Machine Learning for Official Statistics and SDGs

Fourth Live Lecture (Webinar): Starts in **15** minutes

Christophe Bontemps, UN SIAP

UNITED NATIONS siap Statistical Institute for Asia and the Pacific

Machine Learning for Official Statistics and SDGs

Fourth Live Lecture (Webinar):

Starts in **10** minutes

Christophe Bontemps, UN SIAP

UNITED NATIONS
SIAP
Statistical Institute for
Asia and the Pacific

Machine Learning for Official Statistics and SDGs

Fourth Live Lecture (Webinar):

Starts in **5** minutes

Christophe Bontemps, UN SIAP

UNITED NATIONS SIAP
Statistical Institute for Asia and the Pacific

Machine Learning for Official Statistics and SDGs

Tree-based methods: From Trees to (*random*) Forest

UNITED NATIONS siap
Statistical Institute for Asia and the Pacific

[- Agenda -]

- Introduction

[- Agenda -]

- Introduction
- Lecture "*From Trees to Forest*"

[- AGENDA -]

- Introduction
- Lecture "*From Trees to Forest*"
- Q&A

[- Agenda -]

- Introduction
- Lecture "*From Trees to Forest*"
- Q&A
- Next week

[From Trees to Forest]

Trees are methods for classification or regression analysis.

[From Trees to Forest]

Trees are methods for classification or regression analysis.

- Trees are based on recursive binary splits

[From Trees to Forest]

Trees are methods for classification or regression analysis.

- Trees are based on recursive binary splits
- The structure is simple and corresponds to regions in the variable's space

[From Trees to Forest]

Trees are methods for classification or regression analysis.

- Trees are based on recursive binary splits
- The structure is simple and corresponds to regions in the variable's space
- Each node is based on a the value of one variable and a threshold

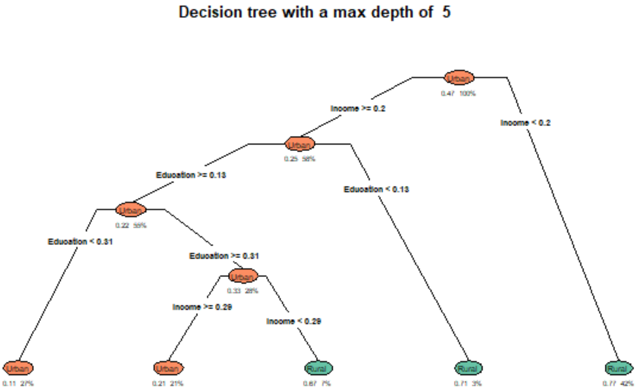
[From Trees to Forest]

Trees are methods for classification or regression analysis.

- Trees are based on recursive binary splits
- The structure is simple and corresponds to regions in the variable's space
- Each node is based on a the value of one variable and a threshold
- Trees can be very detailed and prone to **over fitting**

[Example on a simple tree]

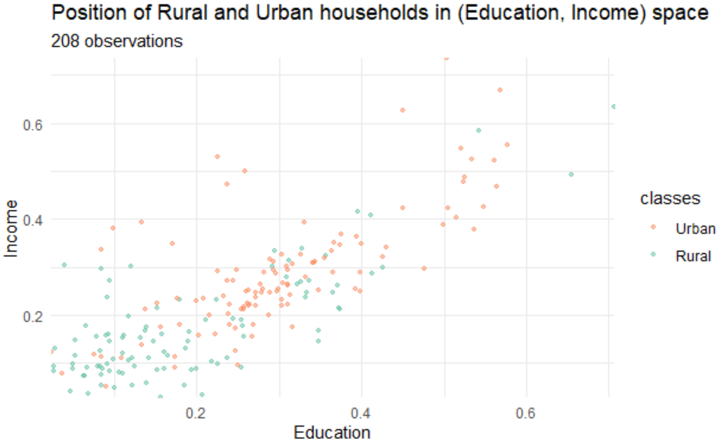
Let us see how this tree is constructed:



4 nodes leading to final leaves $\hookrightarrow$ Depth = 5

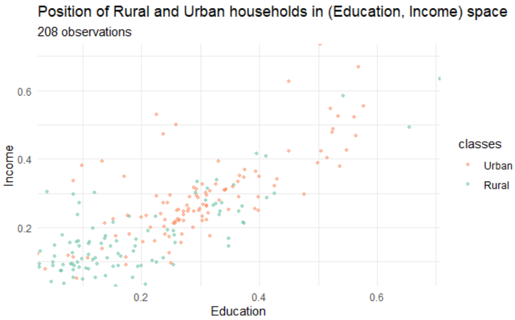
[Example on a simple tree]

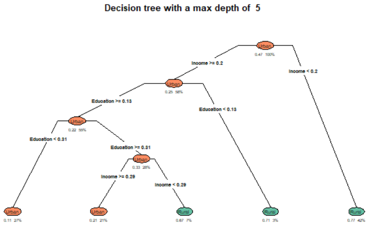
The problem is a 2D space



[Example on a simple tree]

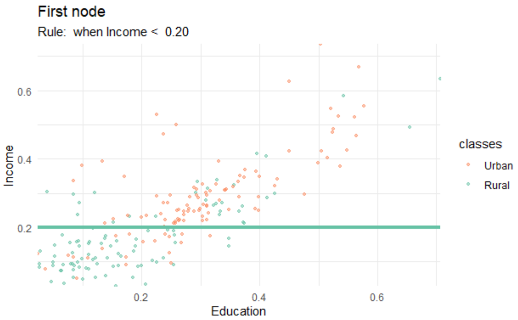
[Example on a simple tree]

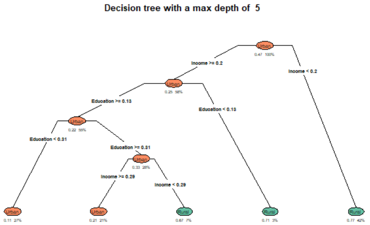




How to split the (Education, Income) space?

