Python :- object oriented, high level programming lang
features.
- integrated
- interactive
- object oriented.
- free & open source
- scripting lang.
- simple & easy to learn
- portable
- small sol. for complex prob.

Python libraries.
①. A library is a collection of files that contains func for other programs.
②. Reusable chunk of code to include in program.

① Numpy. (Numeric python)
- scientific computing & basic & advanced array operations.
- this lib offers many hand features. to perform operation on n-arrays & matrices in Python.
- makes performing math operations on arr easier.

② Pandas
- It provides support for data structures & data analysis tools.
- This lib is optimized to perform data science tasks fast & efficiently.
- Best suited for, structured, labelled data ie. tabular data.
- Pandas core data structure ① Series
② DataFrame

Series - 1D array like structure.
- holds single column

DataFrame - Represents tabular data.
- organized in columns. (col has single datatype)

(3) SciPy.
- contains many different packages & modules to assist in Mathametical & scientific computing
- provides packages like
  (1) Matplotlib for data visulization.
  (2) IPython

(4) Scikit-learn
- contains a lot of efficient tools for ML & statistical modeling.
- used in dimentionality reduction, classification, clustering, & Regression. & unsupervised.
- provides many supervised learning algo's.
- cross validations - to check the result of supervised learning.
- provides toy dataset ie. iris, Boston housing.
- used to extract features from images & Text.

# ~~Data~~ P eg. import numpy as np.
import matplotlib.pyplot as plt.
from sklearn.linear_model import ~~LogisticRegression~~ LinearRegression.
——"—— datasets import make_classification.
~~X,y = make classification (100,2,2,1.)~~
model = ~~Logistic~~ linear Regression()
model.fit(X,y)
plt.scatter(X, y, ~~c=y~~, ~~cmap='bwr'~~)
plt.plot(x, ypred, label="ABC")
plt.xlabel("...")
plt.title(...)
plt.show()

ypred = model.predict(x)
color='blue' label="Actual data"

x = np.array([1], [2], [3], [4], [5])
y = np.array([2, 4, 5, 4, 5])

Data Preprocessing.

- Data mining technique to transform raw data to understandable format.
- reduce data size, find relation b/w data & normalize steps.

① Data cleaning :- fill missing val, smoothing noise, removing inconsistancies.

② Data integration :- collect from diff source, conflict resolved.

③ Data Transformation :- ~~the~~ data transformed to required structure; normalized, aggregated & generalize

④ Data Reduction :- Normalize, aggrigate & generalize

⑤ Data Discretization :- dividing range of val (10-20) in fixed intervals, & assigning each val to one interval

Removing Duplicates.

- in data preprocessing, duplicates are row that appears more than once. causes biased / incorrect results.
- produce false output
- renders ML model useless / not accurate output
- increases datasets size unnecessarily.

code. import pandas as pd.

```
data = { 'Name': ['A','B','C'], 'Age': [10,20,30] }
df = pd.DataFrame(data)
a = df.drop_duplicates()
print(a)
```

Handling Missing Values.

- Missing val = absence of val. in a dataset.
- NaN is its representation.
- ML cant work with missing vals.

ways to handle
① ignore tuple
② use global constant.
③ use Mean / median interpolation.

Eg.
df['A'].fillna(df['~~Sal~~ A'].mean(), inplace=True)

Data Transformation.
- process of changing the format, structure, or values of data to make it suitable for analysis or model traning.

Types of Transformations
1. Normalization / standardization.
2. log transformation.
3. Encoding
4. Scaling.

Eg. Scalar = MinMaxScaler()
df['Normal'] = Scalar.fit_transform(df[['Salary']])
can be done by ETLC Extract transform & load) tools.
either on premise / cloud based.

Analytics Types.
① Predictive analytics
② Descriptive analytics
③ ~~pre~~scriptive / Diagnostic analytics.

① Predictive analytics.
- what could happen in future.
- helps organizations ~~in future~~ to predict future.
- calculates live transactions multiple times to help eval the benifit of a customer transaction

- makes use of vaiity of variable data to make predictions.
- the variability of the component data will have a relationship with what it is likely to predict
- Predictive analytics is covered in following steps.
  1. Project defination → what shal be the outcome of the project
  2. Data collection → various customer interaction at single view
  3. Analysis → data processesd & utility is found.
  4. Statistics → validate (assumption, true output)
  5. modelling → predict multiple models for precise eval
  6. Deployment →
  7. monitoring

Eg :- wether analysis, Helthcare., fraud detection

(2) Descriptive "what has happend"
- looks at past data to understand what has already occured.
- helps in identifying trends patters & summaries
- used in first phase of analytics. involving gathering & organizing data.
- shows relationship b/w product with data
- to organize customer by their preferences
- BI fails to make the data communicate to people. & hence. Descriptive analytics is wed.

