Random Forests

Megan Ruffley (she/her)
Machine Learning Seminar
November 14, 2019

What is random forest?

1. Supervised machine learning
2. The forest is made up of decision trees
3. Random
4. Ensemble approach

Breimen L. (2001) Random Forests. *Machine Learning*, 45, 5-32.

Supervised machine learning

- **Trains** a function that, given a *sample of data* and *desired outputs*, best approximates the relationship between input and output observable in the data.
- Required prior knowledge of what the output should be

Supervised machine learning

- **Trains** a function that, given a *sample of data* and *desired outputs*, best approximates the relationship between input and output observable in the data.
- Required prior knowledge of what the output should be

Movie	My score	Director	Production Co	1st three genres	1st three plot keywords	IMDB score	Meta-score
Star Wars: The Last Jedi	10	Rian Johnson	Lucasfilm	action, adventure, fantasy	wisecrack humor, deception, betrayal	7.5	85
Wonder Woman	9	Patty Jenkins	Warner Bros	action, adventure, fantasy	god, wonder woman, mission	7.6	76
Logan	7	James Mangold	20th Century Fox	action, drama, sci-fi	x-men, marvel, superhero	8.1	77
Zootopia	9	Byron Howard	Walt Disney	animation, adventure, comedy	fox, police, con artist	8	78
Captain America: Civil War	8	Anthony Russo	Marvel Studios	action, adventure, sci-fi	marvel, comic book, superhero	7.8	75
Beauty and the Beast	6	Bill Condon	Mandeville	family, fantasy, musical	beast, fairy tale, disney	7.3	65
Moana	7	Ron Clements	Walt Disney	animation, adventure, comedy	island, ocean, polynesia	7.6	81

Supervised machine learning

- **Trains** a function that, given a *sample of data* and *desired outputs*, best approximates the relationship between input and output observable in the data.
- Required prior knowledge of what the output should be

Movie	My score	Director	Production Co	1st three genres	1st three plot keywords	IMDB score	Meta-score
Star Wars: The Last Jedi	10	Rian Johnson	Lucasfilm	action, adventure, fantasy	wisecrack humor, deception, betrayal	7.5	85
Wonder		Patty			god, wonder woman,		
Logan							
Zootopia							
Captain America: Civil War	8	Anthony Russo	Marvel Studios	action, adventure, sci-fi	marvel, comic book, superhero	7.8	75
Beauty and the Beast	6	Bill Condon	Mandeville	family, fantasy, musical	beast, fairy tale, disney	7.3	65
Moana	7	Ron Clements	Walt Disney	animation, adventure, comedy	island, ocean, polynesia	7.6	81

How would I score The Avengers based on this sample of data?

Supervised machine learning

- Trains a function that, given a sample of data and desired outputs, best approximates the relationship between input and output observable in the data.
- Required prior knowledge of what the output should be
- Two main types of supervised learning….
    - ***Classification (discrete predictions)***
    - ***Regression (continuous predictions)***

Supervised machine learning

- Two main types of supervised learning….
  - ***Classification (discrete predictions)***
  - ***Regression (continuous predictions)***

- Common algorithms include random forests, neural networks, logistic regression, and support vector machines.

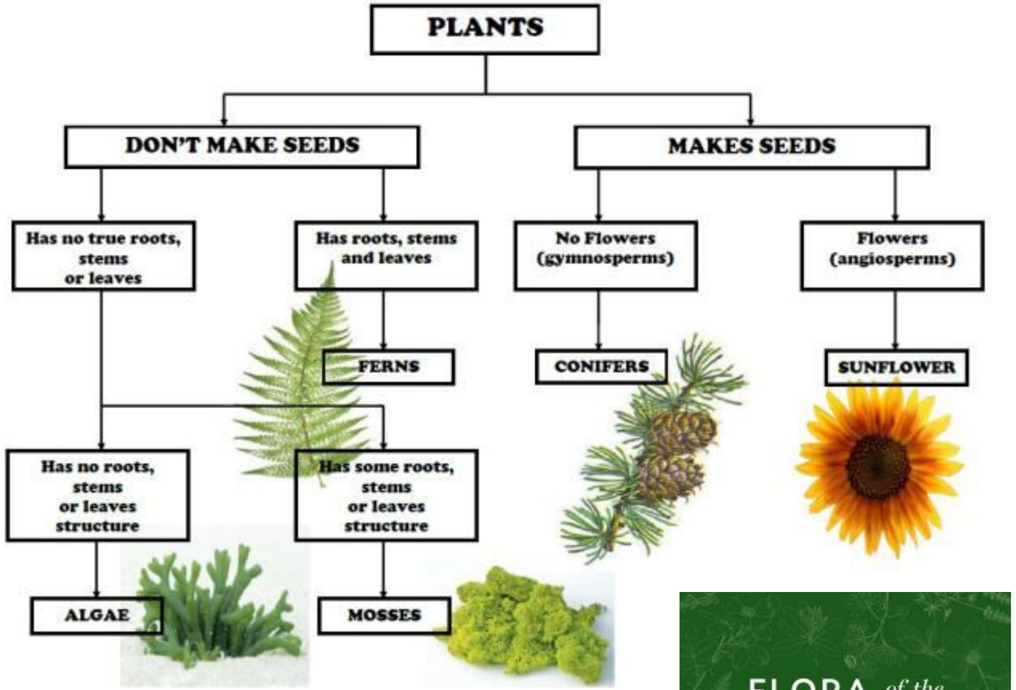
Unsupervised learning, alternatively, identifies previously unknown patterns in a data set without pre-existing labels.

What is random forest?

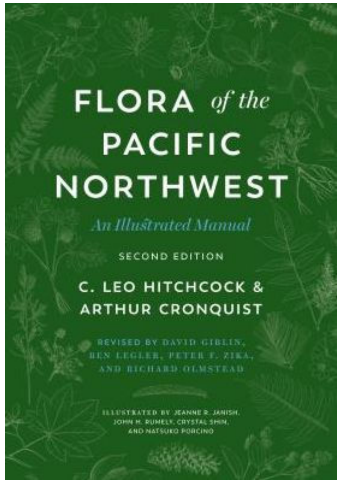
1. Supervised machine learning
2. The forest is made up of decision trees
3. Random
4. Ensemble approach

Breimen L. (2001) Random Forests. *Machine Learning*, 45, 5-32.

