends ko data set mein.

1. Descriptive statistics - like mean, median, mode, range, variance. Standard deviation offer karte hai Quick understanding of central tendencies.
2. correlations - Identify karna Relationship between variables. Pta lagana ki vo kaise ek dusre ko affect karte hai.
3. Trends & patterns - visualization Representation (line charts, histogram / scatter plot) se Revel kar skte hai trend or patterns pta laga skte hai for future.
4. Group comparisons - Data set ko split karna ek Basic criteria pe. jo differences or similarities show karega groups ke beech.
5. Predective modeling - machine learning algorithm predict karte hai outcome or future trends btate hai Based on past
6. Anomalies or outliers - unusual Data points ko detect karna. ~~taki~~ or deviate karna significantly taki yeh impact na kare prediction ko.
7. Missing Data analysis - patterns / Reasons ko samajna Missing values ke peeche ka
8. time series analysis - Examine karna ki kaise variables change ho rhe hai time ke sath, identify karna Seasonality or long term trends.

Data cleaning

↳ Data mein se Duplicate, incorrect, corrupted, incorrectly formatted info ko hatana is Data cleaning

Process



Standardization – Converting format to only one particular for all

Normalization – Sari values ko common scale pe lana adjust karna scale ko.

Dedublication – Extra values ya copies jo hoti hai Data ki unko hatana

verification & Enrichment –

↳ Missing values jo fill ki hai & jo bhi changes kie hai unhe verify karna or store kar kar Rakhana

Data cleaning

Pandas offer karte hai data frames (jo builtin hote hai) Data ko clean karne ke lie.

Renaming variable.

```
Product_df = product_df.rename(columns = {'cost': 'Price',
                                          'shippingLocation': destination'})
```

Insert new column.

```
Product_df['NumberOfProduct'] = 1
```

Checking for NULL values

```
Product_df.isnull()
```

boolen Data frame return karega, agr value null mogi to true(0) varna false (1)

Kisi particular ko check karna ho (column Particular ko)
Ki is column mein koi value null hai ya nhi

```
Product_df['Price'].isnull()
```

Dropping Rows & columns with Null values

Drop all rows having any Null value

```
Product_df.dropna(axis=0)
```

Drop all columns having any null value

```
Product_df.dropna(axis=1)
```

Drop all rows having less than k null values

```
Product_df.dropna(axis=0, thresh=5)
```

Drop all columns having all values

```
Product_df.dropna(axis=0, how="all")
```

Replace values -

Replace all Occurence of 0 & 1.

```
Product_df.replace(0,-1)
```

Filling NA values with some values.

Fill all NA values,

```
Product_df.Fillna(0)
```

Grouping Data

Grouping on data by seller name

```
Product_df.groupby('seller name')
```

Apply statistics on Grouped data -

Getting mean of cost associated from Each seller.

```
Product_df['Price'].groupby(Product_df['sellername'].mean())
```

mean niklega

counting values

```
Product_df['Price'].groupby(Product_df['SellerName'].count())
```

count karega value

Minimum value of cost of Each seller.

```
Product_df['Price'].groupby(Product_df['SellerName'].min())
```

minimum value

Maximum value of cost of Each seller

```
Product_df.['Price'].groupby(Product_df['SellerName'].max())
```

maximum value of

Median value of cost of Each seller -

```
Product_df.['Price'].groupby(Product_df['SellerName'].median())
```

Median nikalega

Std of cost.

```
Product_df.['Price'].groupby(Product_df['SellerName'].Std())
```

standard deviation

Missing Data -

↳ common problem in machine learning.
↳ any information jo absent ho ya unavilable ho Dataset mein.
↳ It occurs ~~where~~ when there is gap or Empty field in collected information.
↳ can happen due to human ① Error during data Entory.
② or Malfunctioning sensors in data collection devices.
③ not choosing any answers during surveys,
④ Or information was not Recorder

Deal karna missing Data ke sath Kafi crucial hai as yeh affect karta hai accuracy, Reliability ya Kisi bhi Conclusion ko jo Data set se nikala jata hai.

Methods to handle missing values

(1) Remove features which contains missing values -
Simplest way is to analyze which feature contains the missing value & delete them., model ko baaki features (like jo ~~Kaaaaa~~ columns bache hai), unke sath train kia ja skta hai
Ye sif tab use ho skta hai jab hama Data set Kaafi acha ho or missing values boht Kam features ki ho, or unka correlation jo ho dependent variables ke sath vo high na ho.

(2) Remove Rows contain missing values -
hum Rows ko delete kar kte hai missing values ko hatane ke lie pr pher agr hamare test data mein vo Rows present hogi to machine learning algo fail ho jayegi correct output predict karne mein,
Or agr large no. of Rows mein missing data hoga to vo training data ko reduce kar dega.

③ Replace missing values with significant data -

↳ well technique.

