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

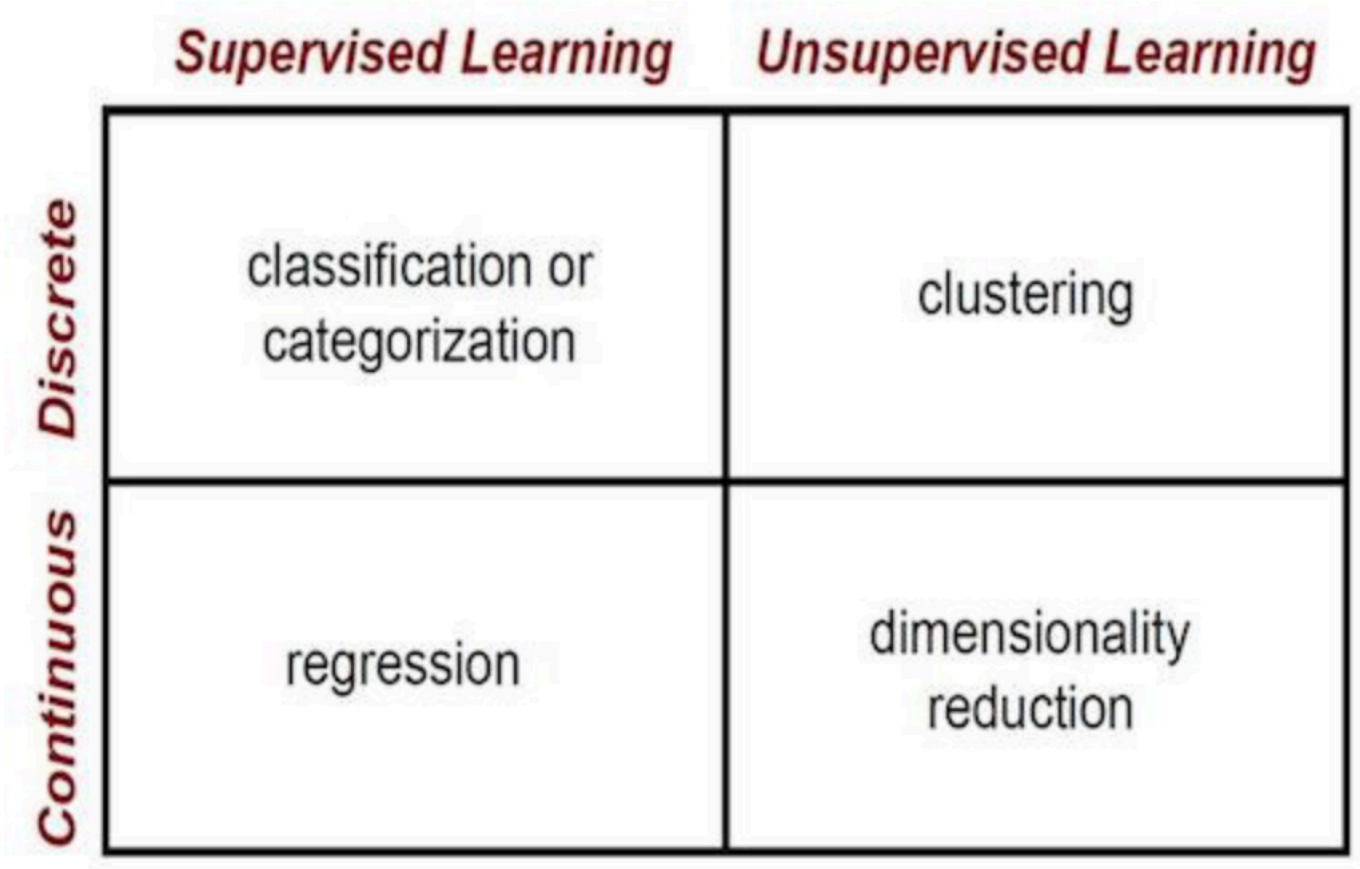
- Two main types of supervised learning….
  - ***Classification***
  - ***Regression***

