FORMERLY Industries Helping Hands

THE iSCALE

Data Science | 30 Days of Machine Learning | Day - 15

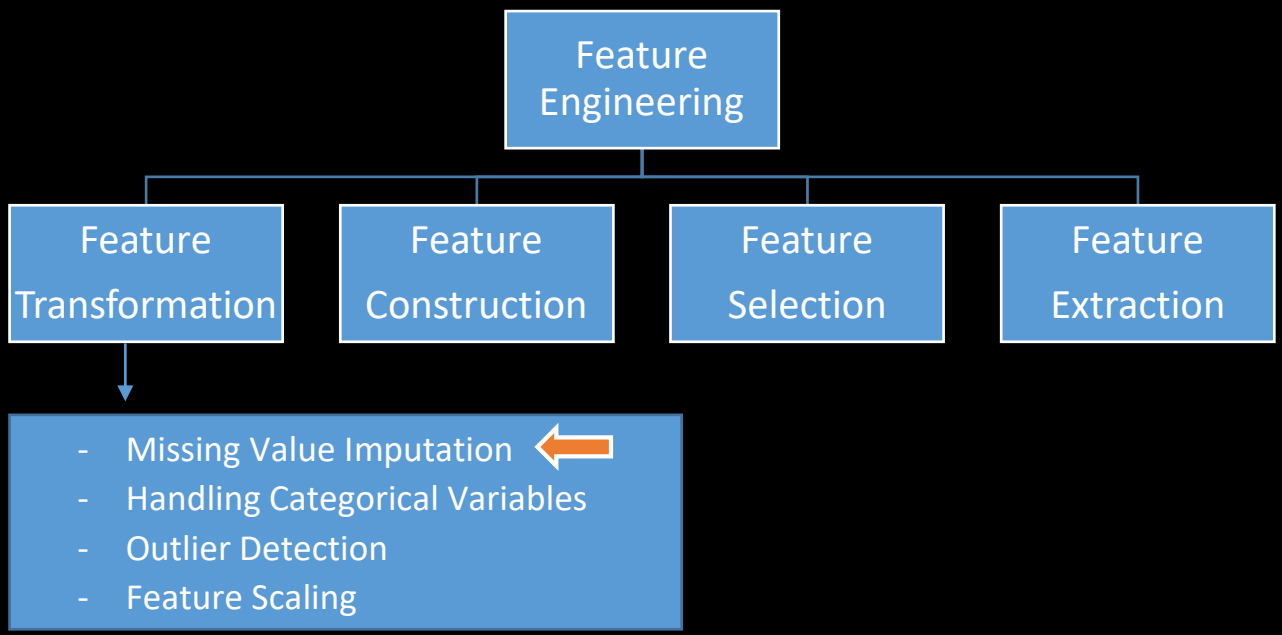
Educator Name: Nishant Dhote
Support Team: **+91-7880-113-112**

----Today Topics | Day 15----

Feature Engineering (Missing Value Imputation)

- KNN Imputer
- K-Nearest Neighbour Calculation Method
- What is Euclidean Distance in Machine Learning?
- How to find K nearest neighbour?
- Find missing imputation value?

