RANDOM FORESTS

Nissy Milcia William
nissymilcia.w@niser.ac.in
Under the guidance of Dr. Subhankar Mishra
MACHINE LEARNING CS-460
National Institute of Science Education and Research Bhubaneswar

RANDOM FOREST

A random forest is a machine learning algorithm consisting of a collection of tree structured classifiers where each tree casts a unit vote for the most popular class at some .

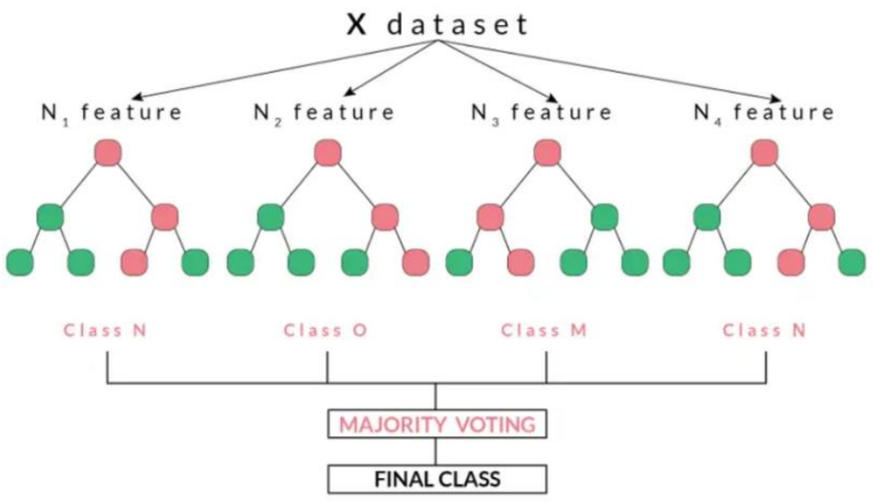
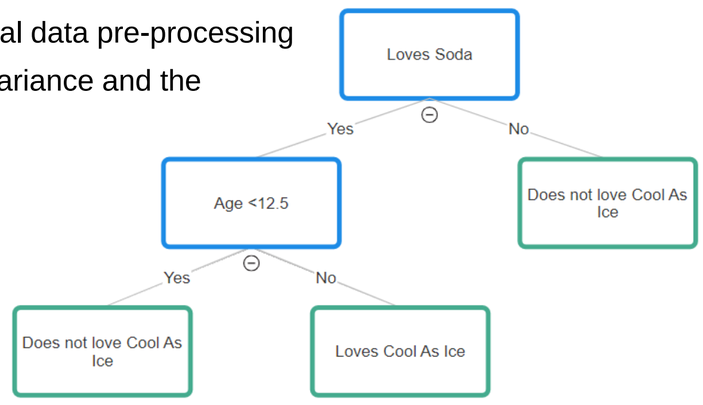


Image from https://images.app.goo.gl/HzKPL4tH5uvjosnD9

DECISION TREES

- Flowchart-like tree structures representing the feature, the rules and the result of the algorithm
- Intuitive approach , requires minimal data pre-processing
- Prone to overfitting as a result of variance and the entire structure of the tree can

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



ENSEMBLE LEARNING: BAGGING

- Group of models (weak learners) work to achieve a final prediction, rather than a single model
- Consists of Bootstrapping, Parallel training and aggregation

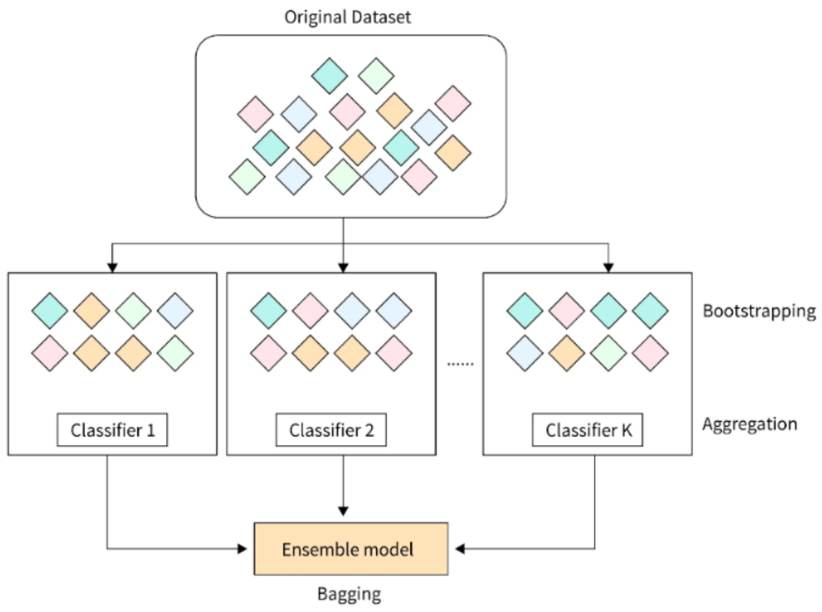


Image from: https://images.app.goo.gl/PupB1pzREbiGnUZJ8

WHY RANDOM FOREST?

