Machine Learning with R: Random Forest Classification

Shaila Jamal

Data Analysis Support Assistant, DASH, McMaster Library

Ph.D. Candidate, School of Earth, Environment and Society, McMaster University

February 21, 2023

McMaster University sits on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the "Dish With One Spoon" wampum agreement.

Session Recording and Privacy

This session is being recorded with the intention of being shared publicly via the web for future audiences.

In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.

Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

Code of Conduct

The DASH program and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information:

scds.ca/events/code-of-conduct/

Certificate Program

The Sherman Centre offers a Certificate of Completion that rewards synchronous participation at 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.

Learn more about the Certificate Program: https://scds.ca/certificate-program

If you would like to be considered for the certificate, verify your participation in this form: https://u.mcmaster.ca/verification

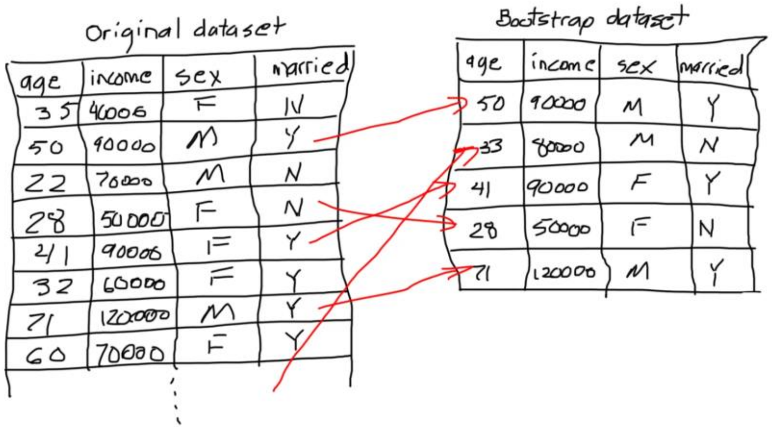
At an unspecified point during the workshop, a code will be read aloud. This is the answer to the third question of the form.

Recordings of the workshops

For workshop recordings, check here: [Search the Online Learning Catalogue | Sherman Centre for Digital Scholarship (scds.ca)]

What is Random Forest?

- Supervised machine learning techniques.
- “a predictive algorithm with higher computational capabilities.”
- widely used machine learning algorithms across multiple disciplines.
- “The random forest algorithm works by **aggregating the predictions** made by **multiple decision trees** of varying depth.”
- “Every decision tree in the forest is trained on a **subset of the dataset called the bootstrapped dataset**”



Original dataset

age	income	sex	married
35	46000	F	N
50	90000	M	Y
22	70000	M	N
28	50000	F	N
41	90000	F	Y
32	60000	F	Y
71	120000	M	Y
60	70000	F	Y

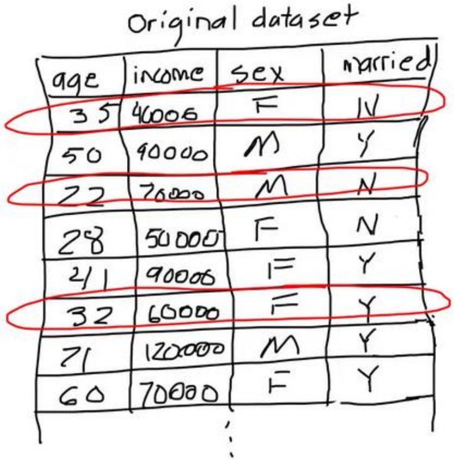
Bootstrap dataset

age	income	sex	married
50	90000	M	Y
33	80000	M	N
41	90000	F	Y
28	50000	F	N
71	120000	M	Y

Source: Maklin, C. “Random Forest In R”. July 30, 2019, accessed on January 29, 2023. https://towardsdatascience.com/random-forest-in-r-f66adf80ec9

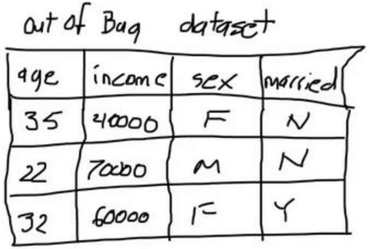
What is Random Forest?