- 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



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)

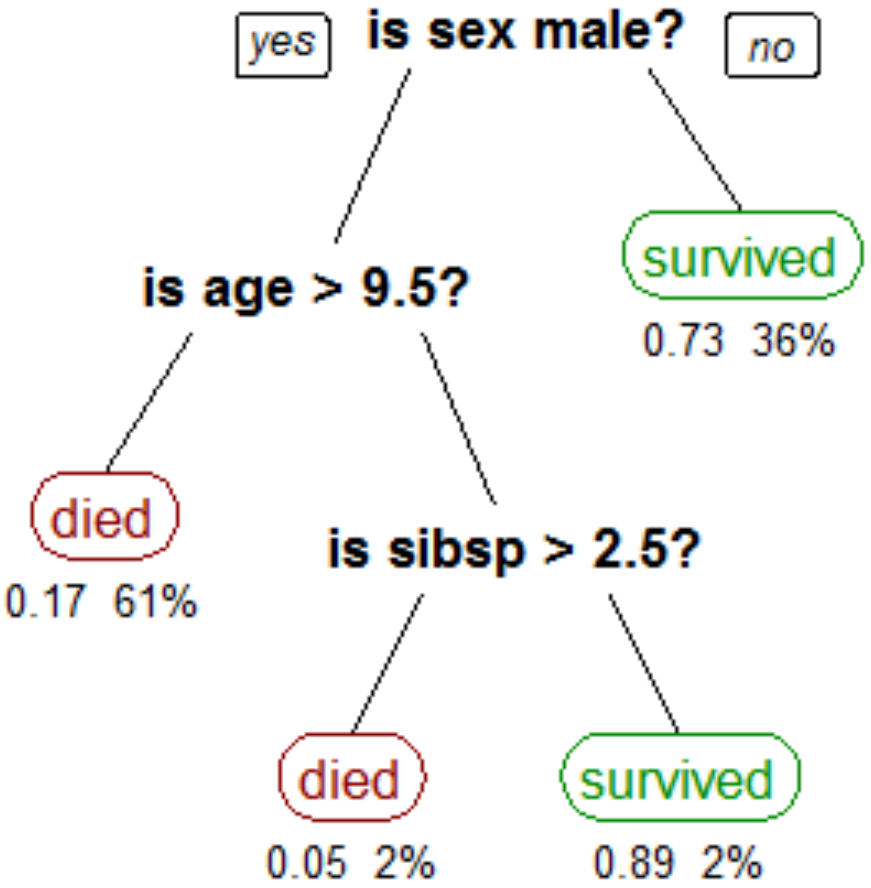
The forest is made up of decision trees

Common examples of decision trees

