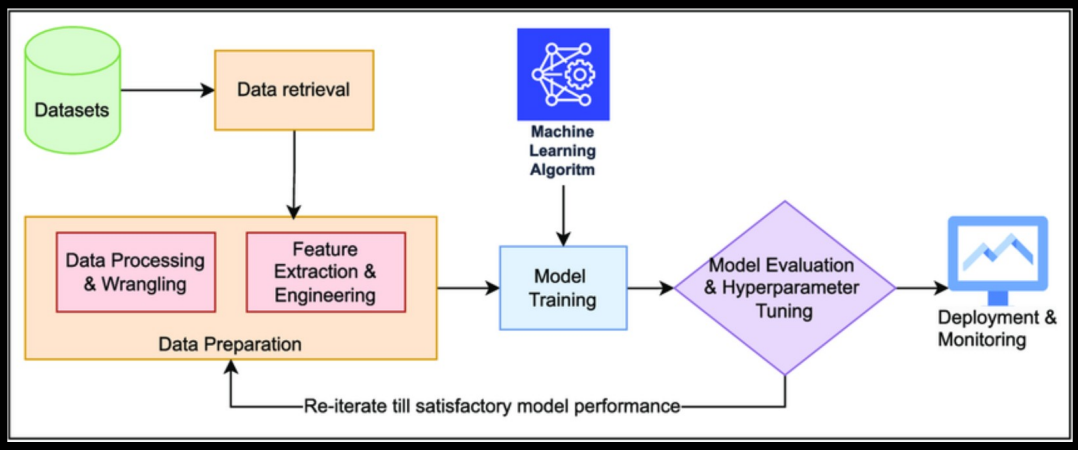
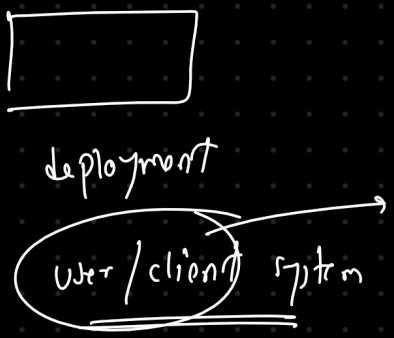
24-08-2025

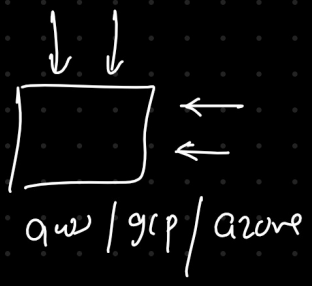
Agenda: feature engineering - 1



sets of step in a sequence : pipeline

website :





Feature engineering :-

→ Handle missing data → None, Null, nan
→ Handling imbalance data
→ Handling outliers
→ Feature scaling
→ Data Encoding
→ Feature selection (Heatmap)

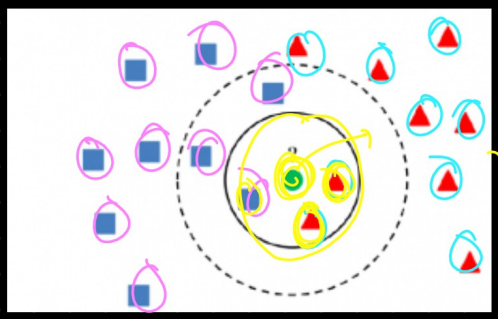
→ Handling missing data
→ Handling outliers

House price dataset: kaggle

Regression: predict numbers (continuous), 27.3, 27.4
classification: labels, categories, x, y, z → tall / pan

Handling missing values

- Delete
- Impute



mean → all the null value have the same value
median →
mode →

100,000 → 30K → null →

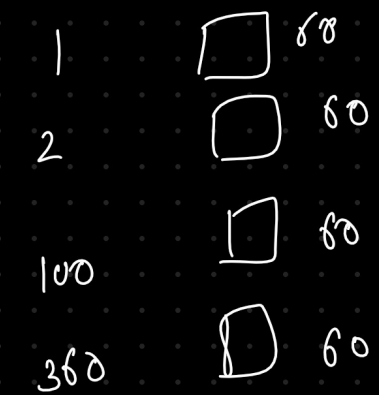
	Attendance (365)	Marks (100)
A	365	100
B	1	1
C	100	20
D	200	60
E	250	80
F	10	5

R1 → 1 - 10
R2 → 10 - 30
R3 → 30 - 50
R4 → 50 - 70
R5 → 70 - 100

principal :-

260 → R5

60 → mean) → scale → 60



4 group → min, max

A → min, max
B → min, max
C → min, max
D → min, max

To revise!

→ pandas → list comprehension
→ numpy → def, lambda
→ stats (mean, median, mode, pdf, pmf, distribution overview)