- “The portion of samples that were left out during the construction of each decision tree in the forest are referred to as the Out-Of-Bag (OOB) dataset”
- While training, “the model automatically evaluates its own performance by running each of the samples in the OOB dataset through the forest”



Original dataset

age	income	sex	married
35	40000	F	N
50	90000	M	Y
22	70000	M	N
28	50000	F	N
21	90000	F	Y
32	60000	F	Y
71	120000	M	Y
60	70000	F	Y
⋮			



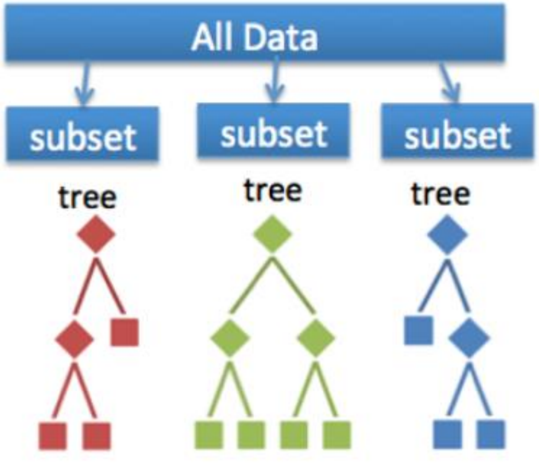
Out of Bag dataset

age	income	sex	married
35	40000	F	N
22	70000	M	N
32	60000	F	Y

Source: Maklin, C. “Random Forest In R”. July 30, 2019, accessed on January 29, 2023. https://towardsdatascience.com/random-forest-in-r-f66adf80ec9

What is Random Forest?

- “This algorithm randomly creates a forest with several trees.”
- “The more trees in the forest the more robust the forest looks like.”
- “The higher the number of trees in the forest, greater is the accuracy of the results.”
- “…builds multiple decision trees (called the forest) and glues them together to get a more accurate and stable prediction”
- “The forest it builds is a collection of Decision Trees”

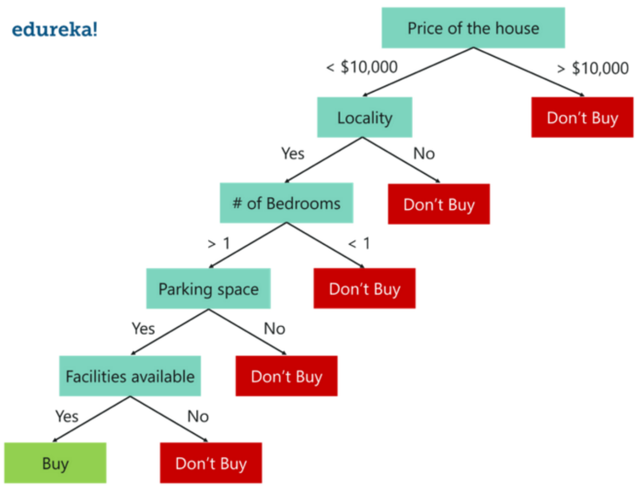


Random Forest – Random Forest In R – Edureka

Source: Lateef, Z. “A Comprehensive Guide To Random Forest In R”. Nov 25, 2020, accessed on January 29, 2023. https://www.edureka.co/blog/random-forest-classifier/

Difference Between Random Forest and Decision Trees

- “a list of parameters that you should consider before buying a house (Predict: buy or Don’t Buy)
  - Price of the house
  - Locality
  - Number of bedrooms
  - Parking space
  - Available facilities

