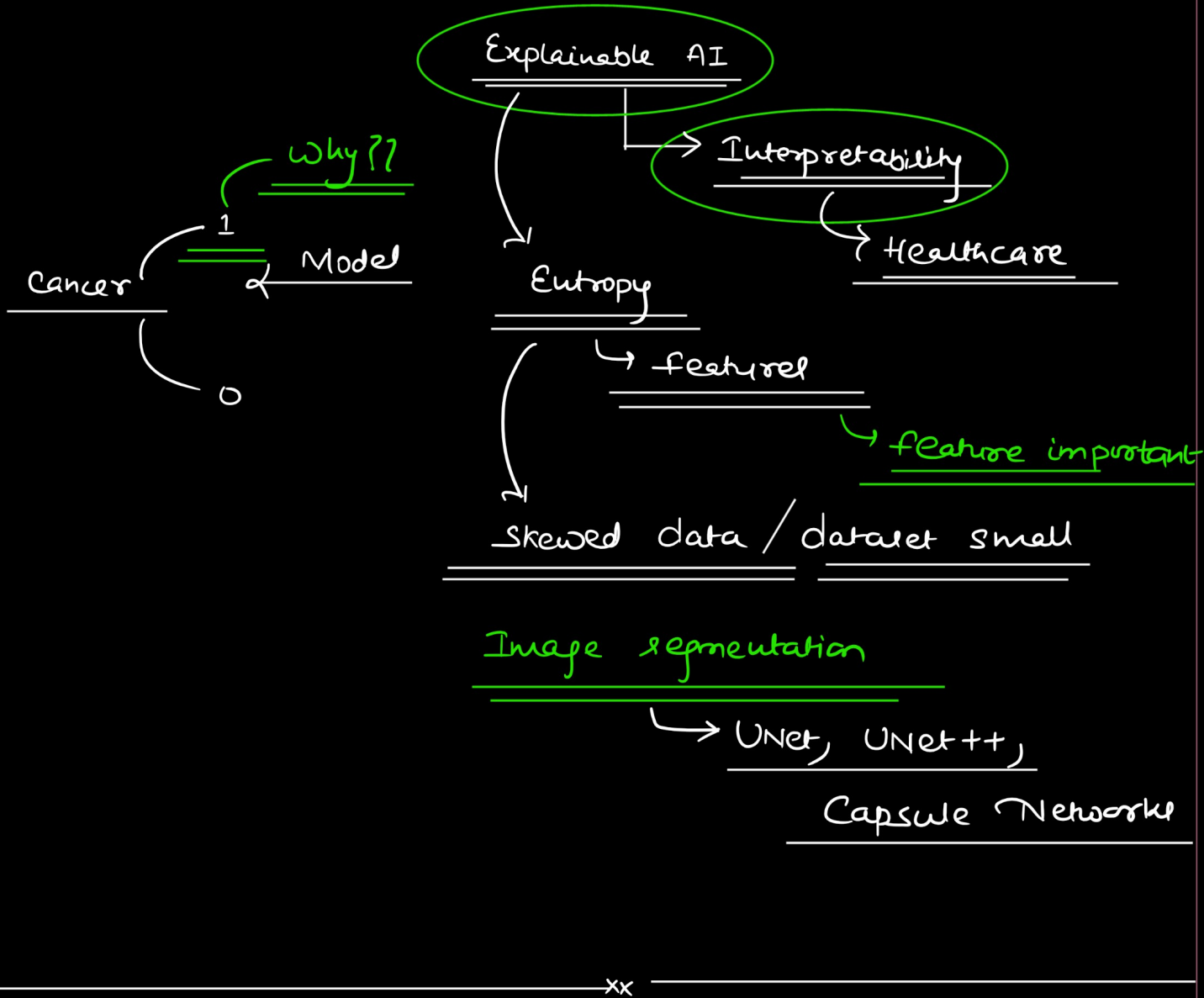


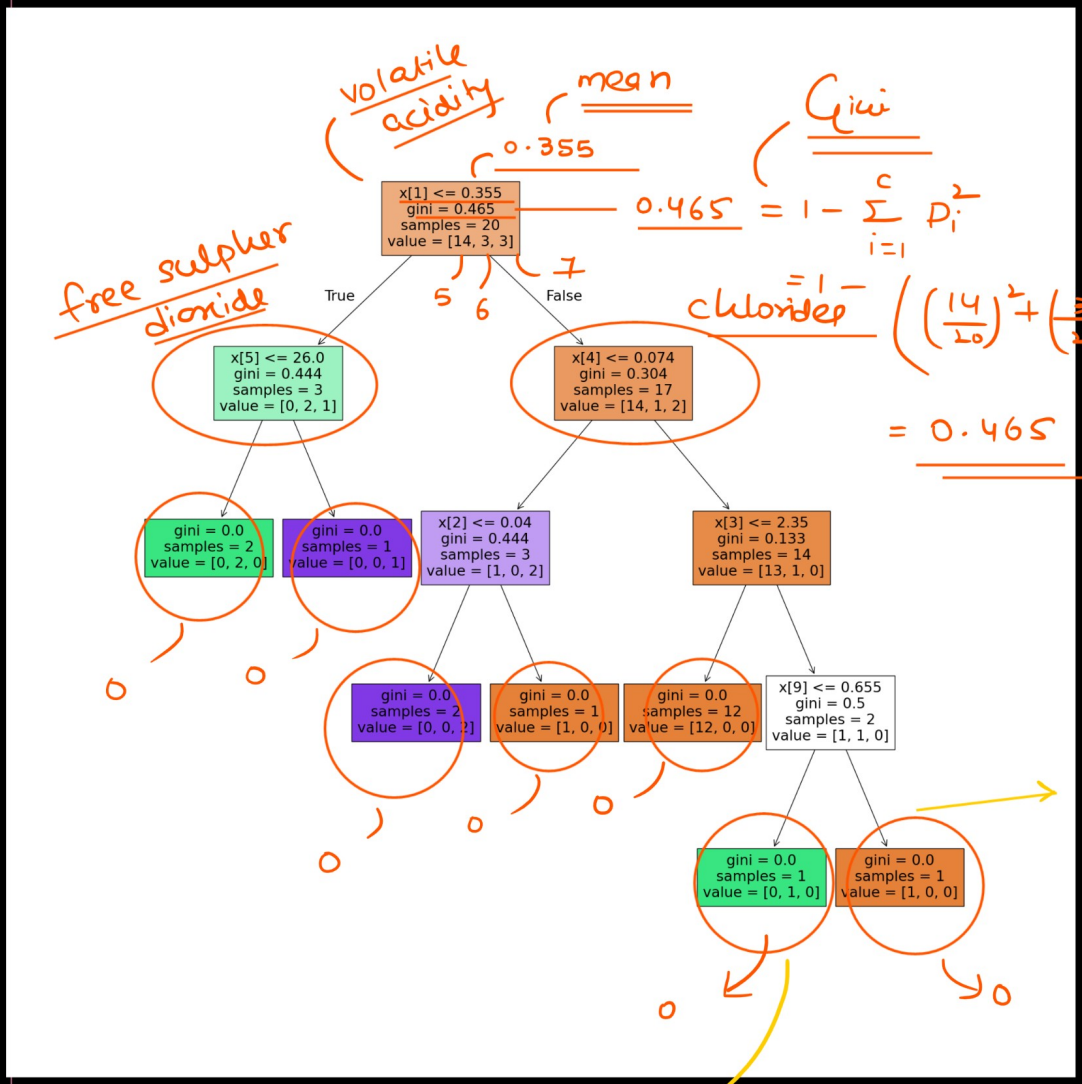

```
Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
       'pH', 'sulphates', 'alcohol'],
      dtype='object')
```