↳ values ko Replace kar dete hai different value ke sath

Numerical values mein Replace kar skte hai

→ Mean ke sath Feature value ki

→ median ke sath Feature value ki

→ zero ke sath.

catagorial ke lie attribute a ek new category "default"

Types of missing valus

MCAR (Missing completly at Random)

- Missingness - completly Random hoti hai, unrelated ho ti hai other variables se, Random subset hote hai Data ke

MAR (Missing at Random)

- Systemetic, but predict ki ja skti hai dependent on other variable.
- khud se missing nhi hoti

MNAR (Missing not At Random)

- missingness dependent hoti hai missing value ya unobserved variables pe
- Significantly impact karti hai, analysis ko

Step ① Identify Karo missing value ko kis type ki hai ~~numo~~. numerical / catogorical (NAN / NULL)

Step ② understand Karo nature ko kya hai (MCAR/MAR/MNAR)

Step ③ (i) kam hai to Row/column ko delete kardo.

(ii) Imputation (Fill Karo values if numerical < Mean / Median / 0

if catogrial - default.

Step 4 jisko data set ko test hoga usme update karo values

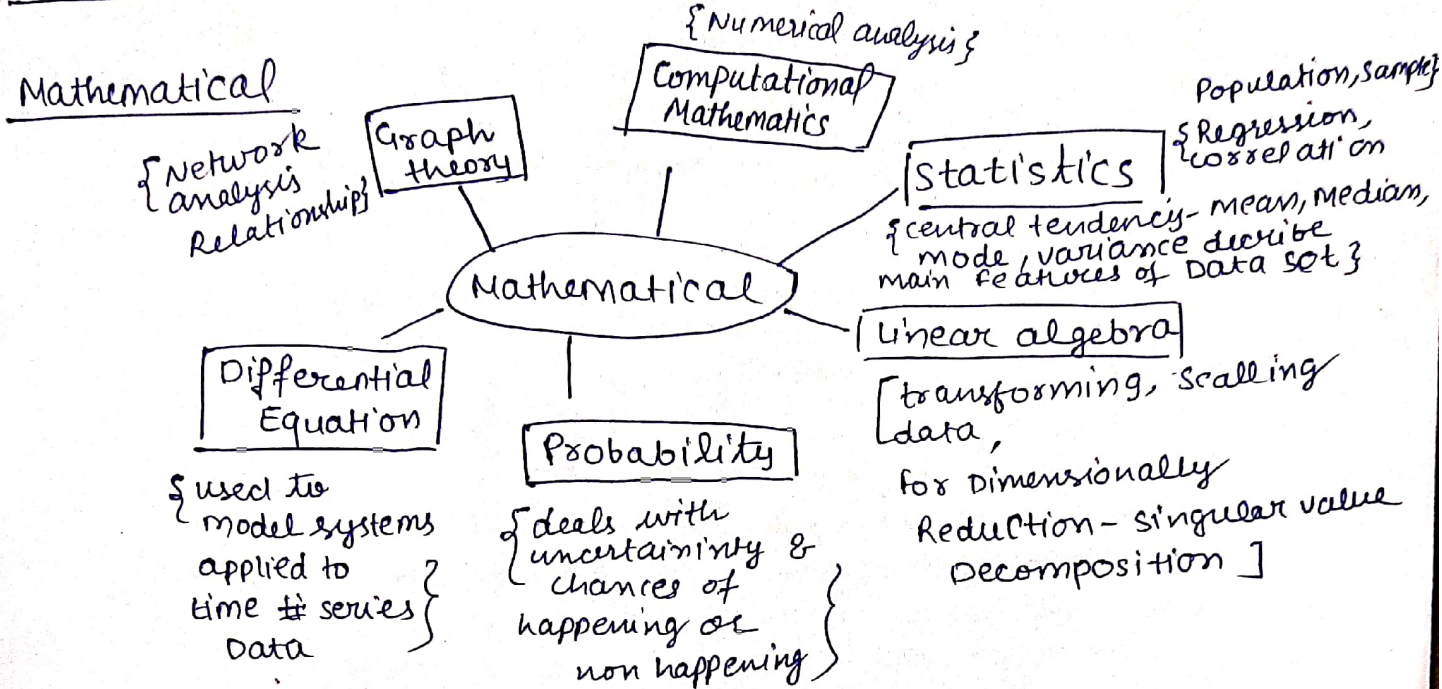
Impact - Impact dekho jo hamari missing values fill ki hai, humne unka impact dekho kya pada. Prediction pe / Kya padega

