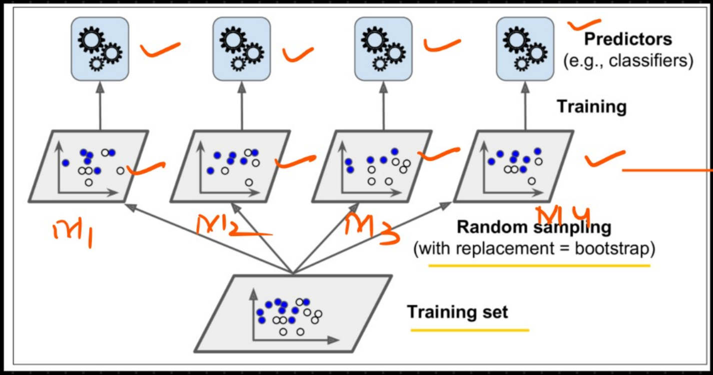
Random forest → Ensemble Technique

↳ Bagging (Bootstrapped Aggregation)



→ Decision Tree

↳ Random forest

Bagging classifier (algorithm)

↱ Real time industry project

Advantage (RF)

- ↳ avoid overfitting (DT) ①
- ↳ increase accuracy ②

M1 — Decision Tree { classification / Regression

↓

overfitting

↳ Pruning

n_estimators = 5, subset of data (Randomly) → by replacement

M1 — DT1 — 0 — 2cr

M2 — DT2 — 0 — 1cr

M3 — DT3 — 0 — 3cr

M4 — DT4 — 1 — 2.5cr

M5 — DT5 — 0 — 4cr

→ Majority voting ⇒ 0

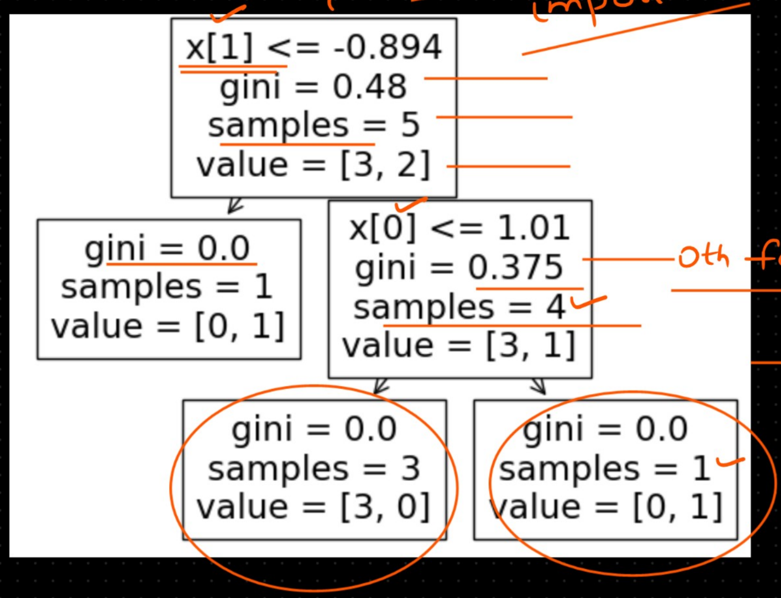
Regression

↳ average of all the given predictions

(2cr + 1cr + 3cr + 2.5cr + 4cr)/5

$$\eta_i \Rightarrow \frac{N\_t}{N}\left(\underline{impurity} - \underbrace{\left(\frac{N\_t\_r}{N\_t} * right\ impurity\right)}_{right\ substree} - \underbrace{\left(\frac{N\_t\_l}{N\_t} * left\ impurity\right)}_{left\ substree}\right)$$

1st feature importance



0th feature importance

$$\eta_1 \Rightarrow \frac{5}{5}\left(0.48 - \frac{4}{5} * 0.375 - \frac{1}{5} * 0\right)$$

$$= 0.18$$

$$\eta_0 \to \frac{4}{5}\left(0.375 - \overset{0}{\frac{1}{4} * 0} - \overset{0}{\frac{3}{4} * 0}\right) = 0.30$$

$$\frac{4}{5} * 0.375$$

Normalization

$$f_0 \to \frac{0.30}{0.30 + 0.18} = 0.625$$

$$f_1 \to \frac{0.18}{0.18 + 0.30} = 0.375$$

$0.48 +$

$= 1 \longleftarrow$

Disadvantage

① More training time

② Interpretability is bit difficult

Customer → { Loan — Yes (1)

Loan — No (0)

↪ ??

③

Advantage → feature importance

↪ Decision Tree

Random forest

feature selection

clf. feature_importances

f_1	f_2 ✓	f_3 ✓	f_4	f_5	target (Loan)
					0
					1
					0
					0
					1