The forest is made up of decision trees



Dichotomous Keys

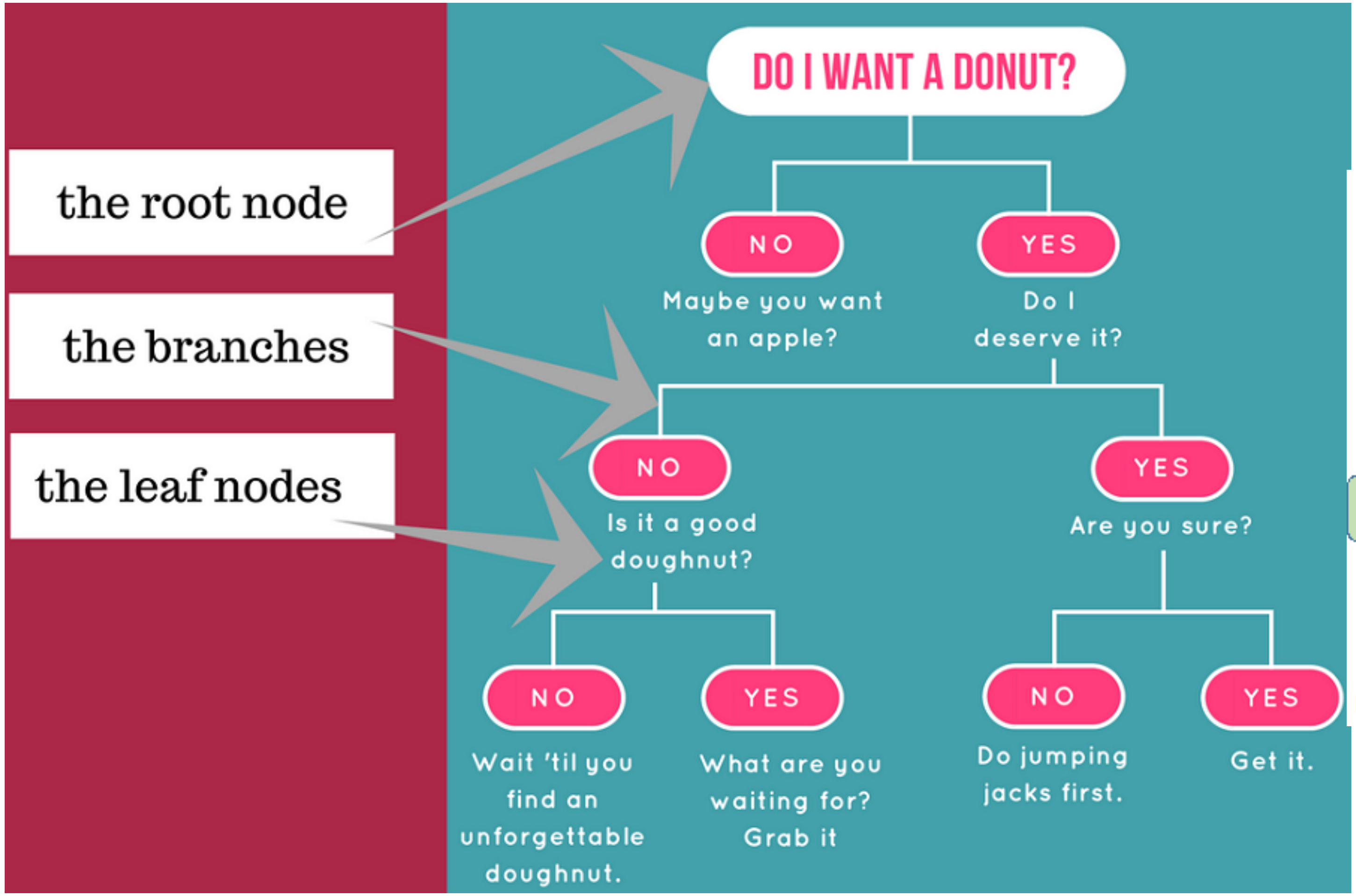


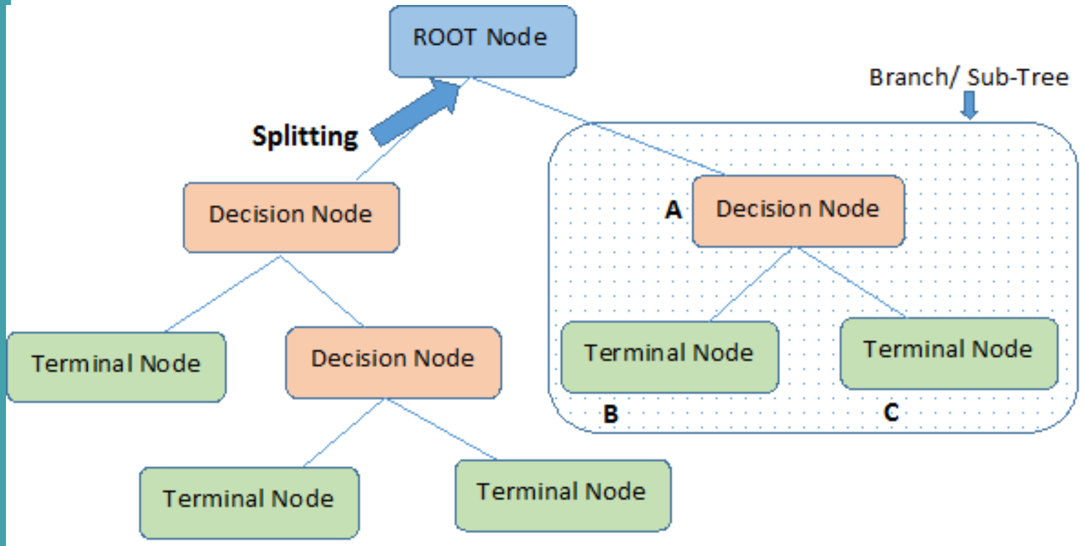
12 *Inclusive Key*

KEY II. PLANTS WITH OPPOSITE OR WHORLED SIMPLE LEAVES

1. Leaves subopposite
   2. Leaves toothed . . . . *Rhamnus*
   2. Leaves entire; southern
      3. Leaves greater than 5 cm long . . . . *Lagerstroemia*
      3. Leaves less than 5 cm long . . . . *Fontanesia*
1. Leaves distinctly opposite or whorled
   4. Leaves lobed
      5. Leaves mostly pinnately lobed
         6. Margin of lobes entire; sap clear; shrubs; fruit a capsule . . . . *Syringa*
         6. Margin of lobes serrate; sap milky or clear; trees or tall shrubs; fruit a capsule or head of achenes
            7. Trees; sap milky; fruit a head of achenes . . . . *Broussonetia*
            7. Shrubs; sap clear or milky; fruit a capsule . . . . *Hydrangea*
      5. Leaves palmately lobed
         8. Leaf blades less than 20 cm long
            9. Petioles with stipules and glands, or if lacking glands, the lower surface of leaf densely pubescent; fruit a drupe . . . . *Viburnum*
            9. Petioles lacking stipules and glands, or if stipules present, the lower surface of leaf glabrous to pubescent, not densely so; fruit a samara . . . . *Acer*
         8. Leaf blades greater than 20 cm long
            10. Leaves with long tapering tip, glabrous or softly pubescent, usually in whorls of 3; pith continuous; fruit a long cylindrical capsule, 20–50 cm long . . . . *Catalpa*

The forest is made up of decision trees



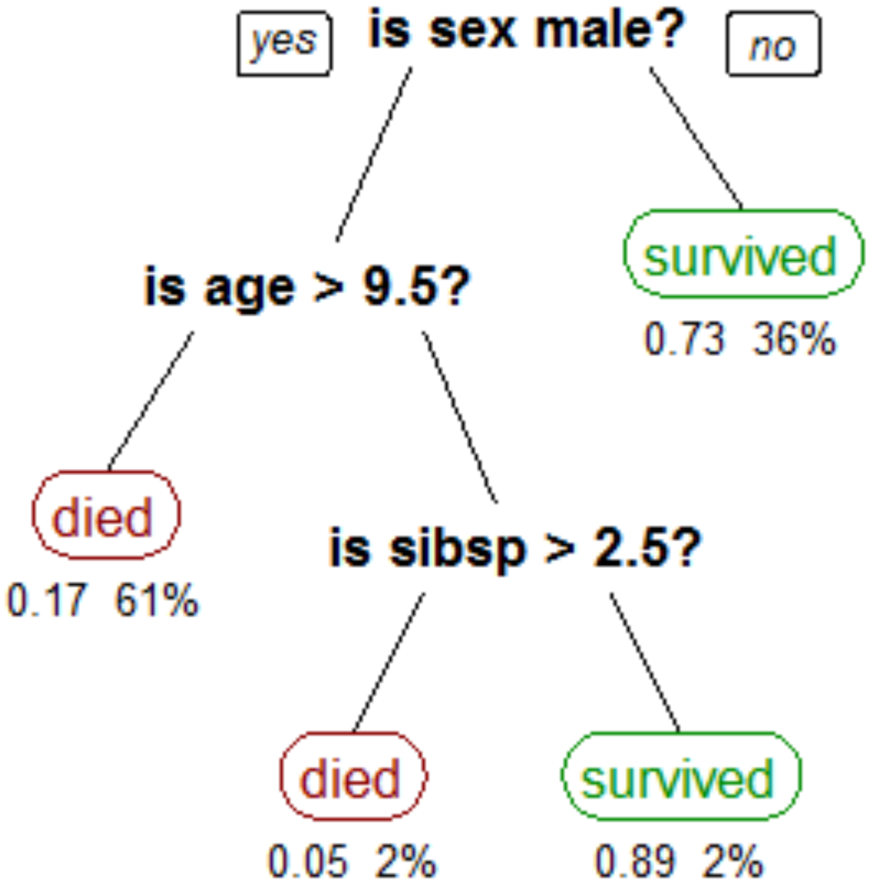


Note:- A is parent node of B and C.

The forest is made up of decision trees

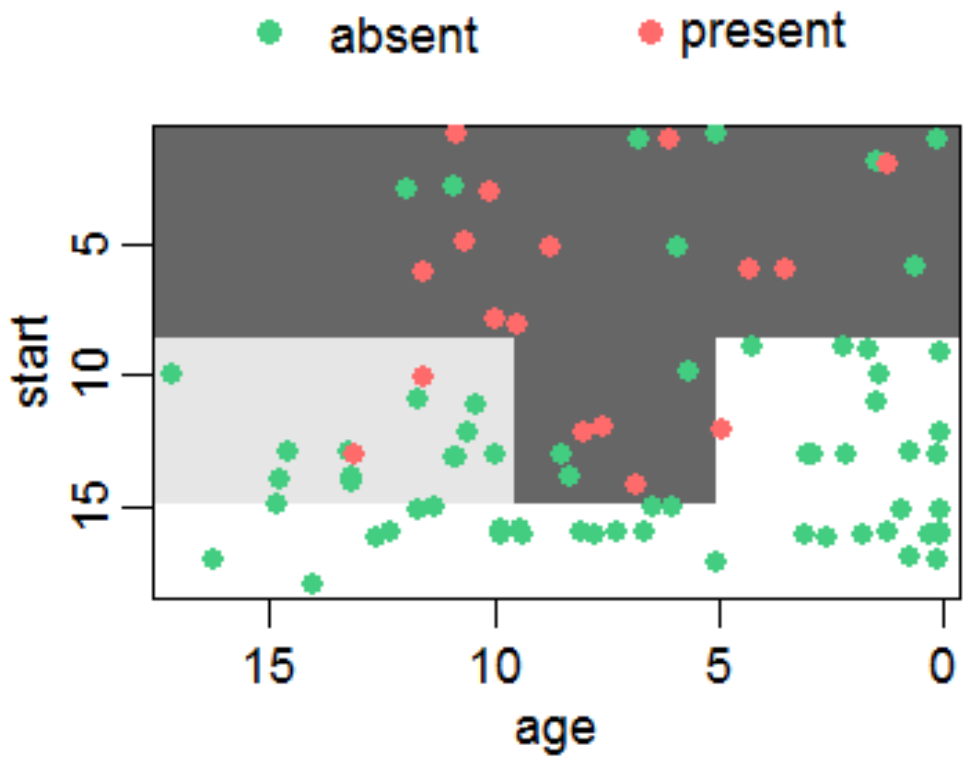
- There are two types of decision trees
  - Classification trees (discrete predictions)
  - Regression trees (continuous predictions)

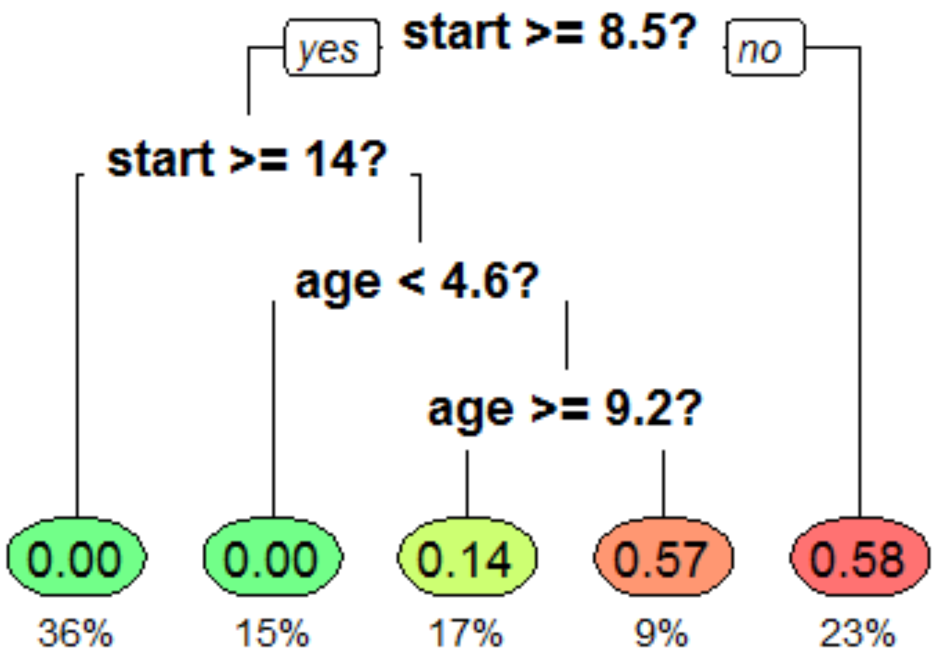
- CART (classification and regression trees)
  - Recursive partitioning algorithm for building these trees.



