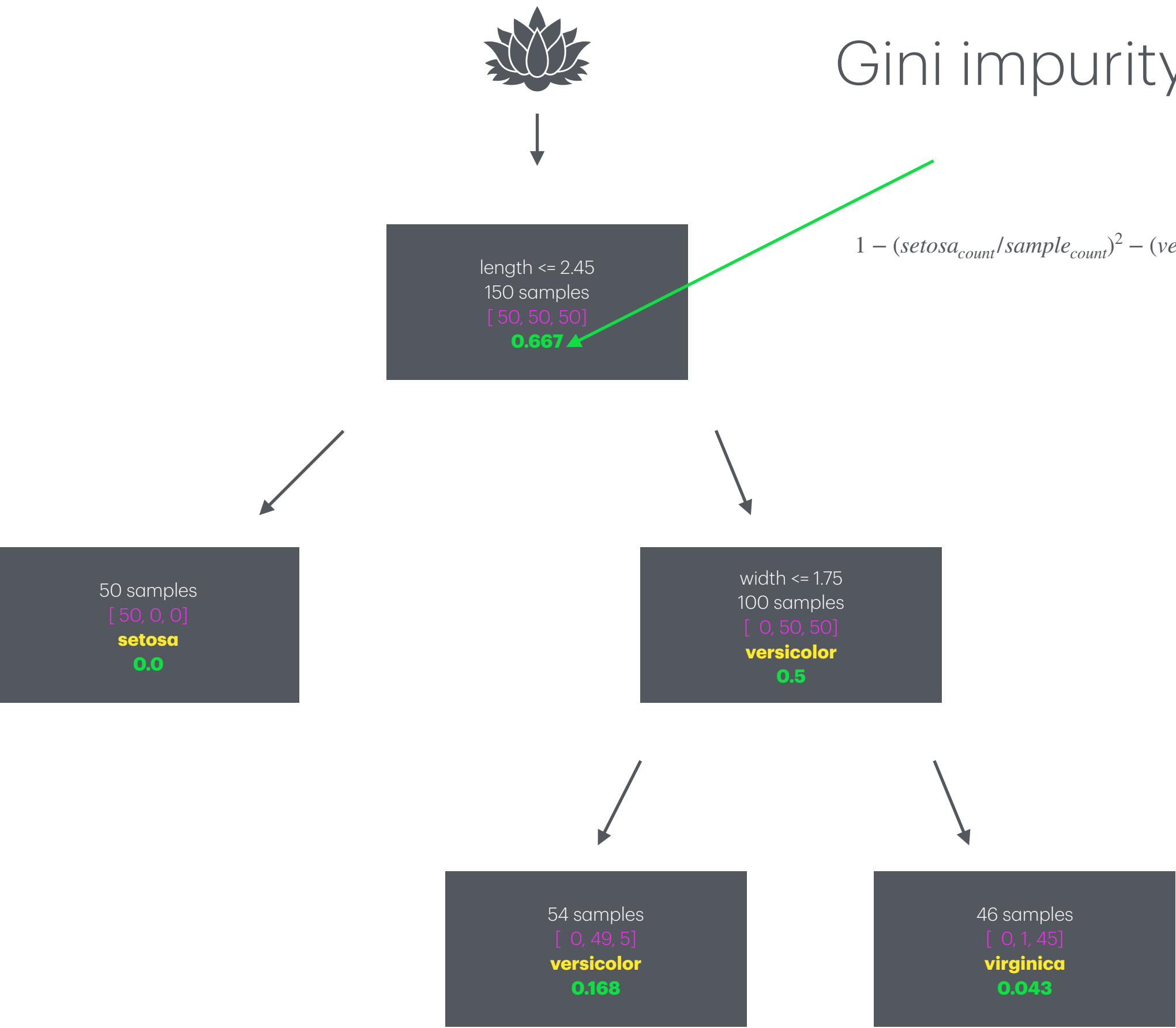


Decision Tree

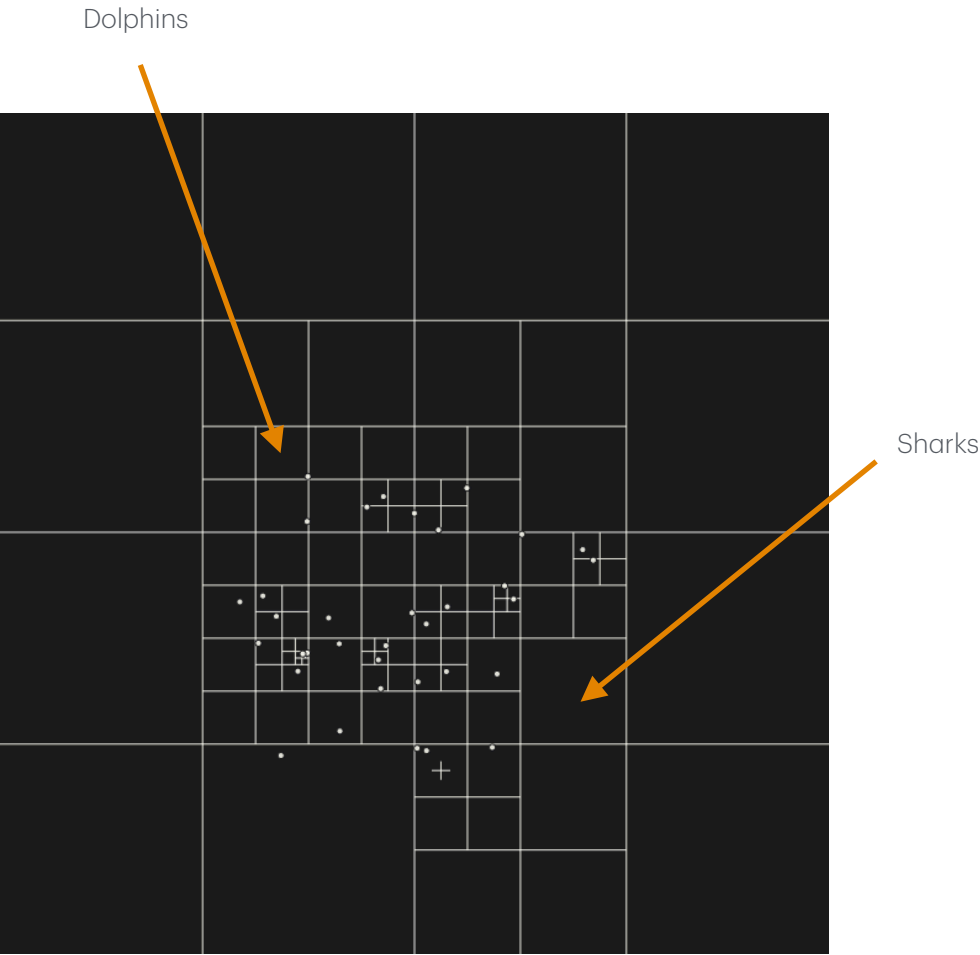
Petal Length	Petal Width	
setosa	versicolor	virginica