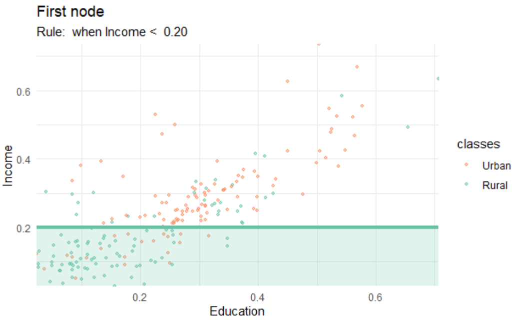
[Example on a simple tree]

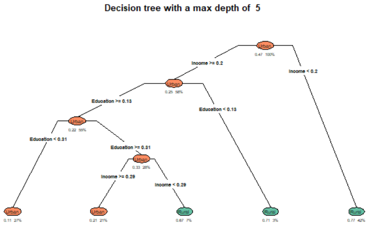




The first boundary decision line

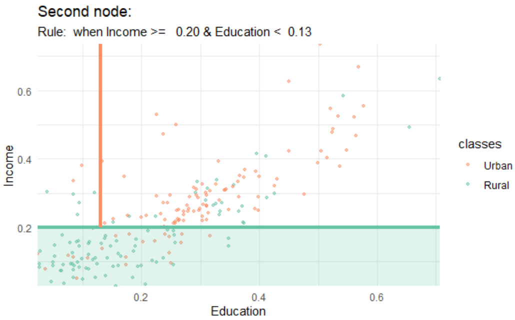
[Example on a simple tree]

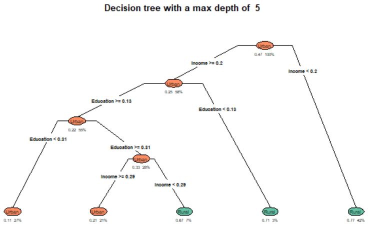




The space below the line is classified as rural

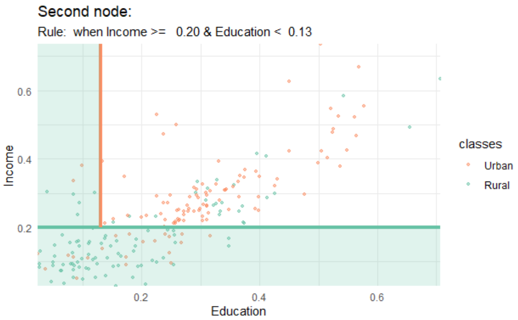
[Example on a simple tree]

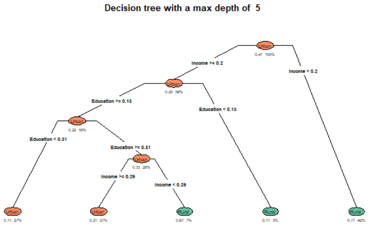




Second boundary decision line

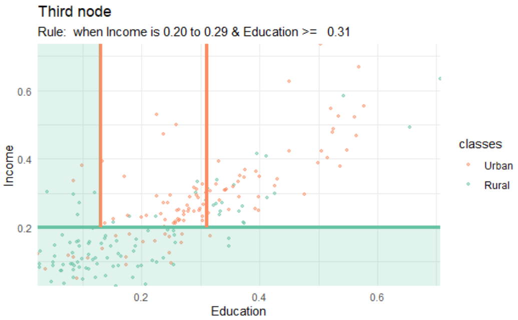
[Example on a simple tree]

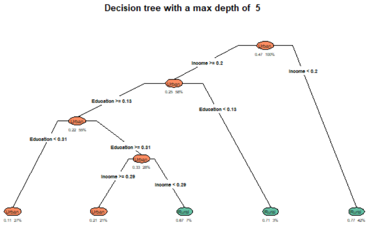




The space on the left of the line is classified as rural

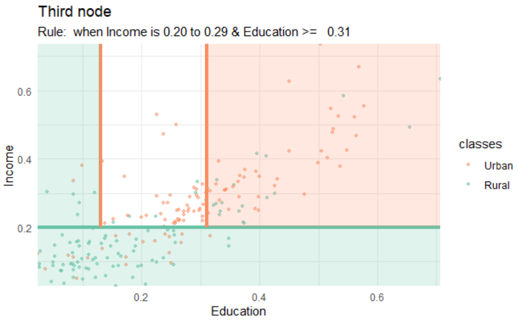
[Example on a simple tree]

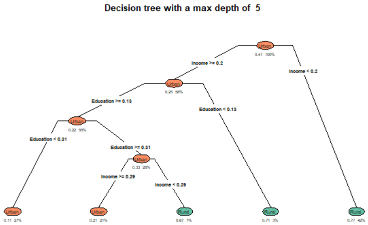




Third boundary decision line

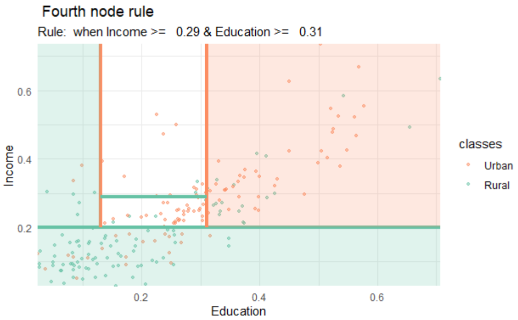
[Example on a simple tree]

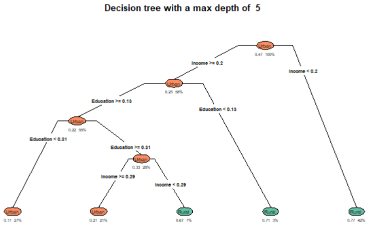




The space on the right of the line is classified as Urban

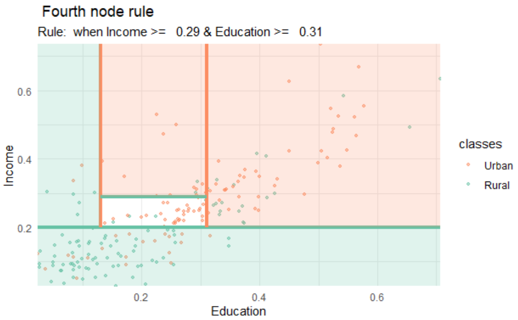
[Example on a simple tree]

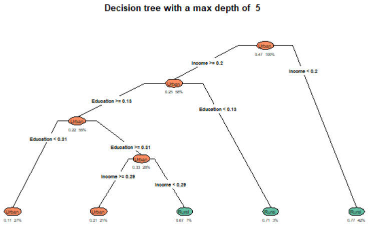




Fourth boundary decision line

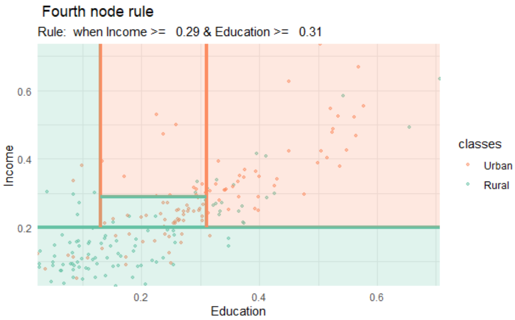
[Example on a simple tree]