The forest is made up of decision trees

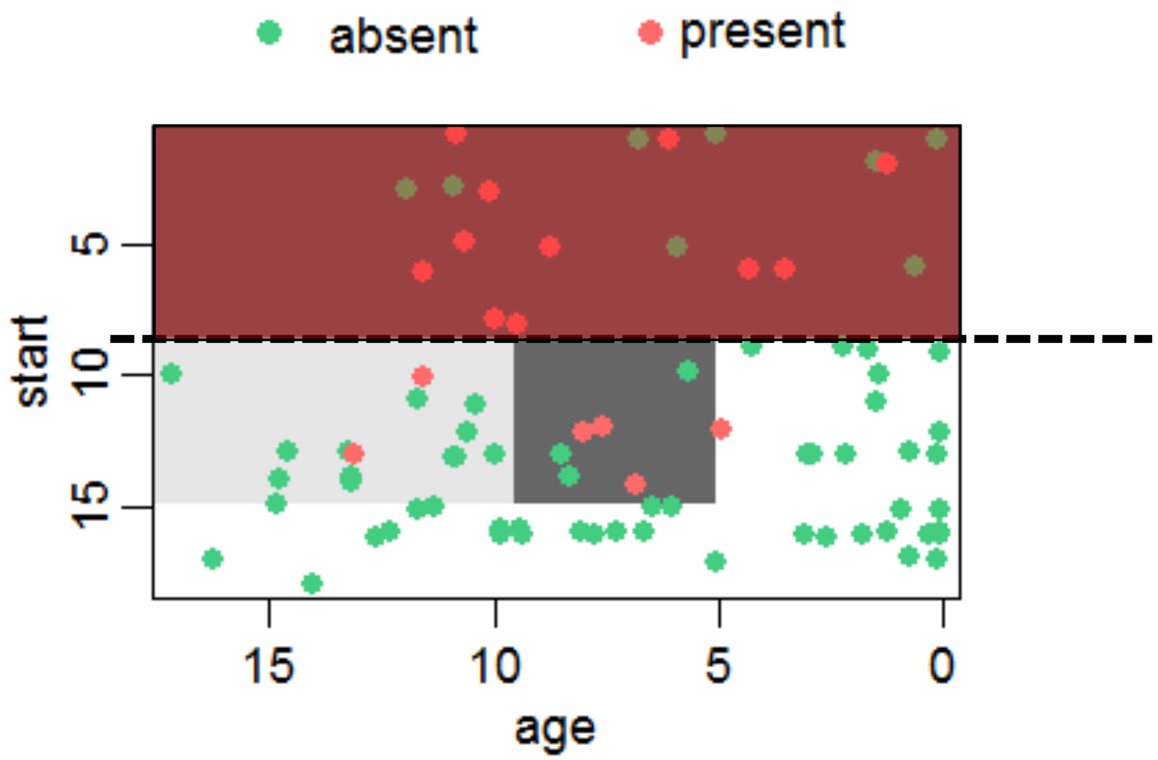
- There are two types of decision trees
  - Classification trees

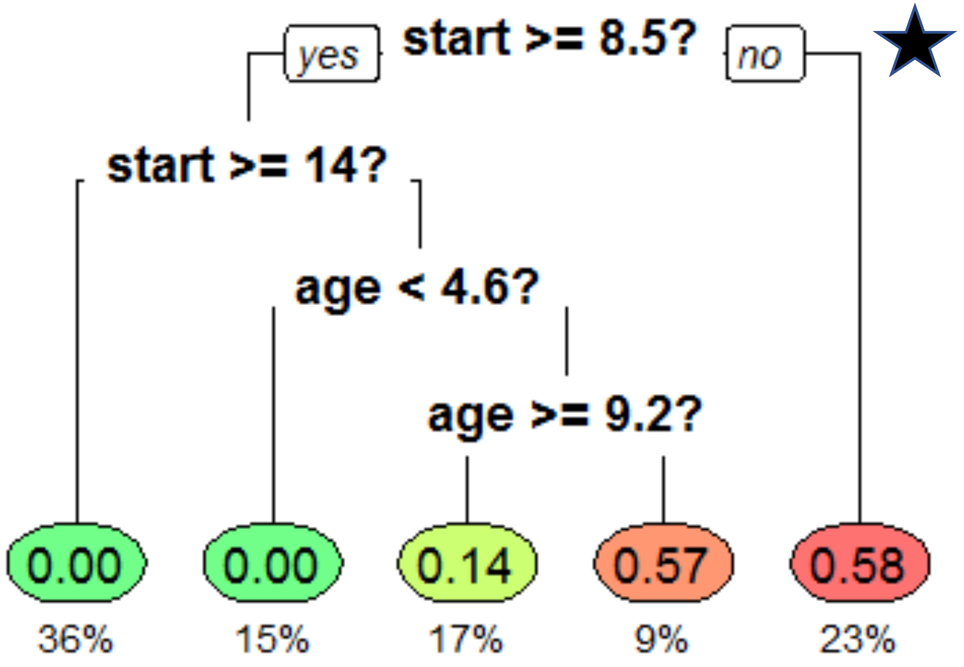




The forest is made up of decision trees

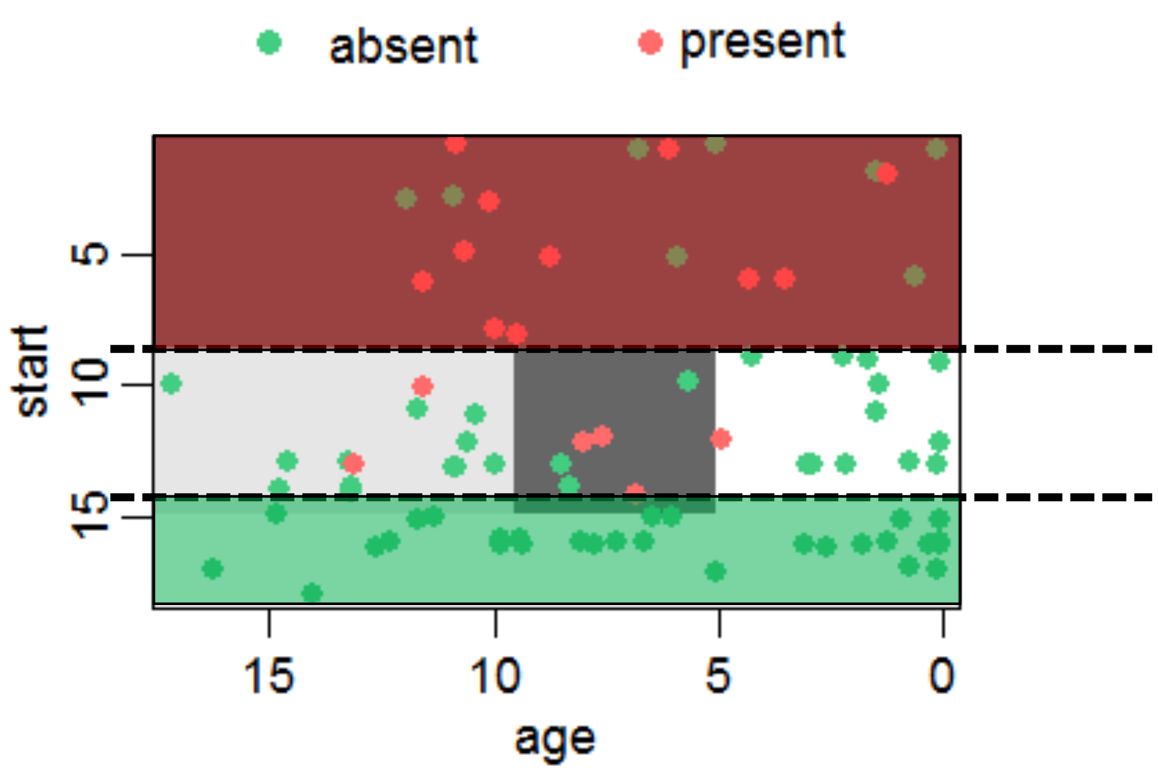
- There are two types of decision trees
  - Classification trees