Gini impurity $G_i = 1 - \sum_1^n P_i, k^2$

$1 - (setosa_{count}/sample_{count})^2 - (vericolor_{count}/sample_{count})^2 - (virginica_{count}/sample_{count})^2 = 0.667$



Goal is to keep branching until the impurity is low for all leaf nodes. The model works to create clustered region of low impurity regions



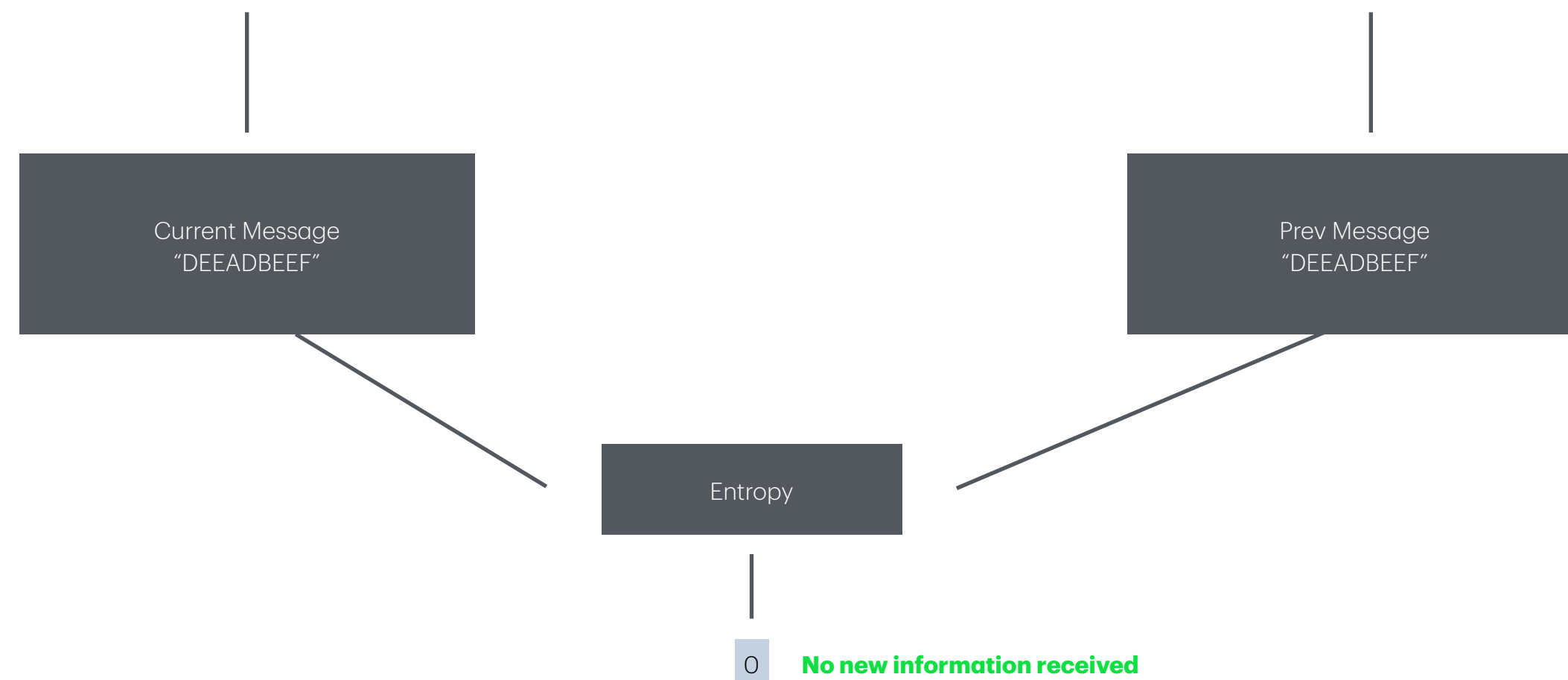
Entropy or Gini

Entropy is the measure of disorder(thermodynamics):

Zero entropy for constant (i.e. still) molecule and well ordered

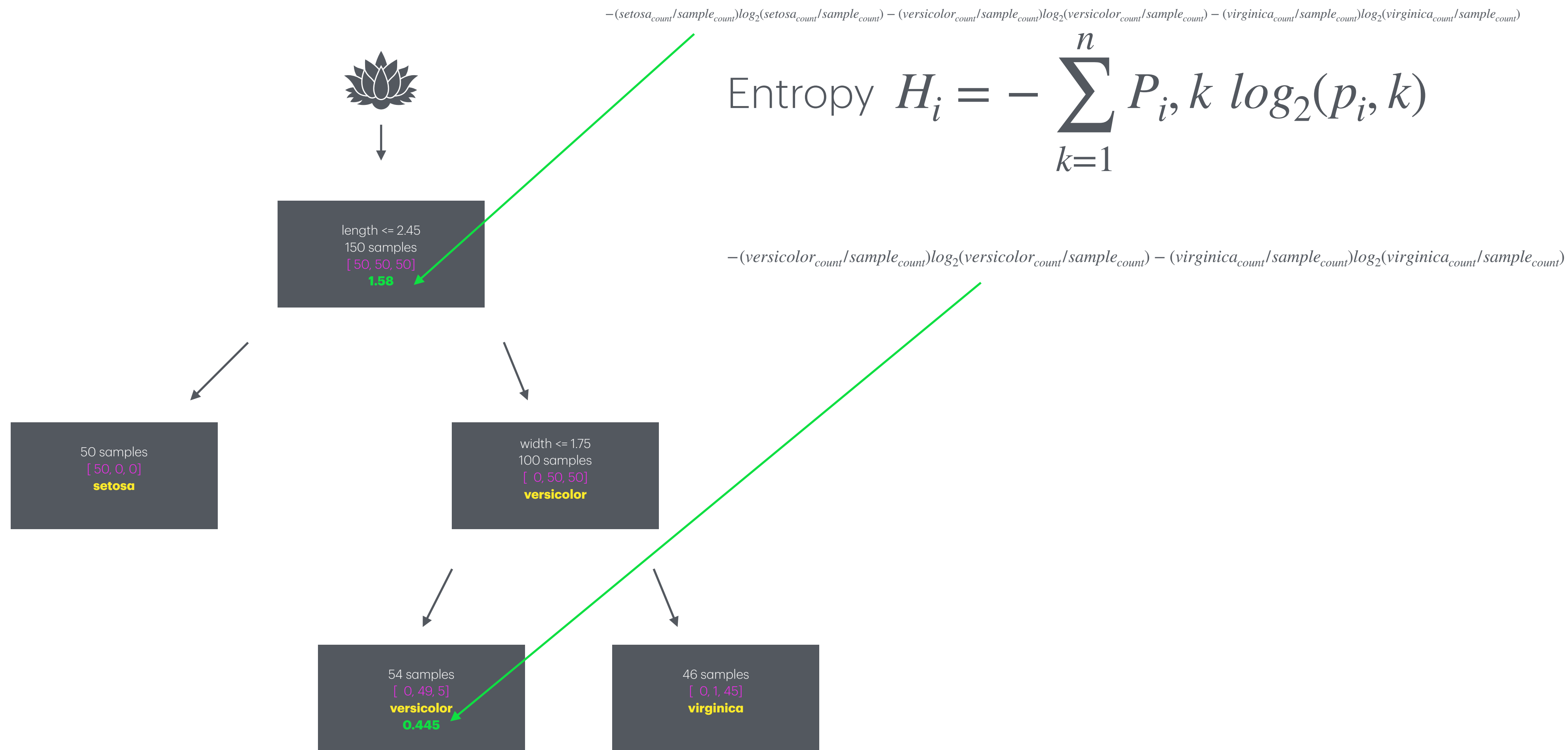
Shannon Information Theory:

Entropy is average information content of message



Machine Learning:

Entropy is zero when a set contains instances of only one class



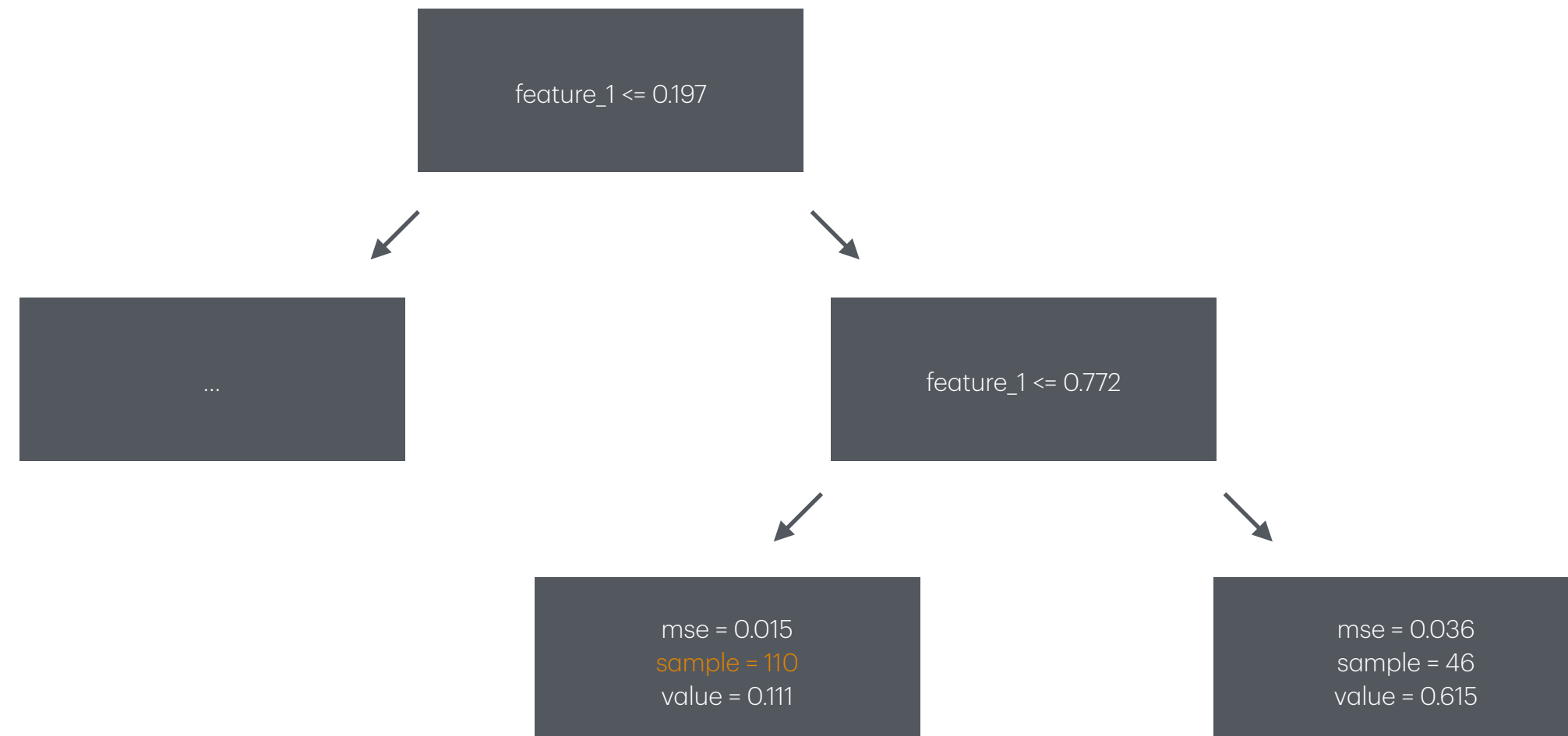
Non-parametric - Number of parameters is not determined prior to training.

Unlike parametric linear models, decision trees are infinite

Decision trees are at risk to overfitting - restricting freedom is how decision tree models are regularized