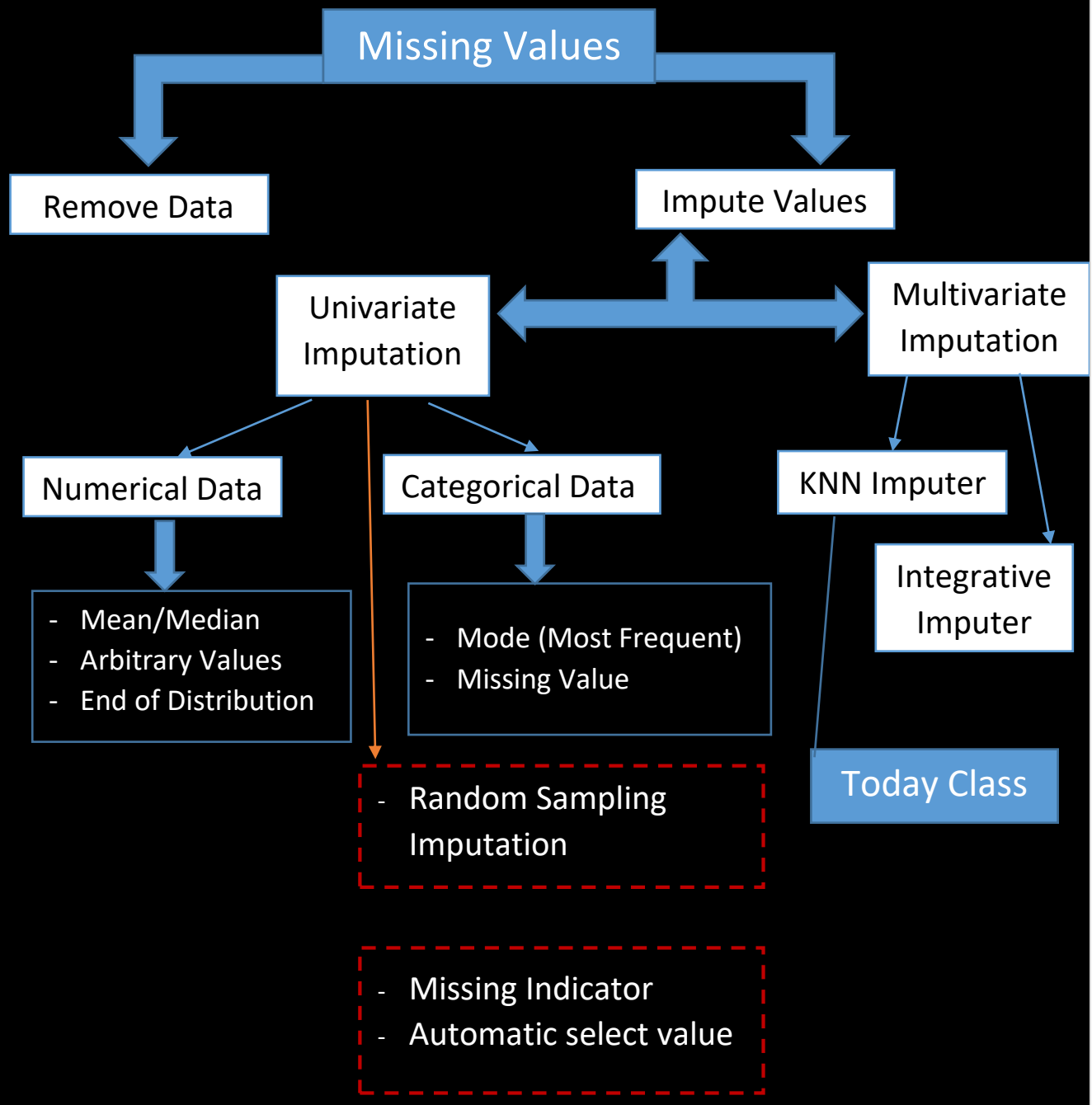
Dataset Link GitHub: https://github.com/TheiScale/30_Days_Machine_Learning/



FORMERLY Industries Helping Hands

THE iSCALE

Today's Topics:



- **KNN Imputer:** KNN imputer is a scikit-learn class used to fill out or predict the missing values in a dataset. It is a more useful method which works on the basic approach of the KNN algorithm rather than the naive approach (straightforward attempt to solve solution) of filling all the values with mean or the median. In this approach, we specify a distance from the missing values which is also known as the K parameter. The missing value will be predicted in reference to the mean of the neighbours.

S.No	Variable 1	Variable 2	Variable 3	Variable 4
1	28	--	48	22
2	--	40	37	24
3	34	22	55	26
4	26	--	30	--
5	50	20	49	--

K-Nearest Neighbour: K= The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.