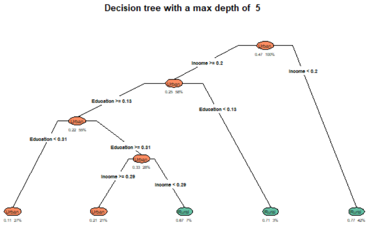




The space above the line is classified as Urban

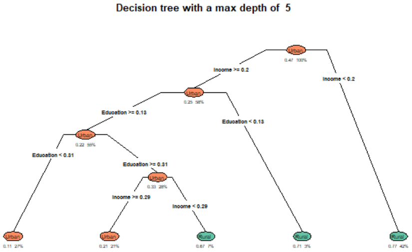
[Example on a simple tree]

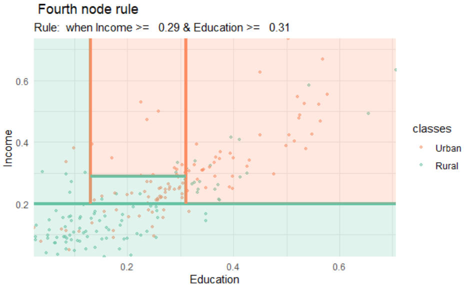




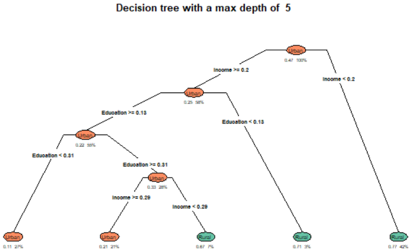
Finally, the remaining space is classified as rural

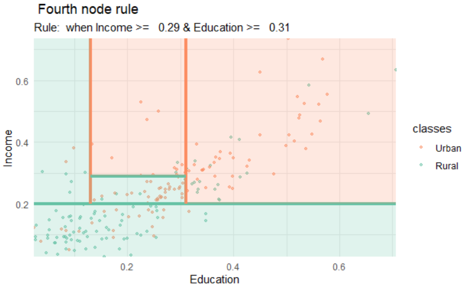
[Problems with trees]





[Problems with trees]

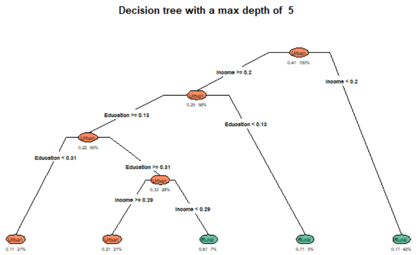


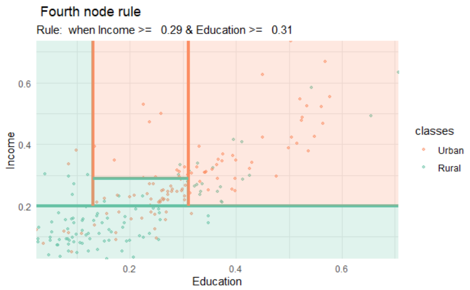


By structure:

- Trees are splitting the variable's space into "rectangles"

[PROBLEMS WITH TREES]

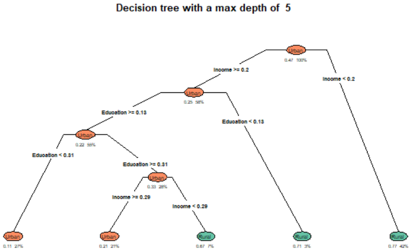


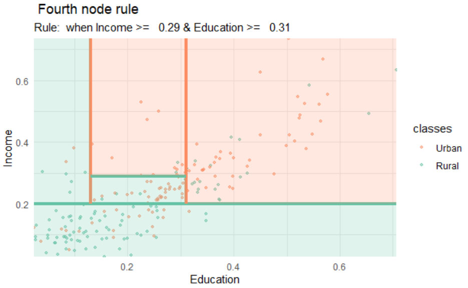


By structure:

- Trees are splitting the variable's space into "rectangles"
- Predictions using a tree may not be very accurate

[Problems with trees]

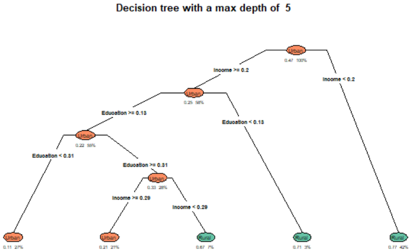


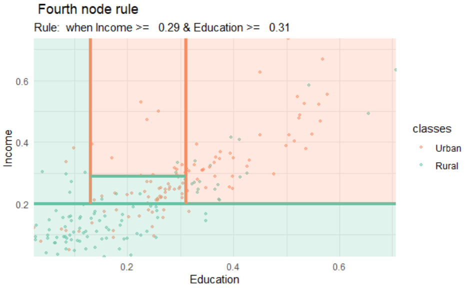


By structure:

- Trees are splitting the variable's space into "rectangles"
- Predictions using a tree may not be very accurate
- Trees are not robust to changes in the data

[Problems with trees]





By structure:

- Trees are splitting the variable's space into "rectangles"
- Predictions using a tree may not be very accurate
- Trees are not robust to changes in the data
- Trees are prone to overfiting

[Problems with trees]

The decision trees suffer from *high variance*:

[Problems with trees]

The decision trees suffer from *high variance*:

- Take some data and build a tree

[Problems with trees]

The decision trees suffer from *high variance*:

- Take some data and build a tree
- Divide the observations in two samples and build a tree on each

[Problems with trees]

The decision trees suffer from *high variance*:

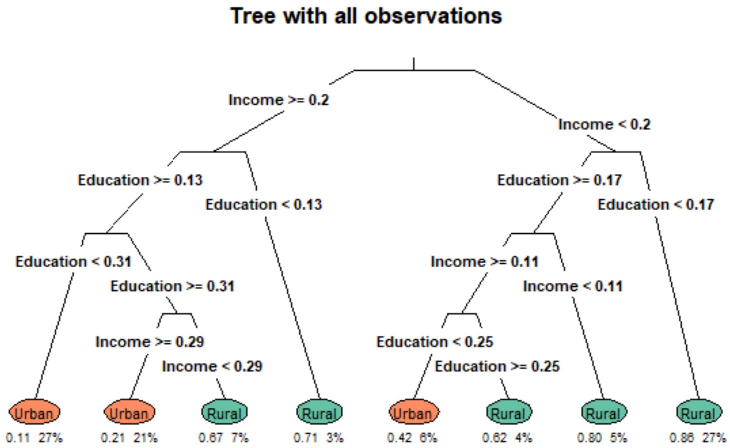
- Take some data and build a tree
- Divide the observations in two samples and build a tree on each

$\hookrightarrow$ The two trees will likely be very different

[Example on a simple tree]

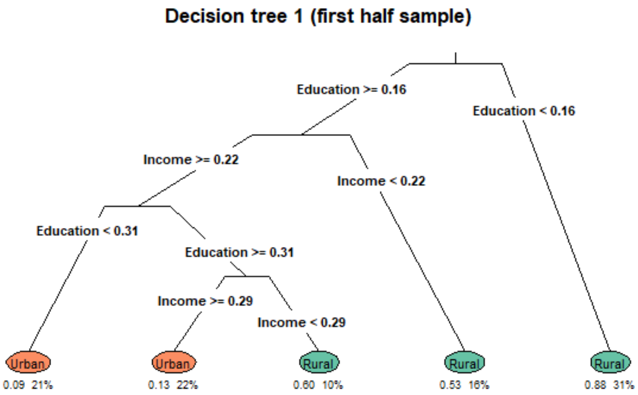
[Example on a simple tree]

Which tree do you trust?