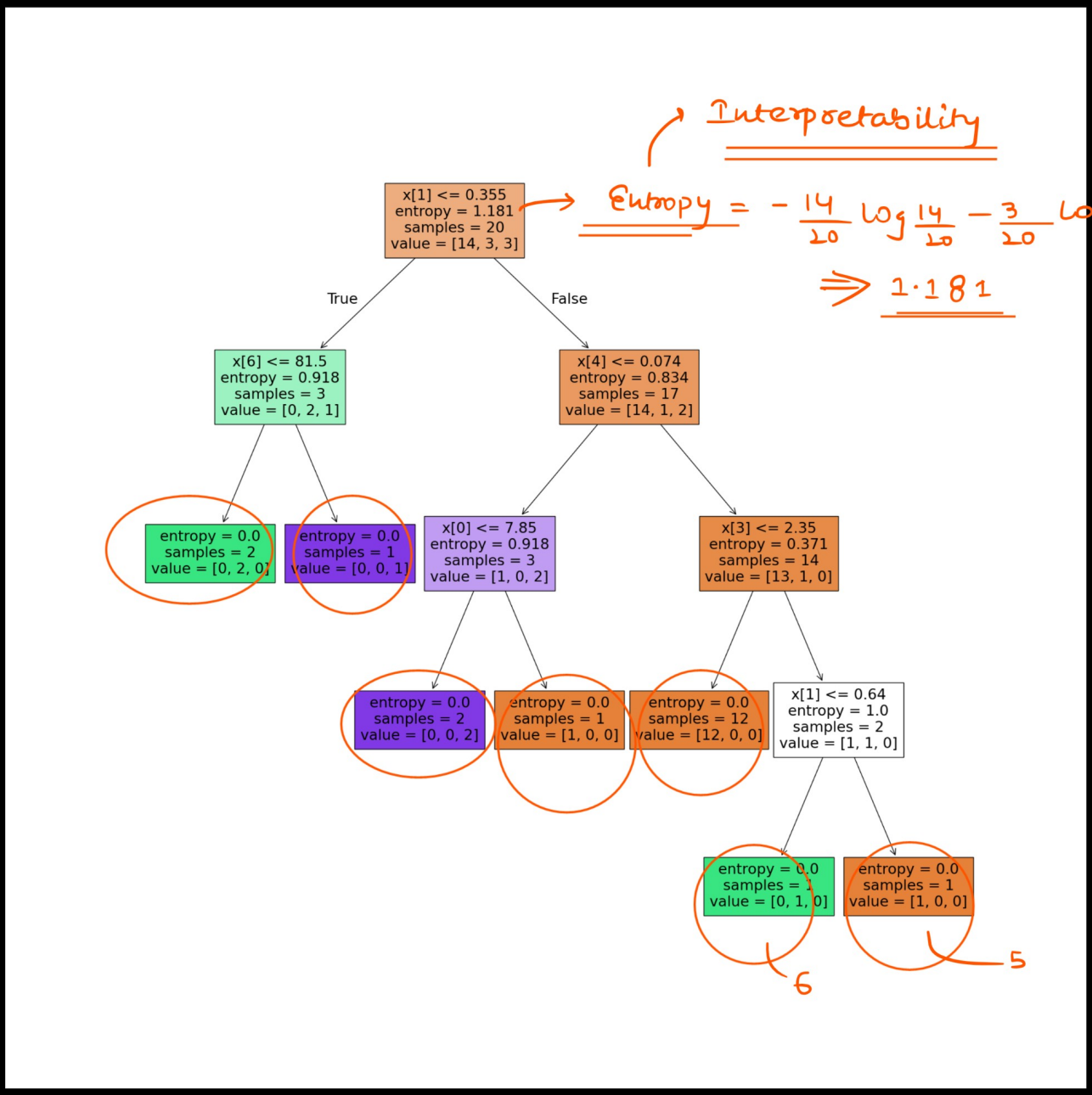
0 1 2 3 4 5 6 7 8 9 10

featured = 11

Counter({5: 14, 6: 3, 7: 3})

5 — 14 ⇒ 14/20
6 — 3 ⇒ 3/20
7 — 3 ⇒ 3/20





Interpretability

Entropy $= -\frac{14}{20}\log\frac{14}{20} - \frac{3}{20}\log\frac{3}{20} - \frac{3}{20}\log\frac{3}{20}$

