Random Forest

Megan Ruffley, Isaac Overcast

CompPhylo Oslo 2019

August 27, 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
- Two main types of supervised learning….

Unsupervised learning, on the other hand, is untrained and infers the natural structure present within a set of data points.

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***
  - ***Regression***

Unsupervised learning, on the other hand, is untrained and infers the natural structure present within a set of data points.

Supervised machine learning

- Two main types of supervised learning….
    - ***Classification***
    - ***Regression***

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

Unsupervised learning, on the other hand, is untrained and infers the natural structure present within a set of data points.

Supervised machine learning

- Two main types of supervised learning….
  - ***Classification***
  - ***Regression***

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

Unsupervised learning, on the other hand, is untrained and infers the natural structure present within a set of data points.

- Mainly for clustering and dimensionality reduction.

Supervised machine learning

	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

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

- There are two types of decision trees
  - Classification trees
  - Regression trees

- CART (classification and regression trees)

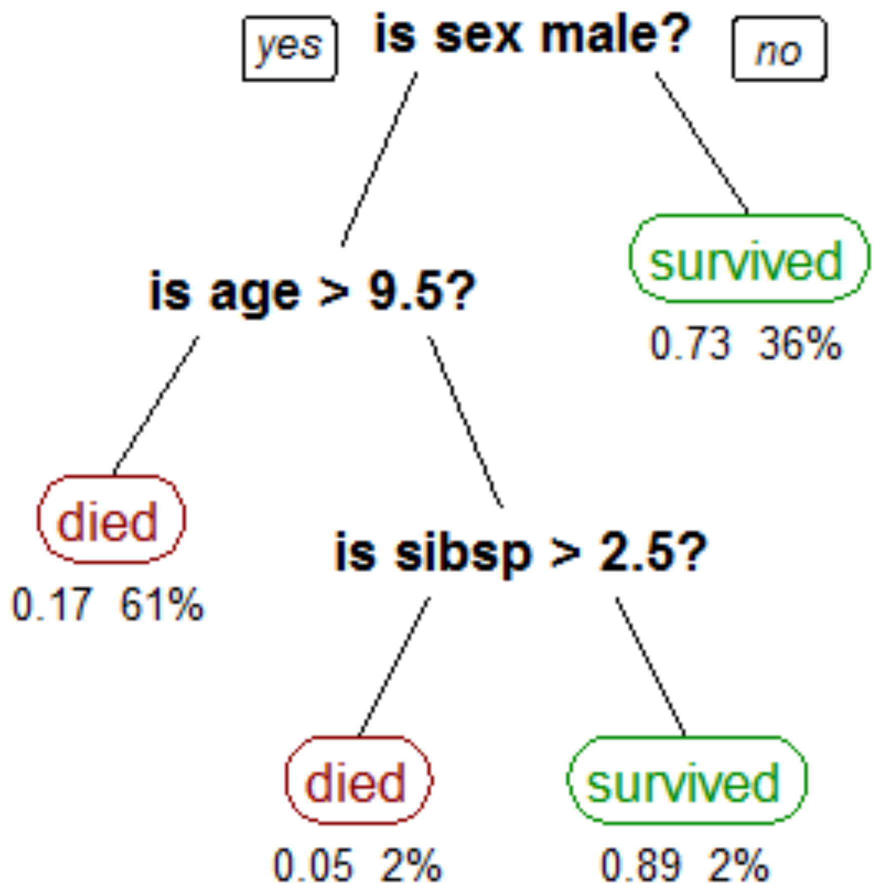
The forest is made up of decision trees

Common examples of decision trees

The forest is made up of decision trees

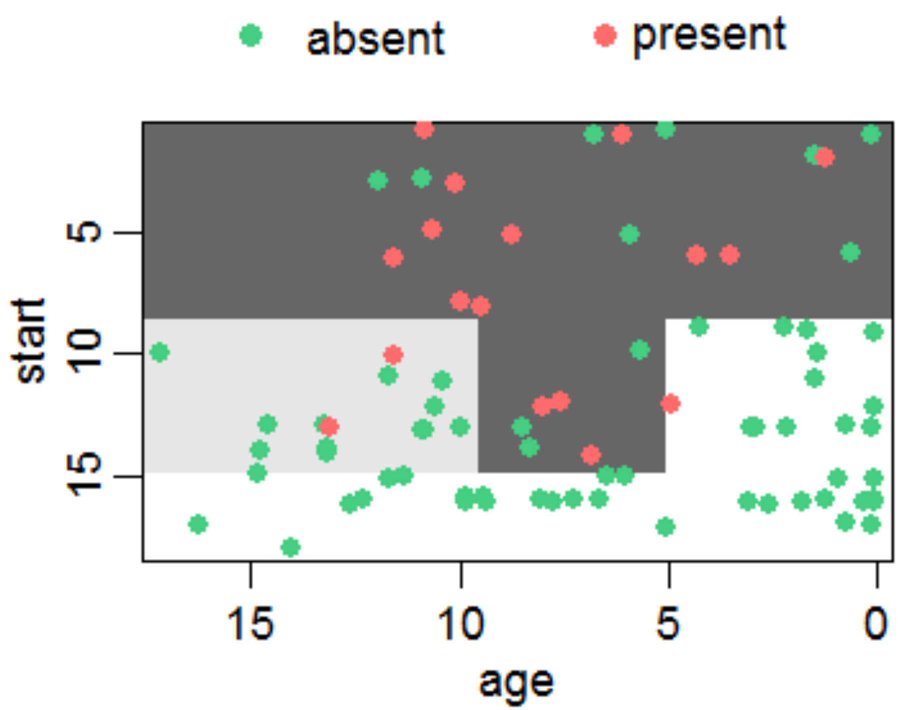
- There are two types of decision trees
  - Classification trees
  - Regression trees