Regression



CART cost function used to measure

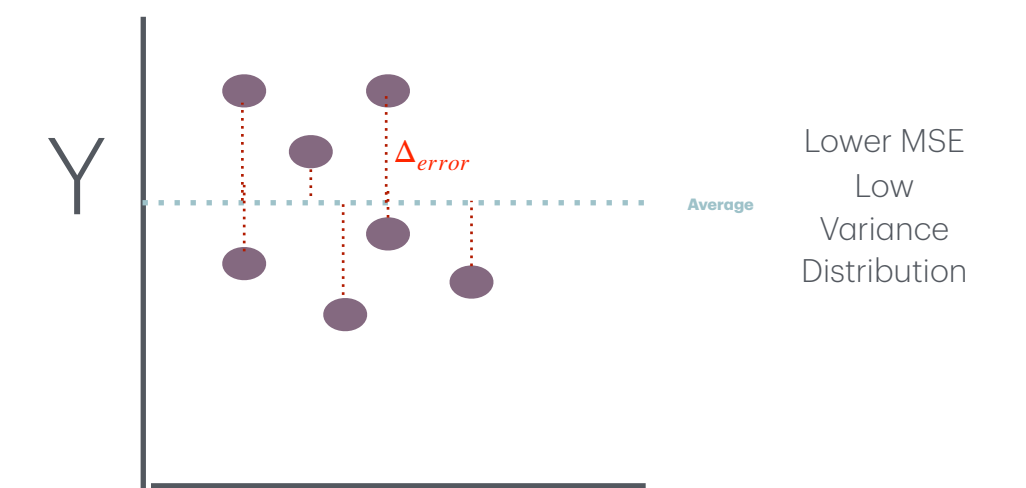
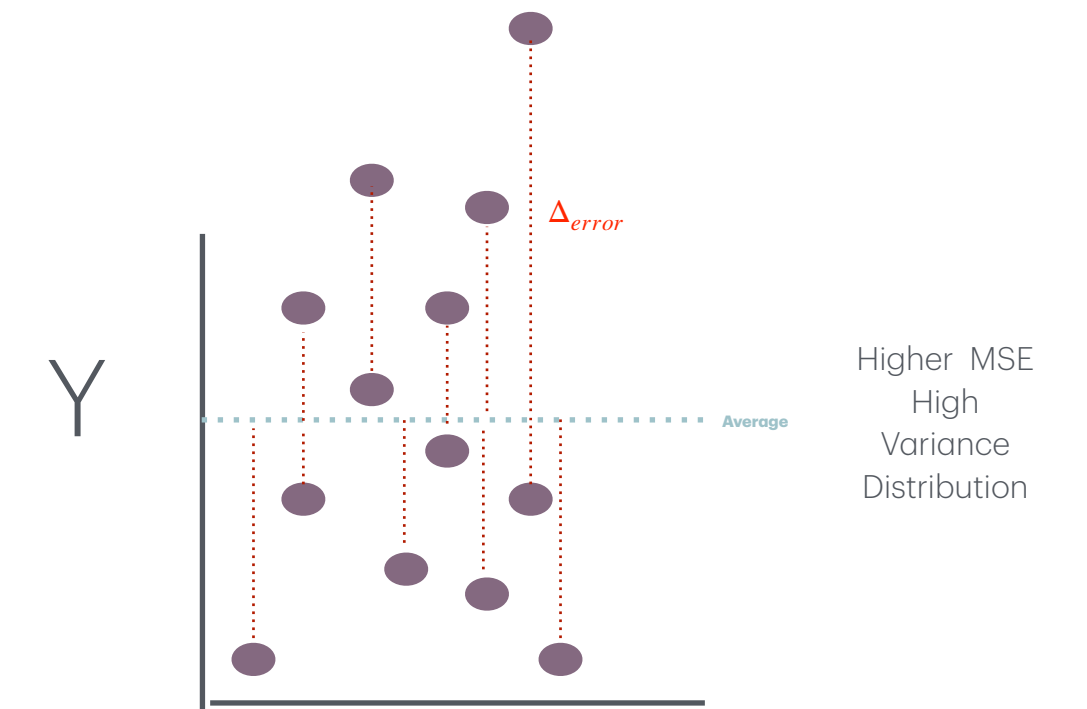
Training samples
feature_1 > 0.197 & feature_1 <= 0.772

[illegible]

