

Revise :-

1. KNN as a Classification & Regression Algo.
2. Templates use for Classification & Reg' Algo.
3. Evaluation Metrics of Classification Algo.
(a) Precision (d) Confusion Matrix
(b) Recall (e) Error Rate
(c) Accuracy (f) Classification Report
4. KNN on a Regression data.

Agenda :-

1. Regression Metrics in Python using scikit learn.
2. K-Means Clustering Algorithm.
3. Implementation of K Mean in python using Sklearn.

* Regression Evaluation Metrics

- (1) Mean Absolute Error
 - (2) Mean Squared Error
 - (3) Root Mean Squared Error
 - (4) R^2 Error ✓ *
 - (5) Adjusted R^2 Error. ✓ *
- } Very Imp

* Euclidean distance calc from Centroid Point to data points
 1st row

$C_3 : - (8, 4)$ $(x_1, y_1) : - (2, 4)$
 (x_2, y_2) (x_1, y_1)

$\underline{ED} : -$

$$\begin{aligned} & \sqrt{(8 - 2)^2 + (4 - 4)^2} \\ &= \sqrt{6^2 + 0^2} = \sqrt{36} = 6 \end{aligned}$$

$$(1) : - \quad \left(\begin{matrix} x_1, y_1 \\ 2, 4 \end{matrix} \right) \quad \left(\begin{matrix} x_2, y_2 \\ 2, 6 \end{matrix} \right) \quad \left(\begin{matrix} x_3, y_3 \\ 4, 7 \end{matrix} \right)$$

$$\underline{\underline{CP}} : - \quad \frac{x_1 + x_2 + x_3 + \dots}{n}, \quad \frac{y_1 + y_2 + y_3 + \dots}{n}$$

$$= \left(\frac{2+2+4}{3}, \quad \frac{4+6+7}{3} \right)$$

$$(1) = (2.67, 5.67)$$

(Ans)

Revision :-

1. Dataset looks in a Unsupervised ML Algo
(Target is absent)
2. Solved a problem using K-Means Clustering.
Bef inc a K-Means Algorithm.
3. K in K-Means represent the number of clusters.
4. Implementation of K-Means in python.

Agenda:-

1. Implementation of k-Means using sklearn in python.
2. Ideal method to select k (no. of clusters)
3. Elbow method.
5. Data preprocessing techniques.

Elbow Method :-

K = 1 \rightarrow C₁

K = 2 \rightarrow C₁, C₂

K = 3 \rightarrow C₁, C₂, C₃

K = 4 \rightarrow C₁, ..., C₄

K = 5 \rightarrow C₁, ..., C₅

K = 20 \rightarrow C₁, ..., C₂₀

WCSS :- Within Cluster
Sum of Squares

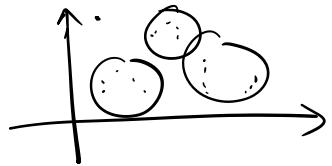
:- used for judging
the optimal k.

Since we have check for each value of k between 1 to 20,

- ① We have use a for loop
- ② $wcss = []$
- ③ Loop summing

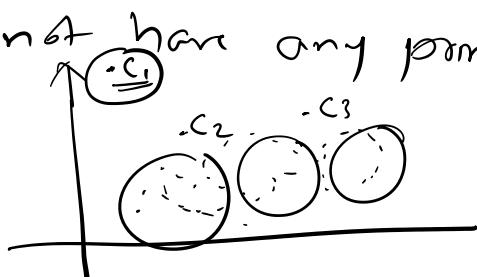
calc wcss and that will be appended to list
④ We create a visual to identify the optimal k.