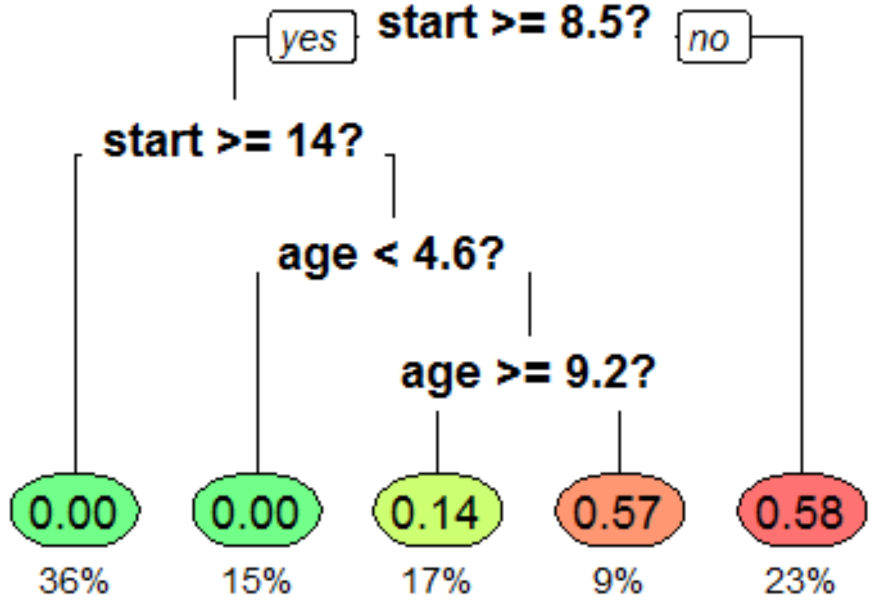
- CART (classification and regression trees)



The forest is made up of decision trees

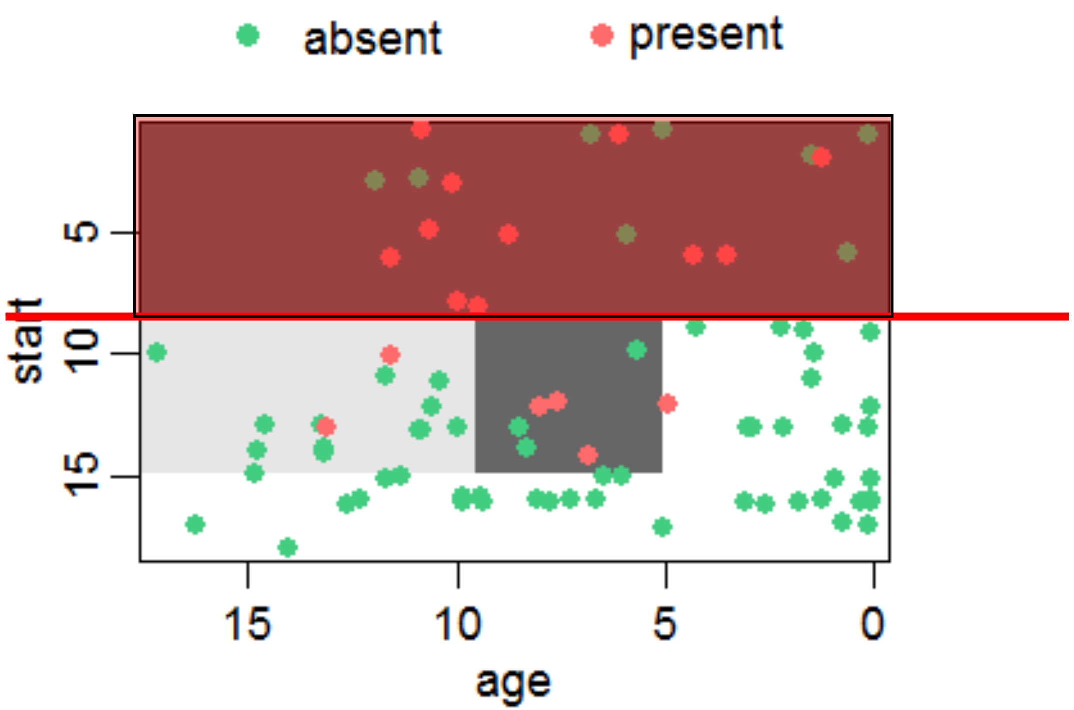
- There are two types of decision trees
  - Classification trees