Purpose :- Summarize historical data

Eg. Monthly salary report

(3) Prescriptive / Diagnostic Analytics :- "Why did it happen"
- investigate cause behind past outcomes
- helps to identify reason or cause of trend or any problem
- assist user to find optimal solution to make right choice

- Uncter utilizes the understanding what, why questions. to help user determine best source of action to take
Eg :- Sales drop in specific region

Market basket analysis
technique that allows us to discover the relationship between products.
makes use of Association Rule. [① Apriori algorithm ② FP growth algorithm.

⇒ Association Rule
- it is useful for discovering interesting relationship hidden in large datasets.
- the uncovered relationships can be represented in the form of association rules on sets of frequent items.
- Association rules are if than statments that helps uncover relationship between seemingly unrelated data.
- eg. "if a customer buys bread (then) he is 80% likely to also purchase milk".
- Association rule mining has 2 steps.
① find all frequent itemsets: (use min support count)
② generate strong association rule from those itemsets. by defination these rules shoud satisfy min support & confidence

Support :- how frequently the total collection of items. {A, B} occur together as a percentage of all transaction

$$\text{Support} = \frac{A + B}{\text{Total}}$$

Confidence :- The ratio of no. of transaction including all items in B & A to the no. of transaction that include all items in A ⇒ $\text{confidence} = \frac{A+B}{A}$

Lift: implies the type of relation those objects have
ie. no relationship
+ve relationship
-ve relationships

if lift val = 1 ⇒ no relationship
lift val > 1 ⇒ +ve relationship
lift val < 1 ⇒ -ve relationship

Given a table eg

~~transaction~~ Item id	list of buy
A1	I1, I2, I5
A2	I3, I2

← Table 1

Steps.

Apriori Algo.

(1). find frequency of occurance of every item (count fi of each item.)

eg.

Item	frequency	support
I1	6	$\frac{6}{\text{total}} \times 100$
I2	8	$\frac{8}{\text{total}} \times 100$.
I3	5	
I4	3	
I5	3	

table 1 has 9 rows in this eg.

(2). use minimum support val ie. 4 to reduce the Itemset.
so. anything less than 4 removed.

Item	fi	support
I1	6	$6/9 \times 100$
I2	8	
I3	5	

(3). make combinations. & find their frequency.

frequent ~~dataset~~ Itemset	create rules	Item	fi	support	
I1 ⇒ I2	⇐	{I1, I2}	5	$\frac{5}{9} \times 100$	← I1 & I2 comes in the. Table 1
I1 ⇒ I3		{I1, I3}	4	$\frac{4}{9} \times 100$	5 times (together)
I2 ⇒ I3		{I2, I3}	4		like in A1 {I1, I2, I5}

(4). make combinations again

{I1, I2, I3}	3

do this till combination cant be done or till ϕ (empty set)

(5). calculate support $\frac{fi(A+B)}{total}$ confidance $\frac{fi(A+B)}{fi(A)}$ lift $\frac{support(X+Y)}{support(X) * support(Y)}$

for x	$\frac{fi(x)}{total}$

$Support(J_1, J_2) = \frac{fi(J_1, J_2)}{total.} \Rightarrow \frac{5}{9}$ ← total no. of rows in table.

$= 0.55$

do same for (J_1, J_3) & (J_2, J_3)

confidance $(J_1, J_2) = \frac{fi(J_1, J_2)}{fi(A)} \Rightarrow \frac{5}{6}$ ← J_1's fi from step 2

OR $= \frac{Support(J_1 \cup J_2)}{Support(J_1)}$ $= 0.83$. [× 100 for %]

$lift(J_1, J_2) = \frac{0.55}{6/9 * 8/9}$ ← support (x+y)

$\frac{fi(x)}{total.} = 0.94 \Rightarrow$ -ve relationship

Consider 50% confidance if not given.

The apriori algo is fundamental algo used for association rule mining in data mining. It discovers frequent itemset & derives association rules.

Steps of algo. ①. create candidate items.
②. Prune infrequent itemsets.
③. Repeat until no new frequent itemsets.
④. Generate association rule.

FP growth - improvement to apriori
- use min support to find frequent data.
- used to find frequent itemsets in a transaction database without candidate generation.
- in tree format.

Eg.

