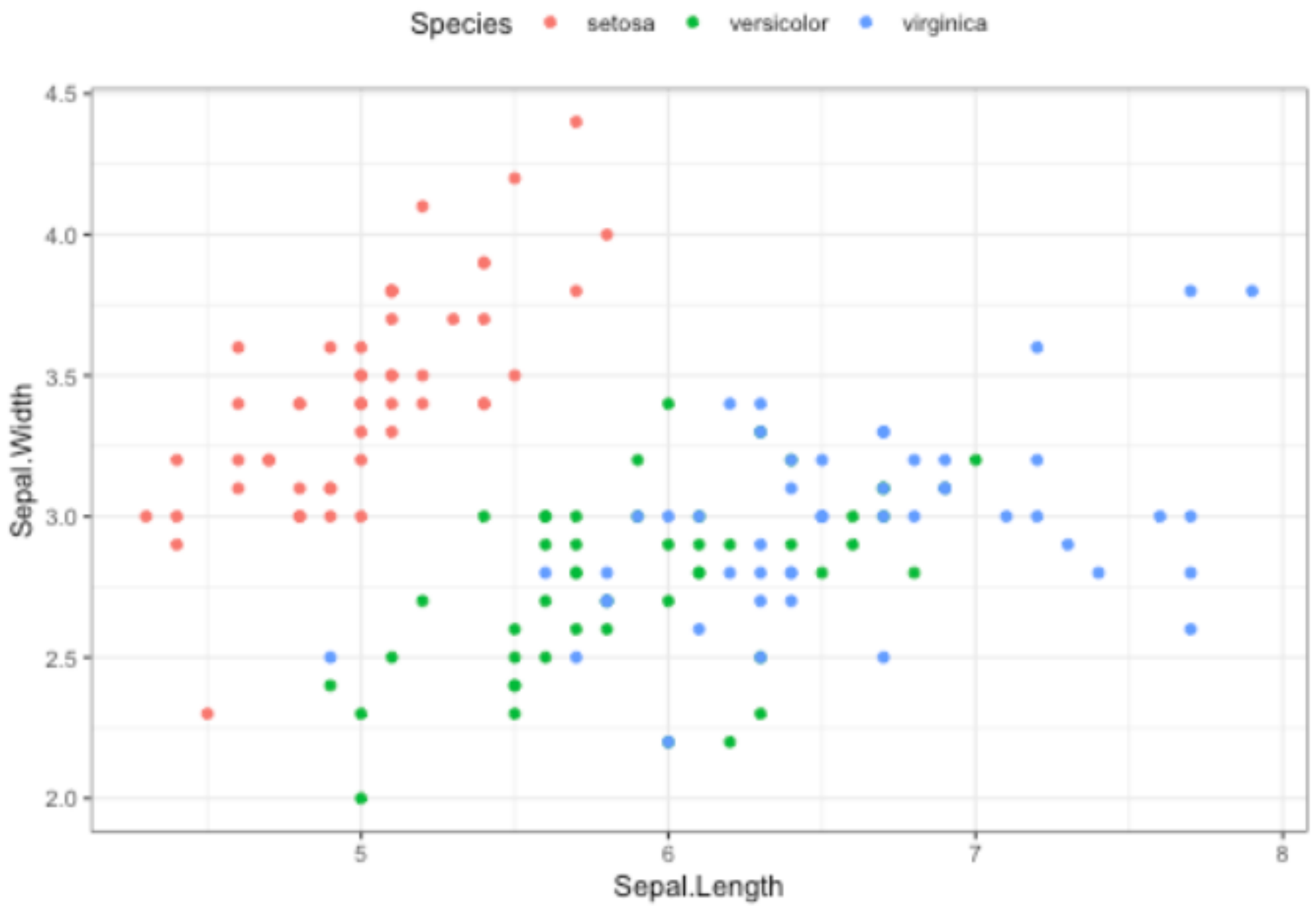
Random Forests

Akash Kharita

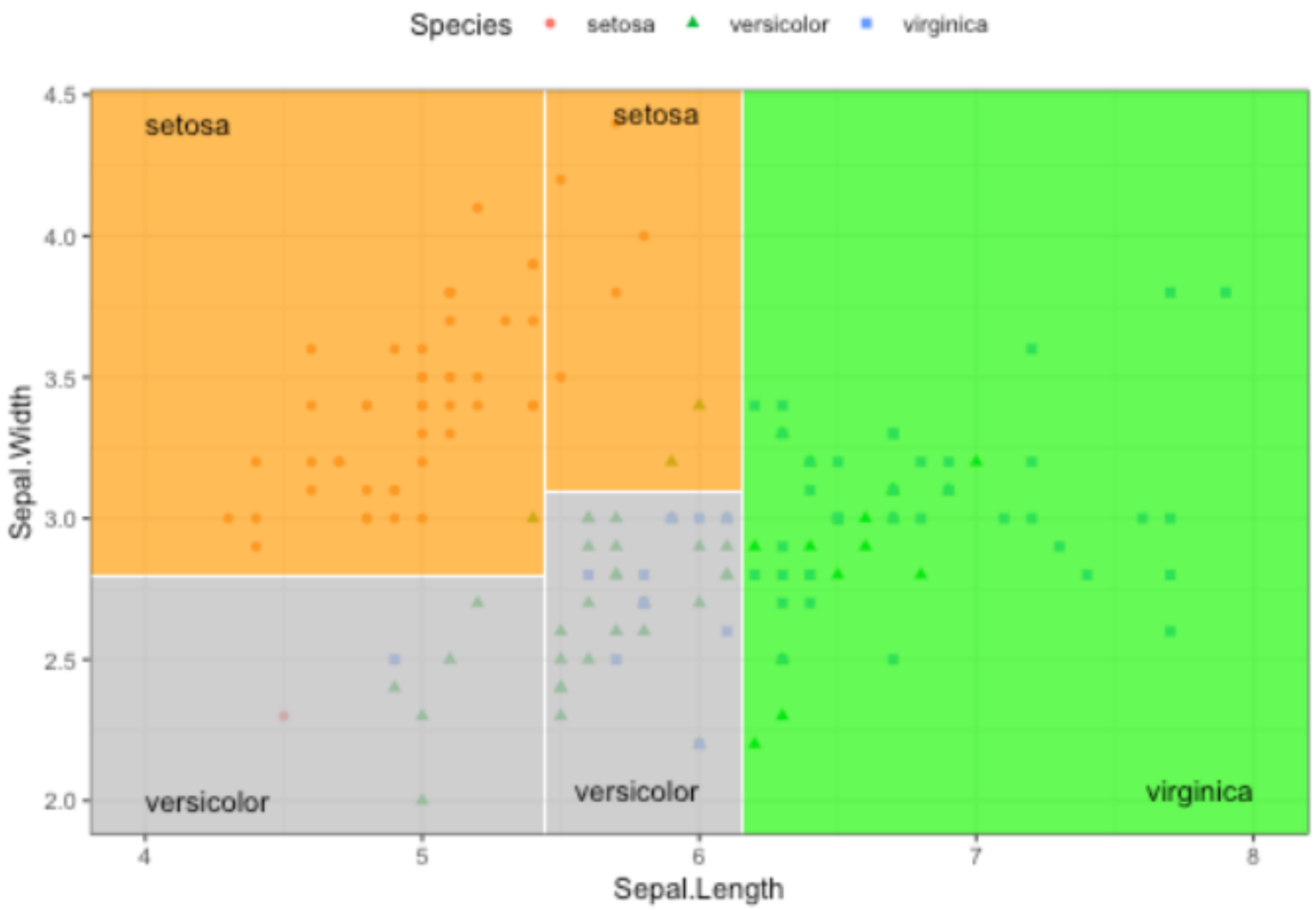
PhD Student, Seismology

University of Washington

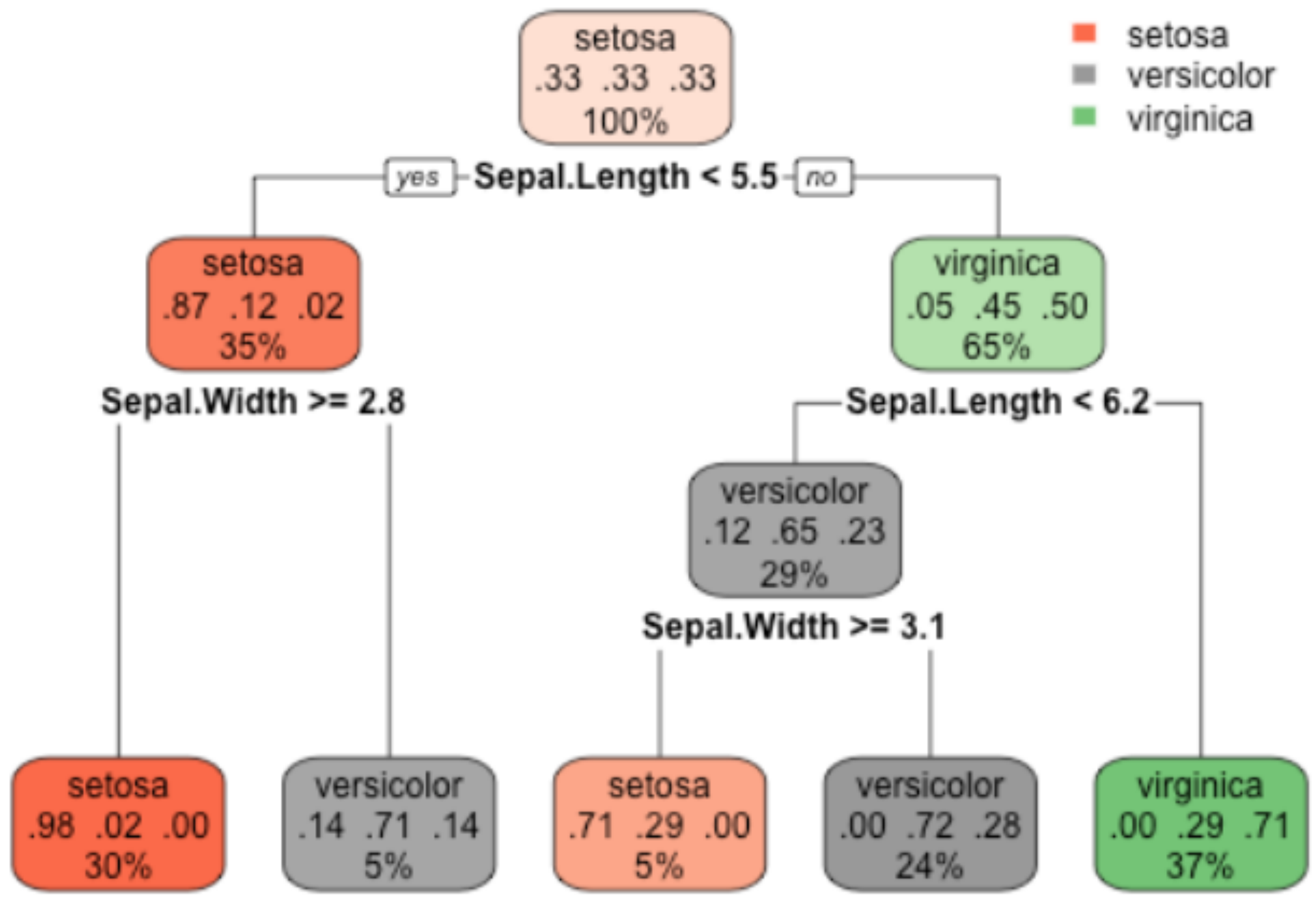
Classification problem: Iris data