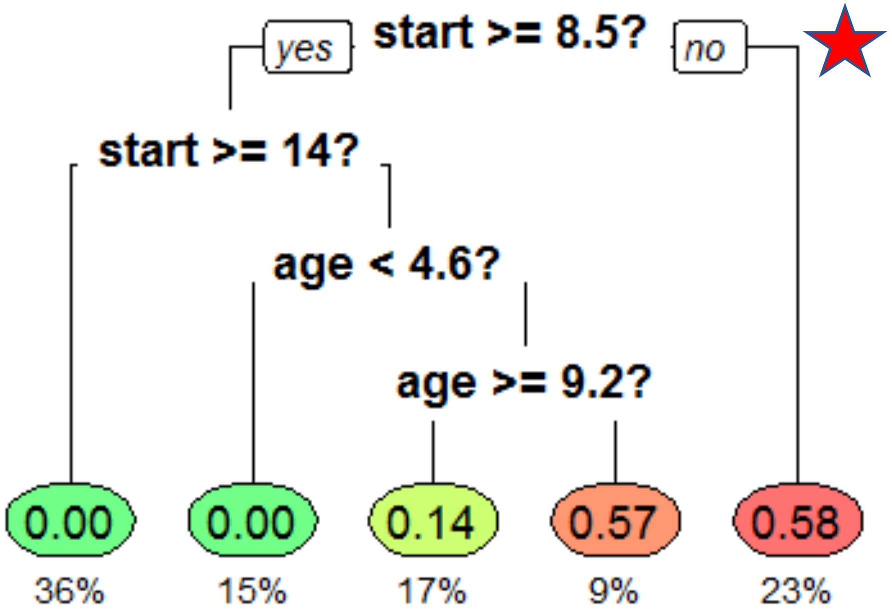




The forest is made up of decision trees

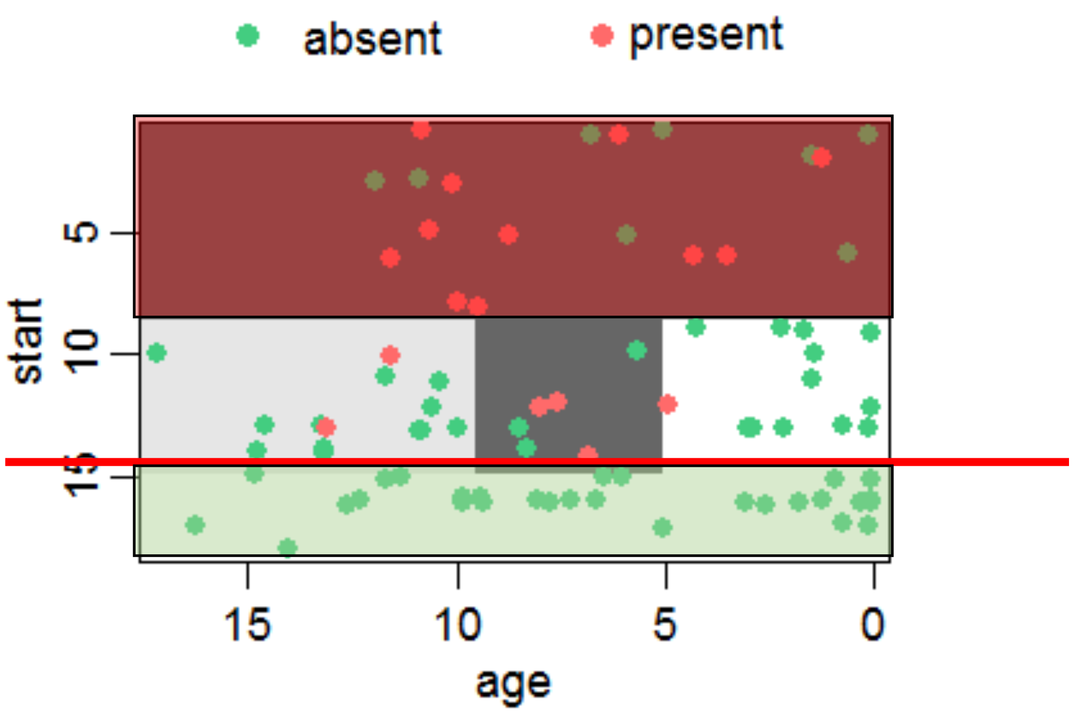
- There are two types of decision trees
  - Classification trees