Init = K - mean s++

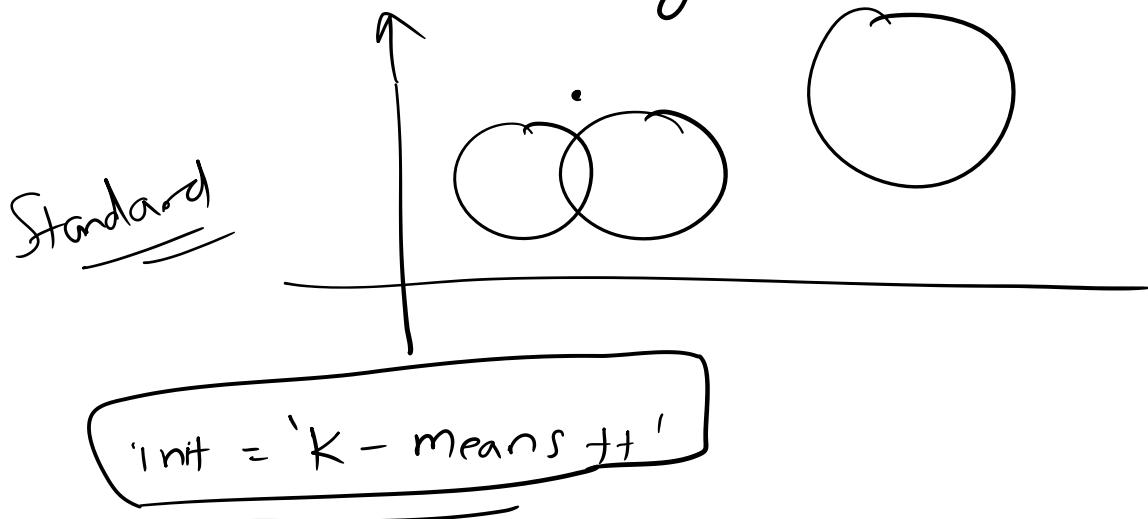


Standard KMeans Algorithm :-

1. Sensitive to the initialization of centroids or the mean points. If the centroid is very far, then that cluster will not have any point associated with it.



2. More than 1 cluster might end up linked with single same cluster.



K-means++

* K-means++ is the standard K-Means Algorithm coupled with smart initialization of centroid points.

* random_state = 42

* Simply used to make the predictions static

$$k = 1$$

133 $\xrightarrow{15}$ 132.5 or 133.5 or 134

ML Algorithms :-

1. You cannot supply a missing data
2. If any column (variable) is having higher values then ML Algo will give importance to that.
3. ML algo do not understand text data.

e.g.:- E-ID E-Name C-Qual. Age Salary Yrs. Exp. ^{Numerical}

Revision :-

1. Implementation of k-Means Algorithm in python using sklearn library
2. Elbow Method for optimal (k) no of clusters.
3. Data Preprocessing :-
 - a. Missing Value Analysis.

Agenda:-

1. Missing Value Treatment techniques.
 1. Mean, Median, Mode, Forward, Backward fill.
 2. Sklearn method for Missing Value imputation.
2. Data Encoding Techniques.

(1) Missing Value Techniques:-

(1) Dropping the rows :-

data =

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40	55000	Yes
France	35	58000	Yes
Spain	32	52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

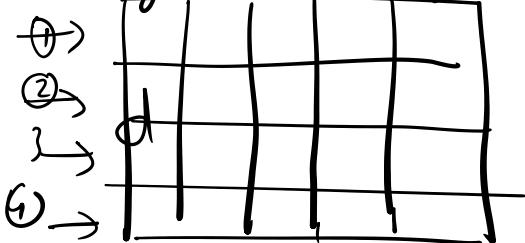
data.dropna()

The entire row gets deleted
 201.

(2) Dropping the column :-

data.drop('Age', axis=1)

* Axis : ↓ ↓ ↓ Data is
Age Sex Non-Gender Classification.



2-D data.

data.drop('Sex')

data.drop('Sex', axis=1)

data.drop('Gender', axis=1)

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40	0	Yes
France	35	58000	Yes
Spain	0	52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Age :- Replace all missing with mean of Age.

Salary :- Replace all missing values with mean of Salary

Mode: Most-frequent value in the dataset

e.g.: [^①Apple, Orange, ^②Apple, Mango, Banana, ^③Apple]

Apple - 3 ← Apple is the mode

Orange - 1

Mango - 1

Banana - 1

data

① Missing Value Analysis.

② It does not accept Categorical (String) data.



Convert Cate data to Num. data



Data Encoding

Data Encoding :- Applicable only on
Categorical data

Nominal (No order)

Ordinal
(Specific Order)

~~Label Encoding~~

One Hot
Encoding

~~One Hot Encoding~~
~~Label Encoding~~

Label Encoding :- For ~~Nominal~~ ^{Ordinal} data.

e.g:-

Country Column in the Nominal data

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Country	Age	Salary	Purchased
0	44	72000	No
1	27	48000	Yes
2	30	54000	No
1	38	61000	No
2	40		Yes
0	35	58000	Yes
1		52000	No
0	48	79000	Yes
2	50	83000	No
0	37	67000	Yes

Label
Enc.