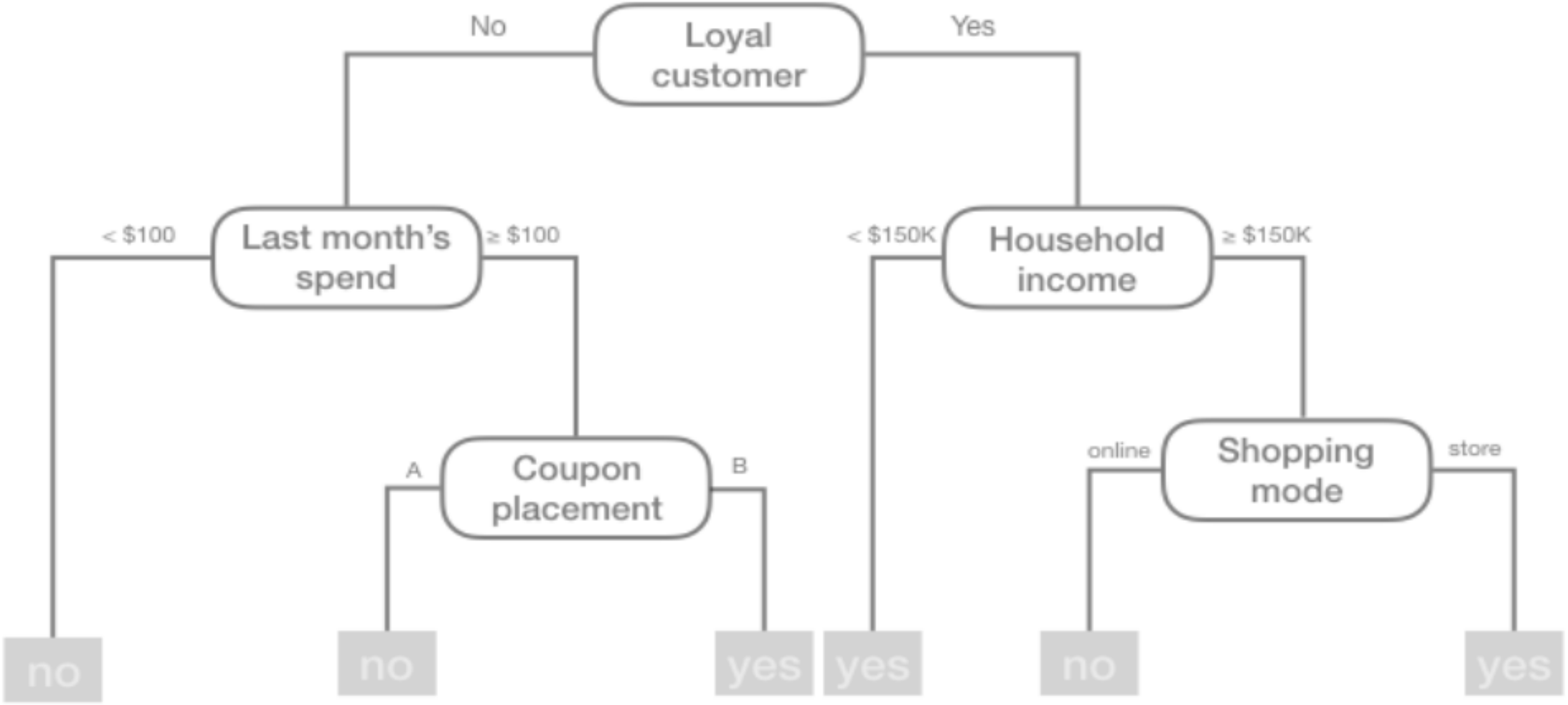
Classification problem: Iris data



Classification tree

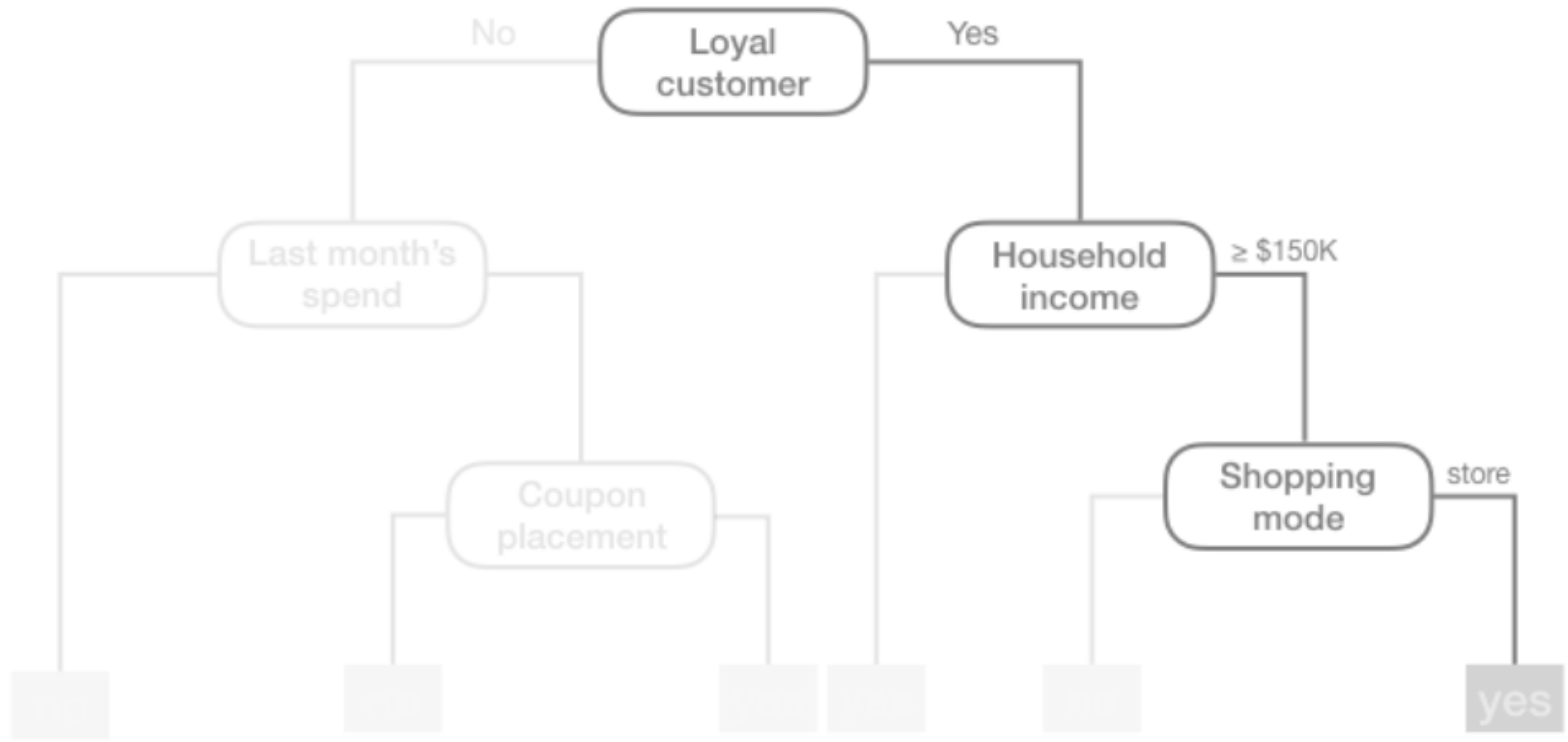


DECISION TREES - EXAMPLE FROM RETAIL DATA



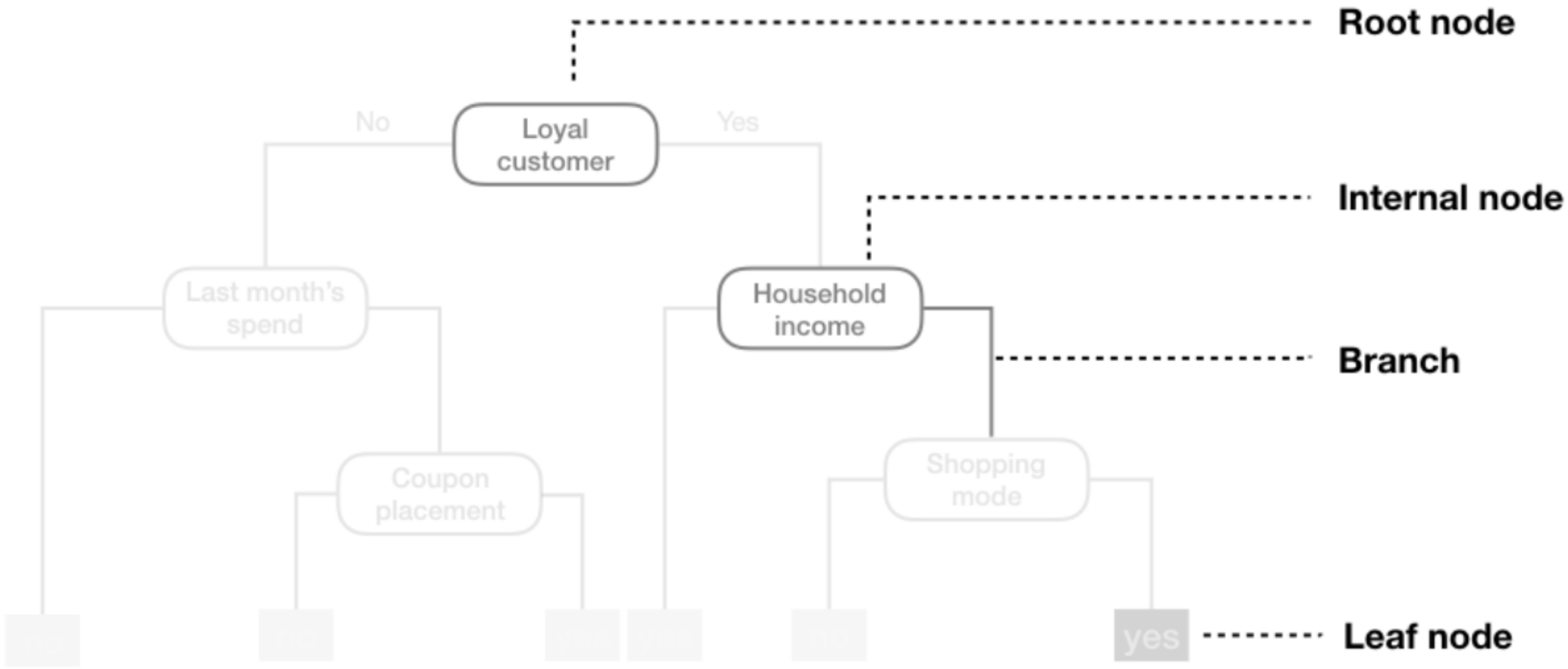
Will a customer redeem a coupon

WHAT FEATURE WILL ALLOW ME TO SPLIT THE OBSERVATIONS AT HAND IN A WAY THAT THE RESULTING GROUPS ARE AS DIFFERENT FROM EACH OTHER AS POSSIBLE (AND THE MEMBERS OF EACH RESULTING SUBGROUP ARE AS SIMILAR TO EACH OTHER AS POSSIBLE)?