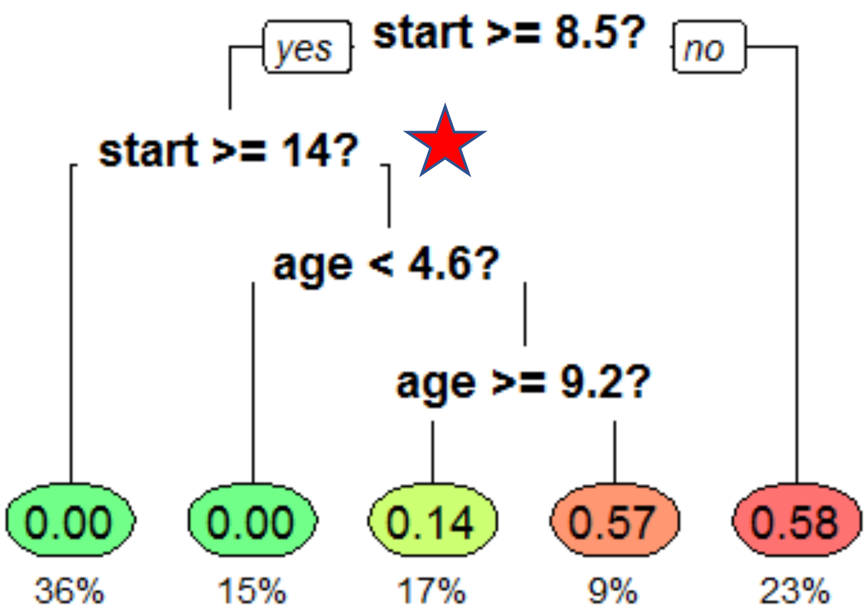




The forest is made up of decision trees

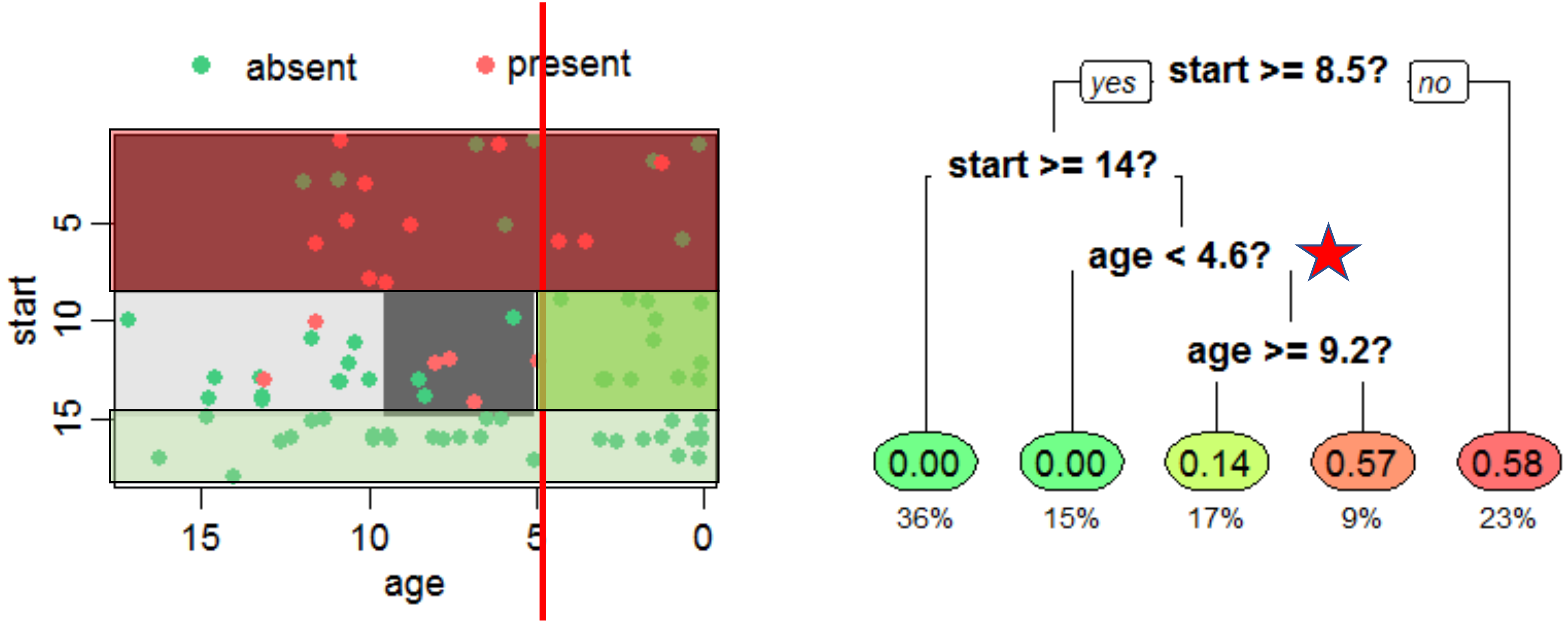
- There are two types of decision trees
  - Classification trees