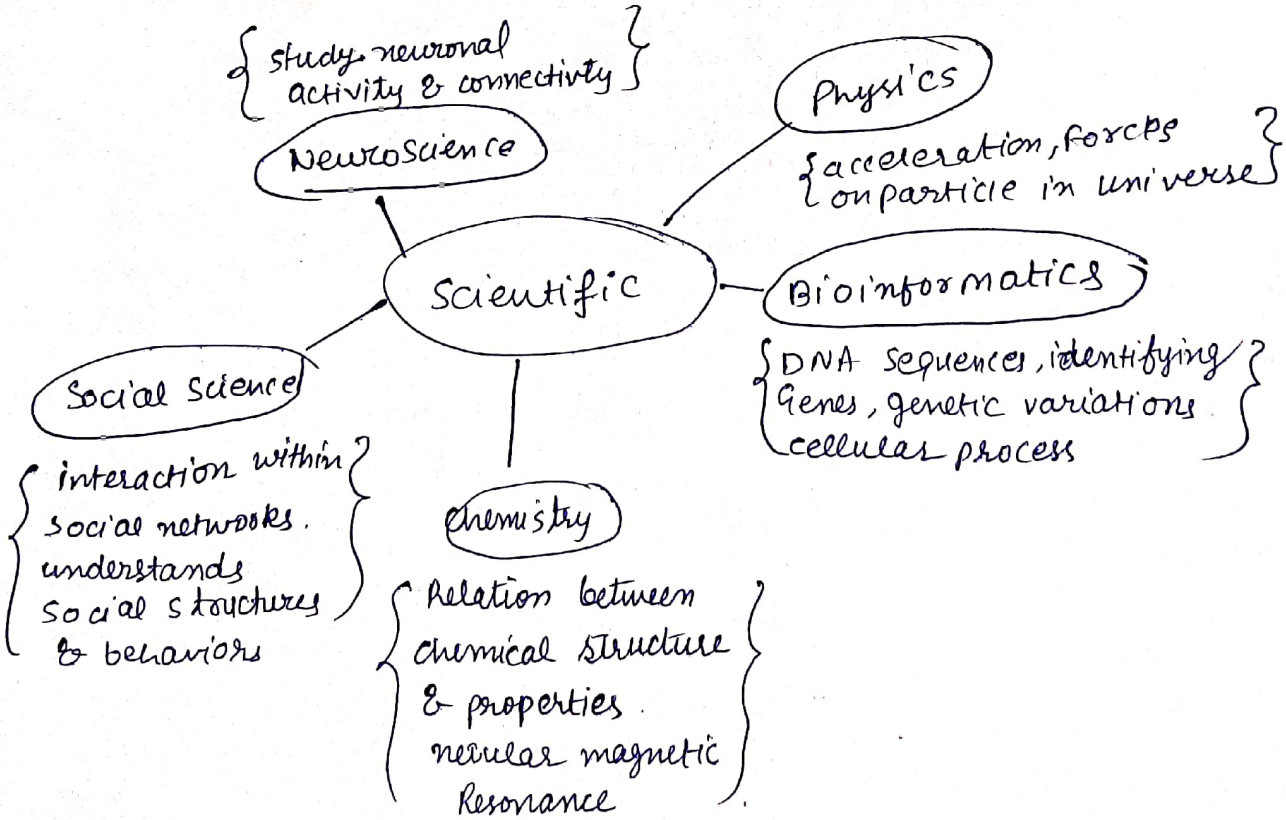
Documentation - Save Karo ki kisme values ~~ko~~ / Kiski values change ki thi as if Baad mein dekhni padi to

Quality / completeness - saari missing values handle hogyi or data set ki Quality check Karni hai

UNIT-IV

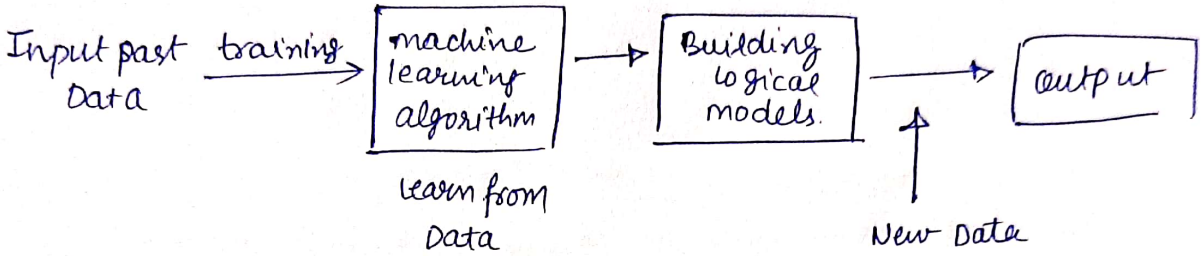
Mathematical & scientific applications for Data analysis





machine learning -

- Subset of artificial Intelligence
- focuses primarily on creation of algorithms, jo computers ko Enable karti hai independently learn karne ke lie data se / Previous Experiences se.
- without being Explicitly programmed, machine learning Enables a machine automatically learn from Data, improve performances from Experiences & predict things.



Features of machine learning -

- uses Data to detect various pattern in given set.
- learn from past Data & improve automatically.
- Data driven technology.
- Deals with huge amount of Data.

