Imbalance

① Split → Stratify = y

② Set Suitable Metric : Recall , precision , f1-score
AUC Score

→ ③ use suitable handling (Resampling)
↳ SMOTE / oversample / undersample

④ use Decision Tree or Ensemble Methods
(less sensitive for imbalance)

⑤ use class_weight (if applicable)

Majority

0 : 500

1 : 5

Imbalance

Minority

1 : 100

1 : 3 (light)

1. undersampling → 0 : 5, 1 : 5 (Random Sampling)

2. oversampling → 0 : 500, 1 : 500 (Resampling with Replacement)

3. SMOTE → 0 : 500, 1 : 500 (synthetic Oversampling)

* Ensemble Methods :

Randomforest → Bagging
↳ Base : DT

- LR, SVR, Knn, DT → Single learner (weak)
- Ensemble → Strong learner

① Voting

Same Dataset
Diff Models

[SVR, LR, DT]

→ Reg : Avg

→ Clf : → hard voting → voting
↳ Soft ~ → Avg (prob)

② Bagging

→ Same Model
Diff Subset ← ③ Boosting
of Data

→ parallel

→ Randomforest

③ Boosting

→ Sequential

→ Xgboost (learning_rate)

→ Gradient Boosting

* Dimension Reduction

(# Rows , # Cols)

↑ observations ↑ features

features → feature Selection

features → feature Extraction → Compression

- Curse of Dimensionality

① Stange

② Time / Response

③ Visualization

Age income

original space $\boxed{f_1 \quad f_2 \quad f_3 \ \_ \ f_{100}}$

⇓ linear Transformation

New space $\boxed{Pc_1 \quad PC_2 \quad PC_3 \ .. \ PC_{100}}$ ordered by variance

↑ indicator for information

Example 50% 20% 15% → select $\boxed{PC1 , PC2 , PC3}$ 85%