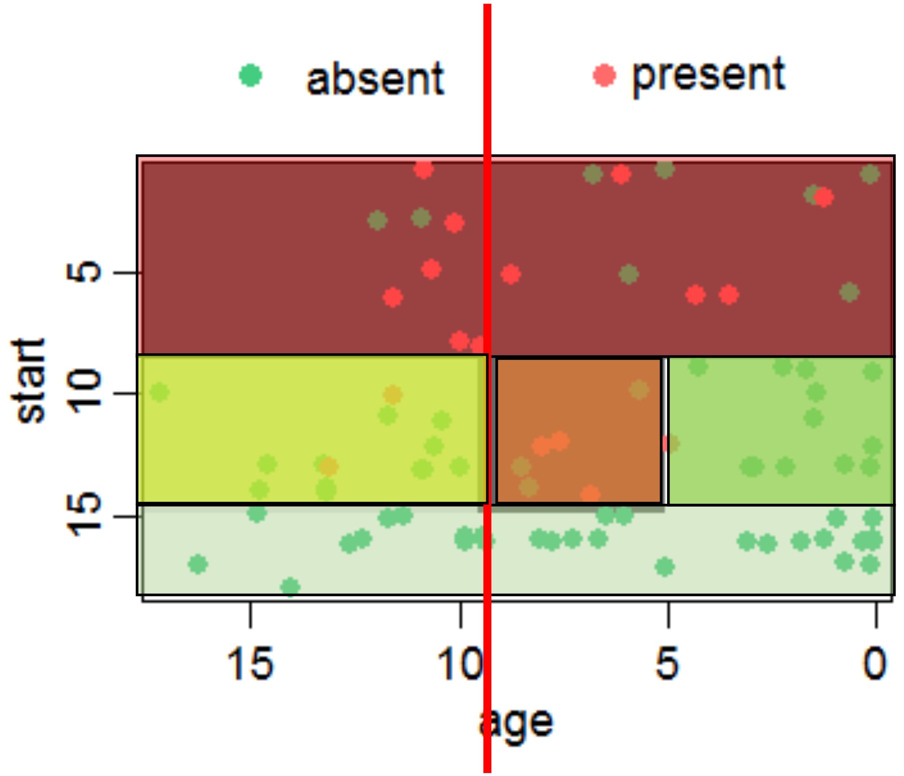
The forest is made up of decision trees

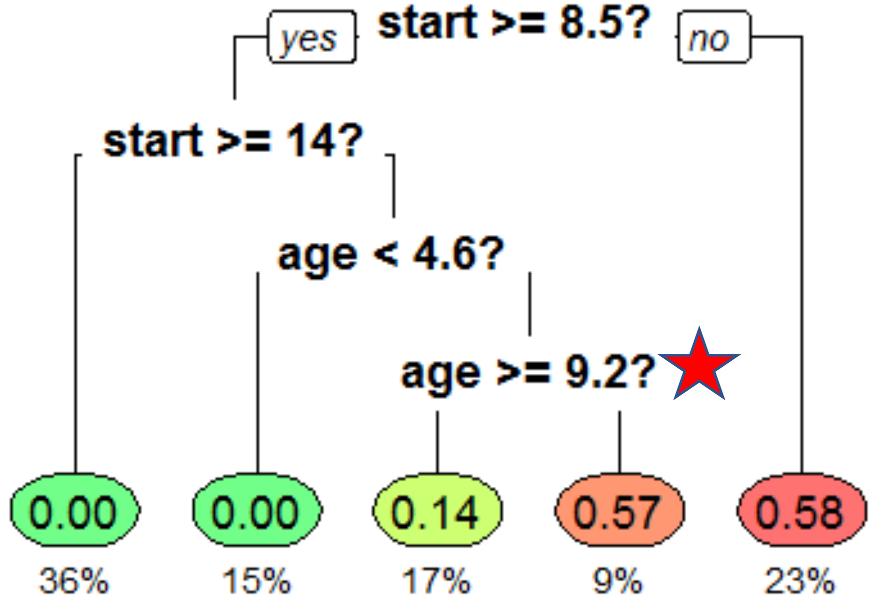
- There are two types of decision trees
  - Classification trees



The forest is made up of decision trees

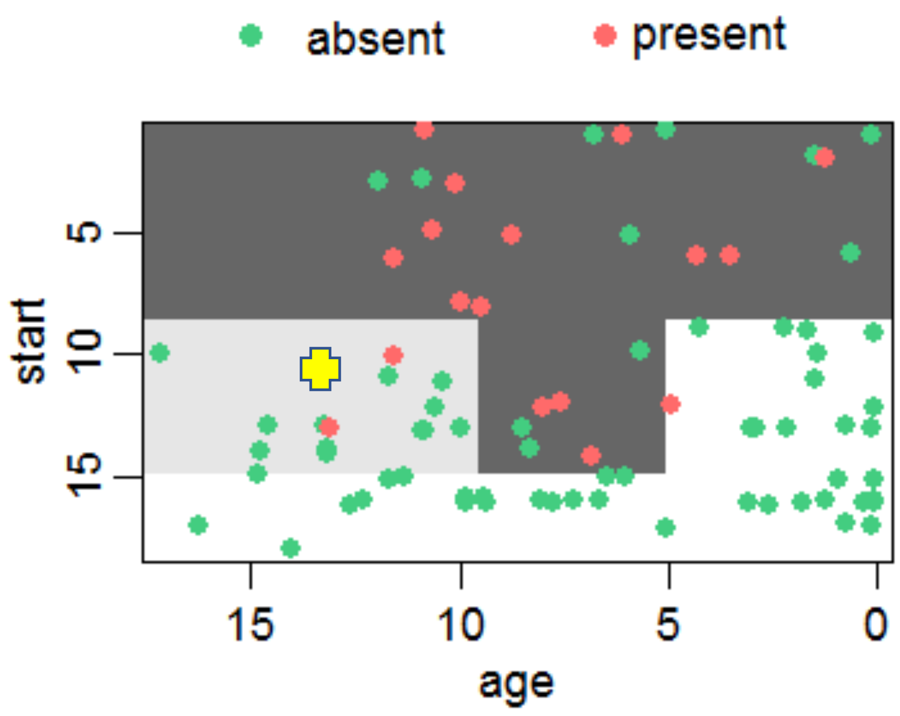
- There are two types of decision trees
  - Classification trees