$$\text{Prediction: } \hat{y}_{node} = \frac{1}{m_{node}} \sum_i y^{(i)}$$

$$\text{MSE Node: } \sum_i (\hat{y}_{node} - y^{(i)})^2$$

MSE Node: $\sum_i (\Delta_{error})^2$

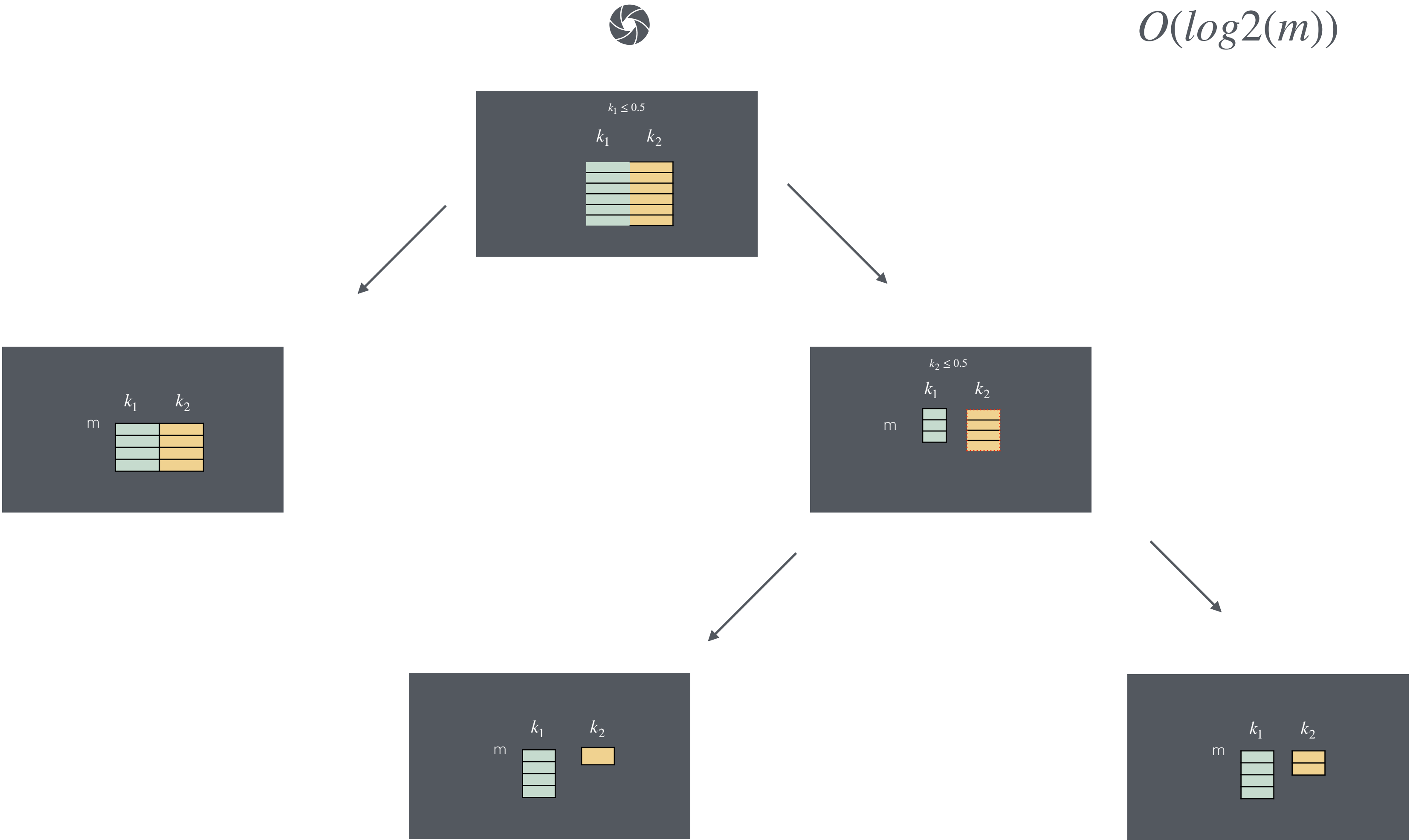


MSE depends on the distribution of samples in node

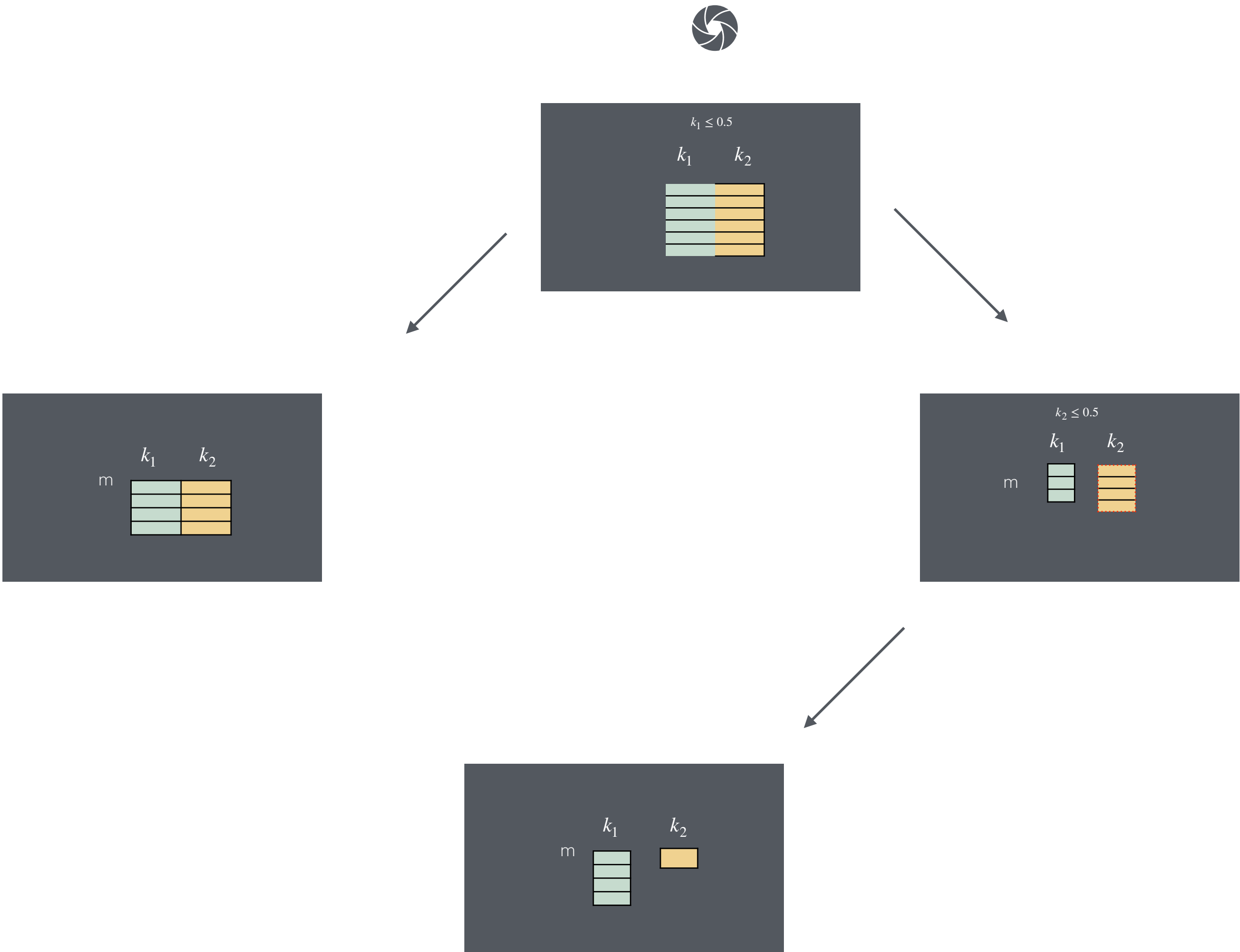
Prediction

Single check on each node is performed:

$$O(\log_2(m))$$



Training



Step to each node in tree fashion:

$O(\log_2(m))$

$n \times m$ compute at each node.
Each sample compares to each feature

Total : $n \times m O(\log_2(m))$

Training: Inside Node



Find optimal split pair (feature and threshold)

Create children nodes with reference to samples

⋮

Exercise 5

If it takes one hour to train a Decision Tree on a training set containing 1 million instances, roughly how much time will it take to train another Decision Tree on a training set containing 10 million instances

n- features in node

m- instances in node

$$n \cdot m \cdot \log_2(m) = 1$$

Calculate Z

$$n \cdot m \cdot 10 \cdot \log_2(m \cdot 10) = Z$$

$$\frac{n \cdot m \cdot 10 \cdot \log_2(m \cdot 10)}{n \cdot m \cdot \log_2(m)} = Z$$

