Machine learning: Random Forest

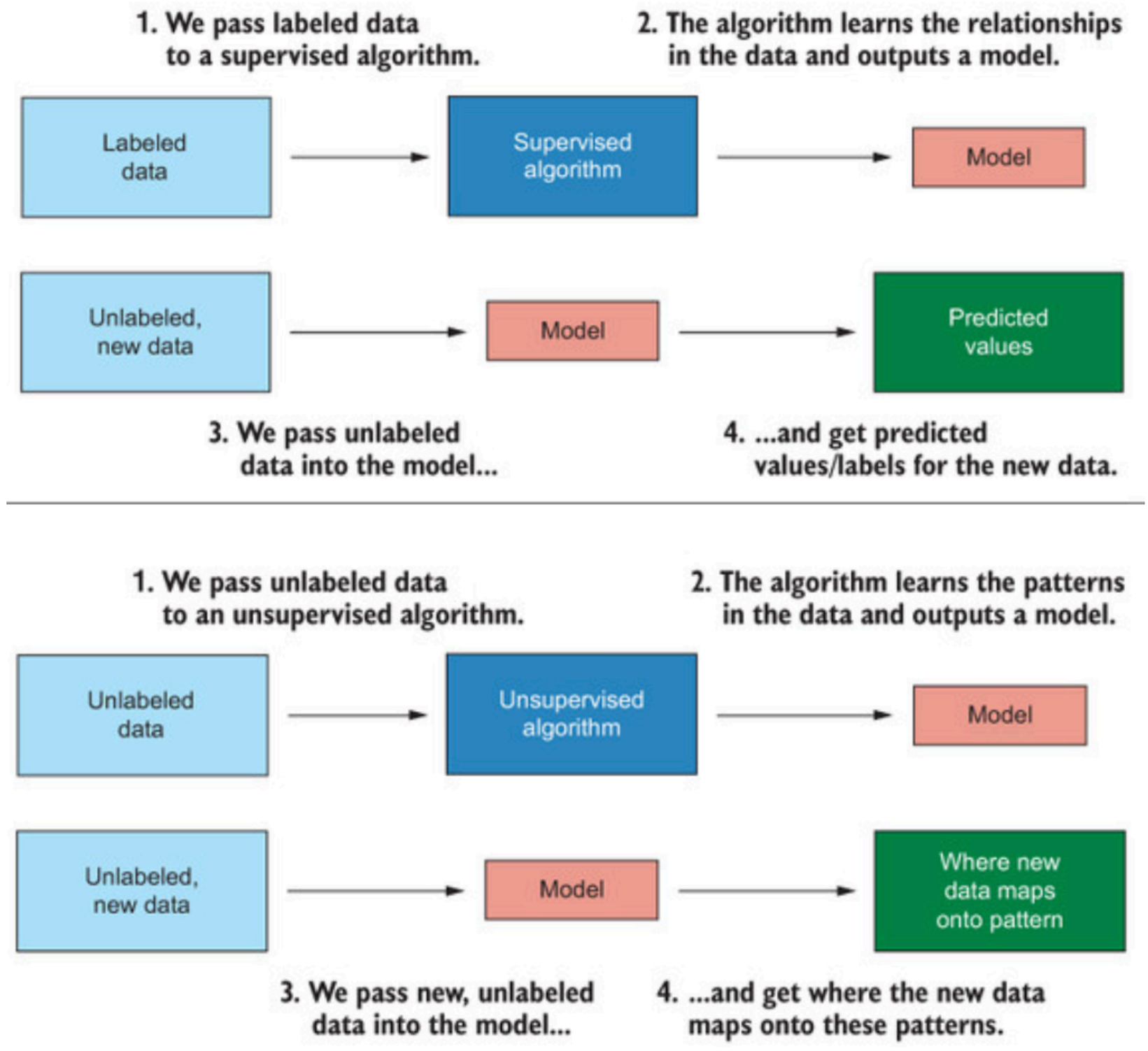
Alex Di Genova

05/07/2024

Machine learning algorithms

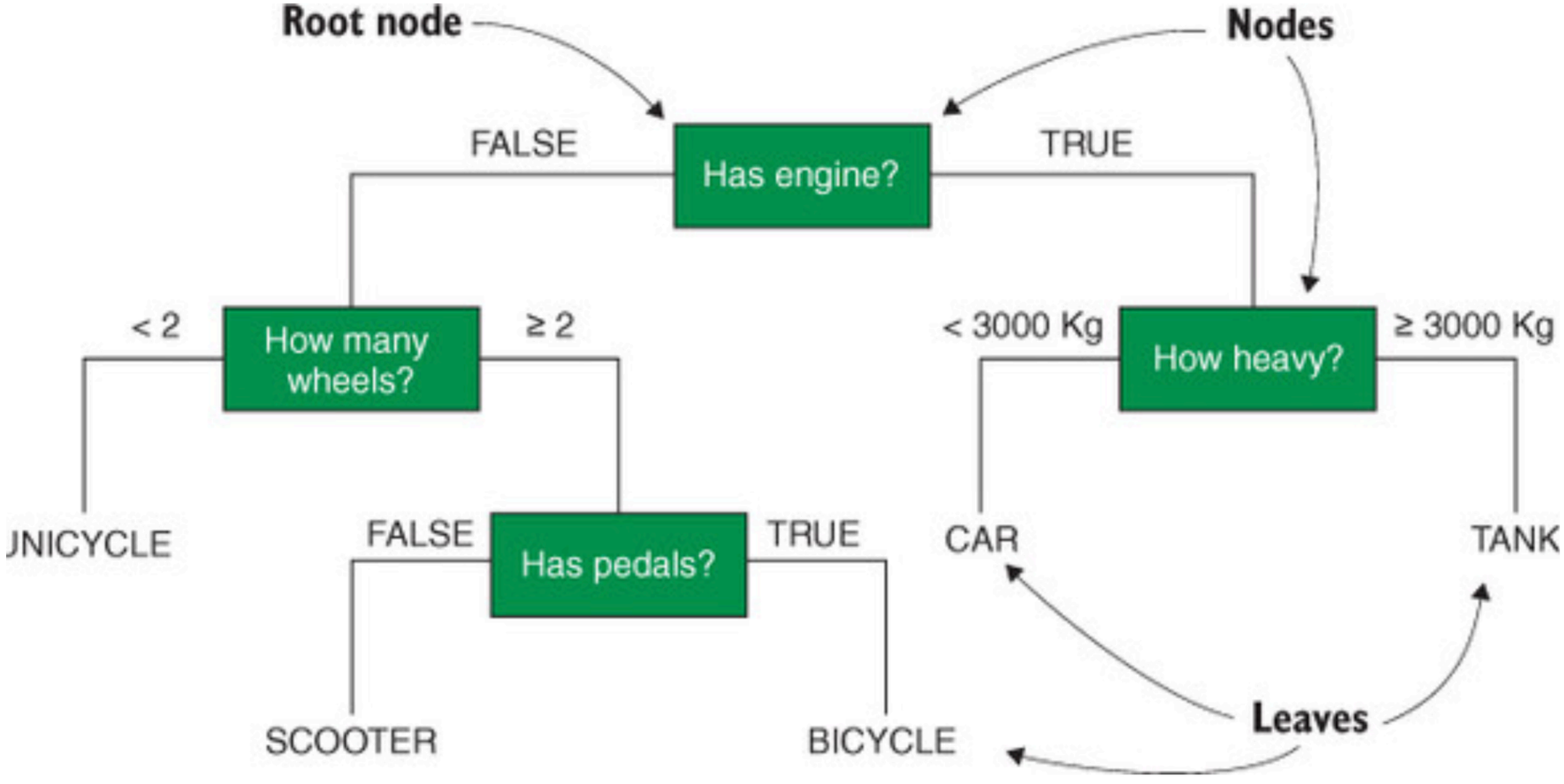
Classes

- Supervised
  - Classification
  - Regression
- Unsupervised
  - Dimension Reduction
  - Clustering
- Semi-supervised



What is Random Forest?