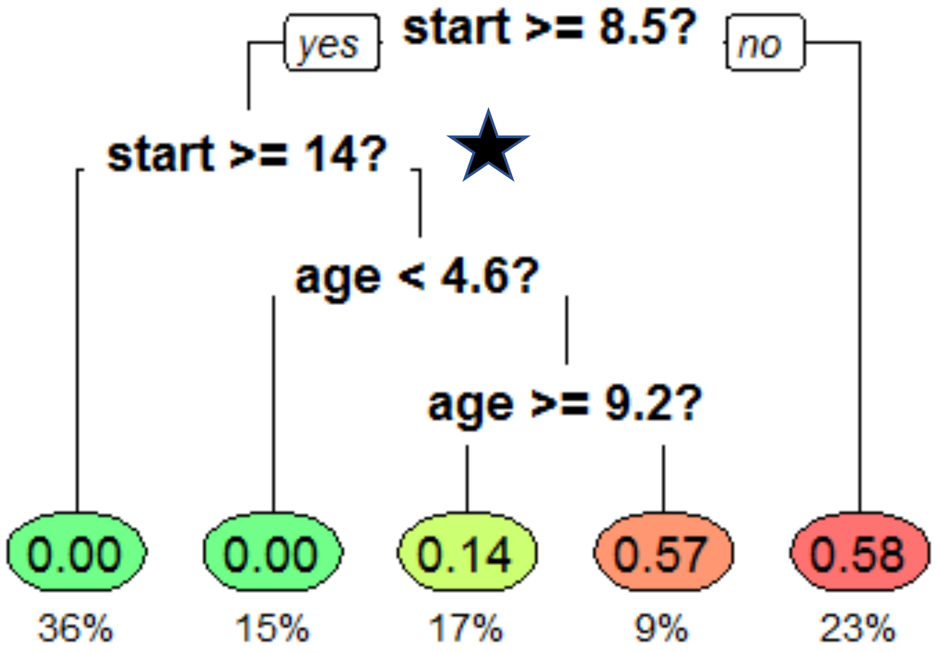




The forest is made up of decision trees

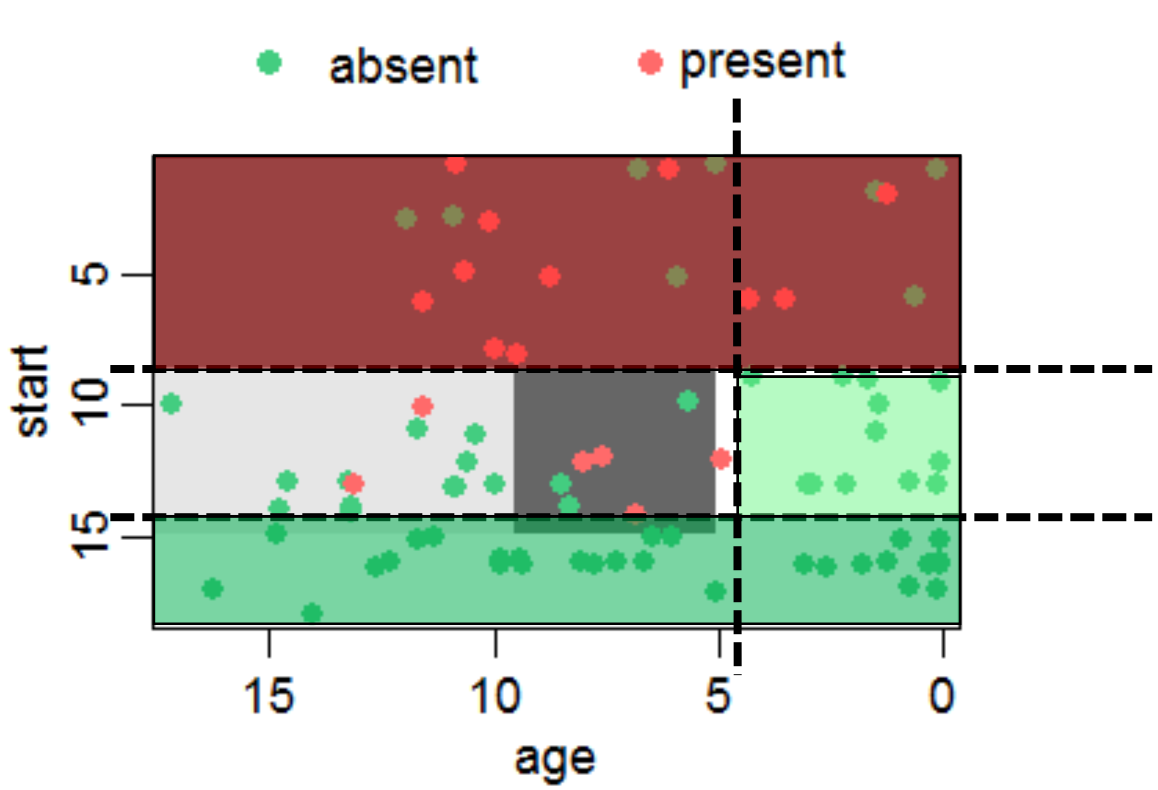
- There are two types of decision trees
  - Classification trees

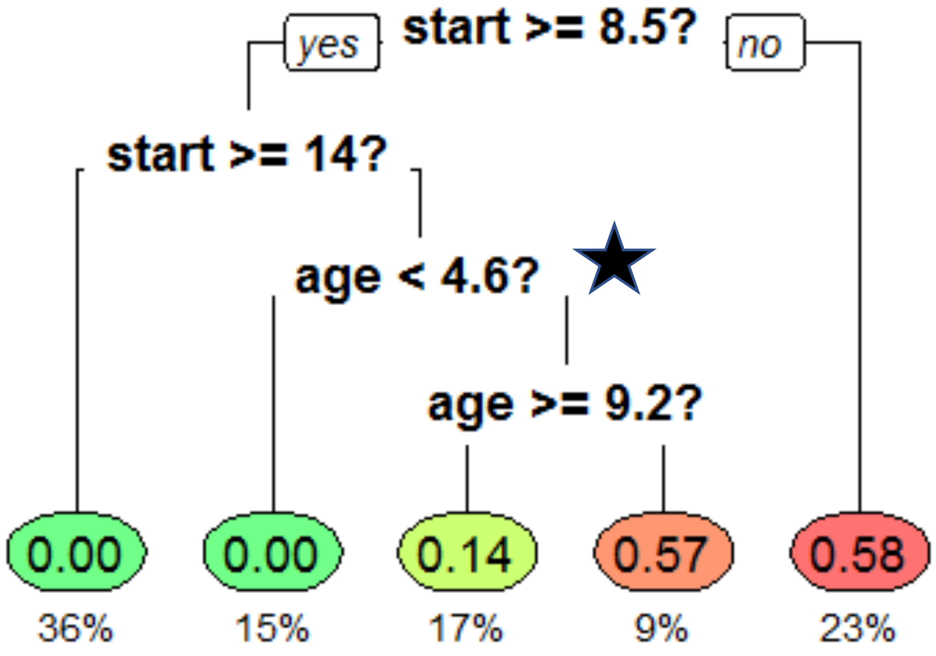




The forest is made up of decision trees

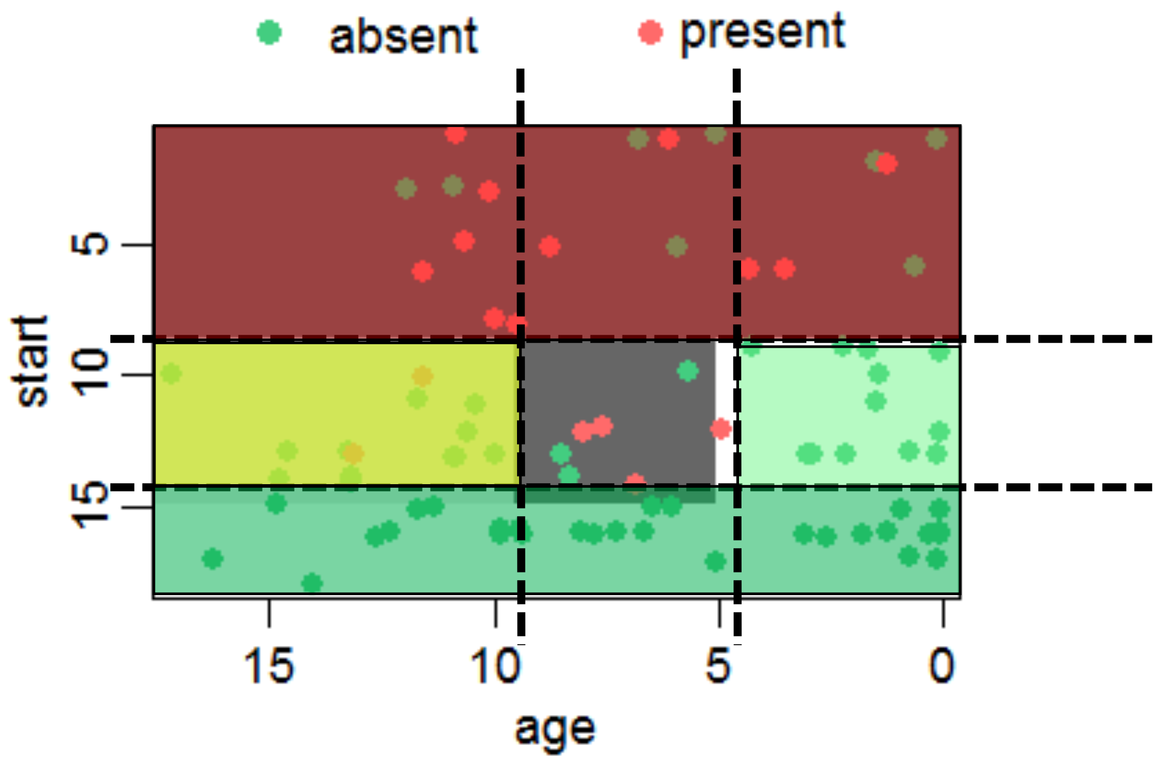
- There are two types of decision trees
  - Classification trees

