Sunday, April 28, 2024 11:25 AM

Data Handling

⇒ Missing Values → > 50% → ignore (drop) column (numerical column)

- not skewed → mean
- skewed → median

categorical column → mode

⇒ Categorical column

exg Gender

Male

Female

(I) if it is nominal category → dummy variables

a) → get_dummies( ) → pandas

b) → Onehot encoding → Sklearn.preprocessing

(II) if it is a ordinal category → label encoding

a) for input → Ordinal Encoder

b) for o/p → label encoder.

⇒ Feature Scaling

1	2	3
0.1	500	10000
0.2	600	30000
0.3	700	40000

to deal with this → Scaling

- Standardization (z-score) ($\mu=0$, $\sigma=1$)
- Normalization (0 − 1)

Sklearn preprocessing

- StandardScaler (Standardization)
- MinMaxScaler (Normalization)

Bollywood - dataset → Assignment

(I) Perform feature Scaling Numerical columns

⇒ Read doc & blog

(II) Make all columns numerical in nature.

⇒ Class imbalance

↳ resampling → Sklearn utils

CGPA	IQ	Placed
10	125	1
9	120	1
8	115	1
7	125	1
6	110	0
6	110	0
6	110	0
6	110	0