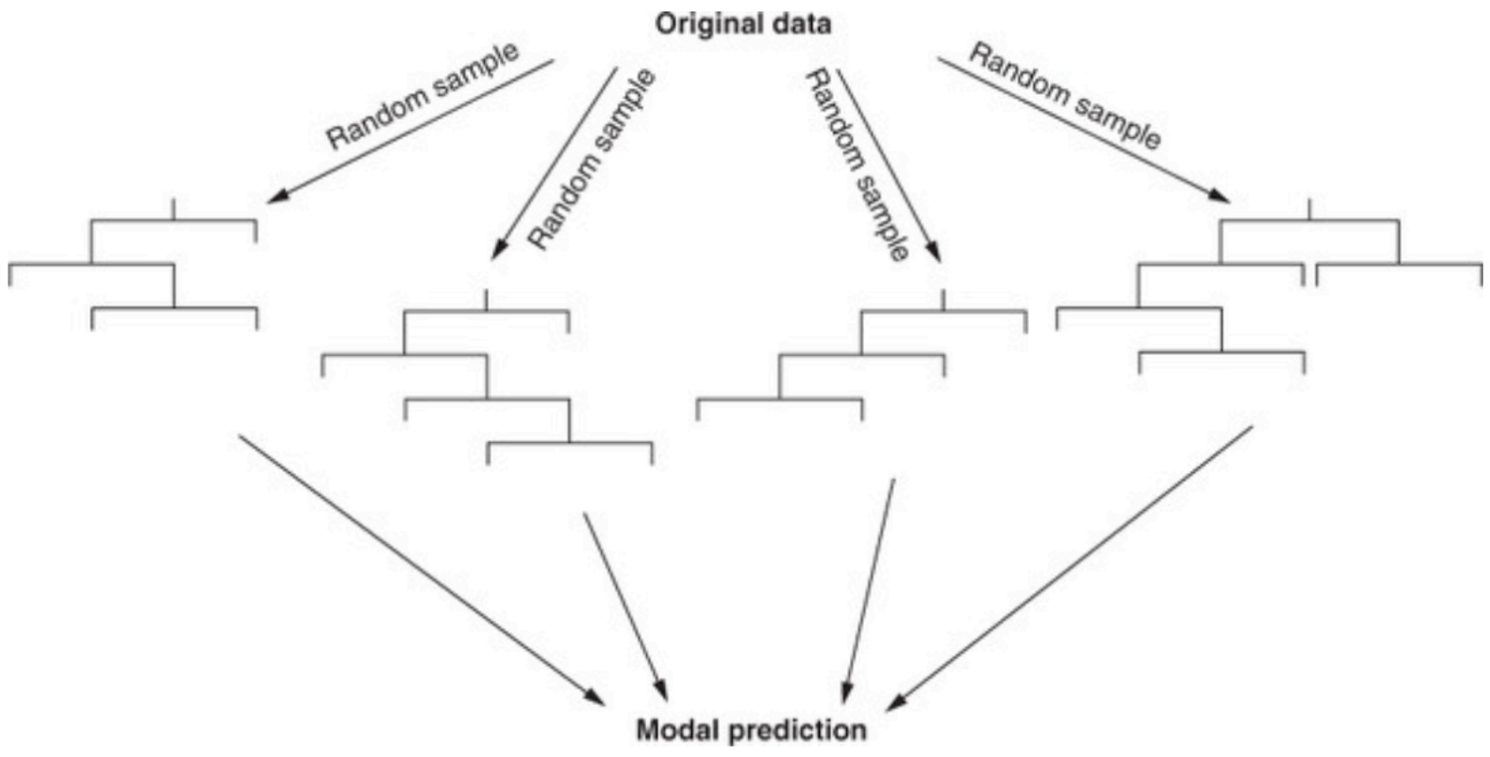
1. Ensemble learning method for classification and regression.
2. Combines multiple decision trees to improve accuracy.
3. Bagging is an ensemble technique that trains multiple sub-models in parallel on bootstrap samples of the training set. Each sub-model then votes on the prediction for new cases. Random forest is an example of a bagging algorithm.