The forest is made up of decision trees

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

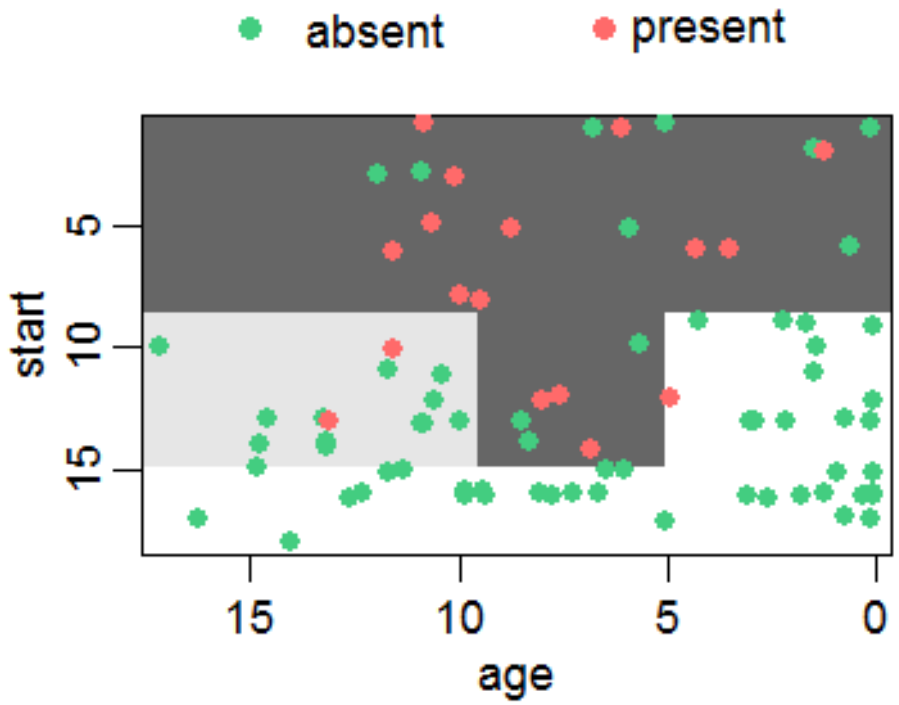
- CART (classification and regression trees)

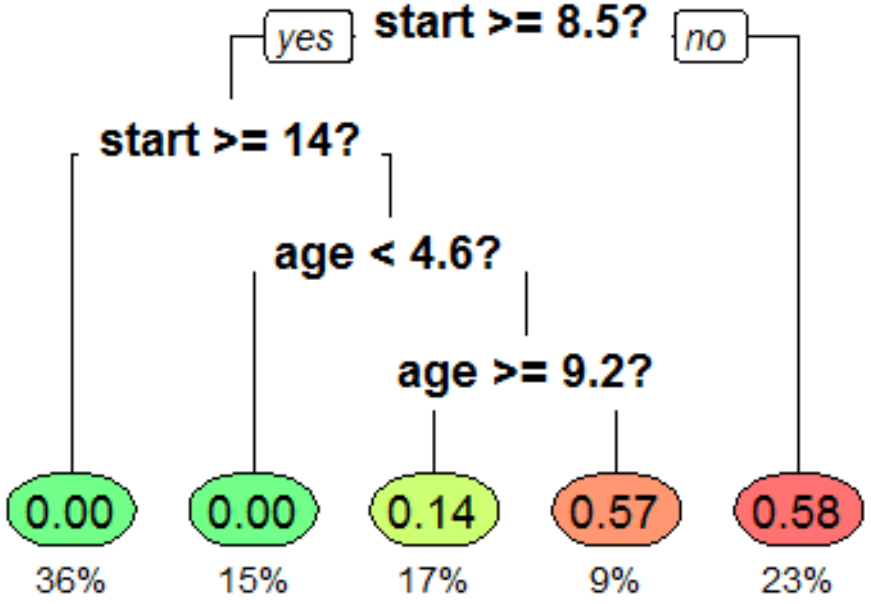


By Stephen Milborrow - CC BY-SA 3.