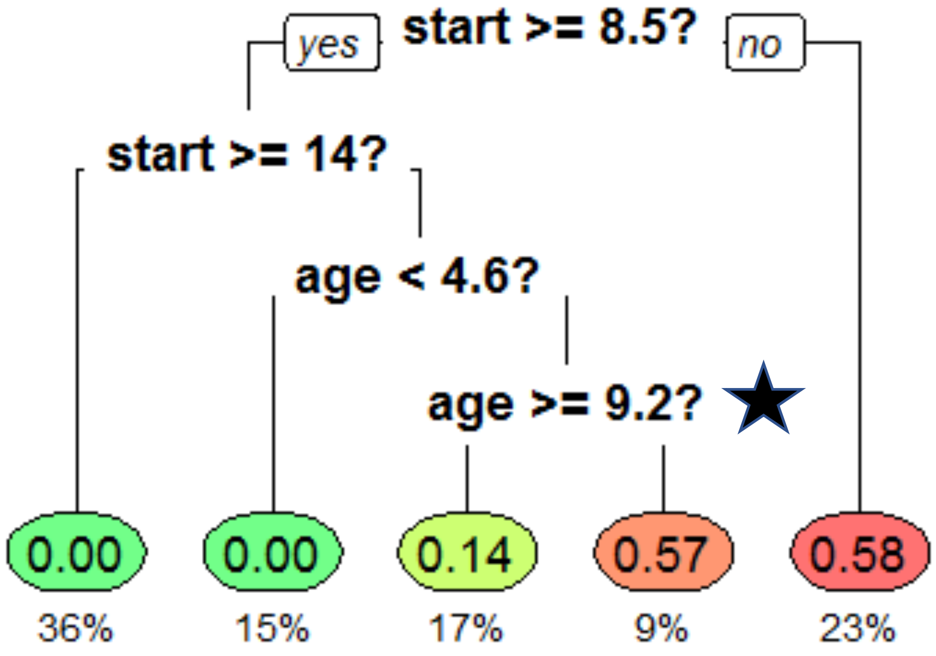




The forest is made up of decision trees

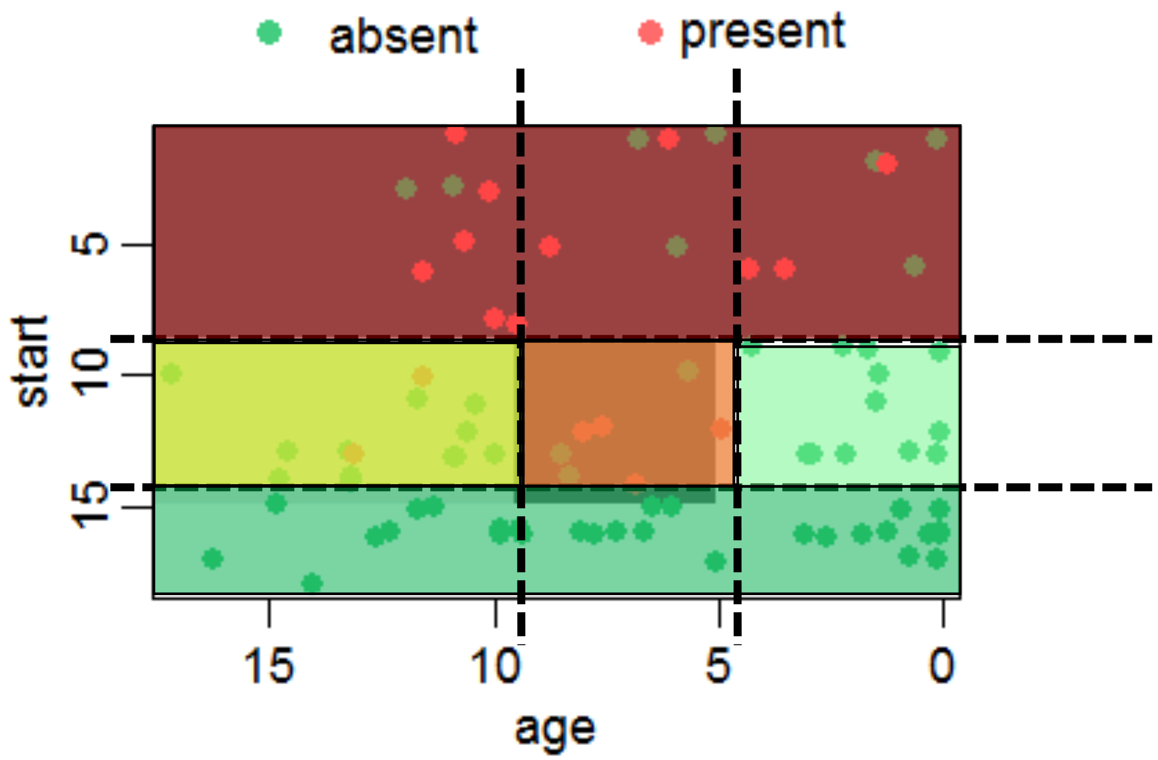
- There are two types of decision trees
  - Classification trees

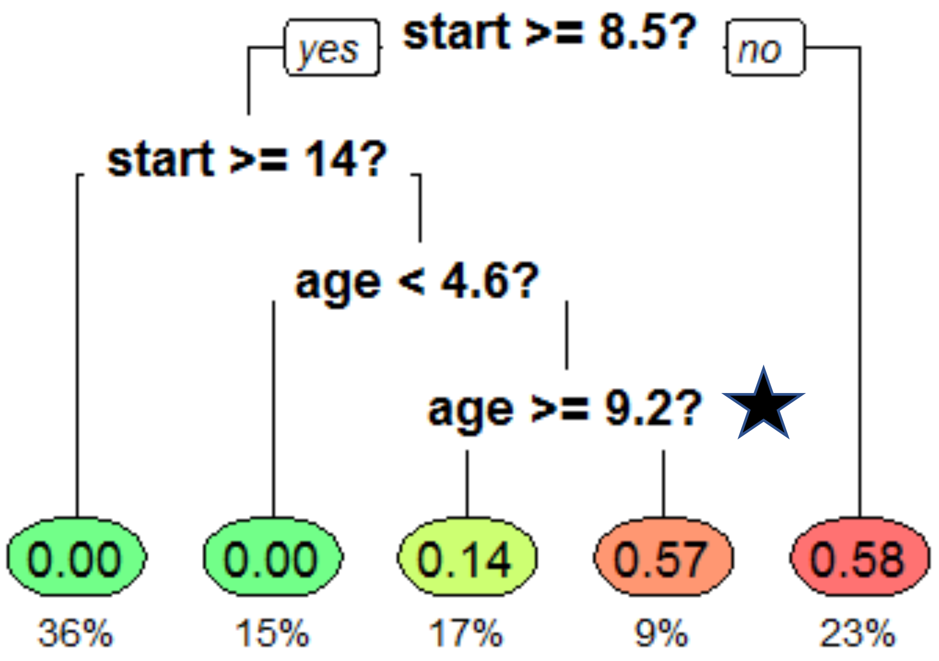




The forest is made up of decision trees

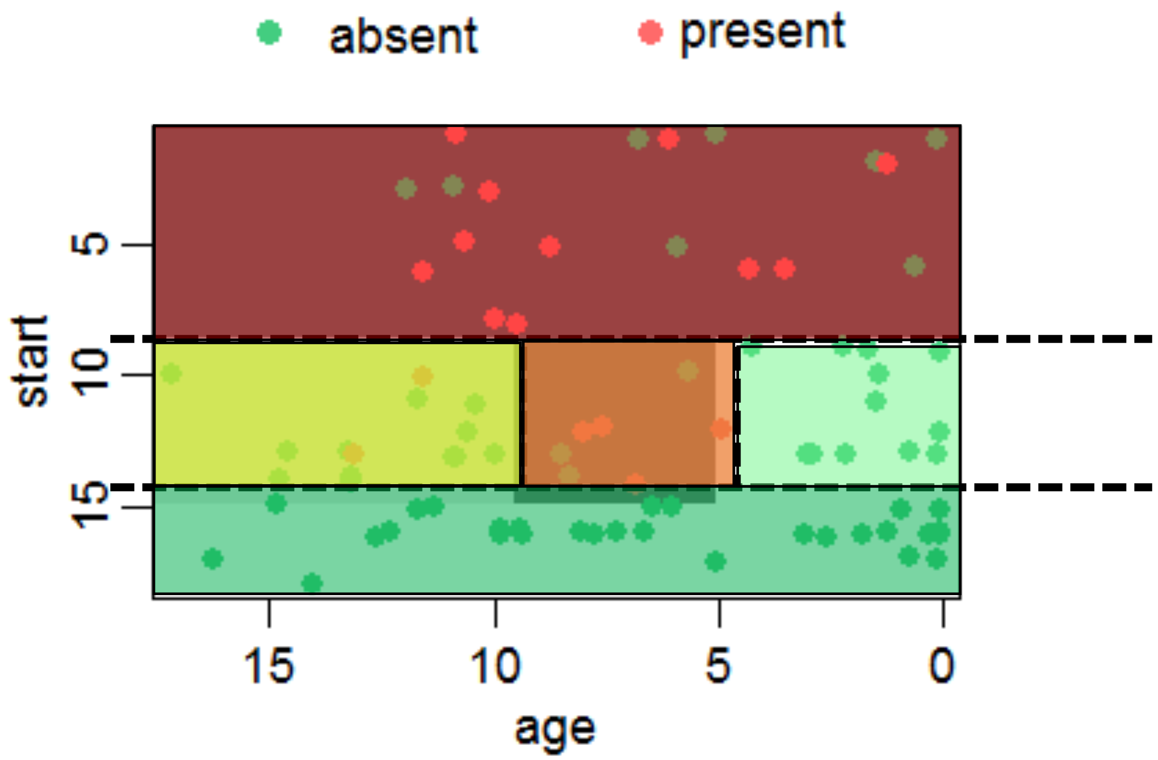
- There are two types of decision trees
  - Classification trees