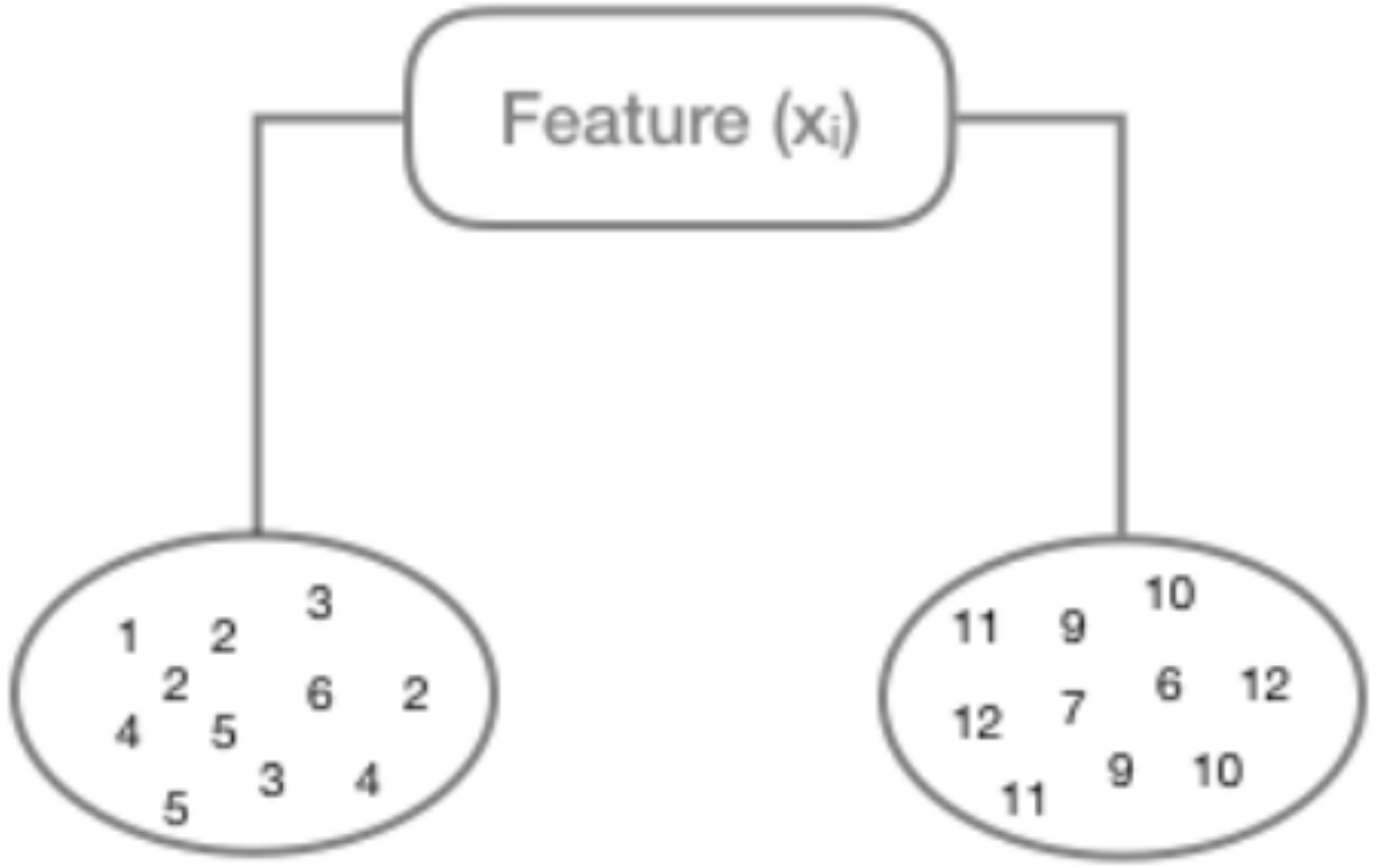



```
if Loyal Customer = Yes and Household income >= $150K and Shopping mode = store then coupon redemption = Yes
```