Need for machine learning

→ Rapid increment in production of Data.
→ Solve complex problems, difficult for a human
→ Decision making in various sector including finance.
→ Find hidden patterns, & Extract useful info from data.

classification

1. Supervised learning < classification, Regression
2. unsupervised learning < clustering, Association
3. Reinforcement learning

applications - Image Recognition, • Medical Diagnosis • automatic language translation, self driving cars.

Supervised learning

Type of machine learning hota hai, jaha algorithm learns karta hai labelled training data se, correct output data bhi milta hai

Meaning : input data ka corresponding output labels or target values hoti hai.

↳ Supervised learning ko Goal hota hai mapping function se learn karna (input se).

Application - Fraud detection, spam filtering

Labelled data - consists of input-output pairs.

Learning algorithm - algorithm learns from labelled data (Patterns & Relationships)

Prediction - Model can predict outputs

Types of Supervised learning

- **Classification** — Predicting categorial labels or class for new inputs
- **Regression** — Predicts continous values or Quantity

workflow

Data collection → Preprocessing → Feature Extraction → Model training → Evaluation → Prediction

Machine learning

- Subset of artificial Intelligence
- focus primarily on the creation of algorithms that enable a computer to independently learn from Data & previous Experiences

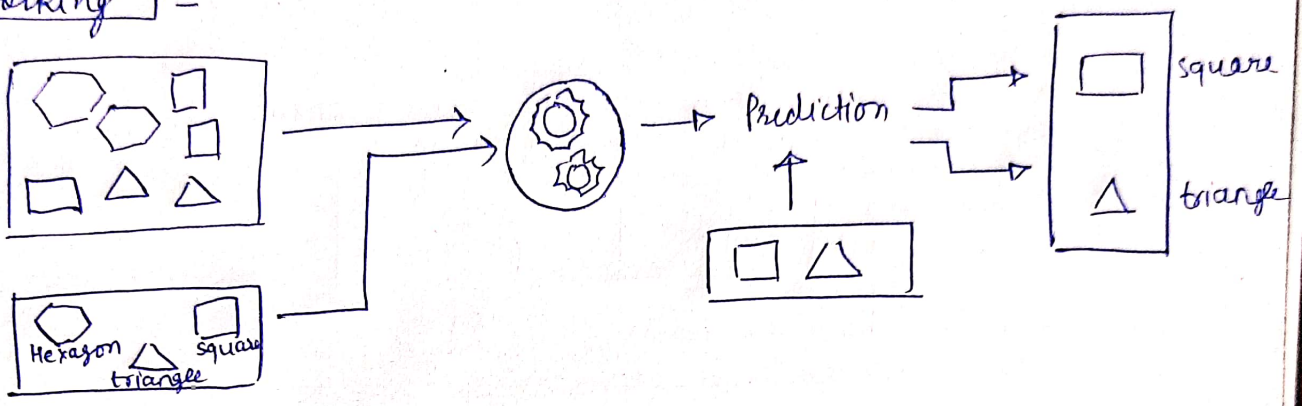
Supervised learning –

↳ type of machine learning jisme machines ko trained karte hai labelled data ke sath, or o correct output predict karte hai.

↳ Aim : To find a mapping function to map input variable (x) and output variable (y)

↳ applications – Risk assessment, Fraud detection, Spam filtering.

working –



steps

1) Determine type of training Data set.
2) collect / Gather the labelled training data.
3) split the training Data set into training dataset, test dataset & Validation Data Set.
4) Determine input feature of training data set.
5) Determine suitable algorithm.
6) Execute algorithm.
7) Evaluate the accuracy of model by providing the test set agr model ne sahi output predict kia, iska matlab model sahi hai

advantages

- model can predict the output on the basis of prior experience.
- Exact idea about the classes of object.
- used to solve Real world problems.

Disadvantages

- not suitable for handling complex tasks.
- Training Required lots of computation time.
- Enough knowledge about classes of object.

Unsupervised learning –

type of machine learning where algorithm is trained on unlabelled data & tasked with finding pattern or Relationship, Structure within Data Bina Explicit Guidance ke

↳ Goal hota hai explore karna or understand karna uske andr ke structure ko or usse extract karna useful insights bina labelled diagram ke.

Types

Clustering

Group similar Data points together Based on some similarity or dissimilarity

Dimensionality Reduction

Reduce no. of features while Retaining important info.