- Ensemble of multiple decision predicts more accurate results, particularly when the individual trees are uncorrelated
- Randomness helps increase tree diversity and is generated by:
  1. **Bagging**: Ensures that every tree is built from a different subset of the original dataset, can give ongoing estimates of the generalization error of the combined ensemble of trees (out-of-bag estimates)
  2. **Random Feature selection**: During the selection of the appropriate variable to split on at each node, a fixed number of random variables are considered, rather than the entire collection of variables
- **Out-of-bag estimates**

ALGORITHM

Precondition: A training set $S := (x_1, y_1), \ldots, (x_n, y_n)$, features F, and number of trees in forest B.

function RandomForest(S, F)
  $H \leftarrow \emptyset$
  for $i \in 1, \ldots, B$ **do**
    $S^{(i)} \leftarrow$ A bootstrap sample from S
    $h_i \leftarrow$ RandomizedTreeLearn($S^{(i)}, F$)
    $H \leftarrow H \cup \{h_i\}$
  end for
  return H
end function
function RandomizedTreeLearn(S, F)
  At each node:
    $f \leftarrow$ very small subset of F
    Split on best feature in f
  return The learned tree
end function

The predictions from each tree are combined either by a majority vote (for classification tasks) or an average (for regression tasks)

id	x0	x1	x2	x3	x4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1
2.8	6.2	4.3	5.3	5.5		

x0, x1

id
2
0
2
4
5
5

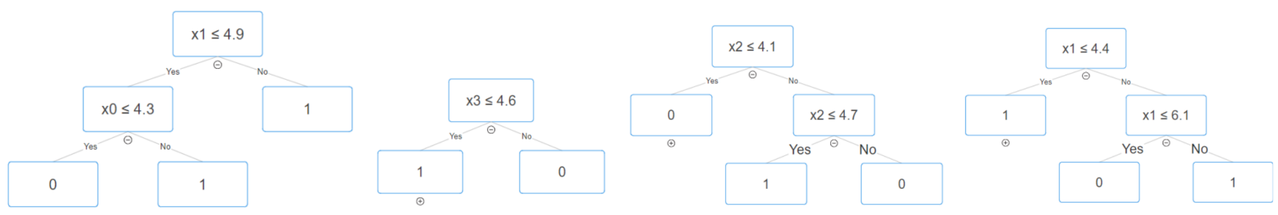
x2, x3

id
2
1
3
1
4
4

x2, x4

id
4
1
3
0
0
2

x1, x3



1 0 1 1

APPLICATIONS

- Random forest algorithms are versatile and can be used for classification as well as regression tasks
- Capable of handling large datasets, missing and complex data. They are regularly applied across fields such as Finance, Healthcare, E-commerce, Manufacturing. Here are some specific examples:

A top-down approach to classify enzyme functional classes and sub-classes using random forest

Chetan Kumar* and Alok Choudhary

Using Random Forest Classifier for Particle Identification in the ALICE Experiment

Tomasz Trzciński[1(✉)], Łukasz Graczykowski[2], and Michał Glinka[1] for the ALICE Collaboration

Random Forest Models To Predict Aqueous Solubility

David S. Palmer, Noel M. O'Boyle,† Robert C. Glen, and John B. O. Mitchell*

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Received May 5, 2006

REFERENCES

1. https://youtu.be/_L39rN6gz7Y?si=ZlVx1xZG3k4iRvFz
2. https://youtu.be/RtrBtAKwcxQ?si=t8MGQuXc5a2jyAja
3. https://www.ibm.com/topics/bagging#:~:text=Bagging%2C%20also%20known%20as%20bootstrap,be%20chosen%20more%20than%20once.
4. https://www.ibm.com/topics/random-forest
5. https://youtu.be/sQ870aTKqiM?si=FJYLdc5s3S-5bUqu
6. https://youtu.be/J4Wdy0Wc_xQ?si=IMa6mX41G9cakktc
7. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf
8. https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/
9. https://link.springer.com/article/10.1023/A:1010933404324
10. https://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests.pdf