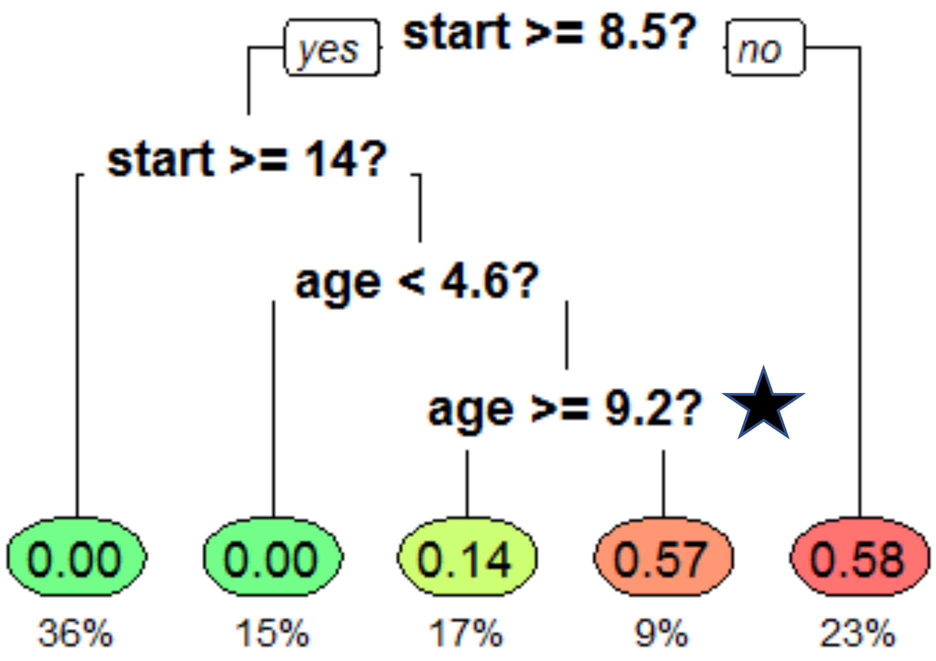




The forest is made up of decision trees

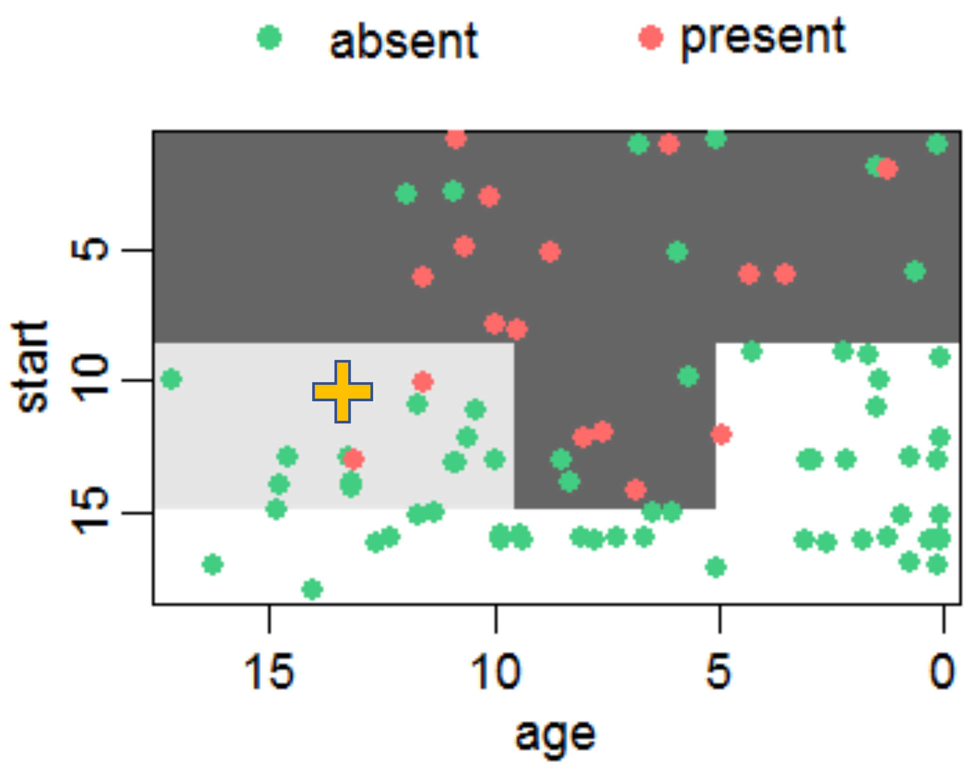
- There are two types of decision trees
  - Classification trees

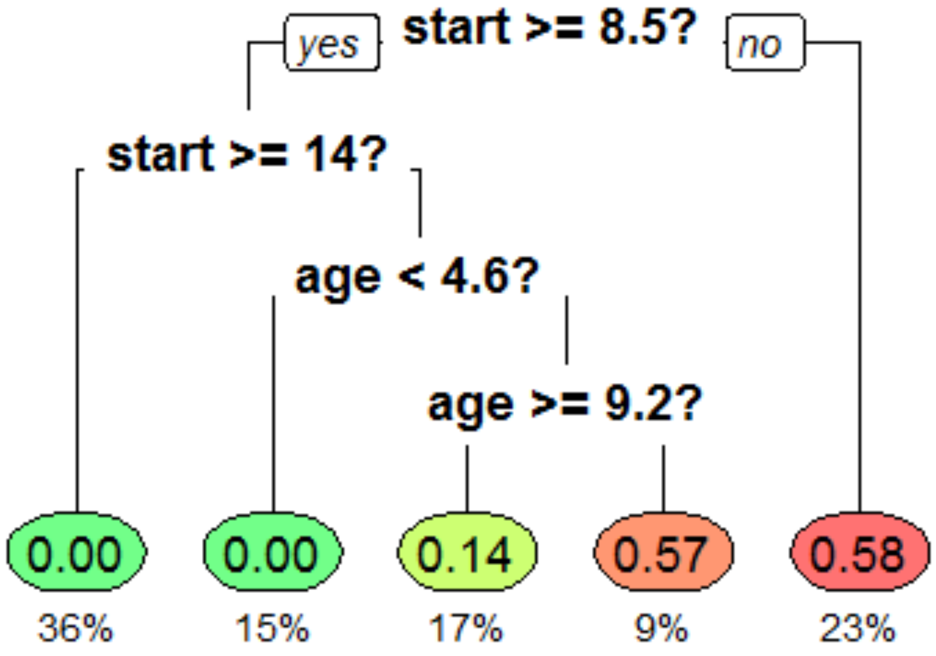




The forest is made up of decision trees

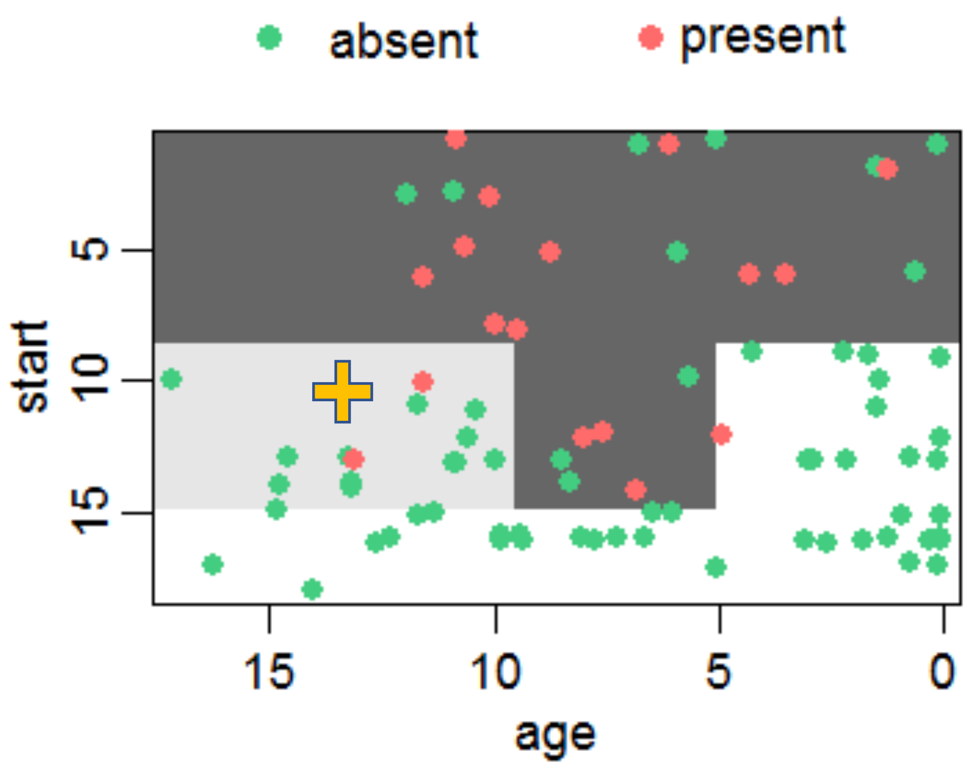
- There are two types of decision trees
  - Classification trees