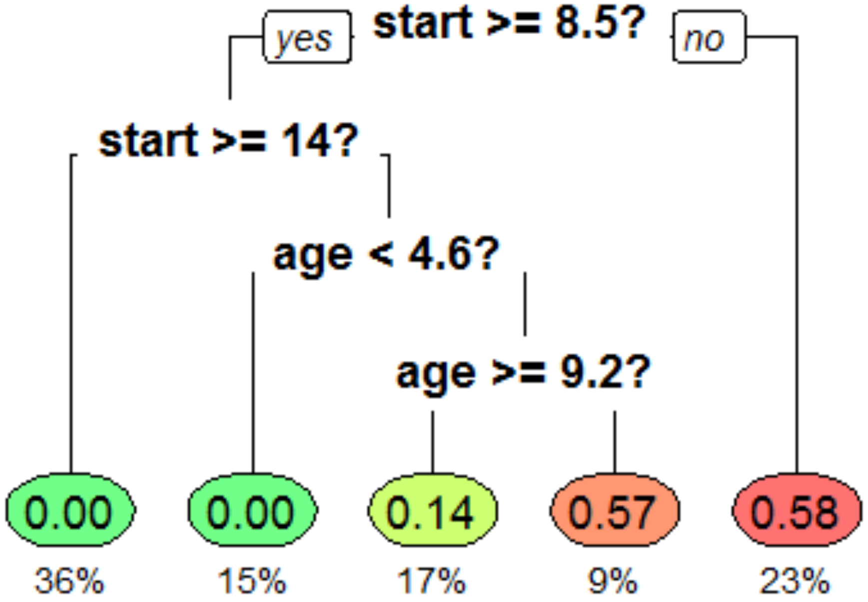




The forest is made up of decision trees

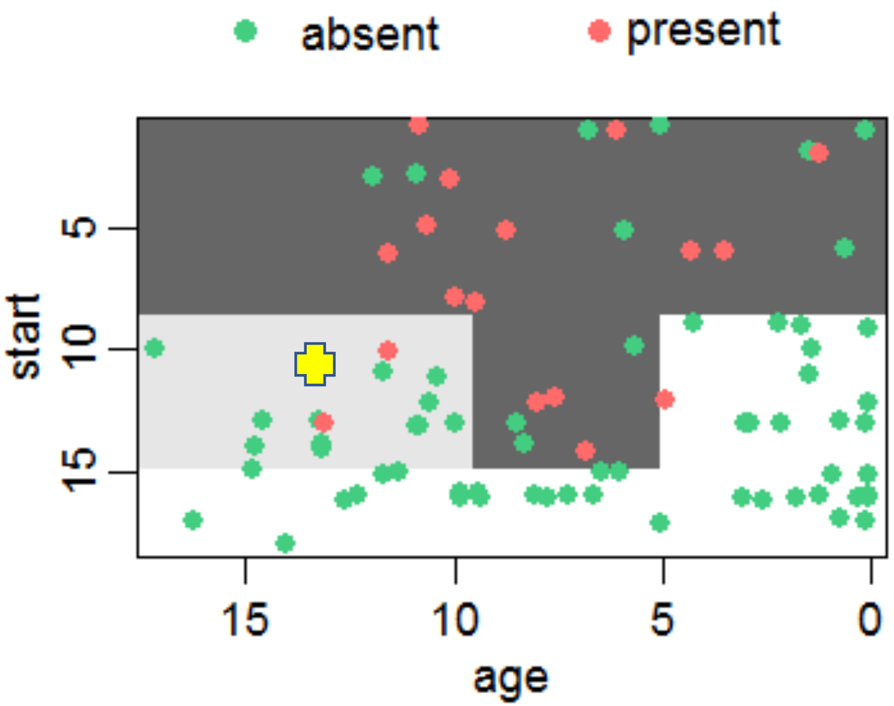
- There are two types of decision trees
  - Classification trees