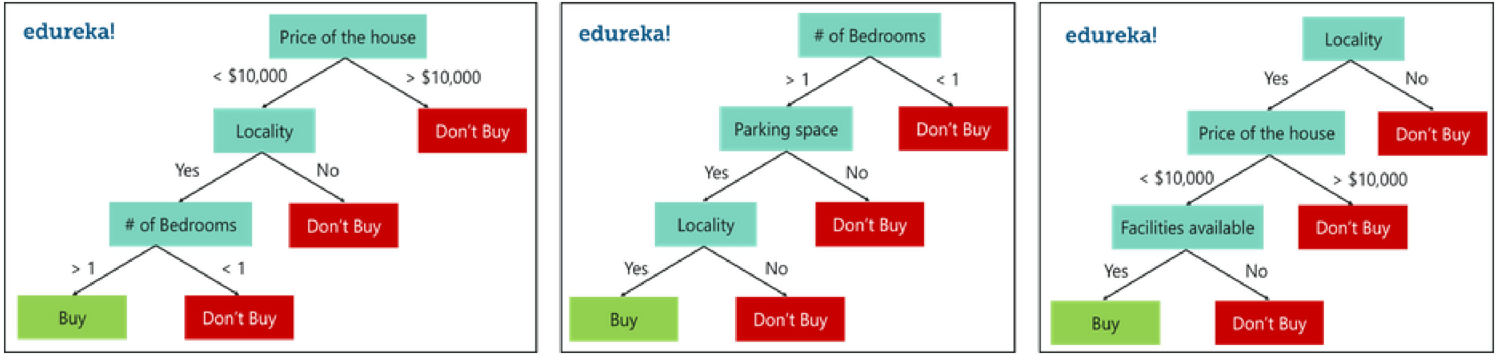


Decision Tree using the entire dataset and all parameters

Source: Lateef, Z. “A Comprehensive Guide To Random Forest In R”. Nov 25, 2020, accessed on January 29, 2023. https://www.edureka.co/blog/random-forest-classifier/

Difference Between Random Forest and Decision Trees

- “Random forest is an ensemble of decision trees, it randomly selects a set of parameters and creates a decision tree for each set of chosen parameters.”

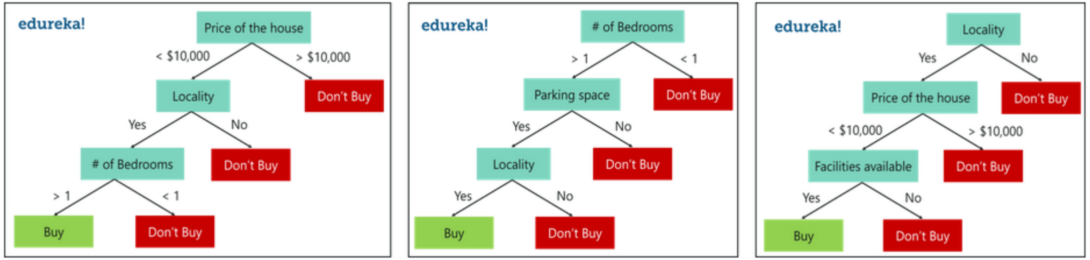


Random Forest With 3 Decision Trees – Random Forest In R – Edureka

Source: Lateef, Z. “A Comprehensive Guide To Random Forest In R”. Nov 25, 2020, accessed on January 29, 2023. https://www.edureka.co/blog/random-forest-classifier/

Difference Between Random Forest And Decision Trees

- In the example, “3 Decision Trees and each Decision Tree is taking only 3 parameters from the entire data set”
- “after creating multiple Decision trees using this method, each tree selects or votes the class (in this case the decision trees will choose whether or not a house is bought), and the class receiving the most votes by a simple majority is termed as the predicted class.



Random Forest With 3 Decision Trees – Random Forest In R – Edureka

Source: Lateef, Z. “A Comprehensive Guide To Random Forest In R”. Nov 25, 2020, accessed on January 29, 2023. https://www.edureka.co/blog/random-forest-classifier/

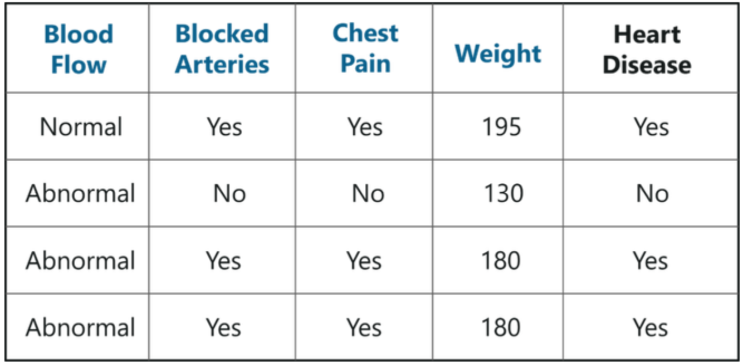
Why use Random Forest?