France - 0

Spain - 1

Germany - 2

One-Hot-Encoding :-

~~Ordinal~~ ~~Nominal~~ data

$\left\{ \begin{array}{l} \text{BS} - \text{Bachelor} \\ \text{MS} - \text{Master} \\ \text{PHD} - \text{Doct.} \end{array} \right.$

Country	Education	Age	Salary	Purchased
France	BS	44	72000	No
Spain	MS	27	48000	Yes
Germany	BS	30	54000	No
Spain	MS	38	61000	No
Germany	PHD	40		Yes
France	PHD	35	58000	Yes
Spain	MS		52000	No
France	BS	48	79000	Yes
Germany	PHD	50	83000	No
France	BS	37	67000	Yes

OHE

Country	Education_MS	Education_PHD	Age	Salary	Purchased
France	0	0	44	72000	No
Spain	1	0	27	48000	Yes
Germany	0	0	30	54000	No
Spain	1	0	38	61000	No
Germany	0	1	40		Yes
France	0	1	35	58000	Yes
Spain	1	0		52000	No
France	0	0	48	79000	Yes
Germany	0	1	50	83000	No
France	0	0	37	67000	Yes

↳ Education / BS

1
 0
 1
 0
 0

↳ Education - MS Education - PHD
 0
 1
 0
 1
 0
 0
 1

OHE increases
 the no. of
 columns in
 the data.

* Whatever we learnt about Label Encoding & One Hot, it is actually is "wrong".

* It is Wrong, because it intensional.

Label
enc

Country	Age	Salary	Purchased
0	44	72000	No
1	27	48000	Yes
2	30	54000	No
1	38	61000	No
2	40		Yes
0	35	58000	Yes
1		52000	No
0	48	79000	Yes
2	50	83000	No
0	37	67000	Yes

Country	Education_Ms	Education_PHD	Age	Salary	Purchased
France	0	0	44	72000	No
Spain	1	0	27	48000	Yes
Germany	0	0	30	54000	No
Spain	1	0	38	61000	No
Germany	0	1	40		Yes
France	0	1	35	58000	Yes
Spain	1	0		52000	No
France	0	0	48	79000	Yes
Germany	0	1	50	83000	No
France	0	0	37	67000	Yes

OH Encoding

Country	Education_Ms	Education_PHD	Age	Salary	Purchased
0	0	0	44	72000	No
1	1	0	27	48000	Yes
2	0	0	30	54000	No
1	1	0	38	61000	No
2	0	1	40	40	Yes
0	0	1	35	58000	Yes
1	1	0		52000	No
0	0	0	48	79000	Yes
2	0	1	50	83000	No
0	0	0	37	67000	Yes

Ordinal

Education :-

eg:-

BS	- 0	2	1	0	Label Enc
MS	- 1				
IT Internship	PHD - 2				
B.Tech					

M.Tech	- 1
Phd	- 0

France - 0
Spain - 1
Germany - 2

Nominal ← One hot Enc

Ind Ctg
B.Tech - 0
M.Tech - 1
Phd - 2
Label Enc

New Table after proper Label encoding

Original data

One Hot encoding & Final table after OHE & LE

Country	Education	Age	Salary	Purchased
France	BS	44	72000	No
Spain	MS	27	48000	Yes
Germany	BS	30	54000	No
Spain	MS	38	61000	No
Germany	PHD	40		Yes
France	PHD	35	58000	Yes
Spain	MS		52000	No
France	BS	48	79000	Yes
Germany	PHD	50	83000	No
France	BS	37	67000	Yes

For Nominal → OHE

For Ordinal → LE

Country_France	Country_Spain	Education	Age	Salary	Purchased
1	0	0	44	72000	No
0	1	1	27	48000	Yes
0	0	0	30	54000	No
0	1	1	38	61000	No
0	0	2	40		Yes
1	0	2	35	58000	Yes
0	1	1		52000	No
1	0	0	48	79000	Yes
0	0	2	50	83000	No
1	0	0	37	67000	Yes



Color - Red	Color - Yellow	Color - Green
1 0	0 1	0 0
0 0	0 0	0 0
0 0	0 1	0 0
0 0	0 0	0 0

*Feature	Scaling :-	Standardization
<u>Age</u> A	<u>Salary</u> β	$[-3, +3]$
R ₁ 25	55000	
R ₂ 35	65000	
R ₃ 40	75000	
R ₄ 45	80000	
R ₅ 50	90000	