DECISION TREE TERMINOLOGY



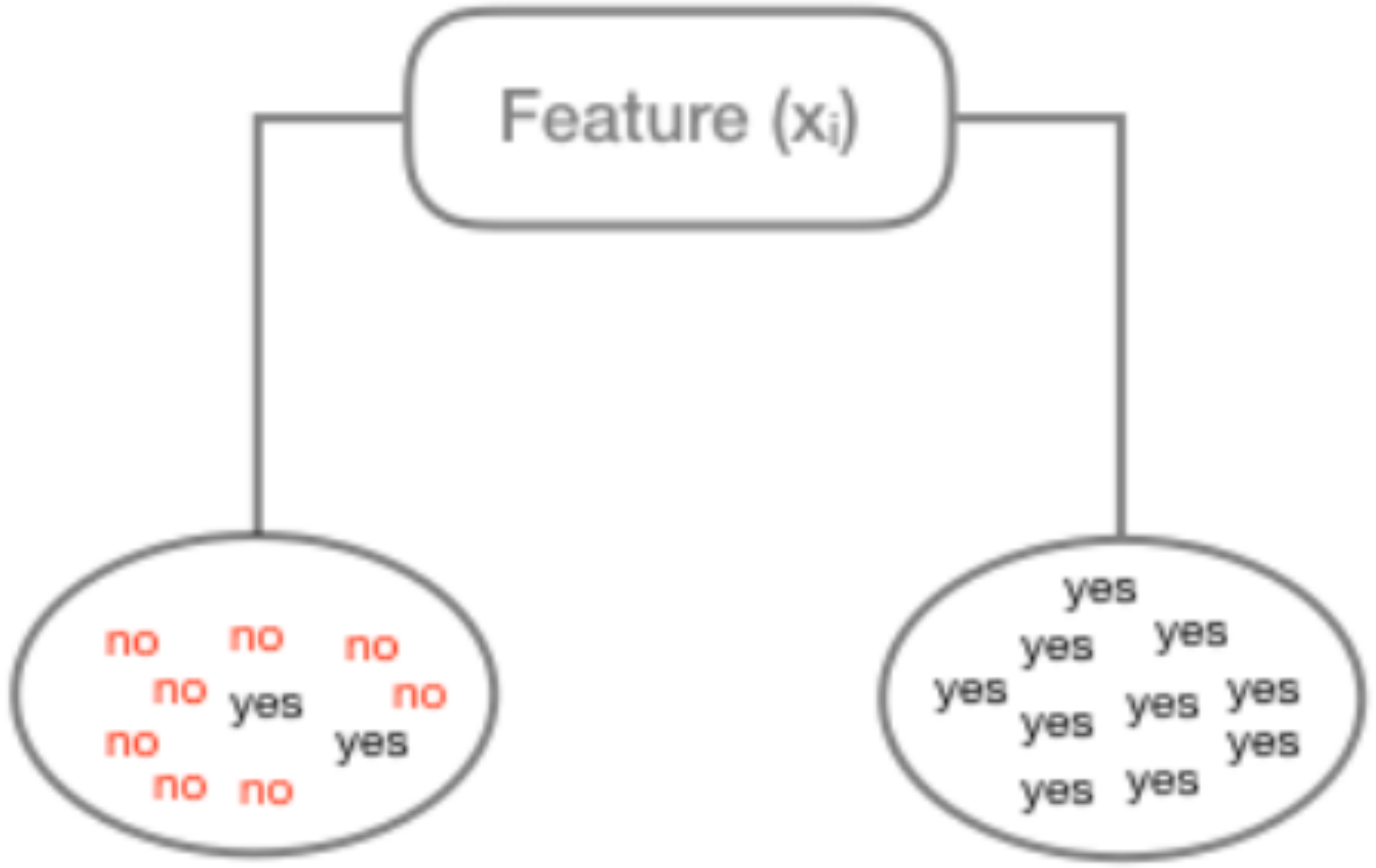
Regression tree



Prediction: 3.36 9.7

SSE: 24.55 36.1

Classification tree



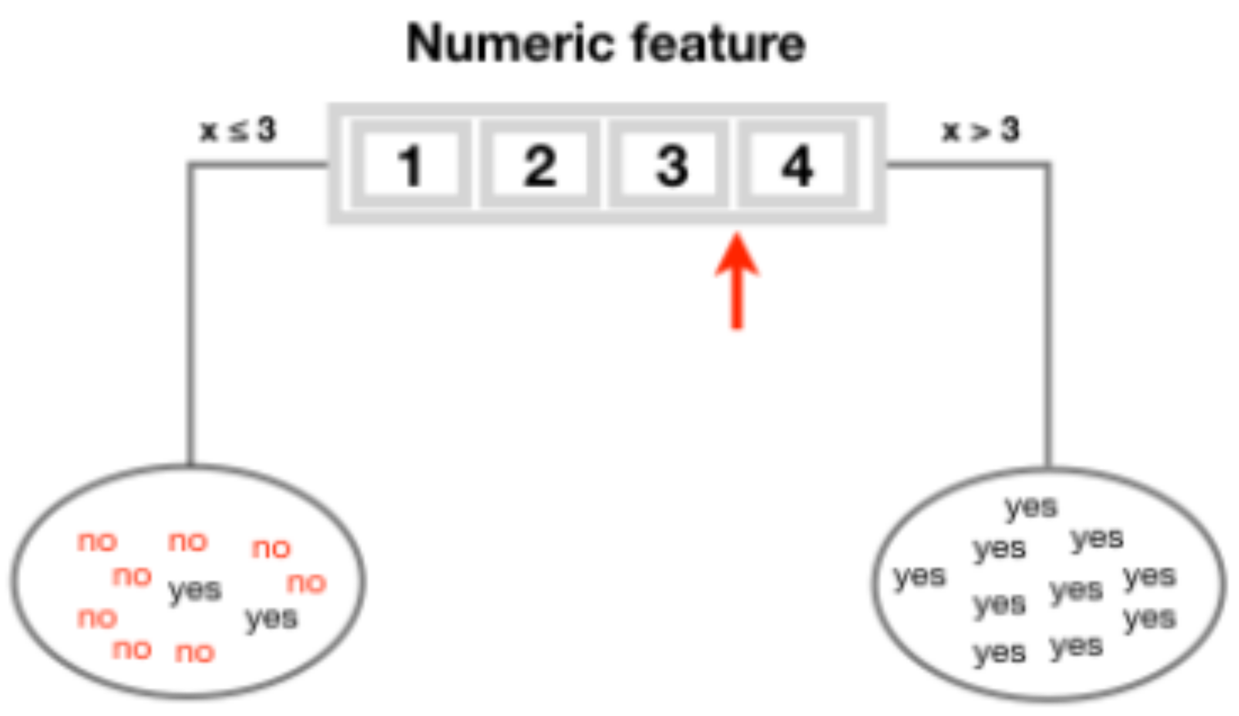
Prediction: no yes

Gini: .16 0

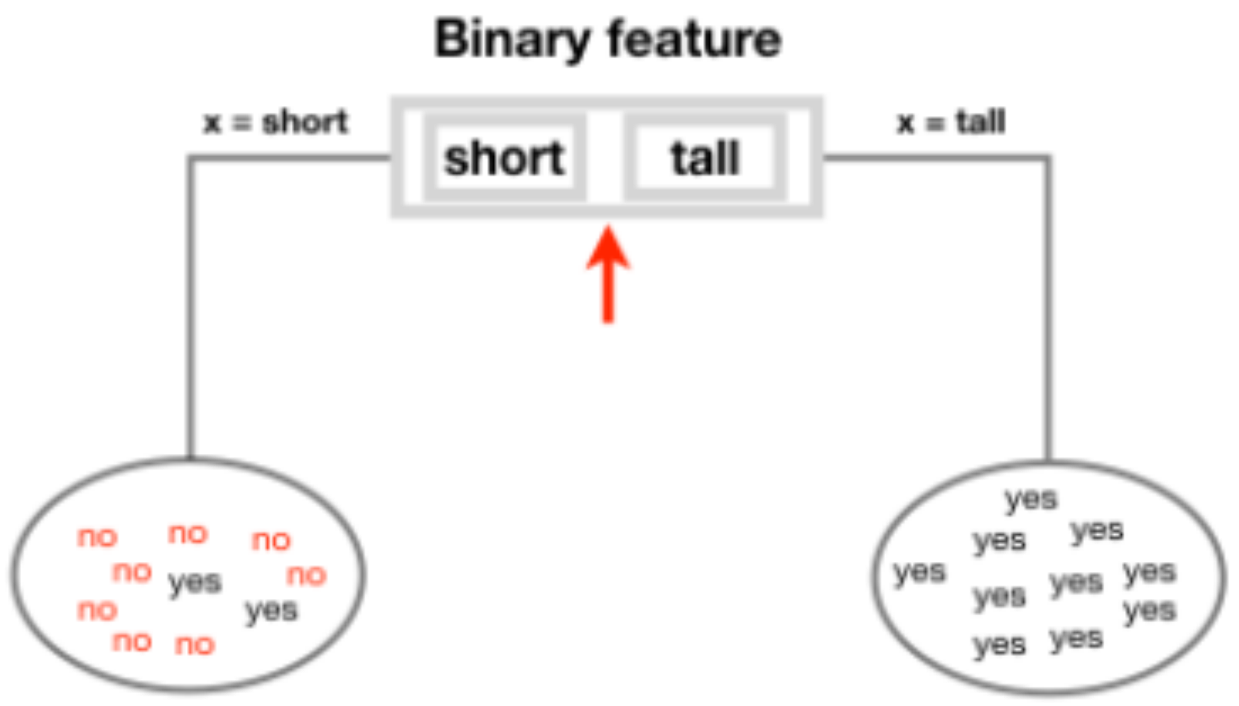
$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$