[Example on a simple tree]

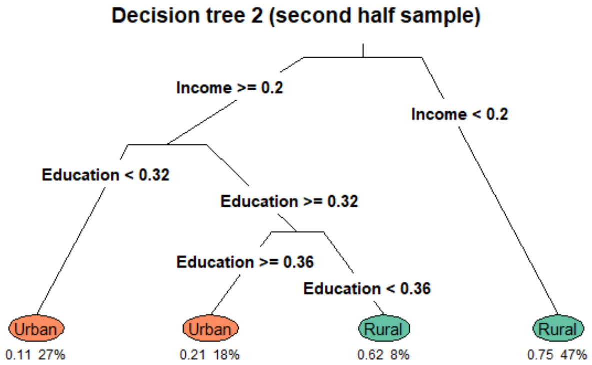
Which tree do you trust?



First tree with 50% of observations

[Example on a simple tree]

Which tree do you trust?

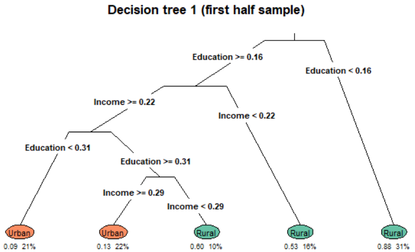


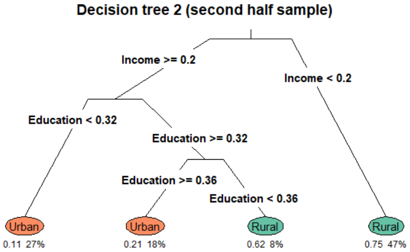
Second tree with 50% of observations

[Example on a simple tree]

Which tree do you trust?

Outcomes are different from one tree to another





$\hookrightarrow$ Instability of the outcome (= high variance)

[How to use trees?]

Alternative to simple tree models:

[How to use trees?]

Alternative to simple tree models:

- Aggregating the results of trees can help

[How to use trees?]

Alternative to simple tree models:

- Aggregating the results of trees can help

↪ "*Bagging*"

[How to use trees?]

Alternative to simple tree models:

- Aggregating the results of trees can help

↪ "*Bagging*"

- Some trees may be too similar (correlated) and have to be "decorrelated" using random draws of the variable used

[How to use trees?]

Alternative to simple tree models:

- Aggregating the results of trees can help

↪ "*Bagging*"

- Some trees may be too similar (correlated) and have to be "decorrelated" using random draws of the variable used

↪ "*Random forest*"

[How to use trees?]

Alternative to simple tree models:

- Aggregating the results of trees can help

↪ "*Bagging*"

- Some trees may be too similar (correlated) and have to be "decorrelated" using random draws of the variable used

↪ "*Random forest*"

- Trees can be constructed *sequentially*

[How to use trees?]

Alternative to simple tree models:

- Aggregating the results of trees can help

↪ "*Bagging*"

- Some trees may be too similar (correlated) and have to be "decorrelated" using random draws of the variable used

↪ "*Random forest*"

- Trees can be constructed *sequentially*

↪ "*Boosting*"

[Bagging]

Bagging stands for *Bootstrap Aggregating* uses a simple logic:

[Bagging]

Bagging stands for *Bootstrap Aggregating* uses a simple logic:

1. Draw B bootstrapped samples from the original sample

[Bagging]

Bagging stands for *Bootstrap Aggregating* uses a simple logic:

1. Draw B bootstrapped samples from the original sample
2. Construct B trees, one on each sample

[Bagging]

Bagging stands for *Bootstrap Aggregating* uses a simple logic:

1. Draw B bootstrapped samples from the original sample
2. Construct B trees, one on each sample
3. *Average* the result to get a prediction with a lower variance

[Bagging]

Bagging stands for *Bootstrap Aggregating* uses a simple logic:

1. Draw B bootstrapped samples from the original sample
2. Construct B trees, one on each sample
3. *Average* the result to get a prediction with a lower variance

[BAGGING]

Bagging stands for *Bootstrap Aggregating* uses a simple logic:

1. Draw B bootstrapped samples from the original sample
2. Construct B trees, one on each sample
3. *Average* the result to get a prediction with a lower variance

In classification, we do not average but take the *majority vote* rule

[*Bagging*]

Bagging scheme:

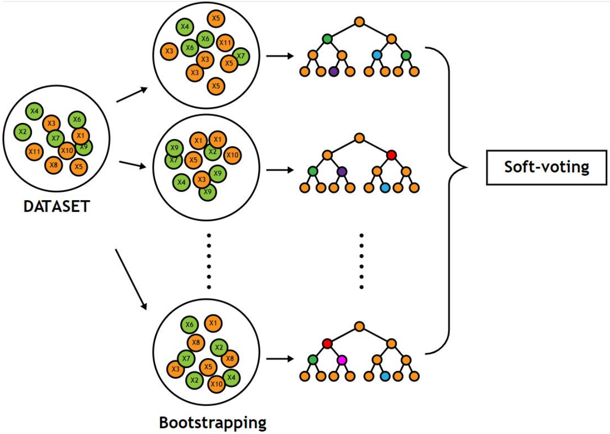


Image from igem.org

[Some simple mathematics...]

Aggregating the results of trees can help decrease the variance

[SOME SIMPLE MATHEMATICS...]

Aggregating the results of trees can help decrease the variance

- Intuitively, if $U_i, \ldots U_B$ are iid with variance σ^2, then the average of these variables has a lower variance

[Some simple mathematics...]

Aggregating the results of trees can help decrease the variance

- Intuitively, if $U_i, \ldots U_B$ are iid with variance σ^2, then the average of these variables has a lower variance

$$\begin{aligned} Var\left(\frac{1}{B}\sum_{i=1}^{B} U_i\right) &= \frac{\sigma^2}{B} \\ &< \sigma^2 \end{aligned}$$

[Some simple mathematics...]