- “Even though Decision trees are convenient and easily implemented, they lack accuracy.
- “Decision trees work very effectively with the training data that was used to build them, but they’re not flexible when it comes to classifying the new sample. Which means that the accuracy during testing phase is very low.”
- Random Forest “….reduce the variation in the predictions by combining the result of multiple Decision trees on different samples of the data set.”
- “Random forest outperforms decision trees as a large number of uncorrelated trees(models) operating as a committee will always outperform the individual constituent models”

Source: Lateef, Z. “A Comprehensive Guide To Random Forest In R”. Nov 25, 2020, accessed on January 29, 2023. https://www.edureka.co/blog/random-forest-classifier/

How Random Forest Works?

- Step 1: Create a Bootstrapped Data Set
  - Bootstrapping indicates re-sampling a dataset. It randomly select samples from the original data set.
  - “A point to note here is that we can select the same sample more than once.”



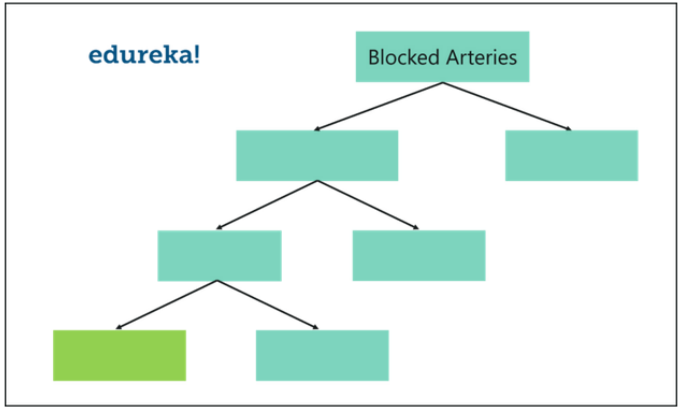
Blood Flow	Blocked Arteries	Chest Pain	Weight	Heart Disease
Normal	Yes	Yes	195	Yes
Abnormal	No	No	130	No
Abnormal	Yes	Yes	180	Yes
Abnormal	Yes	Yes	180	Yes

Bootstrapped Data Set – Random Forest In R – Edureka

Source: Lateef, Z. “A Comprehensive Guide To Random Forest In R”. Nov 25, 2020, accessed on January 29, 2023. https://www.edureka.co/blog/random-forest-classifier/

How Random Forest Works?

- Step 2: Creating Decision Trees
  - build a Decision Tree by using the bootstrapped data set
  - use a random subset of variables at each step