Objective: Minimize disimilarity in terminal nodes

- **Numeric feature**: Numeric split to minimize loss function



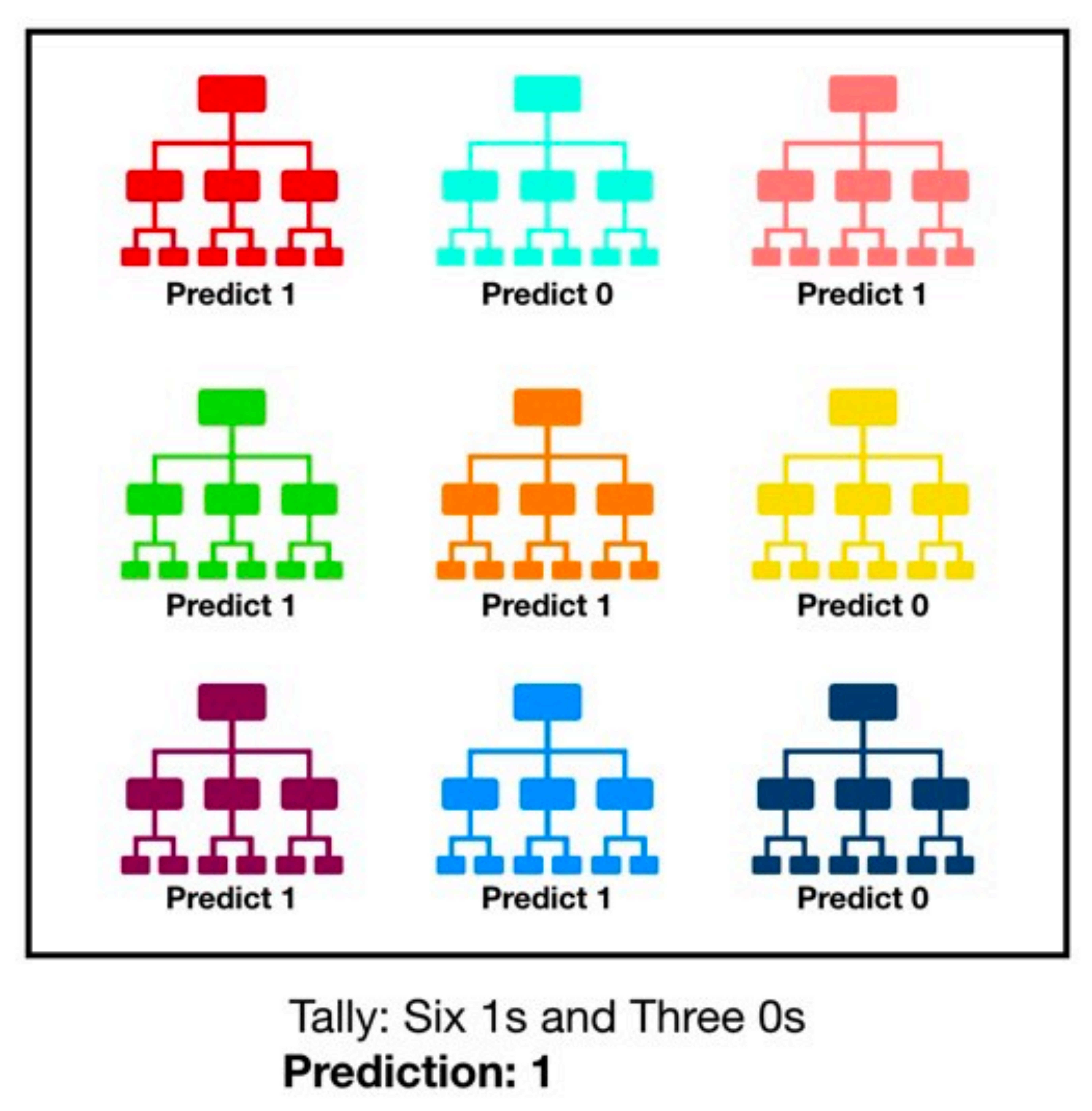
- **Binary feature**: Category split to minimize loss function



- **Multiclass feature**: Order feature classes based on mean target variable (regression) or class proportion (classification) and choose split to minimize loss function

Problem with Individual decision tree

- Tendancy to overfit - Decision trees overfit to the data they are trained on.
- To minimize the cost function, the boundaries that are defined become very specific to the training data.
- If the training data-set is small, or not a well representative of the diversity in entire data, it would produce very poor performance overall.
- Solution 1 - Train on the large dataset.
- Solution 2 - Reduce the complexity of the decision tree model (Constrain the minimum number of elements in the leaf node or depth of the tree)
- Solution 3 –– >



Random Forests

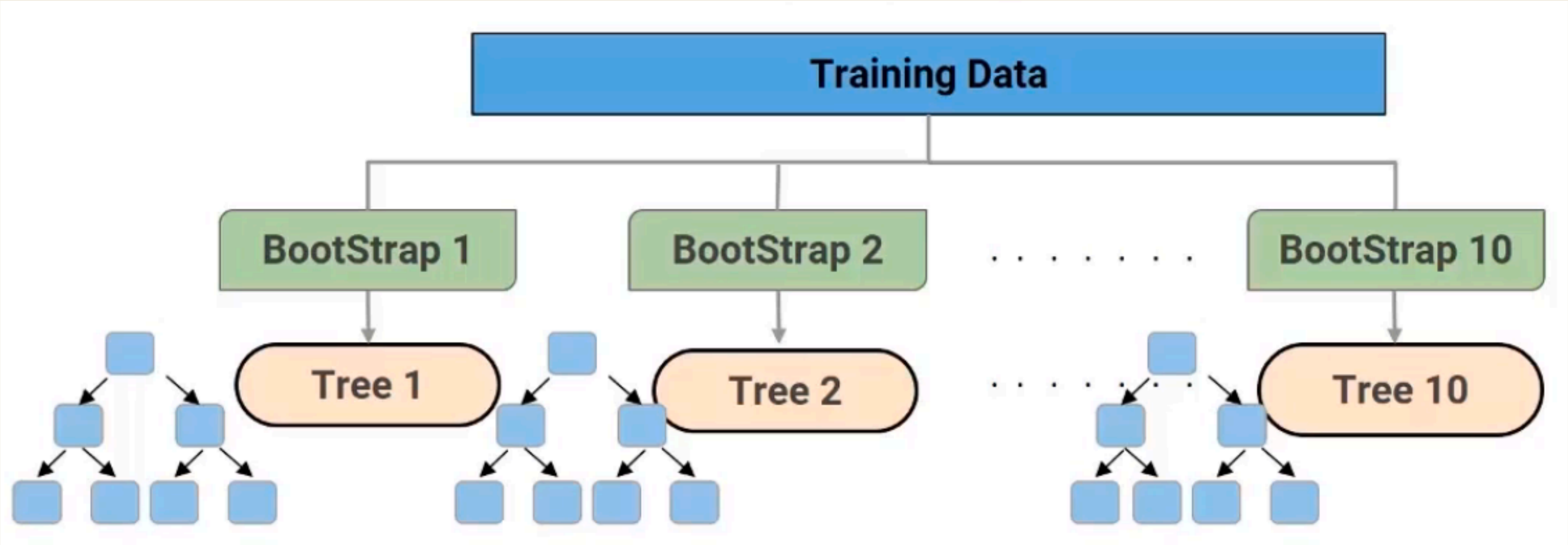
A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

It involves two primary steps that make them more powerful than decision trees (and other ML algorithms -