TID	Item
T_1	{EKMNOY
T_2	DEICNOY
T_3	AEKM
T_4	CKMUY
T_5	CEIKOO

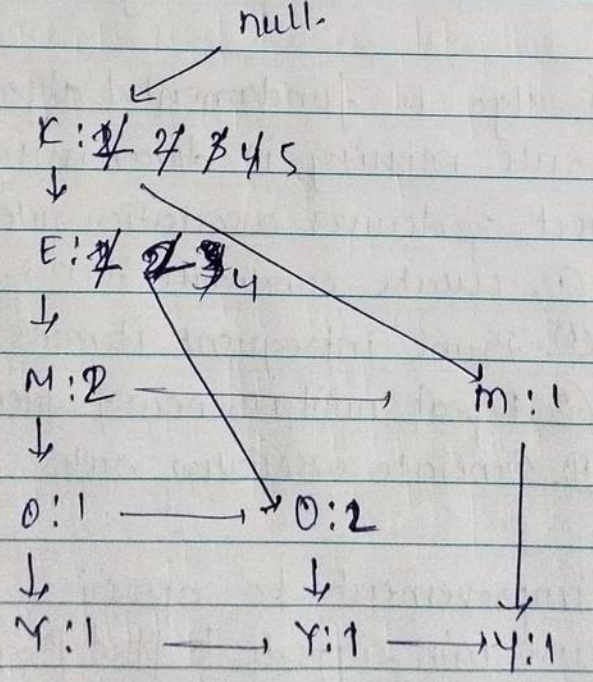
A	1	X
C	2	X
D	1	X
E	4	
I	1	X
K	5	
M	3	
N	2	X
O	4	
U	1	X
Y	3	

min support = 3

arrange in decreasing order of fi

K:5 E:4 O:4 M:3 Y:3

~~Iiii~~ TID	Items.	ordered itemset.
T_1	EKMNOY	KEMOY
T_2	DEICNOY	KEOY
T_3	AEKM	KEM
T_4	CKMUY	KMY
T_5	CEIKOO	KEO



	to each path.	
Y	KEMO, KKO, KM	= 3
O	KEM1 KE 2	= 3
M	KE=2 K=1	= 3
E	K=4	4
K		

Regression.

- data mining func that predicts a number.
- supervised learning technique used to predict continious numerical vals.
- to find relationship b/w independant var(x) & dependant variables(y).

types.

1. Linear regression
2. Logistic regression.

1. Linear Regression

- a supervise learning algo used to predict continious vals.
- models relationship b/w independant & dependant var.

Given by.

$$y = B_0 + B_1 X \quad \text{---(1)}$$

where $B_1 \rightarrow$ slope of line
$B_0 \rightarrow$ y intercept.

$$B_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{(x - \bar{x})^2}$$

$$B_0 = \bar{y} - B_1 \bar{x}$$

put val of x in eq 1. to get new val of Y

table to solve/get B_1 & B_0. Eg.

X	Y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
4	3	1	1	1	1
0	4	-3	2	9	-6
3	0	0	-2	0	0
5	1	2	-1	4	-2
12	10				-7

$\bar{x} = \frac{12}{4} = 3$ $\quad \bar{y} = \frac{10}{4} = 2$

In linear reg. predicted var. are the mean of target var at the given val. of the input var.
- used to solve regression problems.
- graph is a straight line

(2) Logistic Regression.

(2) - form of regression analysis in which the outcome is binary

(1) - Supervised learning algo used for binary classification

(3) - it predicts the posibility of class using a sigmoid function.

$$P(Y=1|X) = \frac{1}{1+e^{-(mx+c)}}$$

eg. spam detection.

(4) - The core component of logistic regression is logistic func. aka Sigmoid func :
which is used to model the probability that a given input belongs to a particular class.

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

where z is the linear combination of input features & their coefficients.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots \beta_n x_n$$

this func. takes a val for z & gives op as 0/1

Why to use it

(1) Probability Output.
- logistic func. maps real val input to 0 or 1 which can be interpreted as probability

② Non lineau tuansfouuation.
logistic uegussion wes logistic func. to squash value to a valid pnobability nauge.

③. Decision Bouudauy
By applying the thueashold (commoonly 0.5) logistic uegussion classifies the output in 0 & 1 classes.

④. Intuepuetability
- helps in undeustauding confidence level of classification

⑤ Mathematical convenience
The logistic func is smoot & diffeuentiable

Role of sigmoid
- conveuts lineau output (z) into a pnobability.
- if Op > 0.5 ⇒ class 1 (+ve class)
- if Op < 0.5 ⇒ class 0. (-ve class).
- helps in binauy classification using thueashold based decision.

	Lineau. Reg.	Logistic. Reg.
Output	Continious value.	Binauy output.
func used	$y = mx + c$	$\hat{o} = \frac{1}{1+e^{-z}}$
pnoblem type	Reguession type	Classification type
O/p nauge	$-\infty$ to ∞	0 to 1
cuuve	stuaite line	S shaped cuuve / non lineau.

Classification.

it predicts categorical labels (classes), prediction models continious valued func.

= considered to be supervised learning

2 types

① Naive Bayes

② Decision tree

① Naive Bayes.