The forest is made up of decision trees

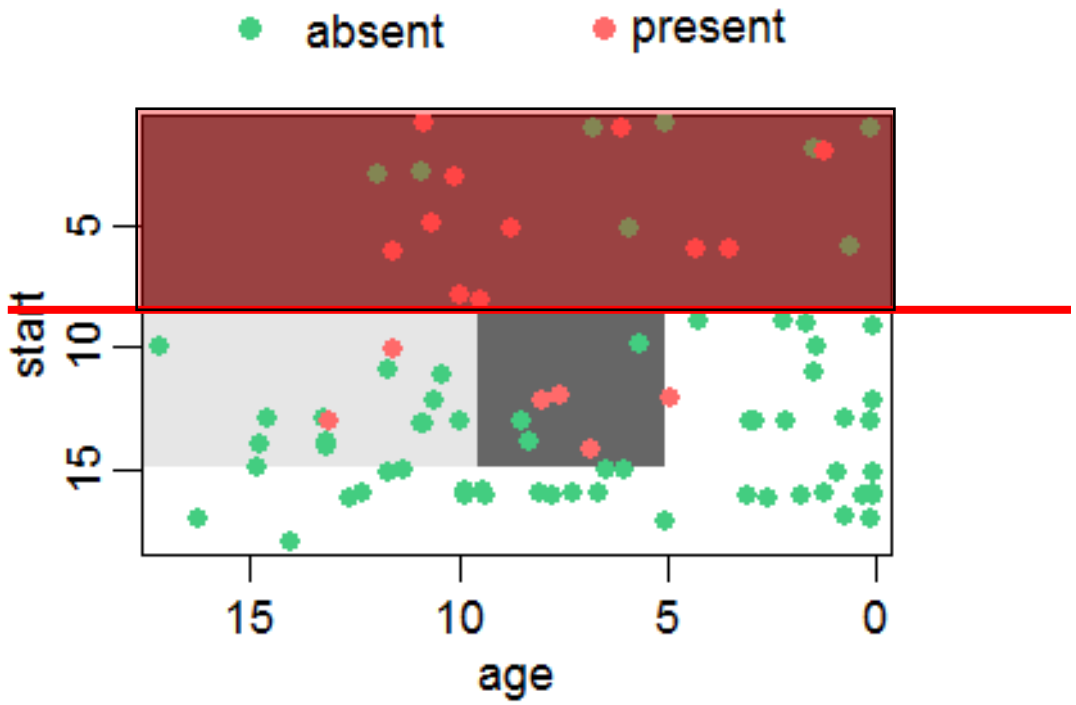
- There are two types of decision trees
  - Classification trees