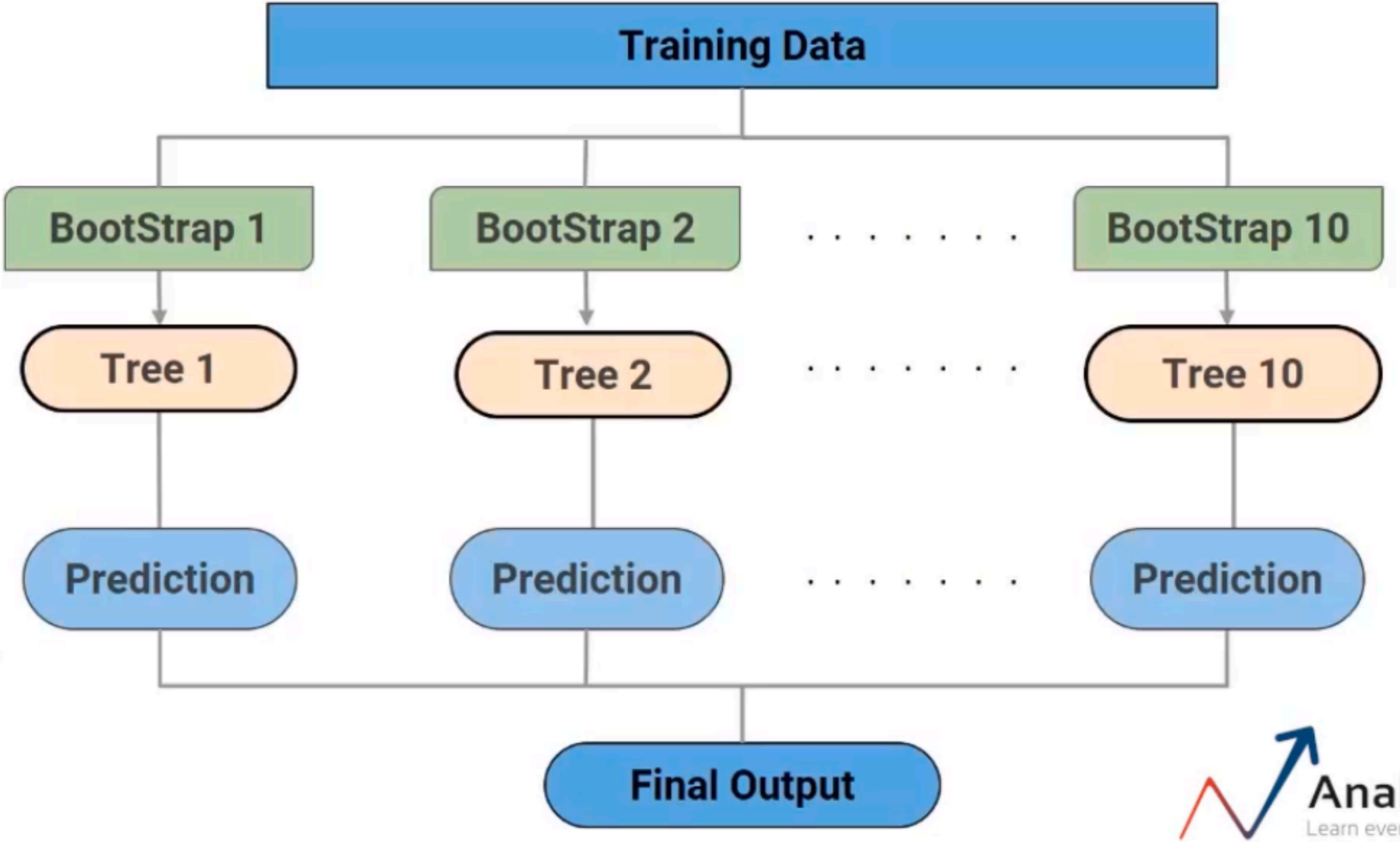
1) An ensemble of trees are trained on randomly subsampled subset of the data. (Bootstrap resampling)

2) The features on which individual trees are trained are constrained.

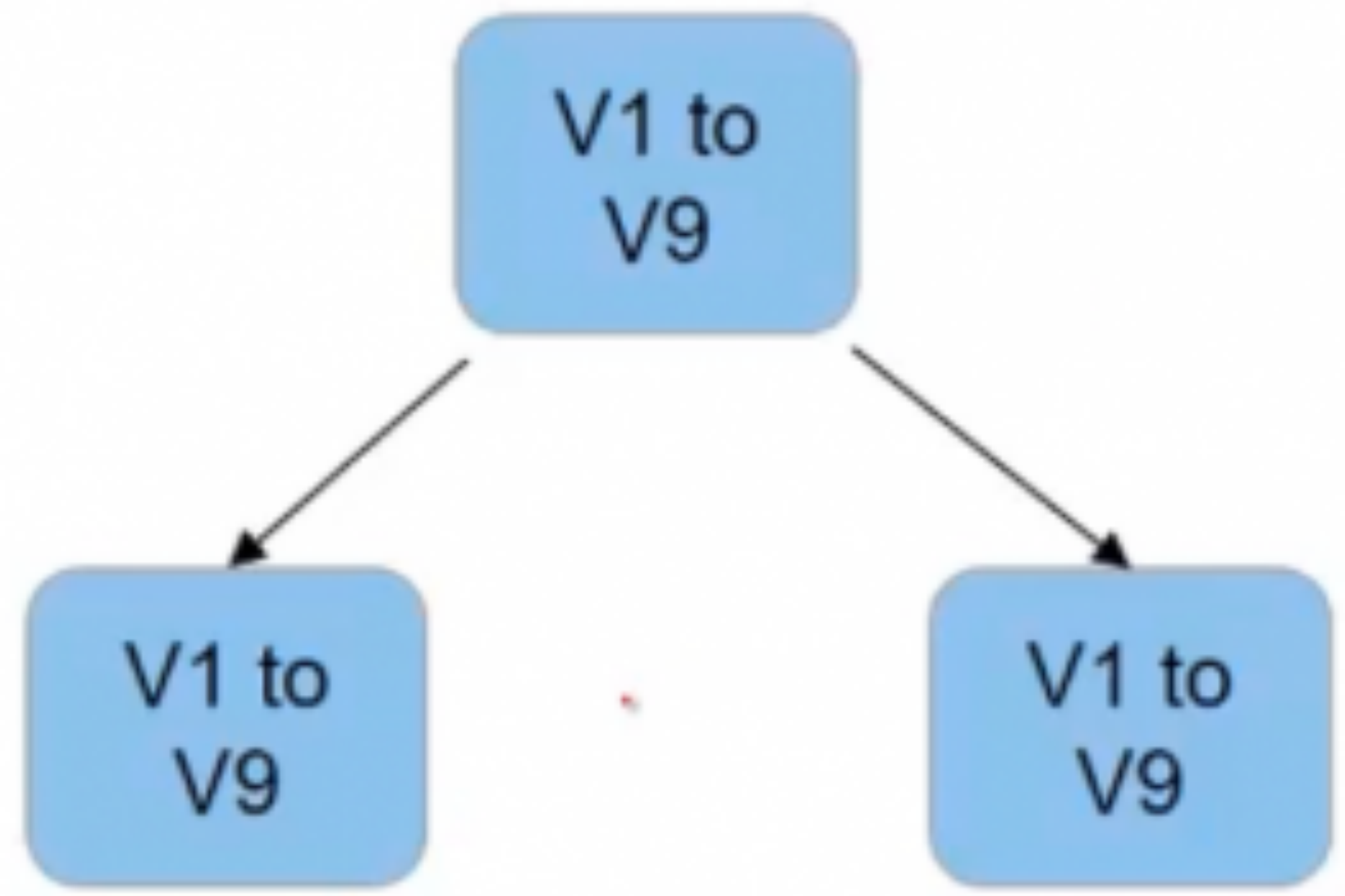
These steps ensure that the individual trees are uncorrelated. Taking an ensemble of decision trees ensure that the overall variance is reduced!