sklearn.metrics.pairwise.nan_euclidean_distances

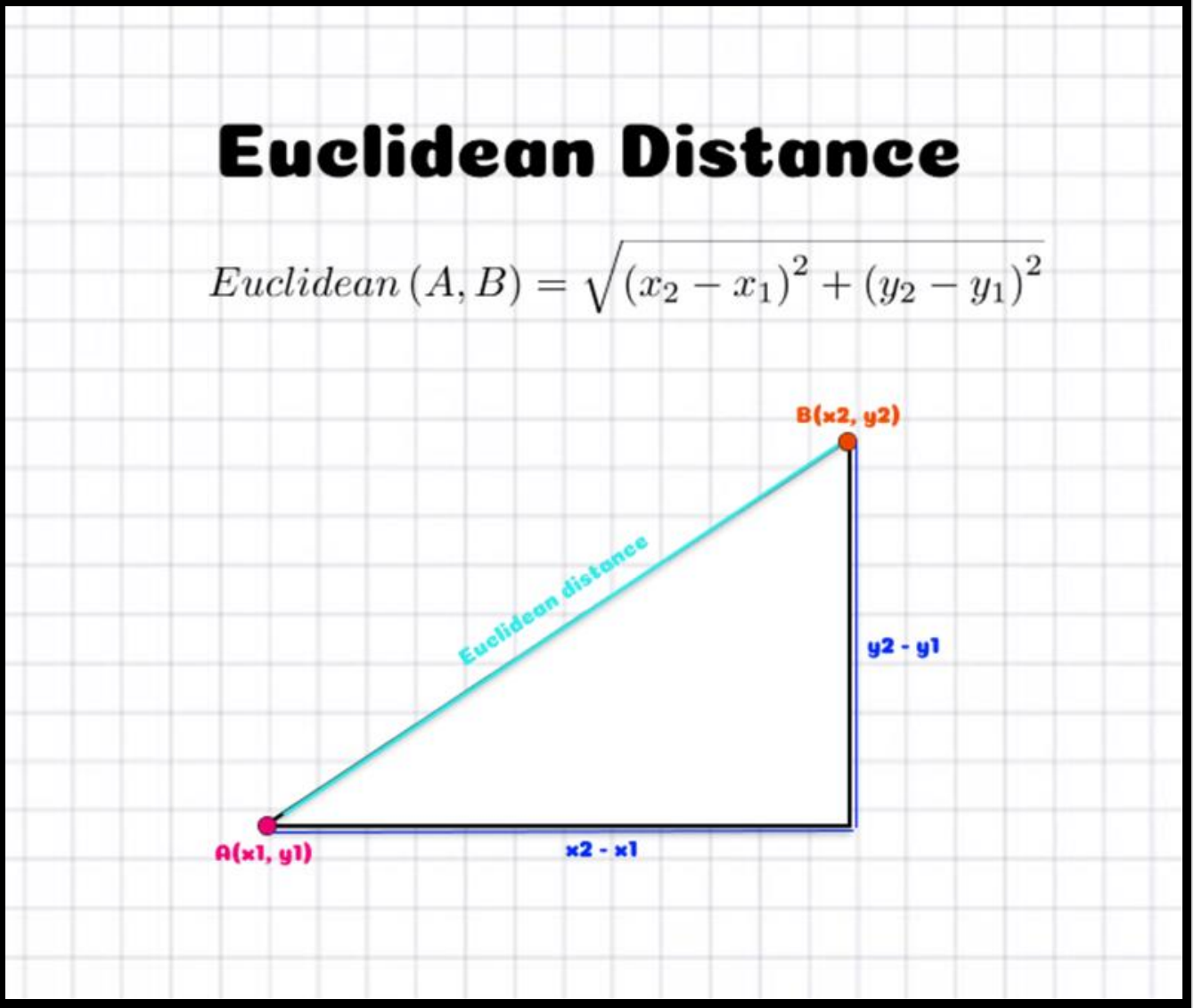
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.nan_euclidean_distances.html#sklearn.metrics.pairwise.nan_euclidean_distances

dist(x,y) = sqrt(weight * sq. distance from present coordinates) where

weight = Total # of coordinates / # of present coordinates

What is Euclidean Distance in Machine Learning?

Euclidean distance is used in many machine learning algorithms as a default distance metric to measure the similarity between two recorded observations. However, the observations to be compared must include features that are continuous and have numeric variables like weight, height, salary, etc.



2 Step We follow:

1. Find "K" nearest neighbour?
2. Find the value?

Example 1: Calculation between Row 1 and Row 2

S.No	Feature 1	Feature 2	Feature 3	Feature 4
1	28	--	48	22
2	--	40	37	24
3	34	22	55	26
4	26	--	30	--
5	50	20	49	--

For this Value ?