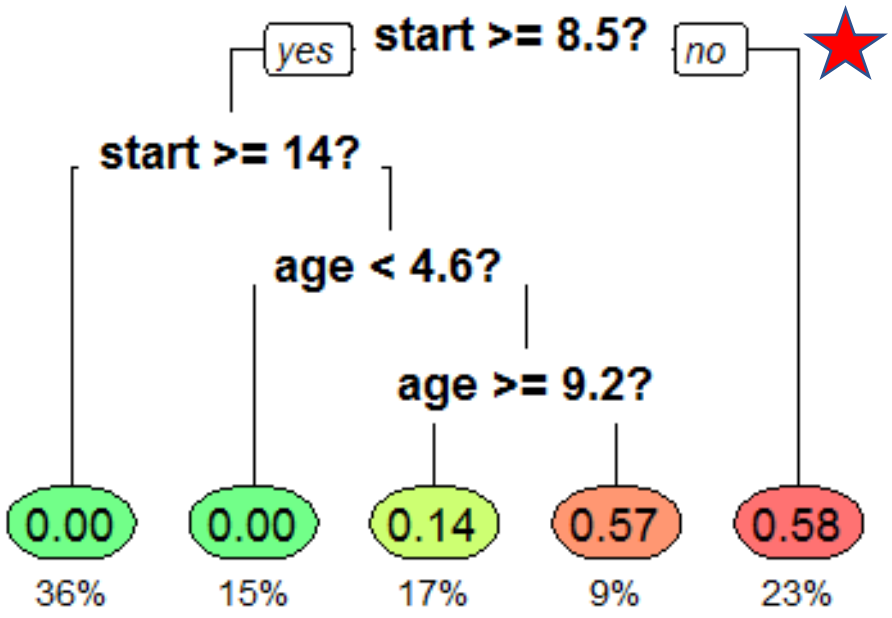




The forest is made up of decision trees

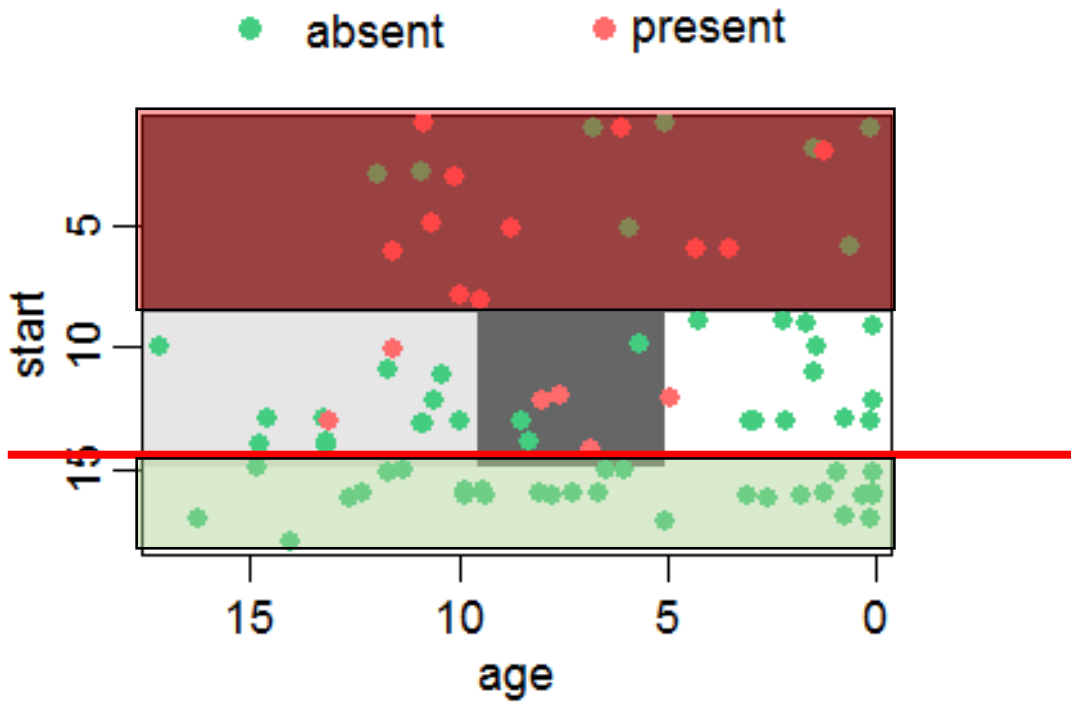
- There are two types of decision trees
  - Classification trees