- Supervised learning algo based on bayes theorem. with a strong assumption of feature independance
- used to classify tasks

$$P(A \mid x_1 \cap x_2 \cap x_3 \cap x_4) = \frac{P(x_1 \mid A) * P(x_2 \mid A) * P(x_3 \mid A) * P(x_4 \mid A) * P(A)}{P(x_1) * P(x_2) * P(x_3) * P(x_4)}$$

Bayes theorem = $P(A \mid B) = \frac{P(B \mid A) . P(A)}{P(B)}$

P(A|B) → Posterior probability

P(B|A) → likelihood.

P(A) → prior. prob of A

P(B) → ——— " ——— B

working.

1. calculate prior probability of each class.
2. calculate likelihood of each feature
3. Apply Bayes theorem. to complete posterior for each class
4. assign the class with higher. posterior probability.

Adv.

Simple & fast.
works well with high dimensional data.
Requires less traning data.

Application.

Spam detection.
Text classification
Sentiment analysis

Eg.

Email	often	free	spam	
1	1	0	No	
2	0	1	yes	
3	1	1	yes	ov sv
4	0	1	No	ox sx
5	1	1	yes.	ov sv

to find. often=1 & free =1 is spam or not

prob(spam is yes) = 3/5

p (spam is no) = 2/5

2/3 ← spam yes for only 3. → ov sv

	yes	No.
often	2/3	1/2
free	3/3	1/2

$$P(yes \mid often, free) = \frac{P(often \mid yes) * P(free \mid yes) * P(yes)}{P(often) * P(free)}$$

$$P(NO \mid often, free) = \frac{P(often \mid no) * P(free \mid no) * P(no)}{P(often) * P(free)}$$

no. need to solve denominators as it is same for both. yes & no.

$P(\text{yes} \mid \text{offer, pain}) = \frac{2}{3} \times \frac{2}{3} \times \frac{3}{5} = \frac{4}{15} = 0.4$

$P(\text{No} \mid \text{offer, pain}) = \frac{1}{2} \times \frac{1}{2} \times \frac{2}{5} = \frac{1}{10} = 0.1$

since the Denominators are same we can say.

as $P(\text{yes} \mid \text{offer, pain}) > P(\text{No} \mid \text{offer, pain})$

∴ the new msj is spam.

Decision Tree

- used to make decisions.
- build by choosing the best attribute to split the at each step. based on information gain, which is calculated using entropy.

steps. yes / No → from last col.

① Calculate the entropy of entire dataset.
ie values {+ for yes, -ve for No}.

eg $\text{val}\{+9, -5\} = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$

② select random attribute. ie wether. it has 3 types. calculate entropy for them. & do the same for other attributes.

Sunny $\{+2, -3\}$ (2 yes, 3 No) $-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

Cloudy $\{+4, -0\}$ $-\frac{4}{4}\log_2\frac{0}{5}$ will lead to $0 = 0$

Rainy $\{+3, -2\}$ $-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97$.

③ Calculate info gain for all attributes. ie. wether, Temp, humiclity & wind.

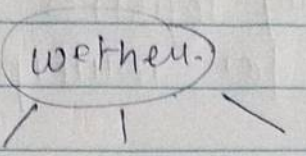
for wether eg. info gain =

$E(\text{whole dataset}) - \sum\left(\frac{\text{no. of Row's (Sunny)}}{\text{total rows in table}} \times E(\text{Sunny})\right) - \frac{\text{no. of Row(cloudy}}{\text{Same}} \times E(\text{C})$

↑ entropy

chance ie, sunny, cloudy & Rainy.

(4). select the highest value of info gain
(5). set that node as root node.
(6) selet the sub val from the root node. i.e.



Select sunny.
& prepare a table with all sunny values.
& find entropy of the entire table
& info gain for sunny.

(7) repeat this for all the attributes. recursively.

entropy
$$H(S) = -\sum p_i \log_2(p_i)$$
p_i = probability of class i

Info gain
$$IG(S,A) = H(S) - \sum\left(\frac{|S_v|}{|S|} \cdot H(S_v)\right)$$

Types of logistic Regression.
(1). Binary logistic Regression.
(2) Multinominal logistic Regression.
(3) Ordinal logistic Regression.

(1) Binary logistic regression.
used when target variable has 2 categories.
eg. Pass y/s fail & yes/no
Output: probablity b/w 0/1 classified using threashold.

(2) Multinomial logistic regression
used when the target variable has more than two unordered categories.
eg. fruit (Apple, Banana, Orange)
- it extends binary logistics regression using a Softmax function for multiple classes.

(3) Ordinal logistic Regression.
- used when the target variable has more than 2 ordered categories.
- eg: poor, avg, good, excellent
- it considers the order / rank of categories.