$$\frac{10 \cdot \log_2(m \cdot 10)}{\log_2(m)} = Z$$

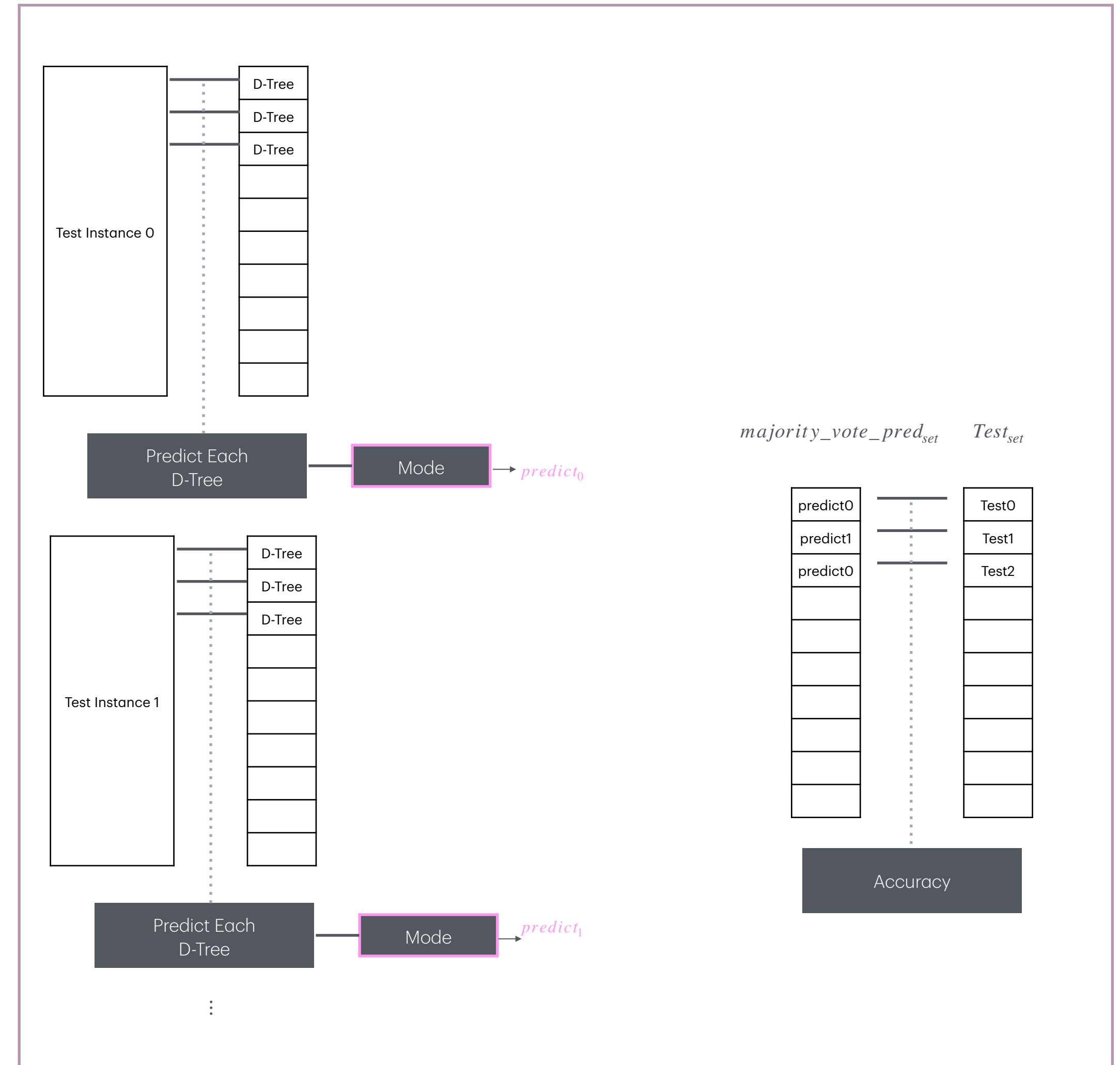
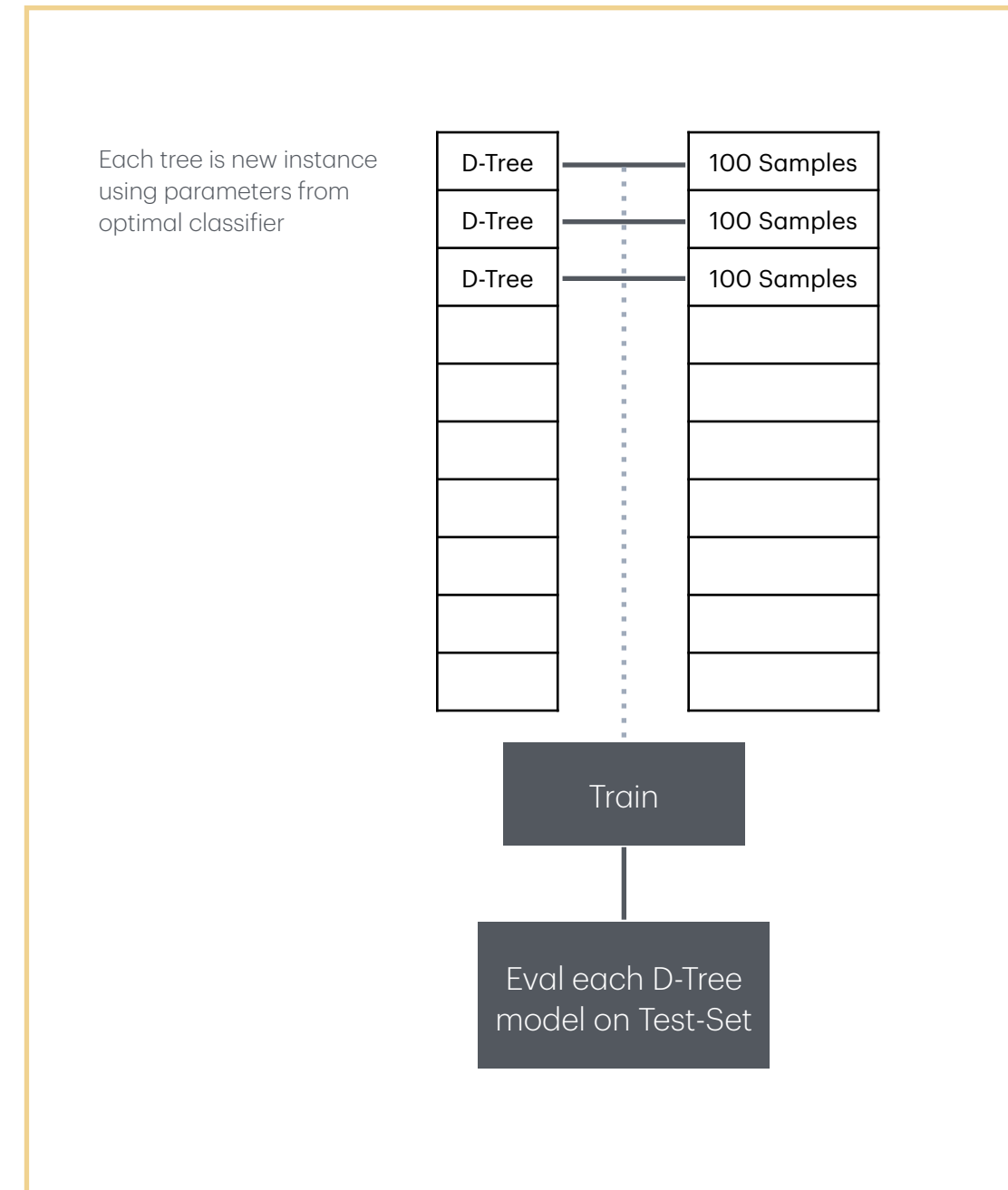
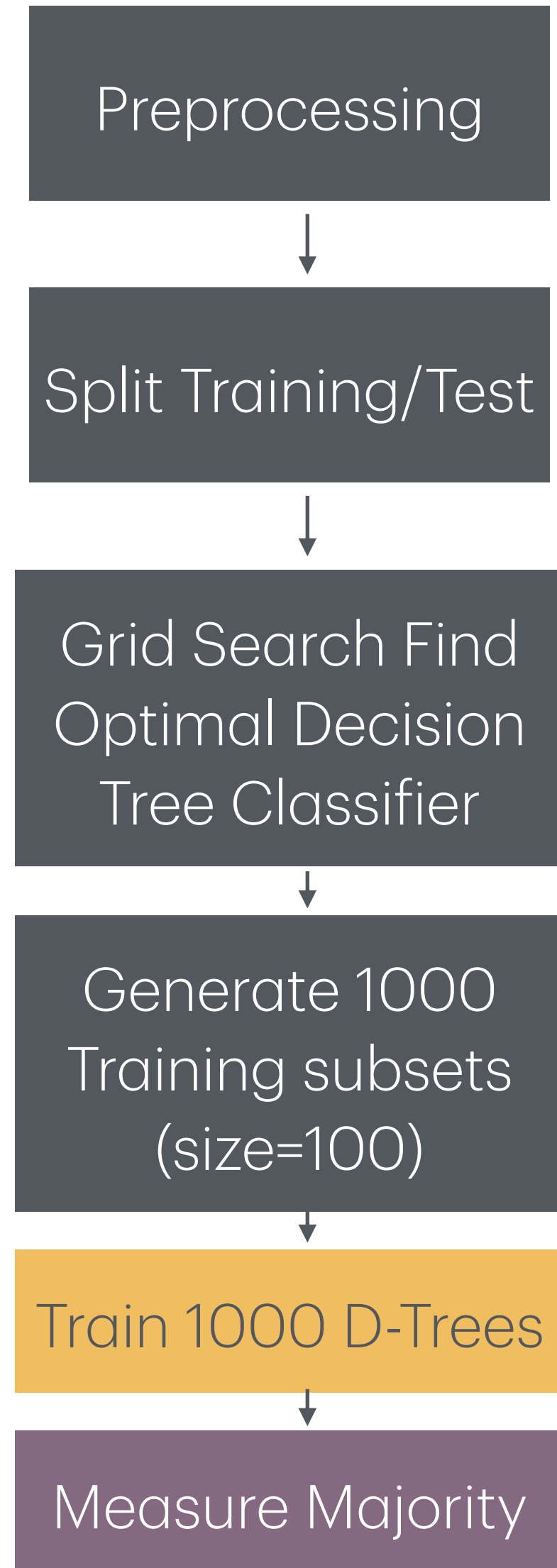
$$\frac{10 \cdot \log(m \cdot 10)}{\log(2)} \cdot \frac{\log(2)}{\log(m)} = Z$$