Workflow –

Data collection → Data Processing → Feature Extracting

↓

Model training → Evaluation

Advantages

- Used for more complex tasks as compared to supervised learning.
- Easy to get unlabelled data

Disadvantages

- Result might be less accurate
- more difficult

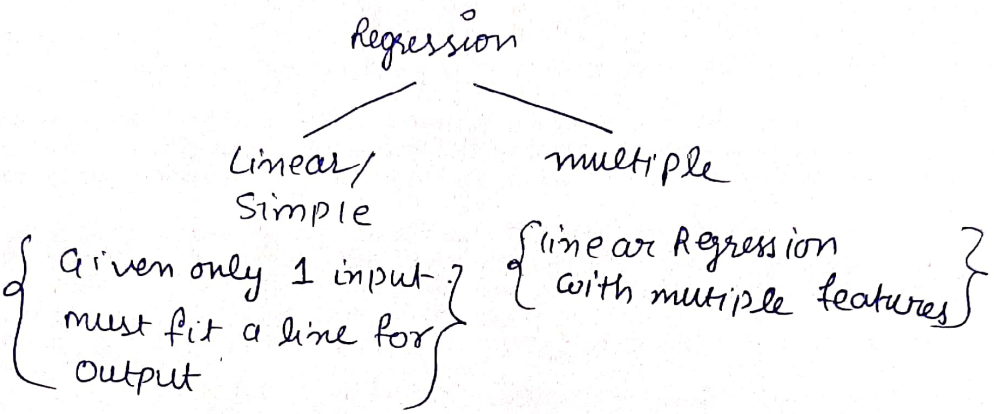
Supervised	unsupervised
• trained using labelled Data	• trained using unlabelled data.
• predicts the output.	• find the hidden pattern
• Produces an accurate result.	• may give less accurate result as compared.
• needs supervison to train the model along with output	• doesn't need any supervision to train the model
• takes direct feedback to check if predicting output correct or not.	• doesn't take any ~~output~~ feedback
• Supervised / Classification, \ Regression.	unsupervised / clusting, \ Dimensionality Reduction
• Data is provided along with the output	only input Data is provided to the model

Regression (technique) (Supervised learning)

↳ Predicting the continous valued output from input.

Example application of Regression

- Predicting stock price
- Forcast sales of moth.



Classification – (Supervised learning technique.

- with the help of features & categorical output, one has to predict the output.
- models are classifier.
  ↳ Role to learn the 'Relationship between input invariable & categorical outcome.
  ↳ classes.

Ex classifying emails – Spam hai ki nhi.

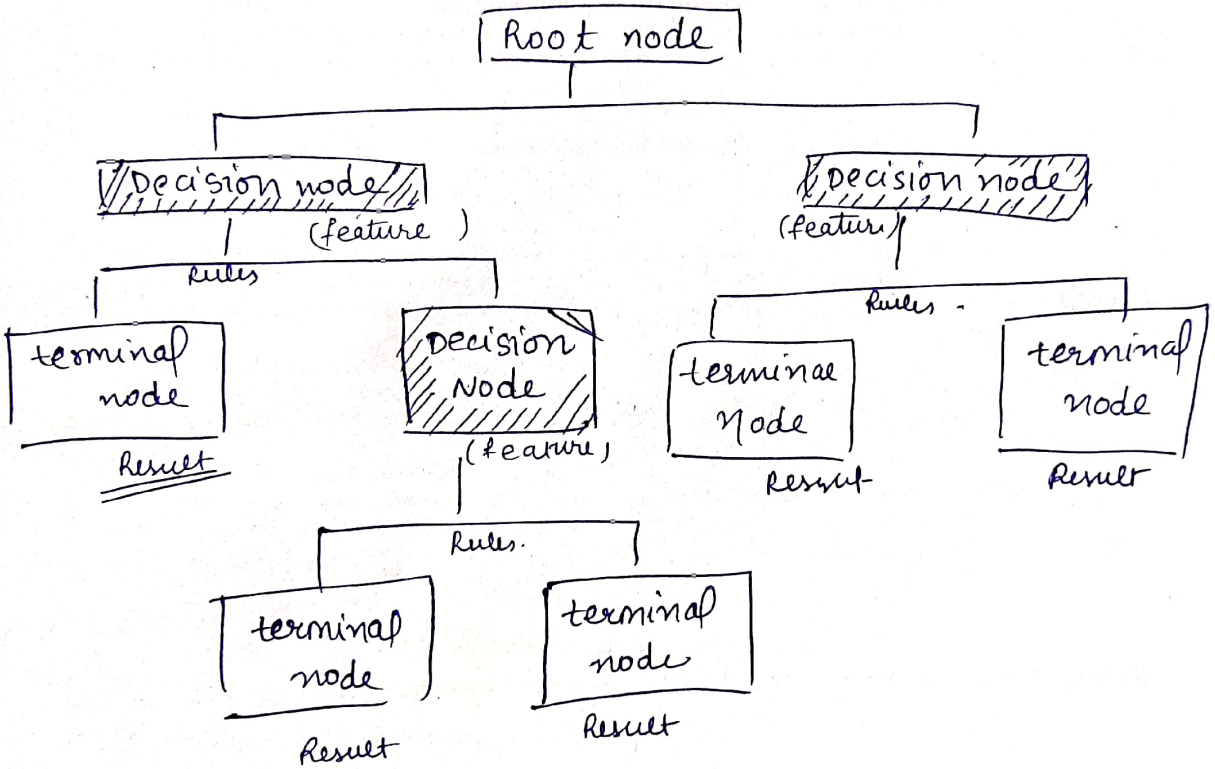
classify Document – from which Gerne a thing belongs.
like – in newspaper it is Related to sports, politics, Education, market etc.