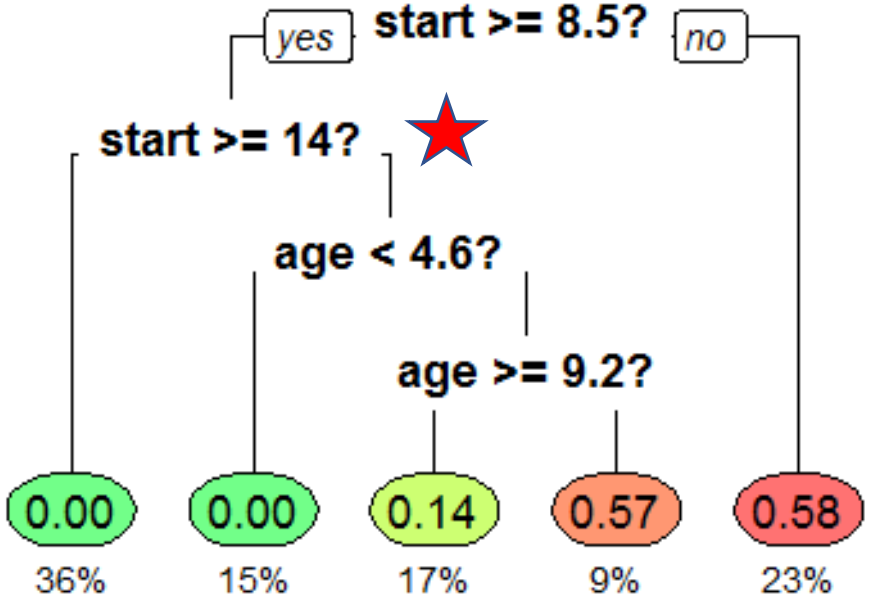




The forest is made up of decision trees

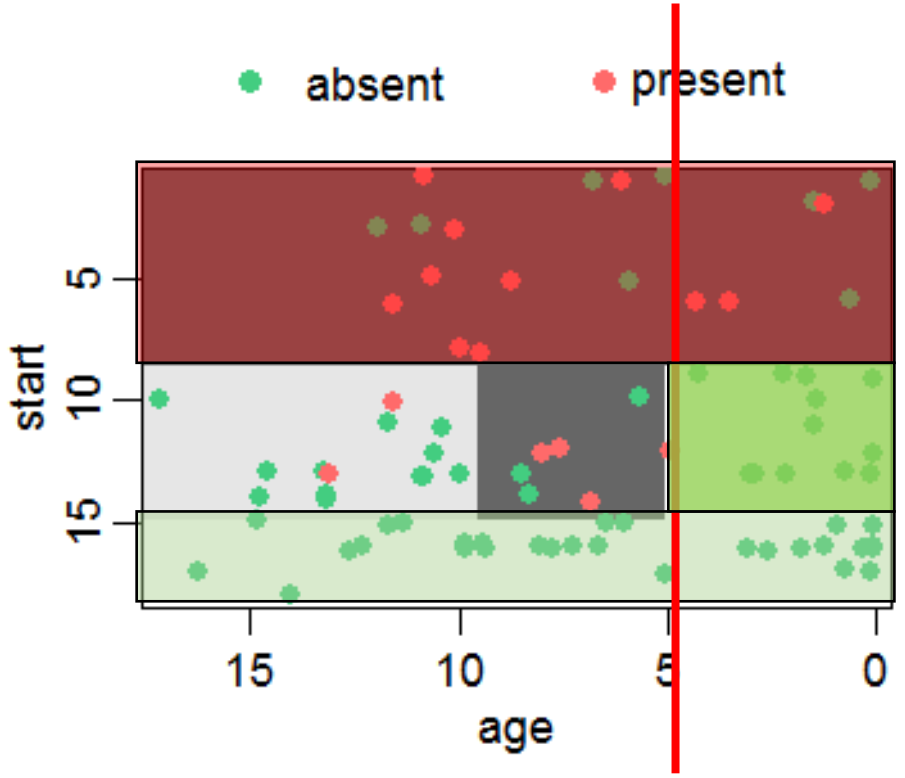
- There are two types of decision trees
  - Classification trees