Give importance

Standardization formula

$$= \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

eg:- C61 A $\frac{40 - 25}{50 - 25} = \frac{15}{25} = 0.6$

$$\frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

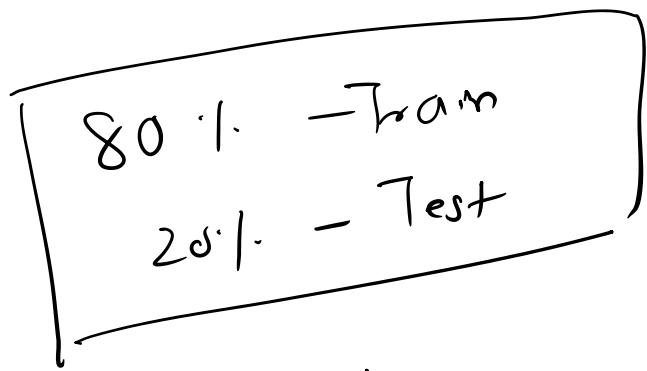
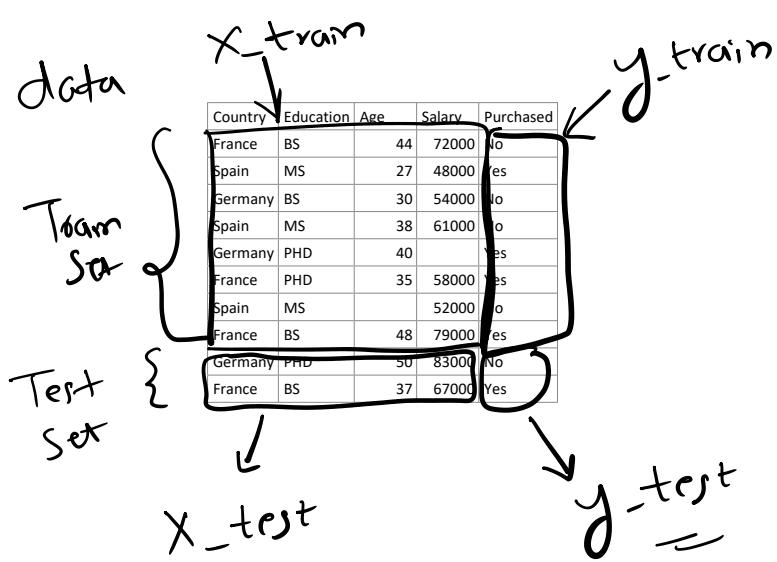
For Col B,

$$\frac{80000 - 55000}{90000 - 55000} = \underline{\underline{0.55}}$$

In Machine Learning Projects, Data preprocessing steps

- ① Missing Value Imputations
- ② Data Encoding
- ③ Divide the data into X & y
- ④ Split the data using train test split
- ⑤ Feature Scaling on X_{train} & X_{test} .

* Train Test Split :-



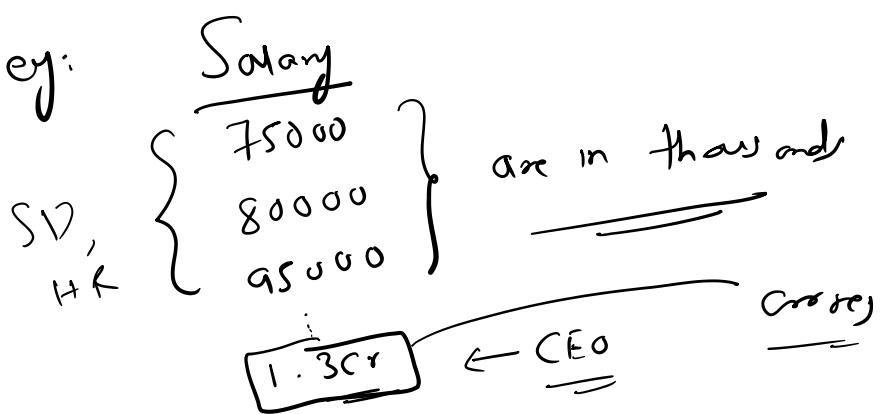
data. shape

(10, 5)

* Outlier Analysis :- Extreme value in any column are outliers.

* We check outliers only in Numerical data.

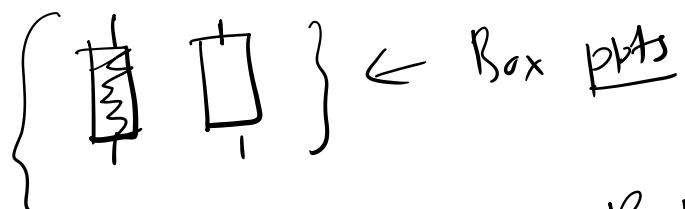
e.g.



e.g. Sachin Tend.

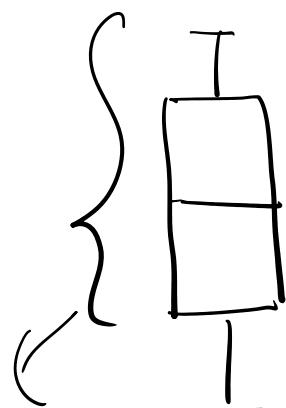
* Outlier Check :- Most popular method is

★ Box plots :-



Box plots

Relation \leftrightarrow Stick \leftarrow Chart



* Mean of the dataset is affected by outliers