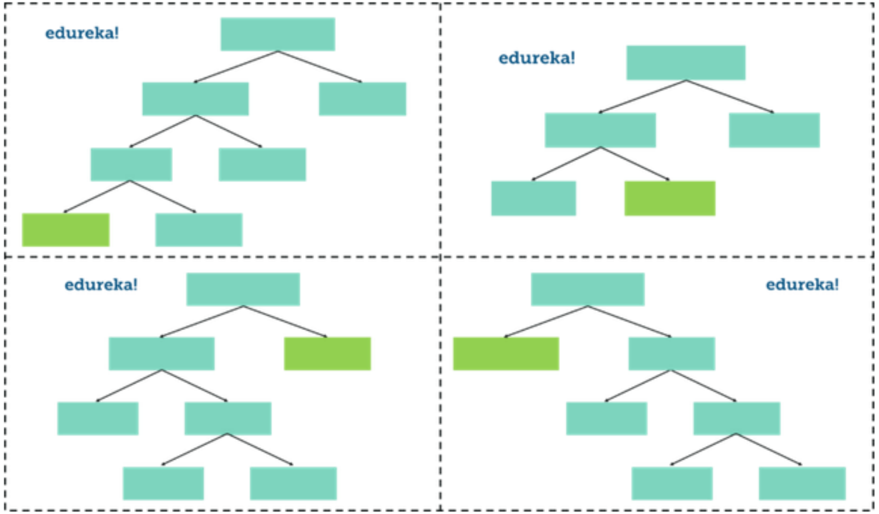


Random Forest Algorithm – Random Forest In R – Edureka

Source: Lateef, Z. "A Comprehensive Guide To Random Forest In R". Nov 25, 2020, accessed on January 29, 2023. https://www.edureka.co/blog/random-forest-classifier/

How Random Forest Works?

- Step 3: Go back to Step 1 and Repeat
  - create more decision trees by considering a subset of random predictor variables at each step.

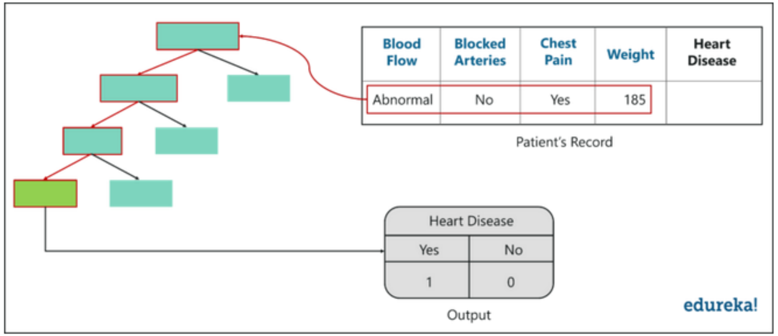


Random Forest – Random Forest In R – Edureka

Source: Lateef, Z. "A Comprehensive Guide To Random Forest In R". Nov 25, 2020, accessed on January 29, 2023. https://www.edureka.co/blog/random-forest-classifier/

How Random Forest Works?

- Step 4: Predicting the outcome of a new data point
  - “The first tree shows that the patient has heart disease, so we keep a track of that in a table as shown in the figure.”
  - “we run this data down the other decision trees and keep a track of the class predicted by each tree. After running the data down all the trees in the Random Forest, we check which class got the majority votes.



