ML Preprocessing

UTKARSH GAIKWAD

CLASS STARTING SHARP AT 4:05 PM
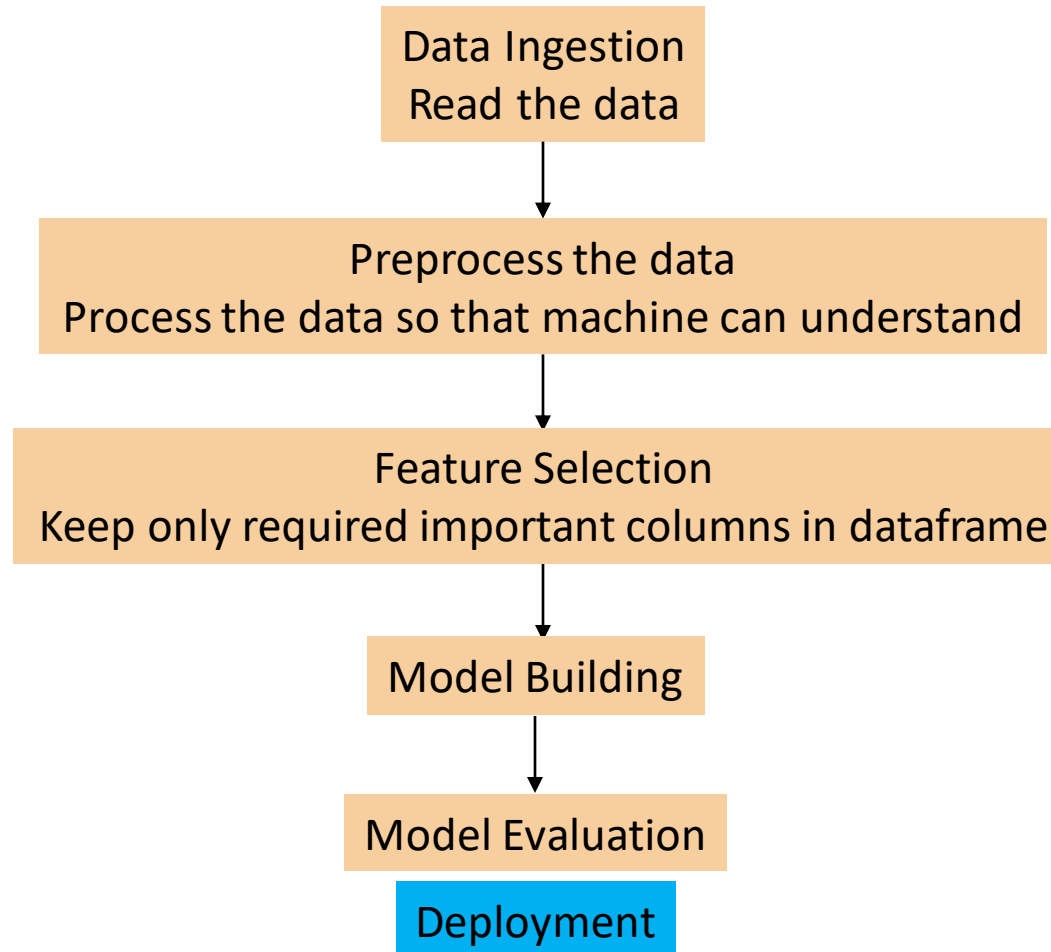
Goal of Data Preprocessing