Aggregating the results of trees can help decrease the variance

- Intuitively, if $U_i, \dots U_B$ are iid with variance σ^2, then the average of these variables has a lower variance

$$
\begin{aligned}
Var\left(\frac{1}{B}\sum_{i=1}^{B} U_i\right) &= \frac{\sigma^2}{B} \\
&< \sigma^2
\end{aligned}
$$

- This is a simple principle that is used here

[*Bagging*]

Bagging scheme:

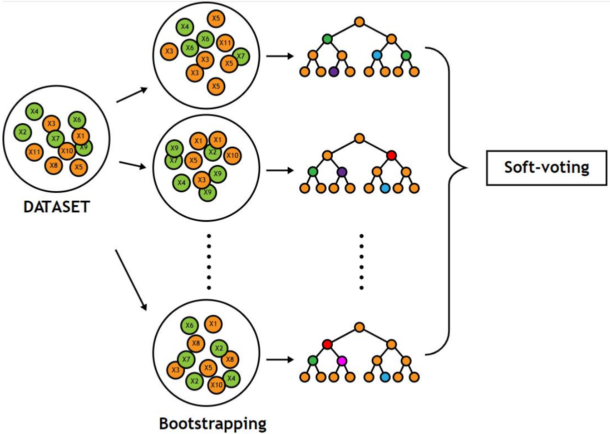


Image from igem.org

[The problem with *Bagging*]

Bootstrapped (*Bagging*) trees use **the same set of variables**

[The problem with *Bagging*]

Bootstrapped (*Bagging*) trees use **the same set of variables**

- Aggregating *correlated* trees will **not** decrease variance

[The problem with *Bagging*]

Bootstrapped (*Bagging*) trees use **the same set of variables**

- Aggregating *correlated* trees will **not** decrease variance
- Intuitively, if $U_i, \ldots U_B$ are with variance σ^2, **and cross-correlation** ρ, then the average has a variance:

$$Var\left(\frac{1}{B}\sum_{i=1}^{B} U_i\right) \quad = \quad \rho \cdot \sigma^2 + \frac{\sigma^2}{B}$$

[The problem with *Bagging*]

Bootstrapped (*Bagging*) trees use **the same set of variables**

- Aggregating *correlated* trees will **not** decrease variance
- Intuitively, if $U_i, \ldots U_B$ are with variance σ^2**, and cross-correlation** ρ, then the average has a variance:

$$Var\left(\frac{1}{B}\sum_{i=1}^{B} U_i\right) \quad = \quad \rho \cdot \sigma^2 + \frac{\sigma^2}{B}$$

- This could be "high" if ρ is "high"

[THE PROBLEM WITH *Bagging*]

Bootstrapped (*Bagging*) trees use **the same set of variables**

- Aggregating *correlated* trees will **not** decrease variance
- Intuitively, if $U_i, \ldots U_B$ are with variance σ^2, **and cross-correlation** ρ, then the average has a variance:

$$Var\left(\frac{1}{B}\sum_{i=1}^{B} U_i\right) \quad = \quad \rho \cdot \sigma^2 + \frac{\sigma^2}{B}$$

- This could be "high" if ρ is "high"

$\hookrightarrow$ *Random forest* use a *decorrelation* algorithm by adding ***randomness*** in the set of variables used

[Random Forest]

Random forest use two mechanisms:

[Random Forest]

Random forest use two mechanisms:

- Bootstrap aggregating (*bagging*) of trees → **Forest**

[Random Forest]

Random forest use two mechanisms:

- Bootstrap aggregating (*bagging*) of trees $\rightarrow$ **Forest**

$\hookrightarrow$ Construct B trees, on B bootstrapped samples

[Random Forest]

Random forest use two mechanisms:

- Bootstrap aggregating (*bagging*) of trees $\rightarrow$ **Forest**
  ↪ Construct B trees, on B bootstrapped samples
- **Random** selection of variables (*Feature sampling*)

[Random Forest]

Random forest use two mechanisms:

- Bootstrap aggregating (*bagging*) of trees $\rightarrow$ **Forest**

$\hookrightarrow$ Construct B trees, on B bootstrapped samples

- **Random** selection of variables (*Feature sampling*)

$\hookrightarrow$ At each node, randomly select only **m** variables

[Random Forest]

Random forest use two mechanisms:

- Bootstrap aggregating (*bagging*) of trees → **Forest**
  - ↪ Construct B trees, on B bootstrapped samples
- **Random** selection of variables (*Feature sampling*)
  - ↪ At each node, randomly select only **m** variables
- The resulting predictor will have a lower variance

$$Var_{Random\ forest} = \rho \cdot \sigma^2 + \frac{\sigma^2}{B}$$

[Random Forest]

Random forest use two mechanisms:

- Bootstrap aggregating (*bagging*) of trees → **Forest**
- ↪ Construct B trees, on B bootstrapped samples
- **Random** selection of variables (*Feature sampling*)
- ↪ At each node, randomly select only **m** variables
- The resulting predictor will have a lower variance

$$Var_{Random\ forest} = \rho \cdot \sigma^2 + \frac{\sigma^2}{B}$$

Common practice (p = *nb of variables*):

[Random Forest]

Random forest use two mechanisms:

- Bootstrap aggregating (*bagging*) of trees $\rightarrow$ **Forest**
  ↪ Construct B trees, on B bootstrapped samples
- **Random** selection of variables (*Feature sampling*)
  ↪ At each node, randomly select only **m** variables
- The resulting predictor will have a lower variance

$$Var_{\text{Random forest}} = \rho \cdot \sigma^2 + \frac{\sigma^2}{B}$$

Common practice (p = *nb of variables*):

- $m \approx \sqrt{p}$ in classification

[Random Forest]

Random forest use two mechanisms:

- Bootstrap aggregating (*bagging*) of trees → **Forest**
- ↪ Construct B trees, on B bootstrapped samples
- **Random** selection of variables (*Feature sampling*)
- ↪ At each node, randomly select only **m** variables
- The resulting predictor will have a lower variance

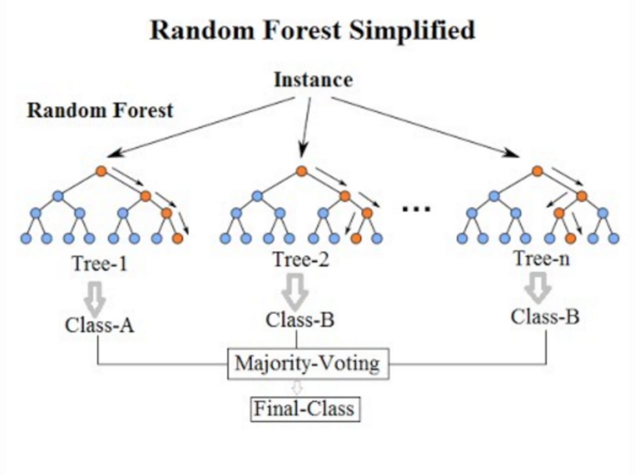
$$Var_{Random\ forest} = \rho \cdot \sigma^2 + \frac{\sigma^2}{B}$$

Common practice (p = *nb of variables*):