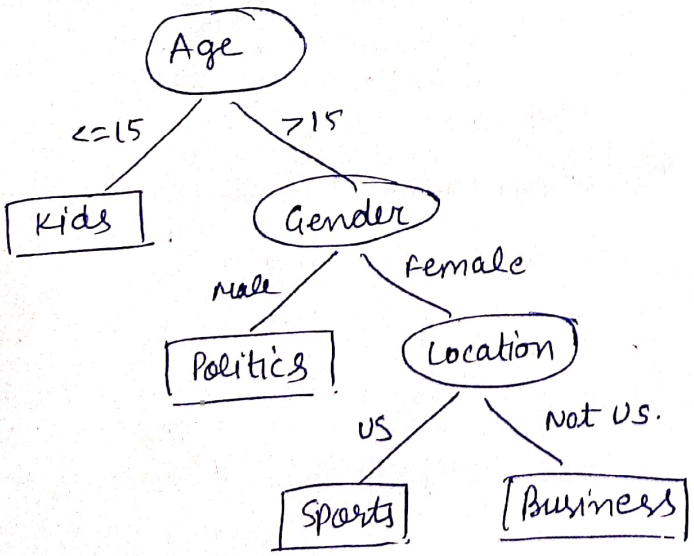
Decision tree -

→ technique of classification.

→ hierarchical tree structure which consists of a root node, branches, internal & leaf nodes.

→ internal nodes of tree denotes feature,
branches denotes - the Rules.
leaf nodes denotes the Result of algorithm.





Advantages of Decision tree

(1) capture non linear Relationship between the Data

(2) Easy to Explain to other. (doesn't need any complex mathematics knowledge to understand the result)

(3) Can handle both numerical & categorical.

(4) Fast process.

(5) handle irrelevent attributes.

(6) not based on assumptions.

(7) cleaning on data not needed.

Simple tree

↳ **Early stopping** - Hum tree ko Banana start karte hai, and stop kar dete hai ek depth pr aakar tree ko Bnana tab stop karte hai ya to jab.
(1) validation Error or decrease na hora ho.
(2) ya pher jab boht kam data Bacha ho kise node ke pass, usse majority ko leke prediction krte hai.

↳ **pruning** - Build karte hai pehle tree ko, or uske Baad tree ki Branches ko cut karna start kar dete hai better generalization Error ke lie.

Decision tree algorithm - operate karta hai data ko analyze karke.
Recursive algorithm., It Repeats until Reach a leaf node.

Step 1. Start karo tree ko Root node ke sath, jo complete dataset contain karta hai.

step 2. Best attribute find karo data set mein. using attribute selection measure.

Step 3. Divide karo Data set ko subset mein, contain kare value According to best attribute.

step 4 Decision tree node Genrate karo.

step 5. Recursively new decision tree Banao, or use karo @ subset Data set ke jo step 3 mein bnae the. or is process ko continue karo jae tk ap data ko or classify karne ke kabil na bacho., or un nodes ko final node / leaf node boldo.

Disadvantages

(1) Kahin na kahin Decision tree ko stop hona padta hai, varna high error milne ke chances hote hai.

(2) faces problem of overfitting.

Decision tree

classification	Regression
Dependent variable - catagorical. Statements ka use nhi hota	Dependent variable - continous. like Ex - salary Based age, years of expirence. Regression tree mein, kisi bhi node pe mean value di jati hai sab values ki, as Prediction.

Decision tree learning -

→ algorithm used to learn the decision tree from training data
→ work karta hai top to bottom
→ best attribute check karta hai kisi bhi point pe, or learn karta hai, create karta hai decision node uske lie.

Decision Boundries -

Gini score –

split of different class attribute at any node
Best value = 0 (when node divide data perfectly)
worst value = 1 (when both branches have classes in same proportion).

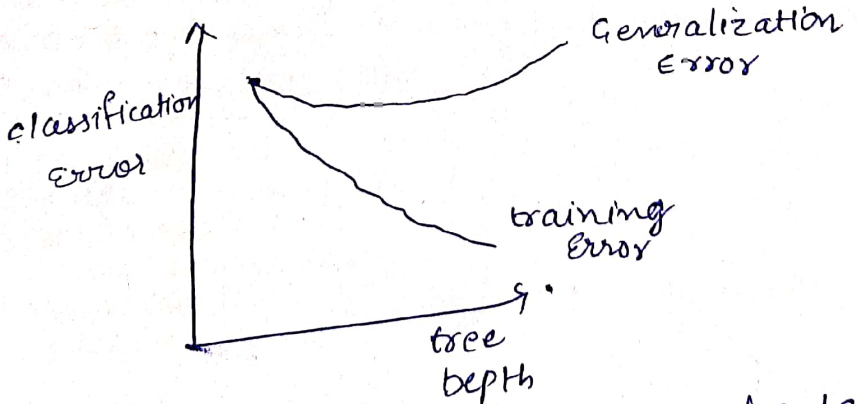
Goal – create a model jo reduce kare Errors ko.

