

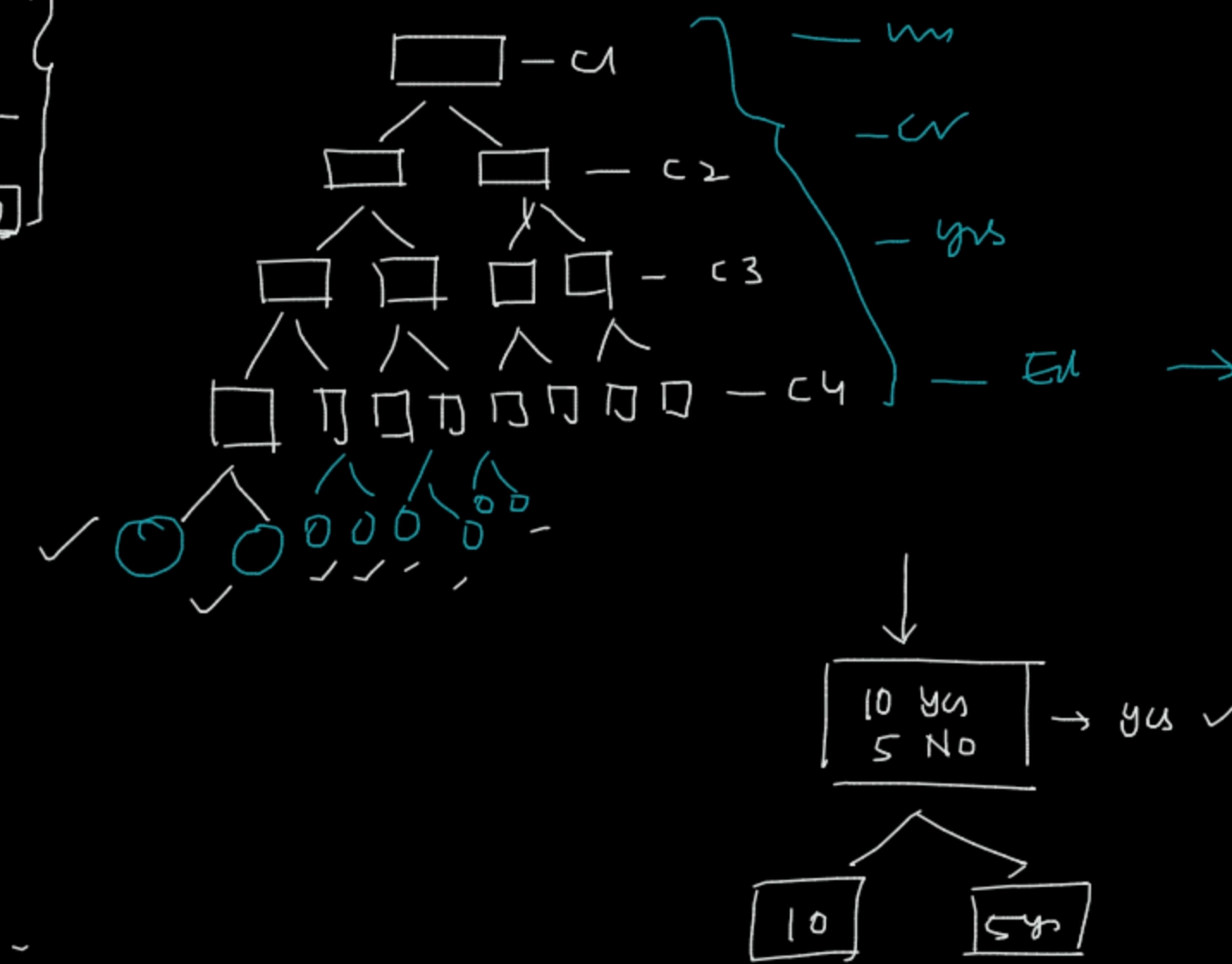
Training Data \rightarrow Extremely well

Testing Data \rightarrow Badly

\uparrow
 $\rightarrow [5 \text{ ppl}]$
 $\underline{3 \text{ Yes}} \rightarrow \text{Yes-} \checkmark$
 $\underline{2 \text{ No}}$

stopping criteria

- 1 max_depth = 4 ✓ }
- 2 max_leafn = 10 ← }
- 3 min_samples = [10]



1 stopping criteria

2 Entropy or Gini Impurity



C5.0

↳ CART

Classification

Entropy & IG $\rightarrow \checkmark$ \leftarrow splitting done
Gini Impurity $\rightarrow \checkmark$