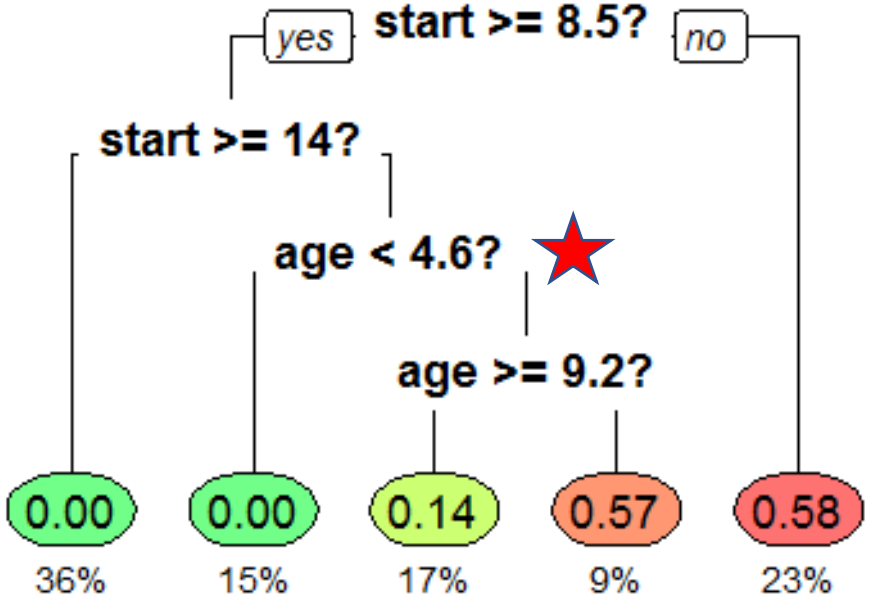




The forest is made up of decision trees

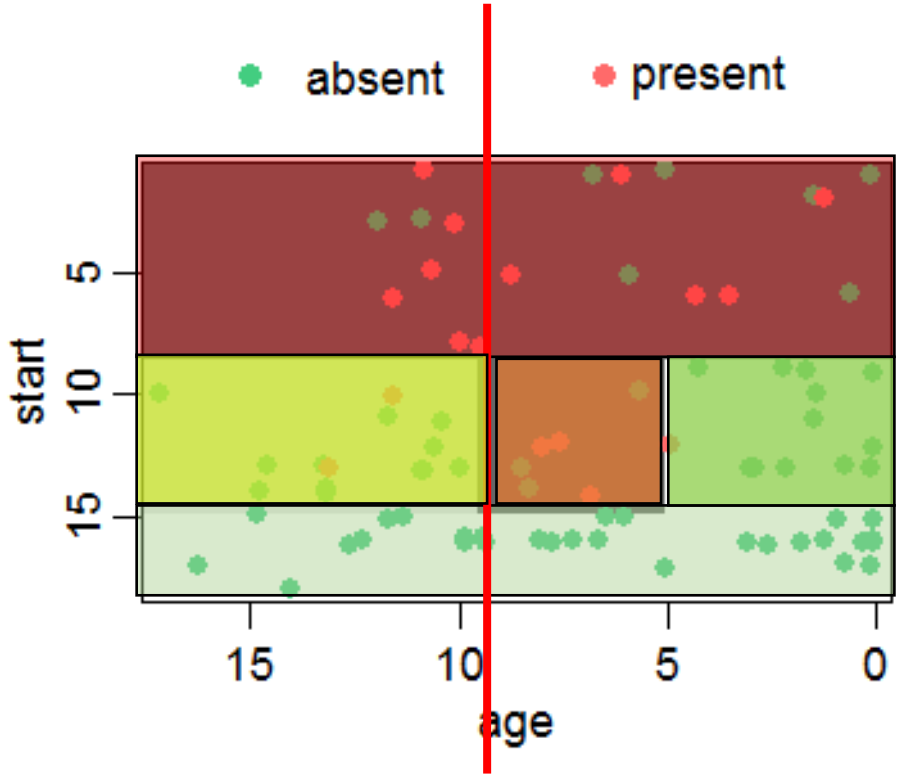
- There are two types of decision trees
  - Classification trees