- $m \approx \sqrt{p}$ in classification
- $m \approx p/3$ in regression

[RANDOM FOREST]

Random forest scheme:



[Optimizing Random Forest]

Random forest are sensitive to several *hyperparameters*

[Optimizing Random Forest]

Random forest are sensitive to several *hyperparameters*

- All the parameters used to build a *tree*

[Optimizing Random Forest]

Random forest are sensitive to several *hyperparameters*

- All the parameters used to build a *tree*
  - *Purity* criterion (Gini *vs* Entropy)

[Optimizing Random Forest]

Random forest are sensitive to several *hyperparameters*

- All the parameters used to build a *tree*
  - *Purity* criterion (Gini *vs* Entropy)
  - Tree *Depth*

[Optimizing Random Forest]

Random forest are sensitive to several *hyperparameters*

- All the parameters used to build a *tree*
  - *Purity* criterion (Gini *vs* Entropy)
  - Tree *Depth*
  - *Complexity* parameter *Cp*

[Optimizing Random Forest]

Random forest are sensitive to several *hyperparameters*

- All the parameters used to build a *tree*
  - *Purity* criterion (Gini *vs* Entropy)
  - Tree *Depth*
  - *Complexity* parameter *Cp*
  - ...

[Optimizing Random Forest]

Random forest are sensitive to several *hyperparameters*

- All the parameters used to build a *tree*
  - *Purity* criterion (Gini *vs* Entropy)
  - Tree *Depth*
  - *Complexity* parameter *Cp*
  - ...
- Parameters used to construct the *Forest*

[Optimizing Random Forest]

Random forest are sensitive to several *hyperparameters*

- All the parameters used to build a *tree*
  - *Purity* criterion (Gini *vs* Entropy)
  - Tree *Depth*
  - *Complexity* parameter *Cp*
  - ...
- Parameters used to construct the *Forest*
  - Number of variables **m** used in each node

[Optimizing Random Forest]

Random forest are sensitive to several *hyperparameters*

- All the parameters used to build a *tree*
  - *Purity* criterion (Gini *vs* Entropy)
  - Tree *Depth*
  - *Complexity* parameter *Cp*
  - ...
- Parameters used to construct the *Forest*
  - Number of variables **m** used in each node
  - Number of trees (**B**)

[Optimizing Random Forest]

Random forest are sensitive to several *hyperparameters*

- All the parameters used to build a *tree*
  - *Purity* criterion (Gini *vs* Entropy)
  - Tree *Depth*
  - *Complexity* parameter *Cp*
  - ...
- Parameters used to construct the *Forest*
  - Number of variables **m** used in each node
  - Number of trees (**B**)
  - Minimum number of observations per leaves

[Optimizing Random Forest]

Random forest are sensitive to several *hyperparameters*

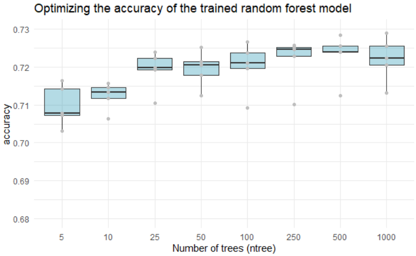
- All the parameters used to build a *tree*
  - *Purity* criterion (Gini *vs* Entropy)
  - Tree *Depth*
  - *Complexity* parameter *Cp*
  - ...
- Parameters used to construct the *Forest*
  - Number of variables **m** used in each node
  - Number of trees (**B**)
  - Minimum number of observations per leaves
  - ...

[Random Forest on an Example]

The impact of the number of trees:

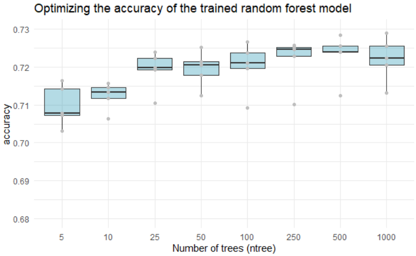
[Random Forest on an Example]

The impact of the number of trees:



[Random Forest on an Example]

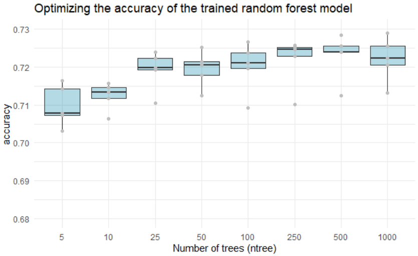
The impact of the number of trees:



↪ Moderate number of trees is OK

[Random Forest on an Example]

The impact of the number of trees:



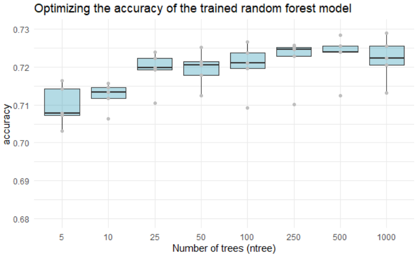
↪ Moderate number of trees is OK

- The variance depends on $\frac{\sigma^2}{B}$

$\nearrow B \Longrightarrow \frac{\sigma^2}{B} \searrow$

[Random Forest on an Example]

The impact of the number of trees:



↪ Moderate number of trees is OK

- The variance depends on $\frac{\sigma^2}{B}$

$\nearrow B \Longrightarrow \frac{\sigma^2}{B} \searrow$

↪ More tree reduces variance (up-to a certain point)

[Improvements]

Trees can also grow differently:

[Improvements]

Trees can also grow differently:

- Random Forest grow *independently*

[Improvements]

Trees can also grow differently:

- Random Forest grow *independently*
- "Vote" or average of the outcome from leaves

[IMPROVEMENTS]

Trees can also grow differently:

- Random Forest grow *independently*
- "Vote" or average of the outcome from leaves
- Trees can be constructed *sequentially*

[Improvements]

Trees can also grow differently:

- Random Forest grow *independently*
- "Vote" or average of the outcome from leaves
- Trees can be constructed *sequentially*

↪ "*Boosting*"

[Boosting]

Boosting helps construct trees sequentially

[Boosting]

Boosting helps construct trees sequentially

- Based on *weak learners*

[Boosting]

Boosting helps construct trees sequentially

- Based on *weak learners*

$\hookrightarrow$ Very simple trees grow on previous *"mistakes"*

[Boosting]

Boosting helps construct trees sequentially

- Based on *weak learners*

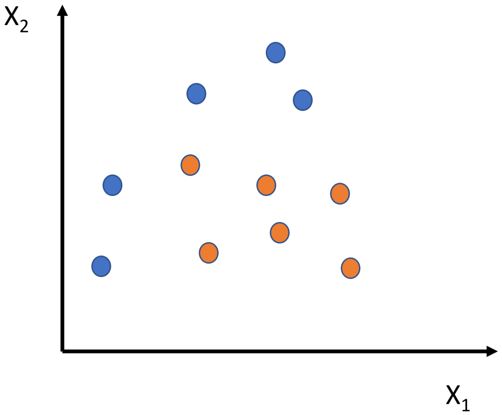
$\hookrightarrow$ Very simple trees grow on previous *"mistakes"*