Blood Flow	Blocked Arteries	Chest Pain	Weight	Heart Disease
Abnormal	No	Yes	185	

Patient's Record

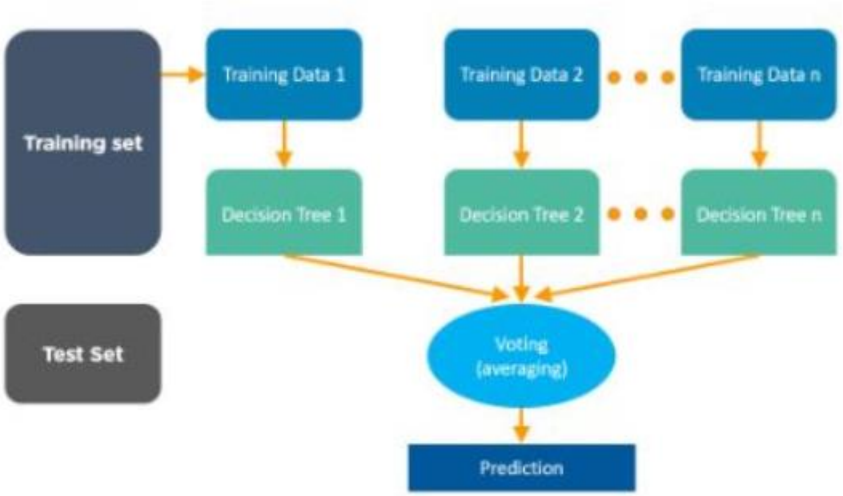
Heart Disease	
Yes	No
1	0

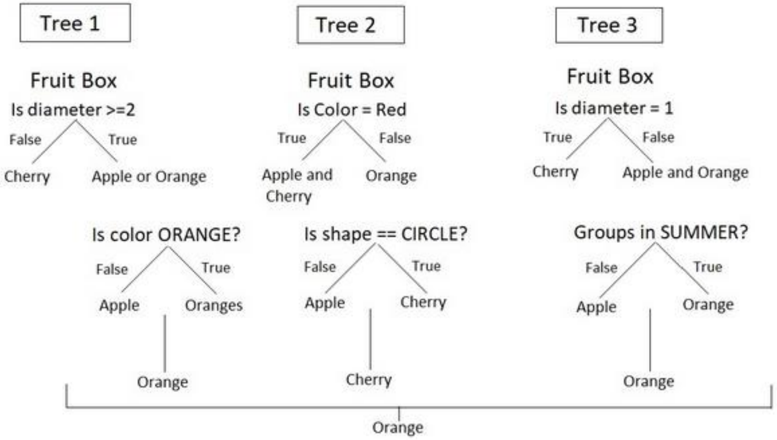
Output

Output – Random Forest In R – Edureka

Source: Lateef, Z. “A Comprehensive Guide To Random Forest In R”. Nov 25, 2020, accessed on January 29, 2023. https://www.edureka.co/blog/random-forest-classifier/

How Random Forest Works?



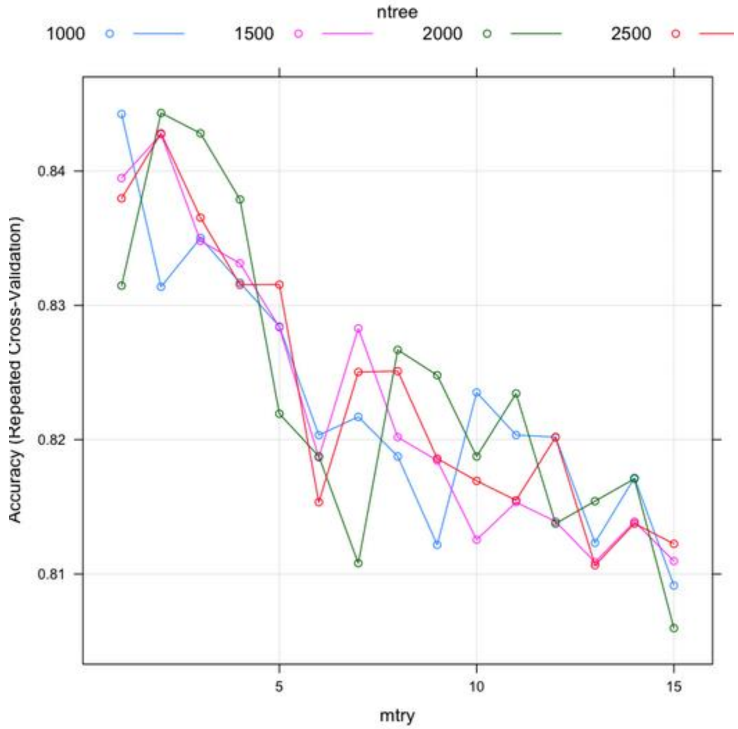


Source: Shruti, M. "Introduction to Random Forest in R". Feb 07, 2023, accessed on Feb 20, 2023. https://www.simplilearn.com/tutorials/data-science-tutorial/random-forest-in-r/
"Random Forest Approach in R Programming". Jun 05, 2020. Accessed on Feb 20, 2023. https://www.geeksforgeeks.org/random-forest-approach-in-r-programming/

Tuning Parameters in Random Forest?

- mtry = the mtry parameter controls how many of the input features a decision tree has available to consider at any given point in time.

- ntree = number of trees. We want enough trees to stabalize the error but using too many trees is unncessarily inefficient, especially when using large data sets.



Source: AFIT Data Science Lab R Programming Guide". N.d. accessed on Feb 20, 2023. https://afit-r.github.io/random_forests

Materials are available at: shorturl.at/E1238

Or

https://drive.google.com/drive/folders/1GpsLuIikhPF0cG_zKtDJpyuI0pxnrWVT?usp=sharing

Thank you!

Questions: jamals16@mcmaster.ca