➢ Main Purpose of Data Preprocessing is to prepare the data for machine to understand
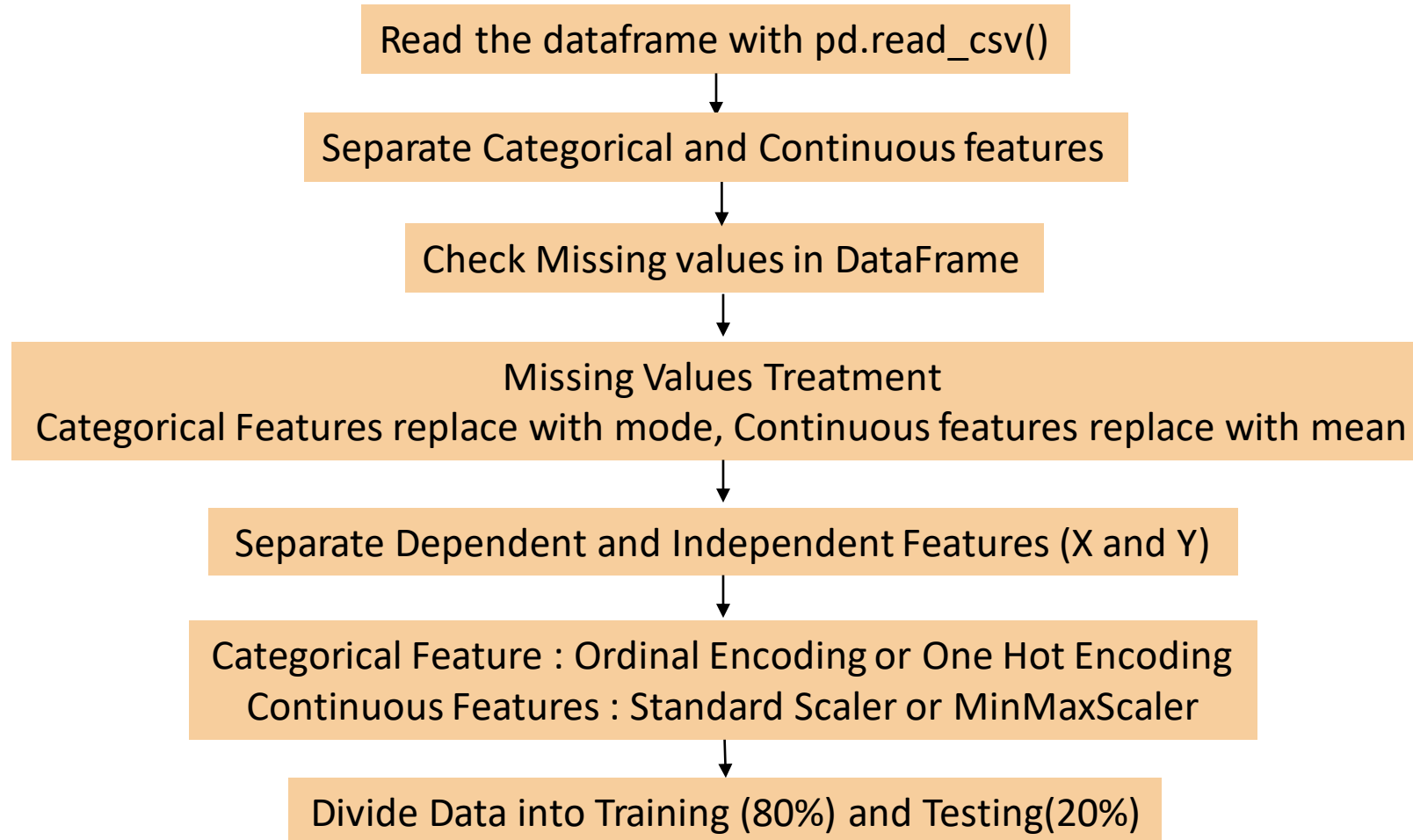
I cannot understand raw data directly
Pre-process the data for me to understand

Machine Learning Process

Data Ingestion
Read the data

Preprocess the data
Process the data so that machine can understand

Feature Selection
Keep only required important columns in dataframe

Model Building

Model Evaluation

Deployment

Basic Steps in creating a Data Preprocessing

Read the dataframe with pd.read_csv()

Separate Categorical and Continuous features

Check Missing values in DataFrame

Missing Values Treatment
Categorical Features replace with mode, Continuous features replace with mean

Separate Dependent and Independent Features (X and Y)

Categorical Feature : Ordinal Encoding or One Hot Encoding
Continuous Features : Standard Scaler or MinMaxScaler

Divide Data into Training (80%) and Testing(20%)

Machine does not understand text directly

Because machine does not understand Text directly we can use 2 approaches to convert categorical features to Numeric data : Label Encoding or One Hot Encoding

Label Encoding Ordinal Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

$\rightarrow$

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

S, M, L, XL ,....

A+ , B+ , AB-,

Need for Scaling of continuous data

Age	Income	Purchase
25	50,000	1,000
30	60,000	2,000
35	70,000	3,000
40	125,000	4,000
45	150,000	5,000

$$Purchase = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Income + \epsilon$$

Higher Error possible

β_2 value is larger compared to Age coefficient

β_1 value becomes lower because of Age is smaller

2 methods to bring down data in same scale (Independent features)

Feature scaling

Normalization   Standardization

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$ $$X' = \frac{X - Mean}{Standard\ deviation}$$

MinMaxScaler
0-1

StandardScaler
-3 to 3

MinMax Scaler

$$Age_{scaled} = \frac{age - \min(age)}{\max(age) - \min(age)} = \frac{age - 25}{45 - 25} = \frac{age - 25}{20}$$

$$Income_{scaled} = \frac{income - \min(income)}{\max(income) - \min(income)} = \frac{income - 50000}{150000 - 50000} = \frac{income - 50000}{100000}$$

Age	Income	Purchase
25	50,000	1,000
30	60,000	2,000
35	70,000	3,000
40	125,000	4,000
45	150,000	5,000

MinMaxScaler →

Age	Income	Purchase
0	0	1,000
0.25	0.1	2,000
0.50	0.2	3,000
0.75	0.75	4,000
1	1	5,000

Standard Scaler - Z scores

$$Z_{age} = \frac{age - mean(age)}{stdev(age)} = \frac{age - 35}{7.9057}$$

Converts all data in
Mean = 0
Stdev = 1

$$Z_{income} = \frac{income - mean(income)}{stdev(income)} = \frac{income - 91000}{43931.7653}$$

Age	Income	Purchase
25	50,000	1,000
30	60,000	2,000
35	70,000	3,000
40	125,000	4,000
45	150,000	5,000

StandardScaler →

Age	Income	Purchase
-1.2649	-0.9333	1,000
-0.6325	-0.7056	2,000
0.0000	-0.4780	3,000
0.6325	0.7739	4,000
1.2649	1.3430	5,000

Thank You

PING ME ON SKYPE FOR ANY QUERIES