- *"Mistakes"* at stage t are overweighed for stage $t + 1$

[Boosting]

[BOOSTING]

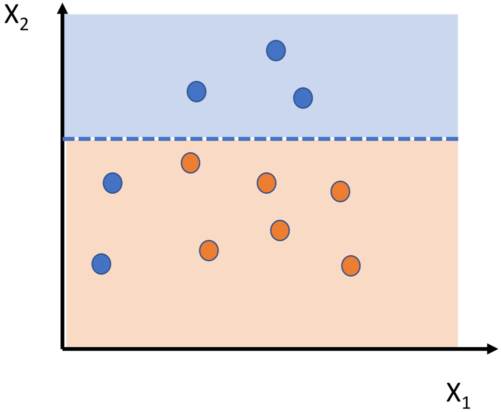
Boosting a tree (iterative process)



Two classes of observations: Orange and blue

[Boosting]

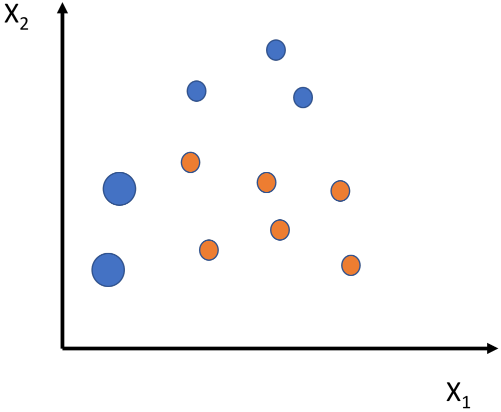
Boosting a tree (iterative process)



First weak learner

[BOOSTING]

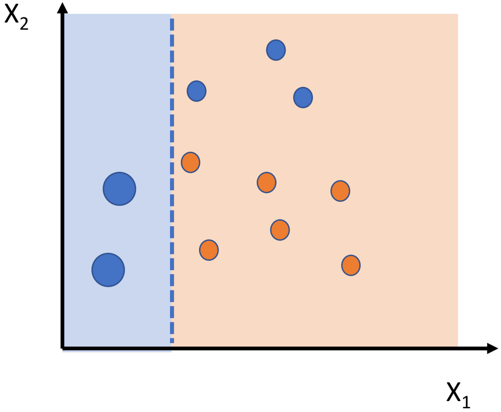
Boosting a tree (iterative process)



Misclassification overweighed

[Boosting]

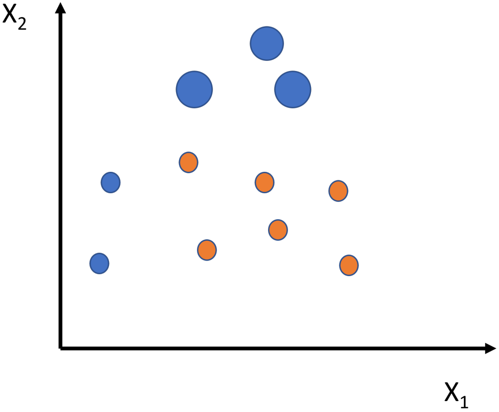
Boosting a tree (iterative process)



Second weak learner with misclassification overweighed

[Boosting]

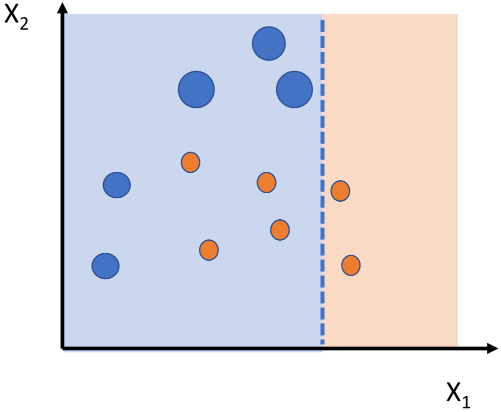
Boosting a tree (iterative process)



New misclassification overweighed

[Boosting]

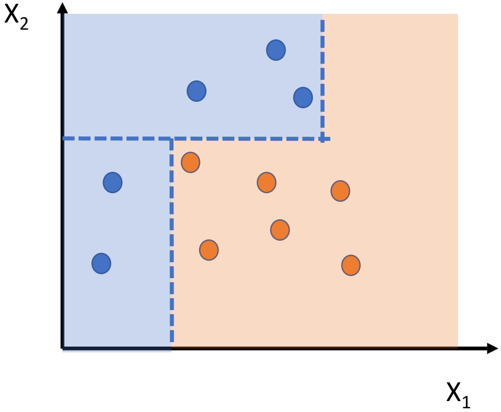
Boosting a tree (iterative process)



Third weak learner with new misclassification overweighed

[Boosting]

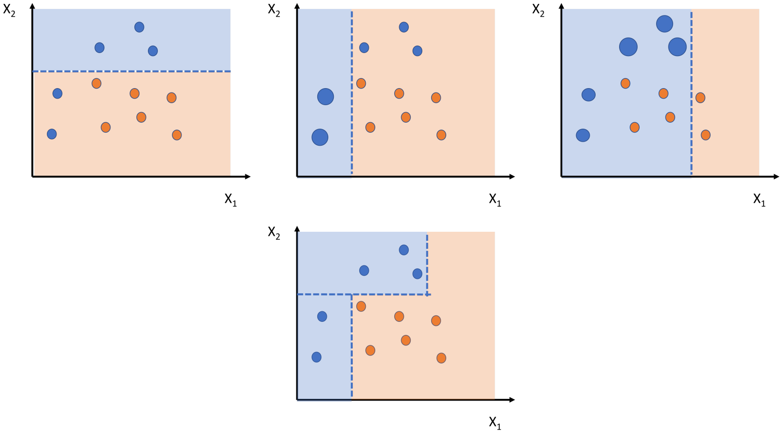
Boosting a tree (iterative process)



Combining weak learners.