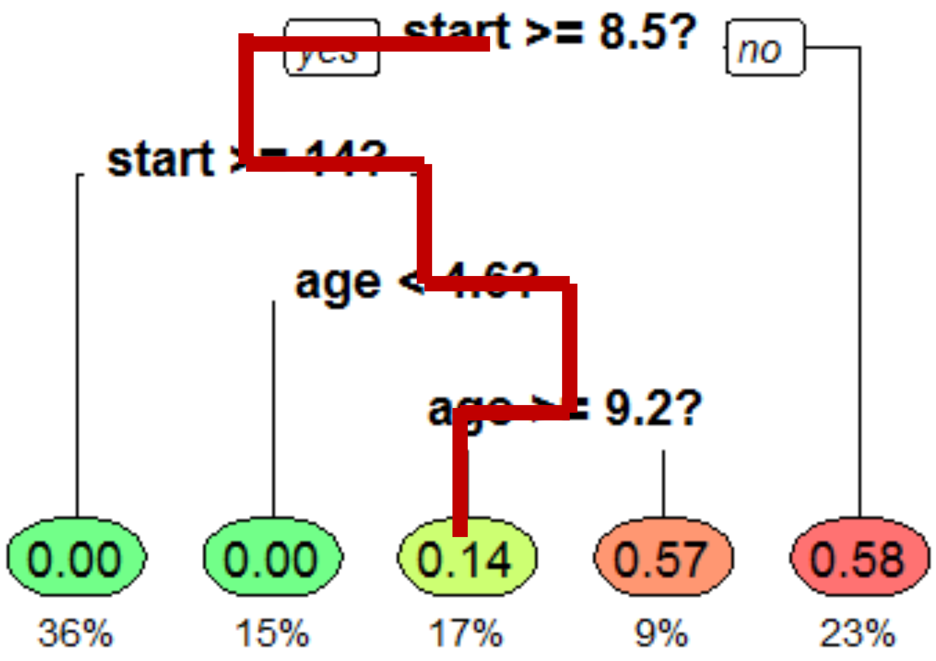




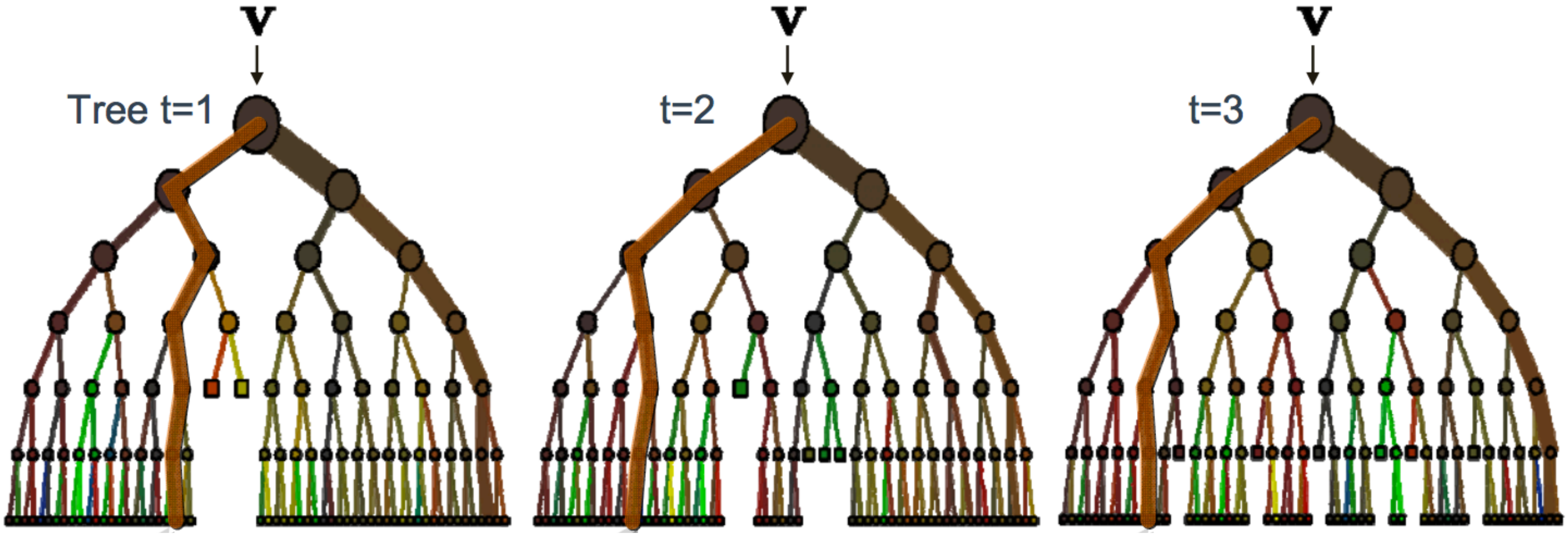
The forest is made up of decision trees

- There are two types of decision trees
    - Classification trees



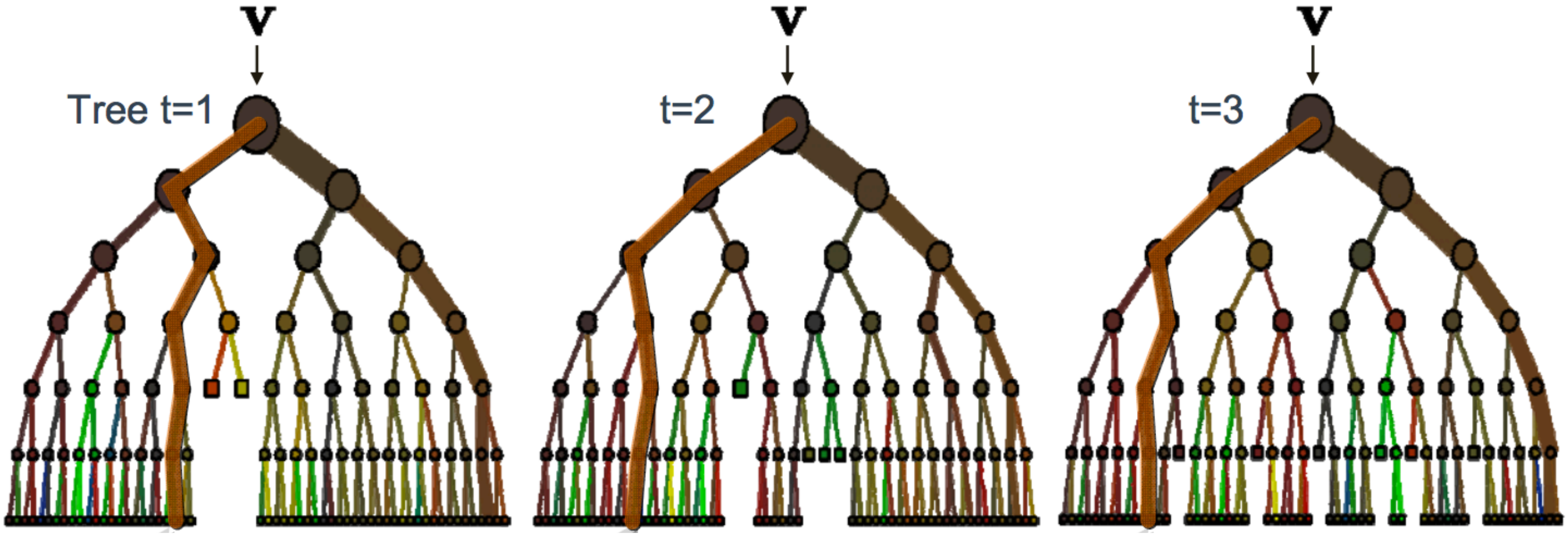


The **forest** is made up of decision trees



Where does an observation with no known response fall in all trees in the forest?

The **forest** is made up of decision trees



Final answer is consensus of all trees

What is random forest?

1. Supervised machine learning
2. The forest is made up of decision trees
3. Random
4. Ensemble approach

Breimen L. (2001) Random Forests. *Machine Learning*, 45, 5-32.

What part is Random?

1. **Random Record Selection :** Each tree is trained using roughly 2/3rd of the total training data drawn at **random with replacement** from the original data. This sample will be the training set for growing the tree.

   *****doing this repeatedly to build trees in the forest is known as Bagging (Bootstrap Aggregating)***

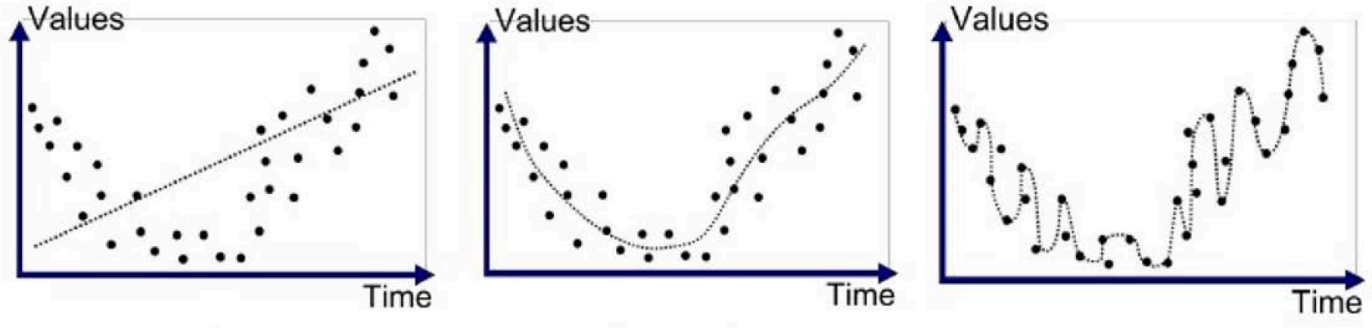
What part is Random?

1. **Random Record Selection :** Each tree is trained using roughly 2/3rd of the total training data drawn at **random with replacement** from the original data. This sample will be the training set for growing the tree.

*****doing this repeatedly to build trees in the forest is known as Bagging (Bootstrap Aggregating)***

**reduced variance amongst the trees in the forest

**avoids overfitting



Underfitted — High bias

Good Fit/Robust

Overfitted — High variance

What part is Random?

1. **Random Record Selection :** Each tree is trained using roughly 2/3rd of the total training data drawn at **random with replacement** from the original data. This sample will be the training set for growing the tree.

*****doing this repeatedly to build trees in the forest is known as Bagging (Bootstrap Aggregating)***

2. **Random Variable Selection :** Some predictor variables (say, m) are selected at **random** out of all the predictor variables and the best split on these m is used to split the node.

*****sometimes referred to as 'feature bagging'***

What part is Random?

1. **Random Record Selection :** Each tree is trained using roughly 2/3rd of the total training data drawn at **random with replacement** from the original data. This sample will be the training set for growing the tree.

doing this repeatedly to build trees in the forest is known as Bagging (Bootstrap Aggregating)

2. **Random Variable Selection :** Some predictor variables (say, *m*) are selected at **random** out of all the predictor variables and the best split on these *m* is used to split the node.

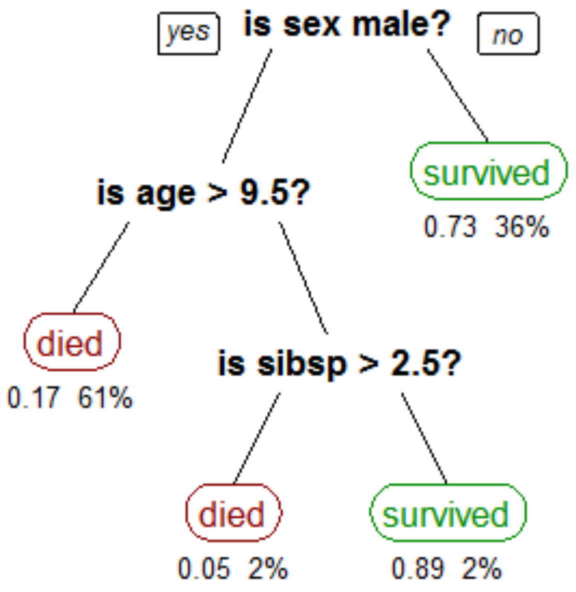
Movie	My score	Director	Production Co	1st three genres	1st three plot keywords	IMDB score	Meta-score
Star Wars: The Last Jedi	10	Rian Johnson	Lucasfilm	action, adventure, fantasy	wisecrack humor, deception, betrayal	7.5	85
Wonder Woman	9	Patty Jenkins	Warner Bros	action, adventure, fantasy	god, wonder woman, mission	7.6	76
Logan	7	James Mangold	20th Century Fox	action, drama, sci-fi	x-men, marvel, superhero	8.1	77
Zootopia	9	Byron Howard	Walt Disney	animation, adventure, comedy	fox, police, con artist	8	78
Captain America: Civil War	8	Anthony Russo	Marvel Studios	action, adventure, sci-fi	marvel, comic book, superhero	7.8	75
Beauty and the Beast	6	Bill Condon	Mandeville	family, fantasy, musical	beast, fairy tale, disney	7.3	65

What part is Random?

1. **Random Record Selection :** Each tree is trained using roughly 2/3rd of the total training data drawn at **random with replacement** from the original data. This sample will be the training set for growing the tree.

***doing this repeatedly to build trees in the forest is known as Bagging (Bootstrap Aggregating)*

2. **Random Variable Selection :** Some predictor variables (say, *m**) are selected at **random** out of all the predictor variables and the *best split*** on these *m* is used to split the node.



**typically, there is an optimal 'm' that reduces correlation amongst the trees without compromising the strength of the classifier*

***recursive binary splitting*

Recursive Binary Splitting

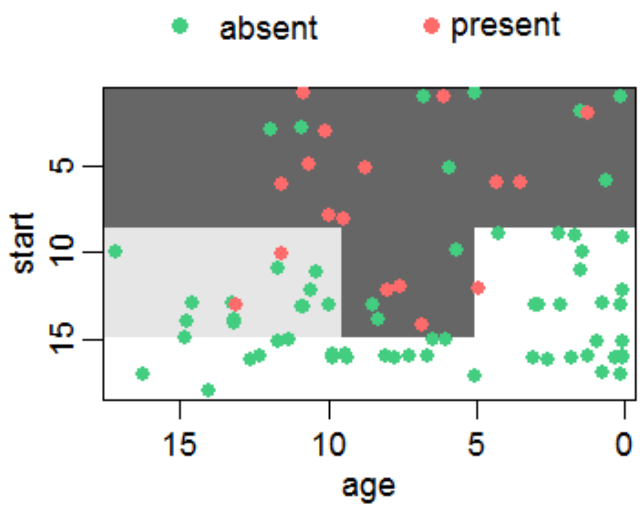
- In this procedure all the features are considered and different split points are tried and tested using a cost function. The split with the best cost (or lowest cost) is selected.