Non-Eclidian = ?

$$= \sqrt{w_j \times (40-?)^2 + (37-49)^2 + (24-22)^2}$$

$$w_j = \frac{\text{total No of Coordinate}}{\text{No of Present Coordinate}} = \frac{3}{2}$$

$$= \sqrt{\frac{3}{2} \times (-11)^2 + (2)^2}$$

$$= \sqrt{1.5 \times [121 + 4]} = \sqrt{1.5 \times 125}$$

$$= \sqrt{187.5}$$

$$= 13.69$$

Example 2: Calculation between Row 2 and Row 3

S.No	Feature 1	Feature 2	Feature 3	Feature 4
1	28	--	48	22
2	--	40	37	24
3	34	22	55	26
4	26	--	30	--
5	50	20	49	--

Nan E. -> Euclidean Distance

$$= \sqrt{\frac{3}{3} \times \left[(22-40)^2 + (55-37)^2 + (26-24)^2\right]}$$

$$= \sqrt{1 \times \left[(-18)^2 + (18)^2 + (2)^2\right]}$$

$$= \sqrt{324 + 225 + 4} = \sqrt{553}$$

$$= 23.5$$

→ Row 2, Row 3

Example 3: Calculation between Row 2 and Row 4

S.No	Feature 1	Feature 2	Feature 3	Feature 4
1	28	--	48	22
2	--	40	37	24
3	34	22	55	26
4	26	--	30	--
5	50	20	49	--

$\Rightarrow x = ?$

$$= \sqrt{wg \times \left((x-40)^2 + (30-37)^2 + (x-24)^2 \right)}$$

$$= \sqrt{\frac{3}{1} \times 49}$$

$$= \sqrt{147}$$

$$= 12.12$$

Example 4: Calculation between Row 2 and Row 5

S.No	Feature 1	Feature 2	Feature 3	Feature 4
1	28	--	48	22
2	--	40	37	24
3	34	22	55	26
4	26	--	30	--
5	50	20	49	--

$nc = 2$

$$= \sqrt{w_2 \times [(20-40)^2 + (49-37)^2 + \alpha}$$

$$= \sqrt{\frac{3}{2} \times [400 + 144]}$$

$$= \sqrt{1.5 \times 544}$$

$$= \sqrt{816} = 28.5$$

FORMERLY Industries Helping Hands

THE iSCALE

All 4 Euclidean Distance Example:

S.No	Feature 1	Feature 2	Feature 3	Feature 4
1	28	--	48	22
2	--	40	37	24
3	34	22	55	26
4	26	--	30	--
5	50	20	49	--

Example → (1) (Row 1 & 2)

$$= \sqrt{\frac{3}{2}(37-48)^2 + (24-22)^2}$$

$$= \sqrt{1.5 \times [(-11)^2 + (2)^2]}$$

$$= \sqrt{1.5 \times [121+4}$$

$$= \sqrt{185} \rightarrow 13.69$$

Example (2) (Row 2 & 3)

$$= \sqrt{\frac{3}{3} \times (22-40)^2 + (55-37)^2 + (26-24)^2}$$

$$= \sqrt{(-18)^2 + (18)^2 + (2)^2}$$

$$= \sqrt{324 + 225 + 4} = 23.5$$

Example - 3 (Row 2 & 4)

$$= \sqrt{\frac{3}{1} \times (30-37)^2}$$

$$= \sqrt{3 \times 49}$$

$$= \sqrt{147}$$

$$= 12.12$$

Example-4 (Row 2 x 5)

$$= \sqrt{\frac{3}{2}[(20-40)^2 + (49-37)^2}$$

$$= \sqrt{1.5[400 + 144]}$$

$$= \sqrt{1.5 \times 544} = \sqrt{816}$$

$$= 28.5$$

S.No	Feature 1	Feature 2	Feature 3	Feature 4
1	28	--	48	22
2	--	40	37	24
3	34	22	55	26
4	26	--	30	--
5	50	20	49	--

2 Step We follow:

Find "K" nearest neighbour?
Find the value?

v

If = k = 2

Row 1 → 13.69

Row 4 → 12.12

$\frac{28+26}{2} = \frac{54}{2} = 27$ → missing value

K = 3

Row 1 → 13.69

Row 4 → 12.12

Row → 23.5

$\frac{28+26+34}{3} = 29.3$

<Start-Coding>

#Import Library

```
import numpy as np
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.impute import KNNImputer,SimpleImputer
from sklearn.linear_model import LogisticRegression

from sklearn.metrics import accuracy_score
```

#Import Dataset

```
df =
pd.read_csv('train.csv')[['Age','Pclass','Fare','Survi
ved']]

----
df.sample(10)
```