[Boosting]

Boosting a tree (iterative process)



Combining *weak learners* creates an efficient tree

[Boosting]

Boosting has several features

[Boosting]

Boosting has several features

- Extremely powerful

[Boosting]

Boosting has several features

- Extremely powerful
- Based on simple trees (weak learners)

[Boosting]

Boosting has several features

- Extremely powerful
- Based on simple trees (weak learners)
- The "weight" placed on "*mistakes*" is important

[Boosting]

Boosting has several features

- Extremely powerful
- Based on simple trees (weak learners)
- The "weight" placed on "*mistakes*" is important

↪ Several choices *gradient boosting*

[Boosting]

Boosting has several features

- Extremely powerful
- Based on simple trees (weak learners)
- The "weight" placed on "*mistakes*" is important

$\hookrightarrow$ Several choices *gradient boosting*

- Sequential procedure (no parallel computation)

[BOOSTING]

Boosting has several features

- Extremely powerful
- Based on simple trees (weak learners)
- The "weight" placed on "*mistakes*" is important

↪ Several choices *gradient boosting*

- Sequential procedure (no parallel computation)

↪ *Xgboosting*

[Takeaways]

- Random forest are simple and easy to interpret

[Takeaways]

- Random forest are simple and easy to interpret
- *Random forest* use two mechanisms:

[TAKEAWAYS]

- Random forest are simple and easy to interpret
- *Random forest* use two mechanisms:
  - Bootstrap aggregating (*bagging*) of trees → **Forest**

[Takeaways]

- Random forest are simple and easy to interpret
- *Random forest* use two mechanisms:
  - Bootstrap aggregating (*bagging*) of trees $\rightarrow$ **Forest**

  $\hookrightarrow$ Construct **B** trees, on **B** bootstrapped samples

[TAKEAWAYS]

- Random forest are simple and easy to interpret
- *Random forest* use two mechanisms:
  - Bootstrap aggregating (*bagging*) of trees $\rightarrow$ **Forest**

  $\hookrightarrow$ Construct **B** trees, on **B** bootstrapped samples
  - **Random** selection of variables (*Feature sampling*)

[Takeaways]

- Random forest are simple and easy to interpret
- *Random forest* use two mechanisms:
  - Bootstrap aggregating (*bagging*) of trees → **Forest**
  - ↪ Construct **B** trees, on **B** bootstrapped samples
  - **Random** selection of variables (*Feature sampling*)
  - ↪ At each node, randomly select only **m** variables

[TAKEAWAYS]

- Random forest are simple and easy to interpret
- *Random forest* use two mechanisms:
  - Bootstrap aggregating (*bagging*) of trees $\rightarrow$ **Forest**
  - $\hookrightarrow$ Construct **B** trees, on **B** bootstrapped samples
  - **Random** selection of variables (*Feature sampling*)
  - $\hookrightarrow$ At each node, randomly select only **m** variables
  - The resulting predictor will have a lower variance

$$Var_{Random\ forest} = \rho_m \cdot \sigma^2 + \frac{\sigma^2}{B}$$

[TAKEAWAYS]

- Random forest are simple and easy to interpret
- *Random forest* use two mechanisms:
  - Bootstrap aggregating (*bagging*) of trees $\rightarrow$ **Forest**
  - $\hookrightarrow$ Construct **B** trees, on **B** bootstrapped samples
  - **Random** selection of variables (*Feature sampling*)
  - $\hookrightarrow$ At each node, randomly select only **m** variables
  - The resulting predictor will have a lower variance

$$Var_{Random\ forest} = \rho_m \cdot \sigma^2 + \frac{\sigma^2}{B}$$

- Many variations in *random forest* exist (*boosting*, *gradient boosting*, *Xgboosting*)

[TAKEAWAYS]

- Random forest are simple and easy to interpret
- *Random forest* use two mechanisms:
  - Bootstrap aggregating (*bagging*) of trees $\rightarrow$ **Forest**
  - $\hookrightarrow$ Construct **B** trees, on **B** bootstrapped samples
  - **Random** selection of variables (*Feature sampling*)
  - $\hookrightarrow$ At each node, randomly select only **m** variables
  - The resulting predictor will have a lower variance

$$Var_{Random\ forest} = \rho_m \cdot \sigma^2 + \frac{\sigma^2}{B}$$

- Many variations in *random forest* exist (*boosting*, *gradient boosting*, *Xgboosting*)
- Several parameters to adjust: Nb of trees, nb of variables in each node, minimum number of obs. in leaves/nodes, tree complexity, stopping rules, etc.

[Takeaways]

- Random forest are simple and easy to interpret
- *Random forest* use two mechanisms:
  - Bootstrap aggregating (*bagging*) of trees $\rightarrow$ **Forest**
  - $\hookrightarrow$ Construct **B** trees, on **B** bootstrapped samples
  - **Random** selection of variables (*Feature sampling*)
  - $\hookrightarrow$ At each node, randomly select only **m** variables
  - The resulting predictor will have a lower variance

$$Var_{\textit{Random forest}} = \rho_m \cdot \sigma^2 + \frac{\sigma^2}{B}$$

- Many variations in *random forest* exist (*boosting*, *gradient boosting*, *Xgboosting*)
- Several parameters to adjust: Nb of trees, nb of variables in each node, minimum number of obs. in leaves/nodes, tree complexity, stopping rules, etc.
- Implemented in many software!

[Q&A]

Write your questions in the chat

[Agenda]

[Agenda]

- Module 5: "*Advanced Methods*" (SVM, K-Means, Neural Networks) is open

[Agenda]

- Module 5: "*Advanced Methods*" (SVM, K-Means, Neural Networks) is open
- The personal data-based ML project is worth a try! (Module 4)

[AGENDA]

- Module 5: "*Advanced Methods*" (SVM, K-Means, Neural Networks) is open
- The personal data-based ML project is worth a try! (Module 4)
- **Next Thursday**, webinar on:

"*The Ethical Considerations of ML for Research and Statistical Purposes*" with
Alice Toms
(Centre for Applied Data Ethics, UK Statistics Authority)

[AGENDA]

- Module 5: "*Advanced Methods*" (SVM, K-Means, Neural Networks) is open
- The personal data-based ML project is worth a try! (Module 4)
- **Next Thursday**, webinar on:

  "*The Ethical Considerations of ML for Research and Statistical Purposes*" with
  Alice Toms
  (Centre for Applied Data Ethics, UK Statistics Authority)
- Enjoy the activities proposed - Last module opens soon.

[AGENDA]

- Module 5: "*Advanced Methods*" (SVM, K-Means, Neural Networks) is open
- The personal data-based ML project is worth a try! (Module 4)
- **Next Thursday**, webinar on:

  "*The Ethical Considerations of ML for Research and Statistical Purposes*" with
  Alice Toms
  (Centre for Applied Data Ethics, UK Statistics Authority)
- Enjoy the activities proposed - Last module opens soon.
- **Continue to post on the forums**

[AGENDA]

- Module 5: "*Advanced Methods*" (SVM, K-Means, Neural Networks) is open
- The personal data-based ML project is worth a try! (Module 4)
- **Next Thursday**, webinar on:

  "*The Ethical Considerations of ML for Research and Statistical Purposes*" with
  Alice Toms
  (Centre for Applied Data Ethics, UK Statistics Authority)
- Enjoy the activities proposed - Last module opens soon.
- **Continue to post on the forums**

Have a nice week and a happy learning!