$$\frac{10 \cdot \log(m \cdot 10)}{\log(m)} = Z$$

m= 1000000

$$11.67 = Z$$

Train Decision Tree



1000 D-Trees/
100 Sample
Predictions
Accuracy

0.8425 0.8265 0.846 0.8485 0.854 0.847 0.848 0.852 0.8455 0.852
0.854 0.8375 0.8435 0.831 0.8395 0.8415 0.837 0.854 0.8575 0.848
0.846 0.848 0.85 0.8215 0.8355 0.8435 0.848 0.849 0.8505 0.856
0.841 0.8385 0.836 0.852 0.854 0.836 0.8515 0.856 0.8535 0.8465
0.845 0.836 0.843 0.8525 0.8515 0.845 0.8585 0.843 0.8515 0.839
0.833 0.85 0.8395 0.8365 0.8485 0.85 0.848 0.8295 0.84 0.849
0.845 0.8525 0.8495 0.851 0.827 0.8475 0.8515 0.827 0.8545 0.849
0.848 0.85 0.842 0.8535 0.849 0.837 0.8485 0.851 0.845 0.832
0.8445 0.854 0.8435 0.8435 0.8555 0.85 0.8445 0.8455 0.8495 0.842
0.8395 0.851 0.8535 0.847 0.8455 0.8515 0.8505 0.854 0.8425 0.8445
0.8355 0.8395 0.856 0.8385 0.8465 0.8545 0.85 0.8375 0.83 0.8525
0.849 0.8415 0.8465 0.851 0.8305 0.8475 0.8415 0.8495 0.849 0.8555
0.8425 0.8335 0.841 0.8415 0.85 0.8505 0.8625 0.847 0.8405 0.846
0.849 0.843 0.837 0.8505 0.836 0.8485 0.847 0.851 0.846 0.8425
0.8435 0.8495 0.837 0.8445 0.8455 0.8435 0.8535 0.8405 0.836 0.8465
0.84 0.842 0.849 0.852 0.8525 0.84 0.8585 0.8525 0.8405 0.845
0.8525 0.8175 0.8385 0.8495 0.8375 0.847 0.8275 0.849 0.851 0.831
0.847 0.844 0.8425 0.8505 0.8525 0.842 0.848 0.8455 0.851 0.8515
0.85 0.843 0.836 0.8435 0.8515 0.847 0.8495 0.8505 0.844 0.85
0.8515 0.846 0.852 0.848 0.8535 0.852 0.8415 0.8475 0.8415 0.855
0.841 0.8485 0.847 0.853 0.8525 0.8495 0.853 0.847 0.8505 0.8525
0.827 0.8515 0.8505 0.8275 0.834 0.8445 0.84 0.85 0.843 0.8565
0.8475 0.8455 0.8545 0.841 0.844 0.853 0.849 0.8435 0.819 0.8415
0.853 0.827 0.838 0.8455 0.846 0.854 0.8525 0.853 0.841 0.85
0.8465 0.849 0.848 0.847 0.8505 0.833 0.8515 0.846 0.8535 0.8445
0.8505 0.8485 0.8585 0.854 0.8525 0.8445 0.853 0.844 0.853 0.85
0.855 0.852 0.8345 0.8355 0.855 0.8495 0.848 0.85 0.8515 0.8405
0.8525 0.8485 0.842 0.854 0.8285 0.848 0.8495 0.846 0.8495 0.8515
0.8475 0.84 0.854 0.844 0.8515 0.8485 0.846 0.845 0.843 0.8505
0.849 0.8575 0.8555 0.8495 0.8535 0.8485 0.849 0.849 0.8525 0.855
0.8525 0.8535 0.8515 0.8335 0.842 0.835 0.843 0.8415 0.8525 0.833
0.8425 0.848 0.8545 0.8545 0.846 0.8365 0.852 0.8435 0.8505 0.8475
0.8545 0.8505 0.841 0.8505 0.8425 0.8425 0.8435 0.845 0.8455 0.847
0.8425 0.846 0.8545 0.845 0.863 0.8525 0.8625 0.8425 0.85 0.8455
0.851 0.82 0.8505 0.837 0.8545 0.8305 0.8525 0.8415 0.8445 0.85
0.8445 0.853 0.851 0.8405 0.8465 0.8515 0.846 0.855 0.8525 0.85
0.847 0.84 0.8455 0.85 0.86 0.842 0.831 0.849 0.839 0.843
0.8405 0.849 0.849 0.8395 0.844 0.837 0.846 0.8485 0.8545 0.8575
0.8525 0.844 0.8535 0.8495 0.8525 0.856 0.8445 0.8505 0.858 0.8405
0.8445 0.8455 0.8385 0.844 0.8455 0.8255 0.8505 0.849 0.847 0.8295
0.85 0.85 0.856 0.8425 0.8565 0.8425 0.856 0.8215 0.837 0.8415
0.8515 0.845 0.841 0.8415 0.8515 0.8405 0.85 0.857 0.847 0.84
0.8435 0.8525 0.843 0.8455 0.8495 0.8555 0.8525 0.833 0.855 0.849
0.8495 0.849 0.844 0.8455 0.844 0.8405 0.846 0.845 0.8465 0.8485
0.851 0.8405 0.8465 0.838 0.849 0.8495 0.8395 0.851 0.846 0.85
0.846 0.851 0.8375 0.8425 0.84 0.845 0.8325 0.841 0.849 0.841
0.854 0.835 0.8445 0.8475 0.8455 0.8525 0.85 0.8305 0.838 0.8465
0.854 0.8455 0.8455 0.846 0.841 0.8395 0.8475 0.8355 0.855 0.846
0.853 0.8525 0.835 0.8355 0.847 0.8445 0.8455 0.8365 0.8495 0.8555