#Check Missing Value

```
df.isnull().mean() * 100
```

#Define X & Y

```
X = df.drop(columns=['Survived'])
y = df['Survived']
```

#Train Test Split

```
X_train,X_test,y_train,y_test =
train_test_split(X,y,test_size=0.2,random_state=2)

----

X_train
```

#Apply KNN Imputer

```
knn = KNNImputer(n_neighbors=1,weights='distance')

X_train_trf = knn.fit_transform(X_train)
X_test_trf = knn.transform(X_test)
```

#Convert in Data Frame

```
pd.DataFrame(X_train_trf,columns=X_train.columns)
```

#Apply Logistic Regression

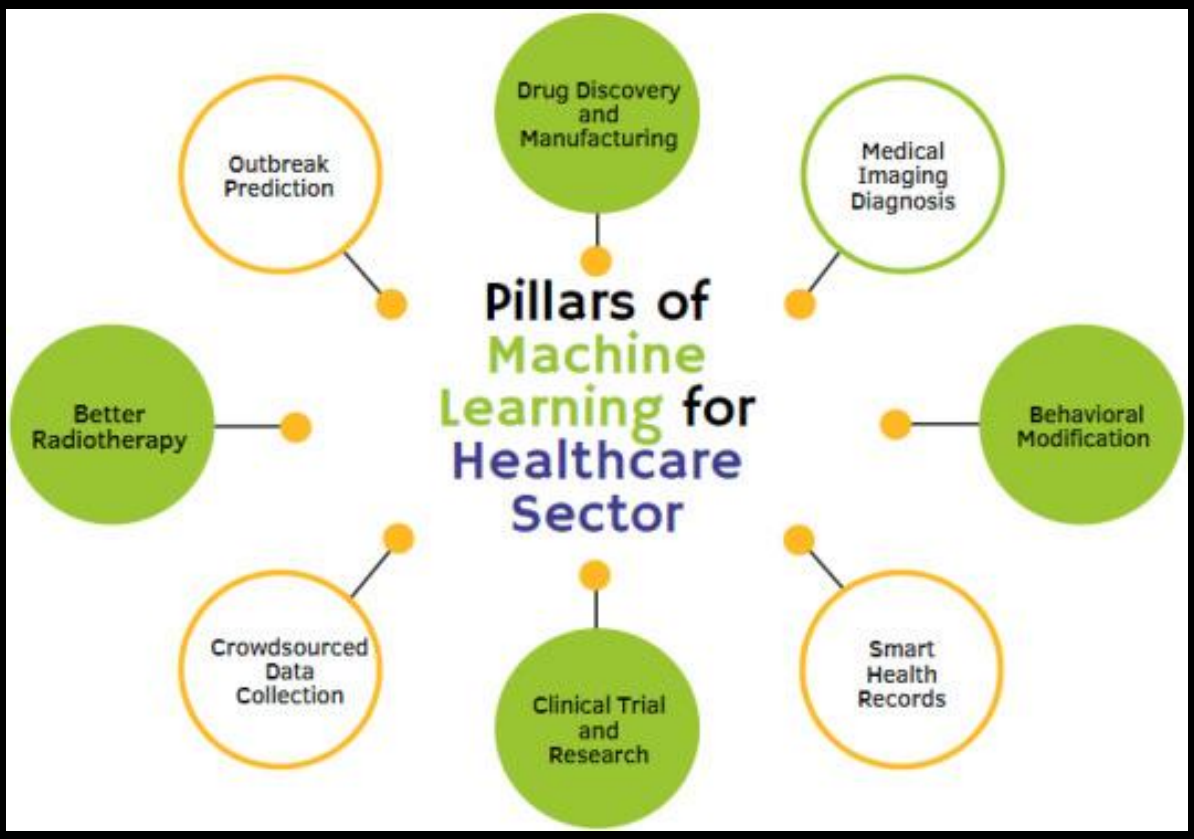
```
lr = LogisticRegression()

lr.fit(X_train_trf,y_train)

y_pred = lr.predict(X_test_trf)

accuracy_score(y_test,y_pred)
```

FORMERLY Industries Helping Hands

THE iSCALE

Day 15: Curious Data Minds

What is a machine learning role in healthcare sector?



Read Blog: https://builtin.com/artificial-intelligence/machine-learning-healthcare

https://www.bluebash.co/blog/ai-healthcare-future-trends-2024/