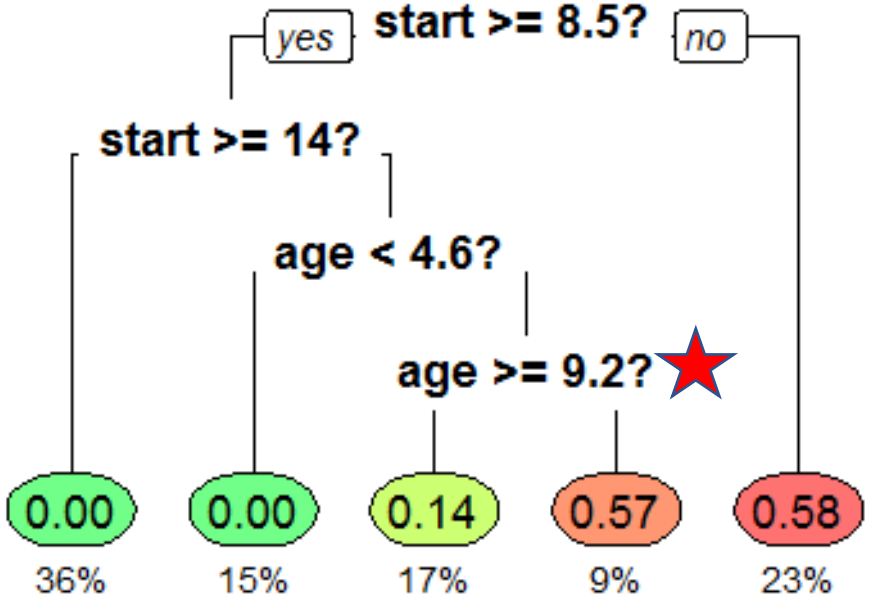




The forest is made up of decision trees

- 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 **Aggregat**ing

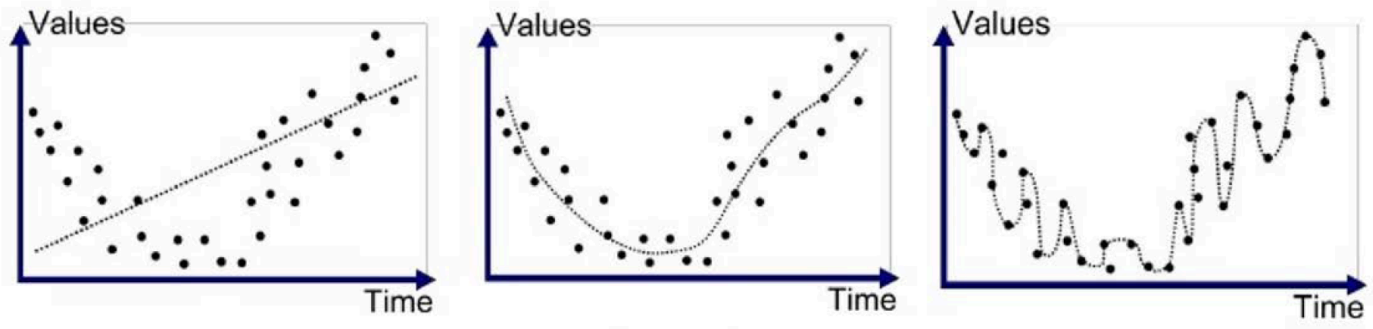
- Generates *m* new training data sets by repeatedly sampling ~2/3 of the data, with replacement.
- Builds *m* decision tress using *m* training data sets.
- *m* Models are combined by averaging (regression) or voting (classification)

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

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.

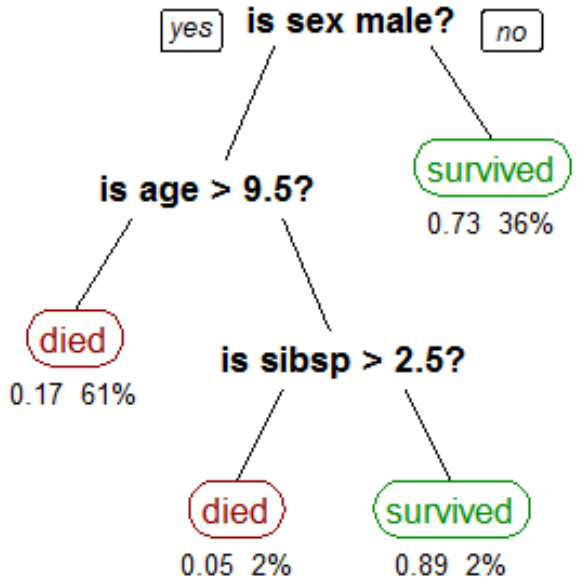
What part is Random?

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***



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!

- [picture of weak tree and strong forest]

- 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.

Error Rates and Validation

Variable Importance