Error – defined as no. of incorrect predictions divide by total no. of examples taken.

Decision tree problem

face karta hai problem of overfitting,
agar hamare pass bohot sare features hai to hum bs split karke Branches hi Bnate rahenge, or jab End mein bohot kam data bachega har grp mein to vo generalized solution hi hoga.

isse Training Error to decrease hoga with increase in depth of tree but generalization Error increase hoga



simple tree better hai overfitting ke term mein isme choose kia jata hai depth or complexity ko decision tree ki uske validation ke Based pe.
agr, simple & complex dono tree's same validation Error denge / approximate denge, to simple ko Choose karenge.

Random forest - (Ensemble technique)

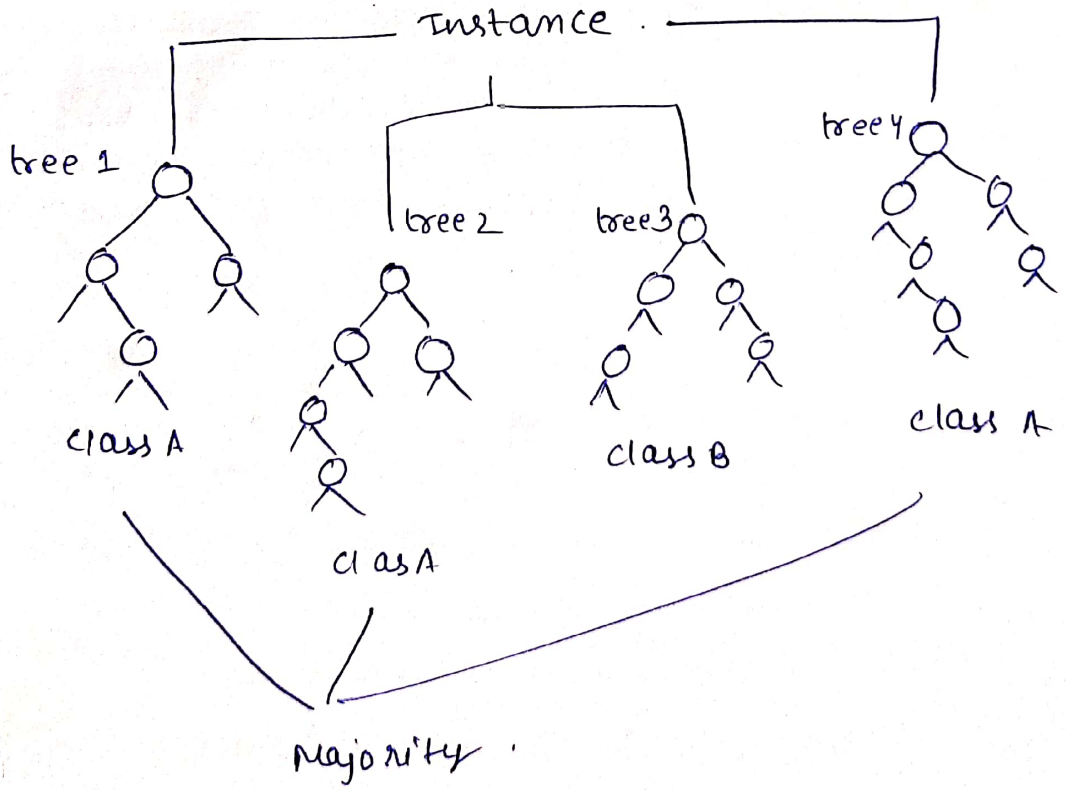
↳ forest of trees.
↳ contains large no. of decision tree jo help karti hai decision lene mein.
↳ tree same strategy se bnae jata hai jaise decision tree bnae jate the. or decide kia jata hai majority votes se.

Random forest advantages over Decision tree

→ Missing data ko handle kar leti hai
→ Regression or classification dono ho jate hai
→ overfitting nhi hoti

Disadvantages -

→ model ke upr kaafi kam control hota hai
→ complex hai
→ Difficult to Explain.



Naïve Bayes –

↳ highly used in text classification.

↳ advantages – Speed ↑, fast & make prediction very easily with high dimension of data.

↳ probabilistic classifier.

↳ naiive because yeh man ke chlta hai EK feature dusre feature ke independent hota hai.