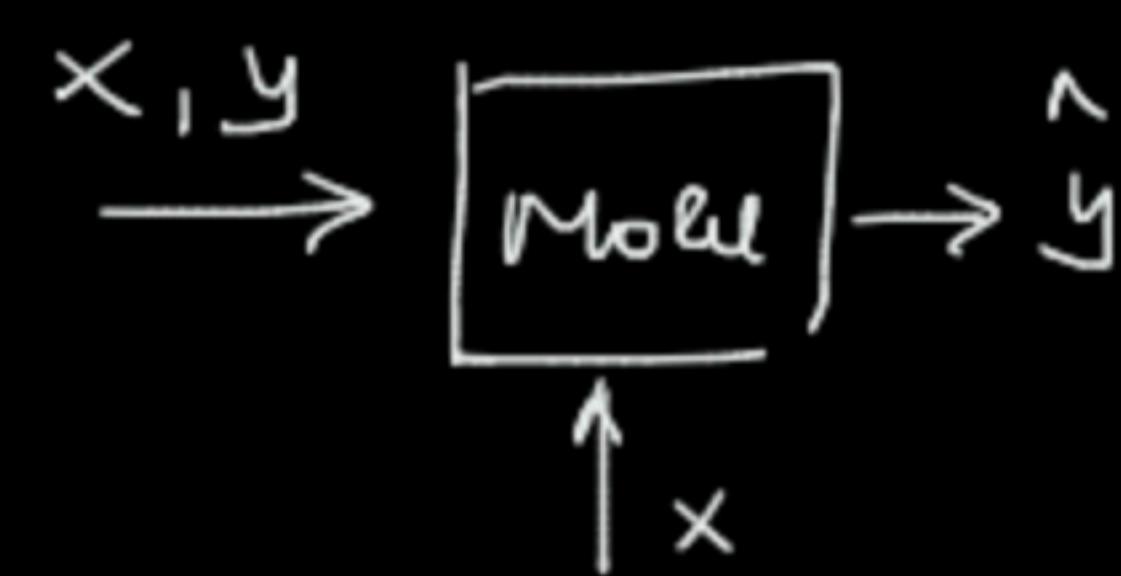
Regression

MSE $\rightarrow \leftarrow$ splitting done



Ensemble Model

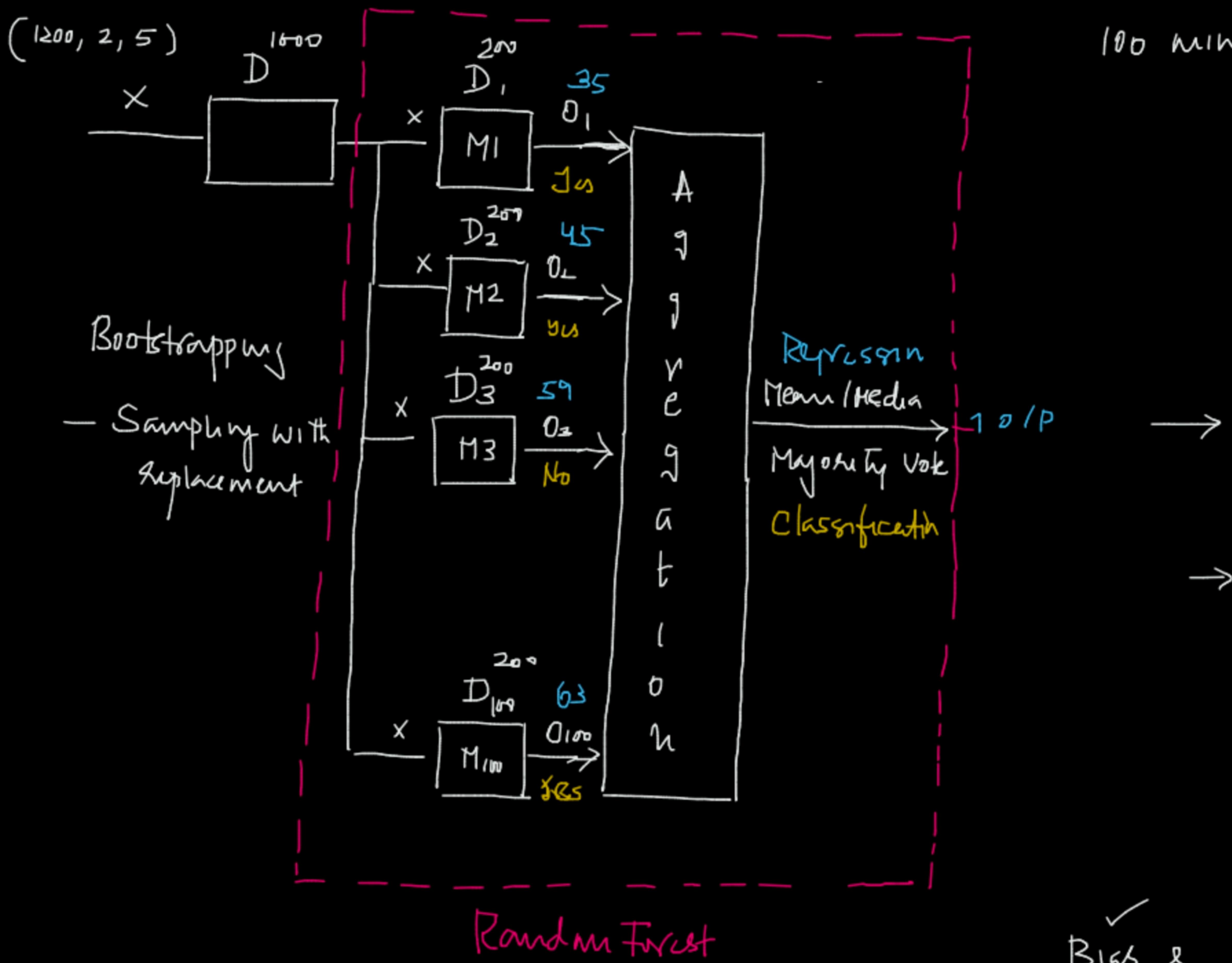
- Collection of models
- Combining multiple models