Training Data
BootStrap 1
BootStrap 2
BootStrap 10
Tree 1
Tree 2
Tree 10

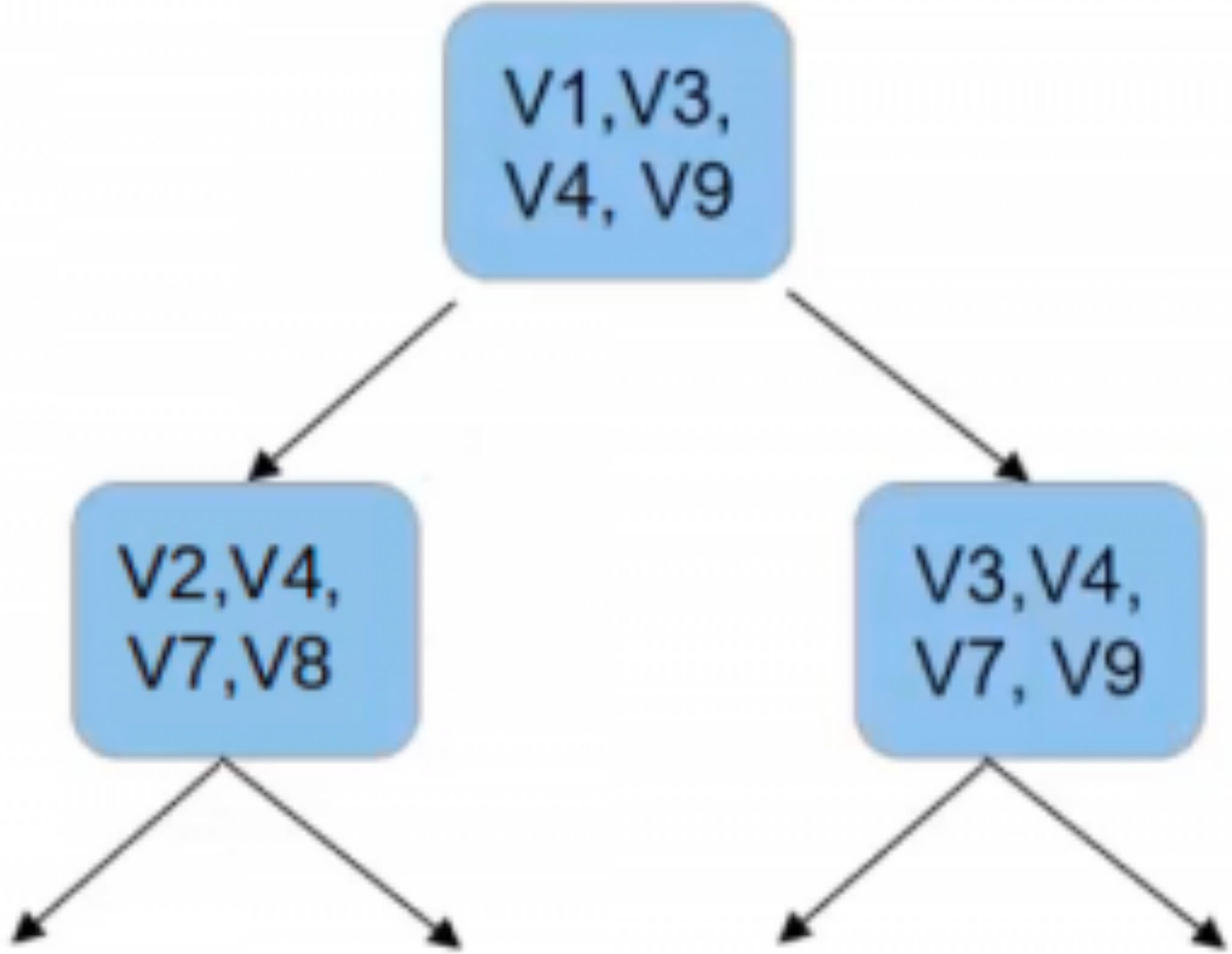
Random Forest

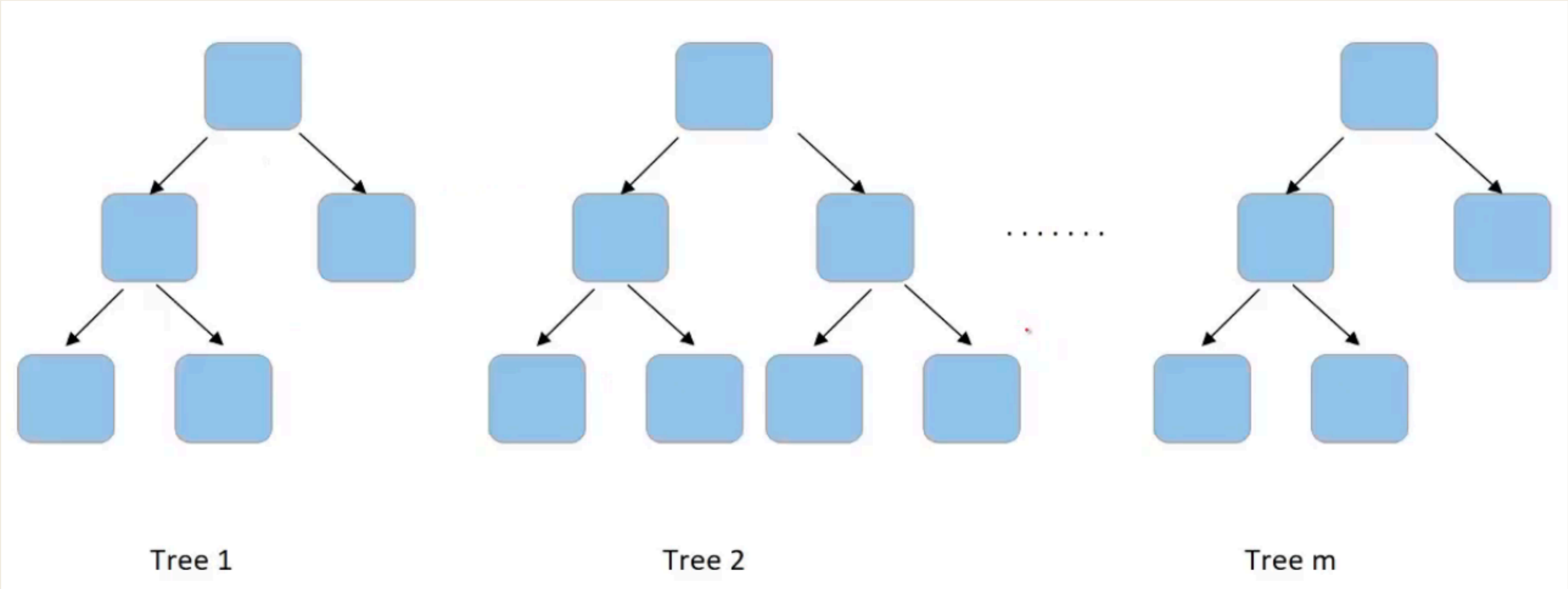


V1 V2 V3 V4 V5 V6 V7 V8 V9



Decision tree





Hyperparameters of Random Forests

- Number of decision trees
- Maximum number of features to consider at splitting (root, log, 1/3)
- Sampling type - with/without replacement.
- Maximum depth of the decision tree.
- Minimum sample for splitting
- Minimum samples per leaf.
- Criteria - Gini, Entropy gain

Advantages of Random Forests

1. It can be used in classification and regression problems.
2. It solves the problem of overfitting as output is based on majority voting or averaging.
3. It performs well even if the data contains null/missing values.
4. Each decision tree created is independent of the other thus it shows the property of parallelization. (n_jobs = -1, for using all the available machines)
5. It is highly stable as the average answers given by a large number of trees are taken.
6. It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.
7. It does not require any prior normalization of the features.

Disadvantages

1) For very large data sets, the size of the trees can take up a lot of memory.
2) It can tend to overfit, so you should tune the hyperparameters.

References -

1) Understanding Random Forest - https://towardsdatascience.com/understanding-random-forest-58381e0602d2

2) Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? - https://www.jmlr.org/papers/volume15/delgado14a/delgado14a.pdf?source=post_page---------------------------

3) Random Forest in Python - https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

4) Analytics Vidhya Random Forests explanation - https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Some Cool Seismological Applications

Dempsey, D. E., Cronin, S. J., Mei, S., & Kempa-Liehr, A. W. (2020). Automatic precursor recognition and real-time forecasting of sudden explosive volcanic eruptions at Whakaari, New Zealand. *Nature communications*, *11*(1), 1-8.

Rouet-Leduc, B., Hulbert, C., & Johnson, P. A. (2019). Continuous chatter of the Cascadia subduction zone revealed by machine learning. *Nature Geoscience*, *12*(1), 75-79.

Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C. J., & Johnson, P. A. (2017). Machine learning predicts laboratory earthquakes. *Geophysical Research Letters*, *44*(18), 9276-9282.

Thanks

Questions?