0.8355 0.852 0.84 0.8375 0.8455 0.846 0.8505 0.8485 0.8525 0.852
0.8395 0.8305 0.847 0.8425 0.836 0.841 0.853 0.8455 0.8485 0.839
0.8375 0.852 0.8495 0.851 0.854 0.8325 0.8495 0.8545 0.859 0.8345
0.8515 0.853 0.8505 0.842 0.858 0.842 0.852 0.845 0.855 0.8465
0.8525 0.8455 0.8535 0.853 0.8535 0.834 0.853 0.839 0.8535 0.857
0.8375 0.8525 0.8475 0.8575 0.8485 0.8465 0.852 0.8575 0.845 0.843
0.8485 0.8485 0.8415 0.852 0.851 0.8365 0.847 0.8405 0.8525 0.8375
0.8435 0.845 0.8465 0.8465 0.8445 0.8465 0.8445 0.8425 0.852 0.847
0.8495 0.847 0.845 0.848 0.8505 0.844 0.8405 0.846 0.8325 0.853
0.8425 0.848 0.8445 0.845 0.8505 0.8525 0.829 0.855 0.8435 0.8405
0.8465 0.8435 0.839 0.8245 0.8475 0.8505 0.8425 0.8455 0.848 0.8425
0.846 0.839 0.843 0.85 0.857 0.8435 0.852 0.849 0.852 0.851
0.8525 0.8475 0.8345 0.8485 0.843 0.85 0.8465 0.8405 0.854 0.8475
0.858 0.85 0.845 0.8505 0.8265 0.85 0.8505 0.847 0.8445 0.8435
0.847 0.8545 0.8375 0.84 0.853 0.827 0.843 0.8355 0.844 0.8515
0.834 0.8405 0.85 0.8385 0.845 0.848 0.8545 0.853 0.848 0.8475
0.8505 0.8455 0.8465 0.8355 0.833 0.8465 0.8345 0.851 0.835 0.847
0.8445 0.857 0.848 0.8395 0.8595 0.855 0.838 0.824 0.854 0.852
0.8455 0.8355 0.8475 0.8485 0.8375 0.85 0.8505 0.8575 0.837 0.8515
0.849 0.8455 0.848 0.836 0.8545 0.8485 0.841 0.848 0.8225 0.847
0.851 0.822 0.844 0.8495 0.837 0.8395 0.846 0.852 0.8525 0.846
0.844 0.8465 0.8365 0.8595 0.85 0.8195 0.8185 0.8545 0.853 0.8415
0.841 0.8315 0.8505 0.8495 0.853 0.841 0.8535 0.846 0.8515 0.842
0.8345 0.8075 0.8505 0.8415 0.8515 0.8505 0.8515 0.853 0.8445 0.8475
0.8495 0.845 0.8405 0.846 0.842 0.823 0.8215 0.842 0.845 0.84
0.8565 0.8475 0.8335 0.8525 0.851 0.8455 0.8425 0.857 0.8485 0.838
0.857 0.8505 0.845 0.8385 0.8565 0.857 0.847 0.842 0.842 0.827
0.838 0.8475 0.8525 0.8435 0.84 0.8545 0.832 0.8505 0.8465 0.845
0.853 0.852 0.8575 0.8315 0.848 0.84 0.848 0.8515 0.834 0.8365
0.847 0.844 0.8545 0.8475 0.856 0.834 0.8505 0.8575 0.8545 0.8545
0.823 0.837 0.853 0.846 0.8505 0.8505 0.8555 0.847 0.84 0.847
0.846 0.842 0.8455 0.845 0.859 0.8495 0.848 0.849 0.8415 0.838
0.8405 0.835 0.8595 0.847 0.848 0.83 0.8505 0.847 0.8435 0.851
0.821 0.861 0.852 0.8425 0.8495 0.853 0.8455 0.849 0.8465 0.8425
0.847 0.856 0.838 0.8445 0.8465 0.8525 0.859 0.826 0.8475 0.853
0.85 0.845 0.839 0.846 0.8435 0.8505 0.846 0.855 0.8495 0.847
0.849 0.845 0.835 0.8395 0.848 0.8505 0.8335 0.8465 0.853 0.8455
0.841 0.8435 0.8475 0.8525 0.8355 0.8445 0.83 0.8455 0.843 0.8445
0.8405 0.8475 0.8565 0.835 0.8445 0.836 0.8475 0.8315 0.8565 0.852
0.8475 0.8455 0.854 0.844 0.8315 0.845 0.8495 0.846 0.8325 0.849
0.8315 0.8465 0.849 0.8335 0.8515 0.842 0.852 0.8415 0.853 0.859
0.85 0.833 0.8385 0.842 0.846 0.849 0.8405 0.843 0.831 0.851
0.839 0.8565 0.843 0.8245 0.8385 0.8555 0.843 0.8475 0.8285 0.8595
0.8575 0.8525 0.8395 0.843 0.844 0.834 0.847 0.857 0.8515 0.8295
0.8515 0.8445 0.85 0.841 0.851 0.846 0.847 0.844 0.828 0.8525
0.8495 0.8365 0.8385 0.8495 0.847 0.8435 0.842 0.824 0.8515 0.853
0.8425 0.849 0.8525 0.8415 0.851 0.849 0.8365 0.852 0.846 0.8595
0.8475 0.846 0.8505 0.8515 0.8475 0.8595 0.8375 0.8465 0.856 0.8455
0.8475 0.844 0.824 0.855 0.844 0.8475 0.8495 0.8425 0.8555 0.8465
0.8505 0.856 0.845 0.852 0.858 0.8455 0.846 0.848 0.853 0.844
0.8455 0.845 0.8515 0.8475 0.852 0.845 0.85 0.845 0.8615 0.85251

Majority - Vote Predictions Accuracy

0.8575