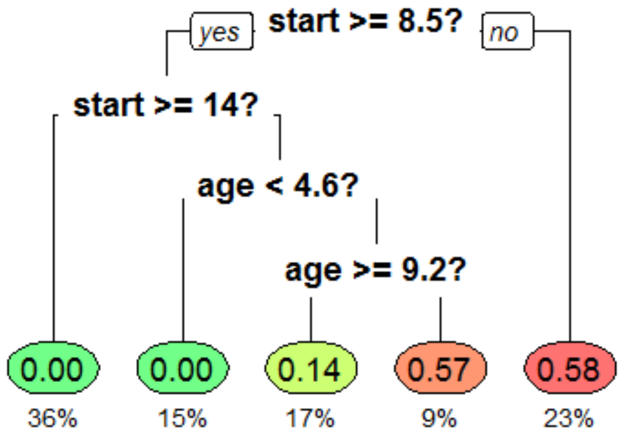
- The cost functions try to find the most homogeneous branches, or branches having groups with similar responses

Recursive Binary Splitting

- In this procedure all the features are considered and different split points are tried and tested using a cost function. The split with the best cost (or lowest cost) is selected.

- The cost functions try to find the most homogeneous branches, or branches having groups with similar responses





Recursive Binary Splitting

- In this procedure all the features are considered and different split points are tried and tested using a cost function. The split with the best cost (or lowest cost) is selected.

- The cost functions try to find the most homogeneous branches, or branches having groups with similar responses

- *When to stop splitting?*
    - **Set minimum number of training inputs to use a leaf; ignore leaves with less or stop**
    - **Set maximum depth: refers to the the length of the longest path from a root to a leaf**

What is random forest?

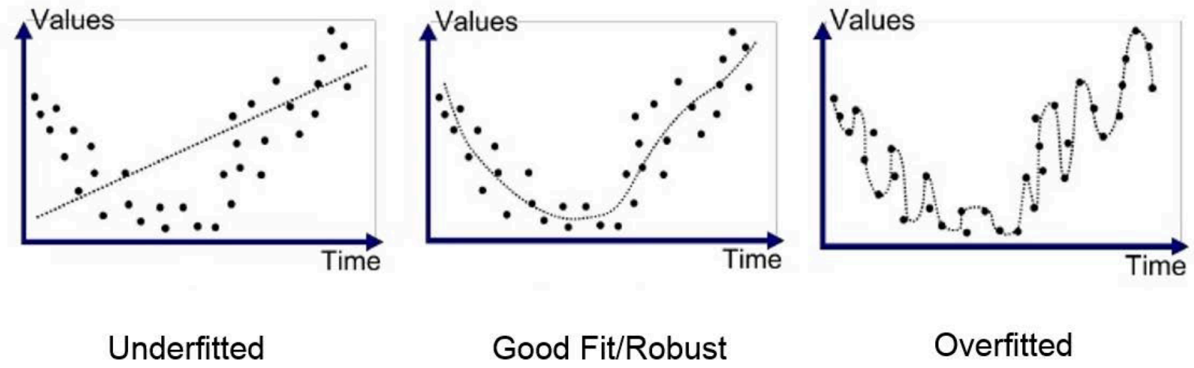
1. Supervised machine learning
2. The forest is made up of decision trees
3. Random
4. Ensemble approach

Breimen L. (2001) Random Forests. *Machine Learning*, 45, 5-32.

Ensemble approach

- The ensemble refers to averaging the predictions across all of the trees. A decision tree alone is a weak predictor, **but together the forest is strong!**

- The trees *must* be constructed using bagging (bootstrap aggregating) and random variable selection in order for the forest to be successful. Otherwise, the trees would be to correlated and have poor predictive power.



Ensemble approach

- The ensemble refers to averaging the predictions across all of the trees. A decision tree alone is a weak predictor, **but together the forest is strong!**

  - **Discrete** response variables: the predictions are "votes" for classes. After all trees in a forest make a prediction, these "votes" are tallied and counted. The proportion of votes for each category is the predicted probability.

  - **Continuous** response variables: the predictions are the average value of the predicted variable.

Ensemble approach

Advantages:

- Handle numerical and categorical predictor and response variables
- Implicitly perform feature importance
- Nonlinear relationships between parameters does not affect tree performance
- Robust to correlated or noisy predictor variables (unlike ABC)**

Disadvantages:

- Create overfit trees that do not generalize well (high variance)
- Create too general of trees with no predictive power (high bias)
- If classes dominate the training data, this can also bias the forest.**

What is random forest?

1. Supervised machine learning
2. The forest is made up of decision trees
3. Random
4. Ensemble approach

Breimen L. (2001) Random Forests. *Machine Learning*, 45, 5-32.

Random Forest procedure

1. **Random Record Selection :** select 2/3 of samples for *n* trees

2. **Random Variable Selection :** select *m* predictor variables for *n* trees

3. **Construct *n* trees** using set parameters of forest and *recursive binary splitting*

4. Using the leftover (1/3) data for each n tree, calculate the misclassification rate - **out of bag (OOB)** error rate for each model

5. Analyze feature importance

Random Forest procedure

1. **Random Record Selection :** select 2/3 of samples for *n* trees

2. **Random Variable Selection :** select *m* predictor variables for *n* trees

3. **Construct *n* trees** using set parameters of forest and *recursive binary splitting*

4. Using the leftover (1/3) data for each n tree, calculate the misclassification rate - **out of bag (OOB)** error rate for each model

5. Analyze feature importance

Random Forest procedure

1. **Random Record Selection :** select 2/3 of samples for *n* trees

2. **Random Variable Selection :** select *m* predictor variables for *n* trees

3. **Construct *n* trees** using set parameters of forest and *recursive binary splitting*

4. Using the leftover (1/3) data for each n tree, calculate the misclassification rate - **out of bag (OOB)** error rate for each model

5. Analyze feature importance

Random Forest procedure

1. **Random Record Selection :** select 2/3 of samples for *n* trees

2. **Random Variable Selection :** select *m* predictor variables for *n* trees

3. **Construct *n* trees** using set parameters of forest and *recursive binary splitting*

4. Using the leftover (1/3) data for each n tree, calculate
the misclassification rate - **out of bag (OOB)** error rate for each model

5. Analyze feature importance

OOB Error Rates

- Using the leftover 1/3 of data (**Out-of-Bag data**) that was not used to build a particular decision tree, validate the decision trees.

- If we grow 1000 trees in our forest, then a record will be OOB for roughly (.37*1000) 370 trees.

- Using these 370 trees the data was not used in, estimate the correct response variable for the data.

  ****For a discrete dependent variable, the vote will be tallied and counted. This is the RF score and the proportion of votes for each category is the predicted probability.**

  ****In a continuous case, it is average value of the predicted variable.**

- Aggregate error from all trees to determine **overall OOB error rate** for the classification.

Random Forest procedure

1. **Random Record Selection :** select 2/3 of samples for *n* trees

2. **Random Variable Selection :** select *m* predictor variables for *n* trees

3. **Construct *n* trees** using set parameters of forest and *recursive binary splitting*

4. Using the leftover (1/3) data for each n tree, calculate the misclassification rate - **out of bag (OOB)** error rate for each model

5. Analyze feature importance

Feature Importance

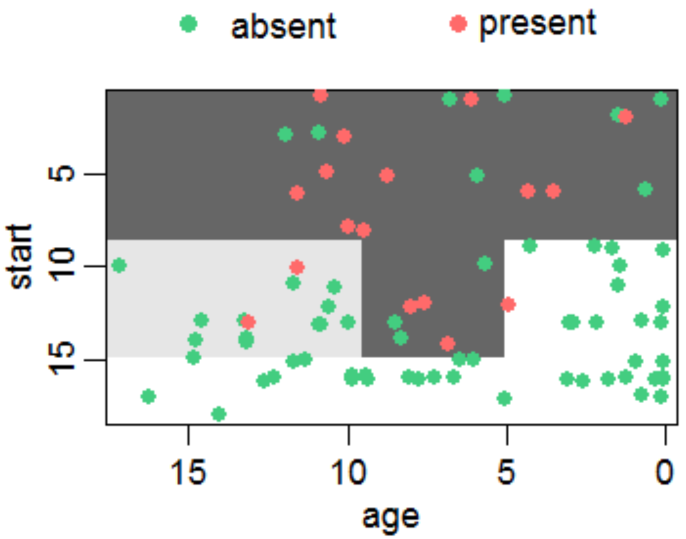
- Can sometimes provide the "*why*" in "*why is this working so well?*"

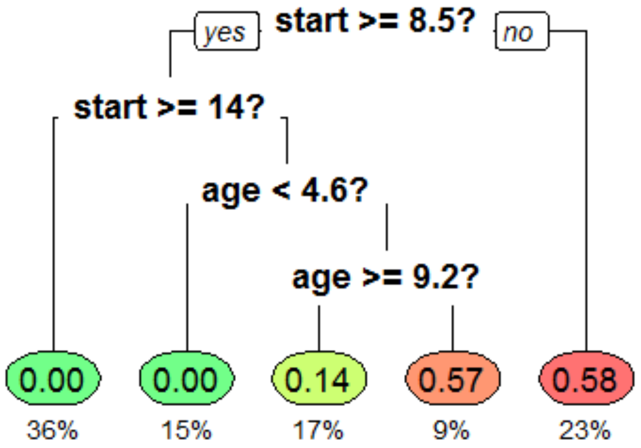
- How well are the feature (predictor) variables splitting the data at each node?

Gini impurity: GINI

- Measures feature importance based on how variables contribute to ***node purity.***

- In other words, if, when used, a feature results in splits that generally split between, not within, classes, then that variable increases node purity.





Other Feature Importance Measures

- **Mean Decrease Accuracy (Permutation Feature Importance) -** How much the model accuracy decreases if we drop that variable.
  - We don't quite "drop" it, but rather, permute the data to become random.
  - Re-estimate the forest, with this variable as "random"
  - Compare the change in error rates between "real" and "random" data
  - *****ONE FEATURE AT A TIME*****

- **Mean Decrease Gini**