- (a) Bagging ✓ → Technique — Random Forest
- (b) Boosting ✓ → Technique — AdaBoost, XGBoost
- (c) stacking
- (d) cascading

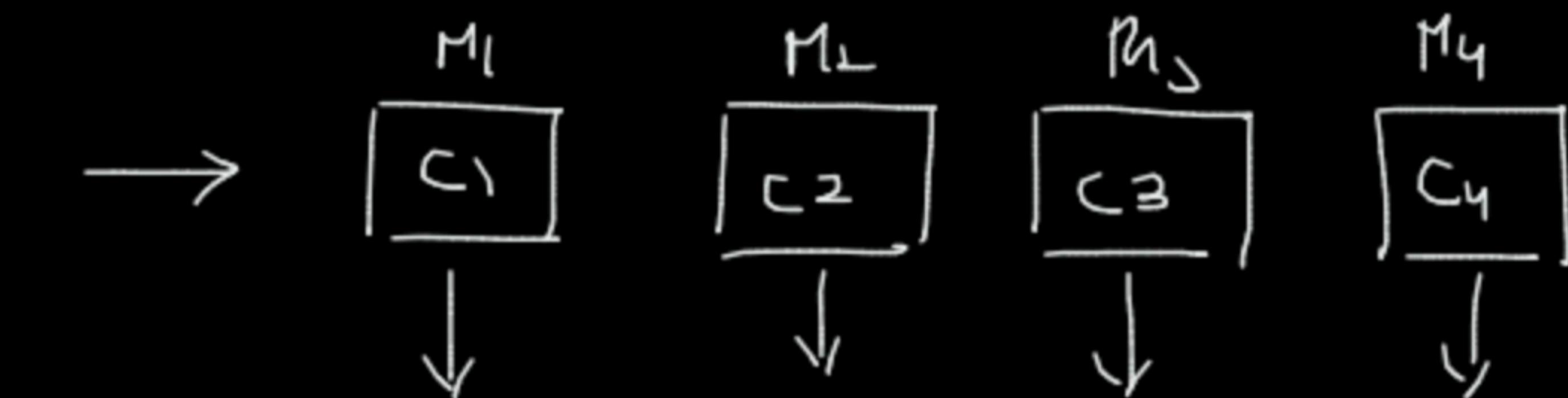
/

Bagging - Bootstrapped Aggregation



1 Models are indep of each other

Parallelisation is possible



$\rightarrow 25$ min

2 Homogeneous Models

3 Every model is given equal importance

✓
Bias & Variance \rightarrow ✓

10.37 cm

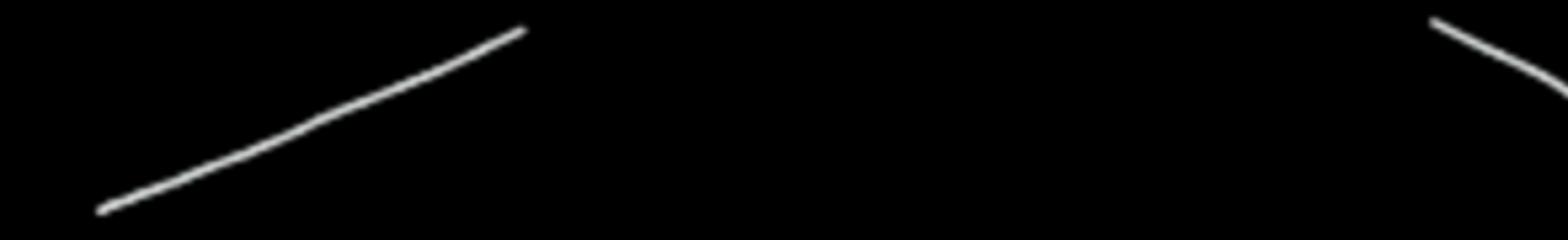
← Data Encoding →

le = LabelEncoder()

le fit-transform(^{col})

Label Encoding

— Convert categorical data to numeric data



Species

0	setosa
1	Versicolor
2	Virginica

Alphabetical

order

one-hot Encoding

← dummy columns →
Species_Set Species_Ver Species_Virg

1	0 ✓	0 ✓
0	1	0 ✓
0	0 ✓	1

Sparse matrix

| ohe = OneHotEncoder() →

ohe fit-transform() ↴

2 Pandas → get_dummies()

S. Josa 1 0 0 ✓

0 0 0 - 1

✓or 0 1 0 ✓

1 0 0 - 2

✓✓ 0 0 1 ✓

0 1 0 - 3

0 0 1 - 4

3 - 2w

(0 - 9w