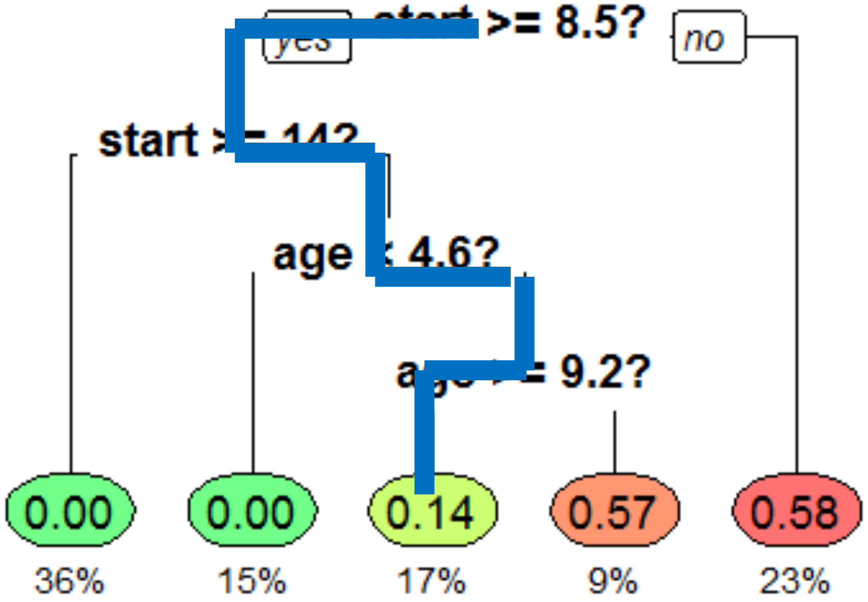




The forest is made up of decision trees

- There are two types of decision trees
  - Classification trees





The forest is made up of decision trees

- There are two types of decision trees
  - Regression trees
  - These are a little bit more complicated. We will get into them later.

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)***

Bagging = **B**ootstrap **Agg**regat**ing**

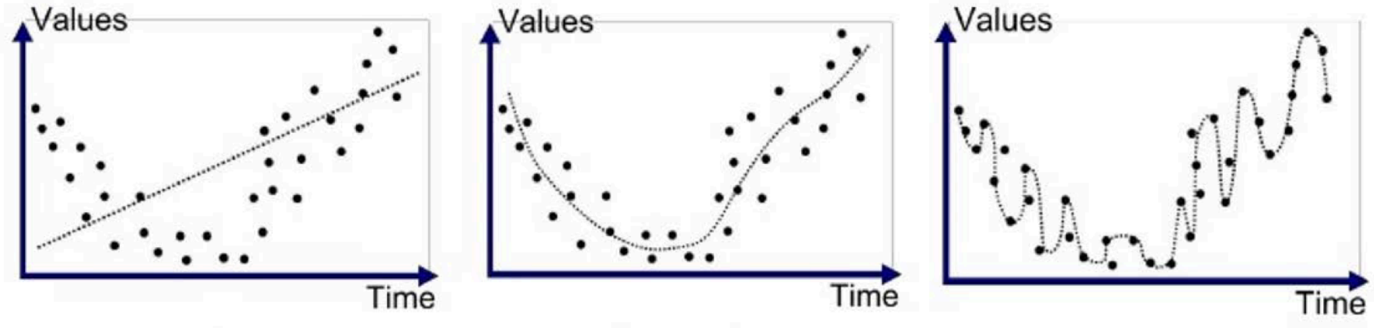
- Generates *m* new training data sets by repeatedly sampling ~2/3 of the data, with replacement.
- Builds *m* decision tress using *m* training data sets.
- *m* Models are combined by averaging (regression) or voting (classification)

Bagging = Bootstrap **Aggregating**

- Generates *m* new training data sets by repeatedly sampling ~2/3 of the data, with replacement.
- Builds *m* decision tress using *m* training data sets.
- *m* Models are combined by averaging (regression) or voting (classification)

**reduced variance amongst the trees in the forest

**avoids overfitting



Underfitted | Good Fit/Robust | Overfitted

High bias | | 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.

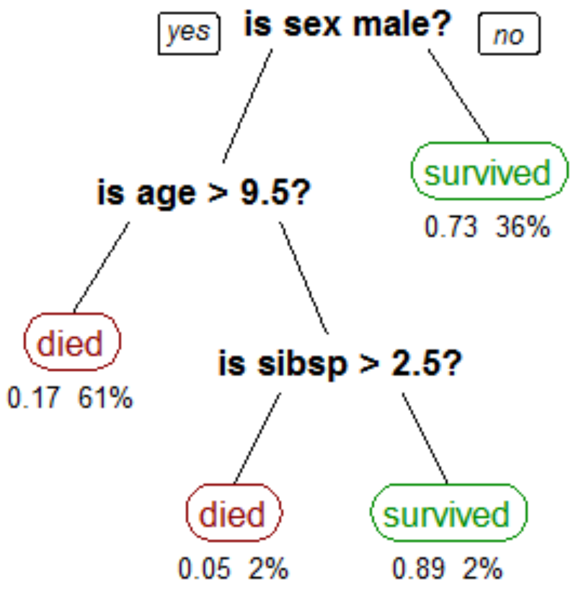
*****sometimes referred to as 'feature bagging'***

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.