~~fcat~~

↳ Bayes thm se apply hota hai

$$P\left(\frac{X}{y}\right) = \frac{P\left(\frac{y}{X}\right) P(X)}{P(y)}$$

$P\left(\frac{y}{X}\right)$ → likelihood

$P(X)$ → Prior

$P(y)$ → evidence

$P\left(\frac{X}{y}\right)$ → Posterior Probab.

Clustering – way to group the objects in a way that

→ similar object single group mein Rahe

→ Dissimilar alg alg.

⇒ Interclustur similarity ↑
Intraclustur Disimilarity ↑

Feature	Predictive	Trends
Objective	future outcome predict karta hai based on historical Data.	Identify karta hai patterns & tendencies over time.
Nature	future focused	past focused.
Data usage	historical Data build karke prediction karta hai.	Examine karta hai historical data identify karta hai patterns.
techniques	machine learning methods	time series analysis
Decision making	Aims to inform decision making for future Events.	Assist in making Decision Based on historical Data.
Example	weather Forcasting.	seasonal variations

Predictive mining

↳ **Prediction + Data mining**

Practice of using Data, Statistical algorithm and machine learning techniques, Pattern identification & predict future outcomes based on historical Data

Exploration & analysis of large Data sets to discover meaning ful patterns, Relationship or insights, (meaningful understanding / Perception)

Process

Data collection & prepration

collect & prepare Relevent data, jo include karta ha historical records sensor data, customer info ya koi bhi Dataset

Data Exploration & Feature Selection

EDA conducts kia jata hai to patterns Relationship pta lagane ke lie or Feature Selection involve karta hai identify karna attributes + variables Prediction ke lie

model building & training

machine learning algo, Regression, decission tree neural network, model training with patterns & Relationship

(Models - Evaluate kie jate hai, accuracy, Reliability dekhne ke lie)

model Evaluation & validation

(Model performance dekhte hai)

Prediction & Deployment

Model validate ho gya, use kia ja skta Prediction karne ke lie,
↳ Sales forcast
↳ customer behavior
or any other

Iterative Process & Refinement

Predective mining iterative process hai, Models might be Refined, Retrained & updated Regularly as new Data Roz update hota hai jisme predictions karni hoti hai

Trend analysis -

Examining data over time jisse identify karte hai consistent patterns or tendencies ko.

↳ used in finance, Economics, marketing, technologies. to make informed disicision Based on historical data.

↳ It Reveals insights to direction & behavior of variable or set of variable Ek specified period ke lie.

Key aspects -

1. Data collection :- Relevent Data collect karta hai over a specific time period., measurable variable se hota hai

2. time series - Data ko chronologically arrange karte hai, time series create karne ke lie. Each data point series ka Repersent karta hai variable ki value, Ek specific time ke lie.

3. visualization - plotting graphs jisse trends easily find out kar sake.

4. Identification of patterns - data ko analyze karenge or recurring patterns nikalenge.

5. statistical technique - Regression analysis, mean, median, mode & all.

6. Forecasting - planning & decision making, trend identify hone ke Baad.

7. factors influencing trends - dhundo un factors ko jo trends ko affect karte hai.

8. long term / short term trends - kitne time tk ka trend hai.

Recommendation system -

Applications or platforms jo Recommend karte hai useful things to the user according to interest.

Goal of Recommendation system -

Present karna items user ko jisme vo intrested hai.
Yeh selection or increase karane mein help karta hai sales ki

Recommendation Engine -

1. Popularity Based.
2. Content Based
3. Classification Based.
4. Collaborative filtering.
5. Model Based.

Popularity Based -

→ Based on usage of items by all users
(jo item sbse jyada use kia jata hai, sbse popular mana jata h) isslie vo sbko suggest kia jata hai

→ Based hota hai past usage count pe.

→ yeh Recommentation system used hota hai news article (jisme Recommend kia jata hai item new user ko)

→ iska drawback hai, yeh assume karta hai Har user same type ka hai, or sbko same Recommendation deta ~~is~~ hai without even knowing about the profile geography or other attributes.

Important

Content Based Recommendation Engine -

→ Based hota hai feature pe item ke,
→ items ko compare karte hai or Recommend karte hai, jiske features mil rhe hote hai us item se jisko user ne higher Rate kia hota hai.
→ Memory Based system hota hai kaunki it memorizes the items.
→ but inme training Data large hota hai & fit nhi kar pata memory mein islie preffer nhi karte

Imp

Collaborative filtering –

↳ Based on the idea of similar interest

↳ it depends on the choices of two users

collaborative

user-user

Based on similar user, on their item purchase

user ek hi hoga pr uski choices dekhi jayengi kya-kya item usko pasand aa rhe hai

{ agar ek item pasand aa rha hai to usi se jaise item Recommend honge }

item-item

→ Based on similar items purchased by users.

user do honge but unme similarity dekhi jayegi, ki kya kya item dono ko pasand aa rhe hai, ya ek ko acha nai to dusre user ko vhi Recommend hoga