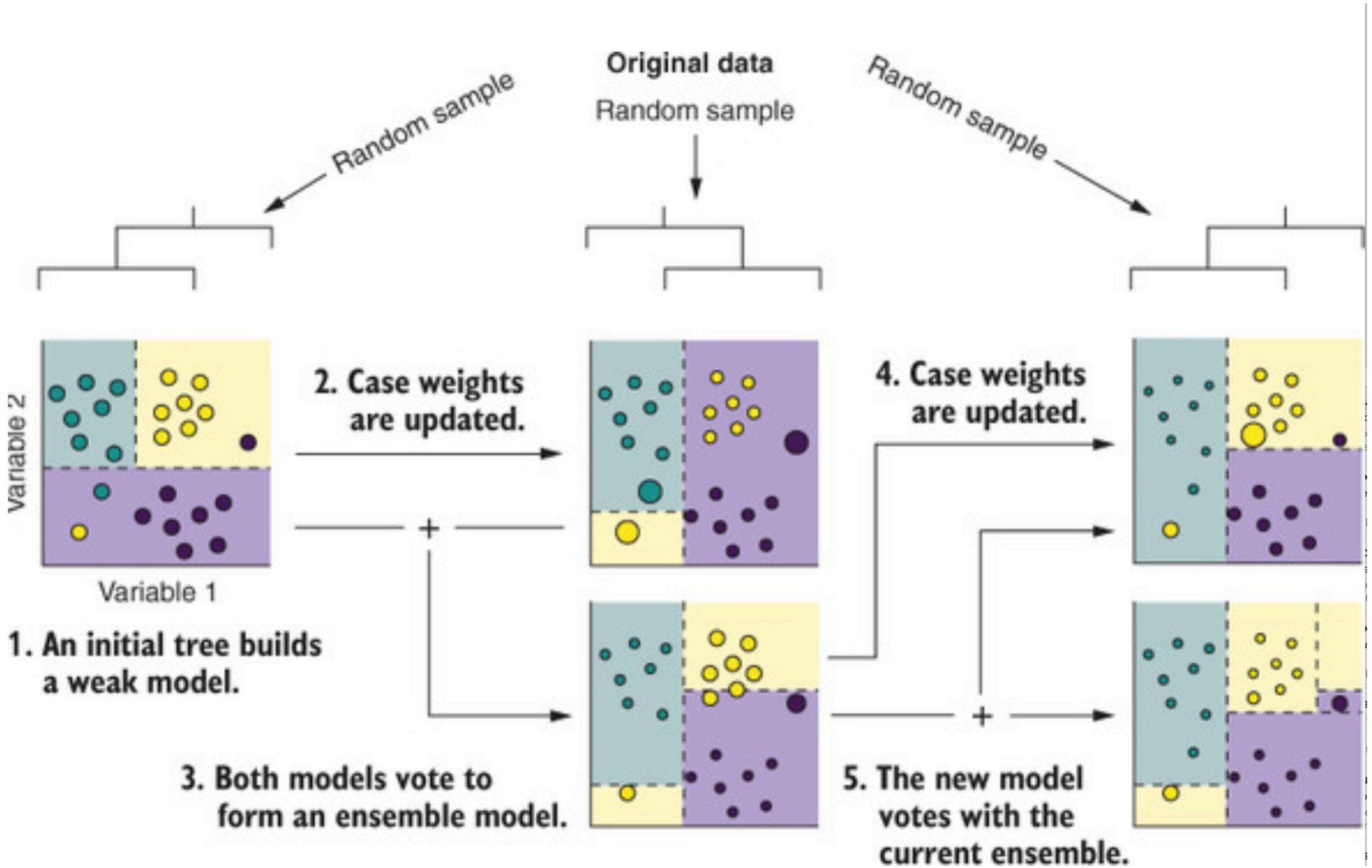
Random forest

- Multiple decision trees are learned in parallel, each one trained on a bootstrap sample of cases from the training set.
- When predicting new data, each tree makes a prediction, and the modal (most frequent) prediction wins.

•Voting:
  - Majority vote for classification
  - Averaging for regression



Random forest



$$\text{model weight} = 0.5 \times \ln\left(\frac{1 - p(\text{incorrect})}{p(\text{incorrect})}\right)$$

$$\text{case weight} = \begin{cases} \text{initial weight} \times e^{-\text{model weight}}; \text{ if correctly classified} \\ \text{initial weight} \times e^{\text{model weight}}; \text{ if incorrectly classified} \end{cases}$$

Random Forest parameters

- ntree — The number of individual trees in the forest
- mtry— The number of features to randomly sample at each node
- nodesize— The number of cases allowed in a leaf
- maxnodes— The maximum number of leaves allowed
- The more trees we have, the better.

A real example

- Using random forest to determine genomic haplotypes.

•Implementation Steps:

  •- Data preparation

  •- Model training

  •- Model evaluation

•Tips:

  •- Feature importance analysis

  •- Handling class imbalance

  •- Cross-validation

Book:

Machine learning with R

Questions?

Practice!!!