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

Ensemble approach

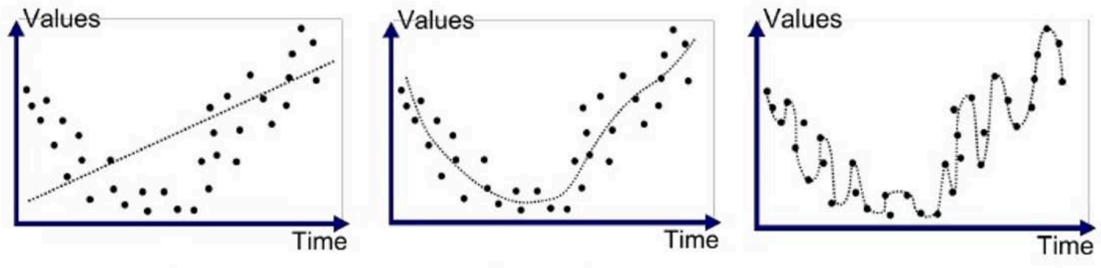
- The ensemble refers to averaging the predictions across all of the trees. A decision tree alone is a weak predictor, but together the forest is strong!
    - **For a discrete dependent variable, the predictions are "votes" for models. After all trees in a forest make a prediction, these "votes" are tallied and counted. The proportion of votes for each category is the predicted probability.**
    - **In a continuous case, it is average value of the predicted variable.**

- 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!
  - **For a discrete dependent variable, the predictions are "votes" for models. 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.**
  - **In a continuous case, it is average value of the predicted variable.**

- 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.



Underfitted Good Fit/Robust Overfitted

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 :** Each tree is trained on roughly 2/3rd of the total training data that is drawn at random with replacement from the original data.

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

Random Forest procedure

1. **Random Record Selection :** Each tree is trained on roughly 2/3rd of the total training data that is drawn at random with replacement from the original data.

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

3. Construct all decision trees in the forest.

Random Forest procedure

1. **Random Record Selection :** Each tree is trained on roughly 2/3rd of the total training data that is drawn at random with replacement from the original data.

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

3. Construct all decision trees in the forest.

4. For each tree, using the leftover (1/3) data, calculate the misclassification rate - **out of bag (OOB)** error rate, for each model and then the overall OOB error rate.

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.

- Each of these trees gives a classification on leftover data (OOB), and we say the tree "votes" for that class. The forest chooses the classification having the most votes over all the trees in the forest.
  - ****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 :** Each tree is trained on roughly 2/3rd of the total training data that is drawn at random with replacement from the original data.

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

3. Construct all decision trees in the forest.

4. For each tree, using the leftover (1/3) data, calculate the misclassification rate - **out of bag (OOB)** error rate, for each model and then the overall OOB error rate.

5. Analyze feature importance

Feature Importance

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

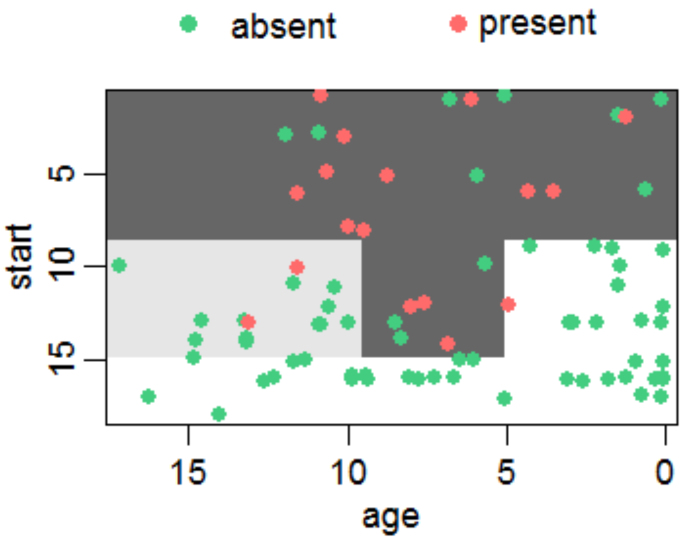
- How well are the feature (predictor) variables splitting the data at each node?

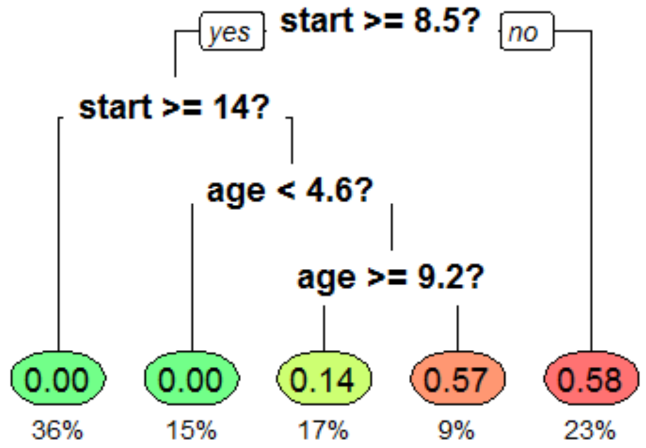
- Gini impurity/information gain (entropy)

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.





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.

- The more “impure” of a predictor, or higher the GINI, the less important the feature is for RF.

Feature Importance in R

- **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**