$\Rightarrow 1.181$

Pruning → Removing

Prepruning

Post pruning

hyperparameter

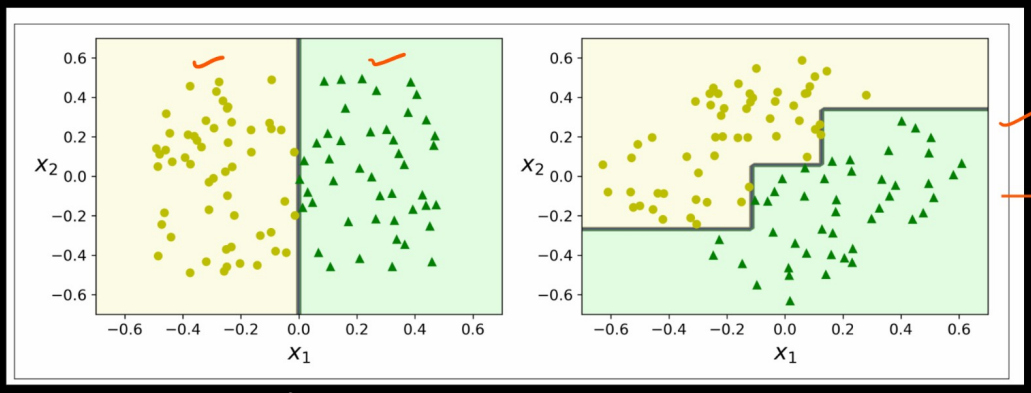
- max_depth = 4,
- min_samples_split

GridSearchCV,

RandomizedSearchCV

Create the entire DT first & then it starts cutting

Limitations in Decision Tree



sensitive to the data points

→ overfitting

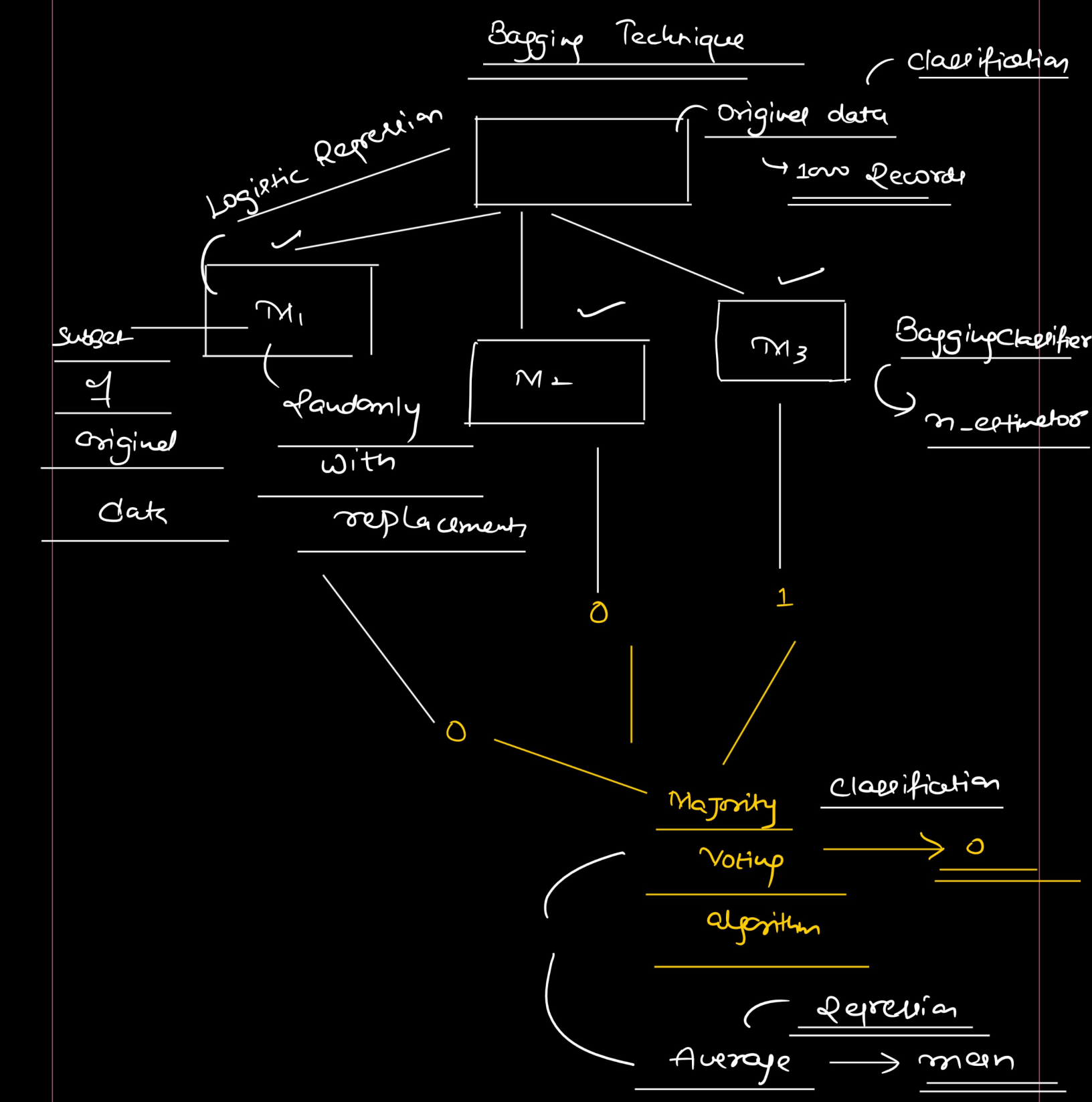
↳ doesn't generalized well

45°

Ensemble Techniques

Real time industries

1. Bagging ← RandomForest
2. Boosting
3. Stacking


Bagging Technique
Classification
Original data
1000 Records
Logistic Regression
M1
M2
M3
Subset of Original Data
Randomly with replacements
BaggingClassifier
n_estimators
0
1
0
Majority Voting algorithm
Classification
0
Regression
Average
mean