March 8, 2024 | 4:30-6:10pm

Virtual Workshop

Machine Learning with R: Random Forest Classification Approach

u.mcmaster.ca/scds-events

Data Analysis Support Hub

SCDS

Library

McMaster University

Machine learning with R: Random Forest Classification Approach

Amirreza Mousavi
Master’s student at McMaster University

DASH: Data Analysis Support Hub Workshop Series
8/3/2024

McMaster University is located on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the “Dish With One Spoon” wampum agreement.

Laslovarga, “Webster Falls in Winter, Waterdown, Hamilton, Ontario, Canada - Spencer Gorge / Webster's Falls Conservation Area,” 23 January 2011, Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Waterdawn_Webster_Falls_in_Winter8.jpg

Code of Conduct

The Sherman Centre 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 Attendance 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

Verify your participation at a session: 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.

DASH: Data Analysis Support Hub Workshops

Register for upcoming DASH events: u.mcmaster.ca/scds-events

March 14. 2024: “Web Scraping with Python’s Beautiful Soup” – John Fink

March 22, 2024 : “Machine Learning with R: Logistic Regression” - Humayun Kabir

March 28, 2024: “Intermediate Python programming” – Amirreza Mousavi

Apr 30, 2024: “Survival Analysis with R” – Humayun Kabir

Etc

Book an Appointment with the DASH Team

Receive help from a member of the DASH team! DASH can assist with the following topics:

- Creating data visualizations, including charts, graphs, and scatter plots
- Figuring out which statistical tests to run (e.g., t-test, chi-square, etc.).
- Analyzing data with software including SPSS, Python, R, SAS, ArcGIS, MATLAB, and Excel
- Choosing which software package to use, including free and open-source software
- Troubleshooting problems related to file formats, data retrieval, and download
- Selecting methodology and type of data analysis to use in a thesis project

Book an appointment: https://library.mcmaster.ca/services/dash

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.

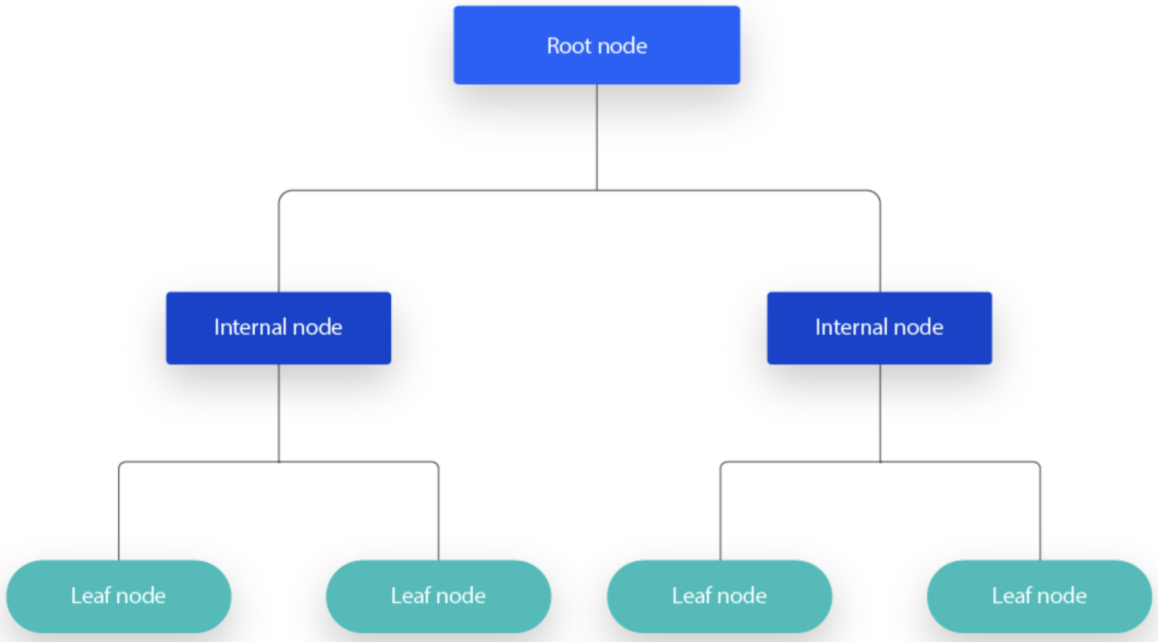
Random Forest

What is Random Forest?

Random forests, also known as random decision forests, are an ensemble learning technique used for classification, regression, and other tasks. This method works by creating numerous decision trees during training. In classification tasks, the random forest outputs the class that is chosen by the majority of the trees. In regression tasks, it returns the mean or average prediction from the individual trees. Random decision forests address the tendency of decision trees to overfit to their training data.

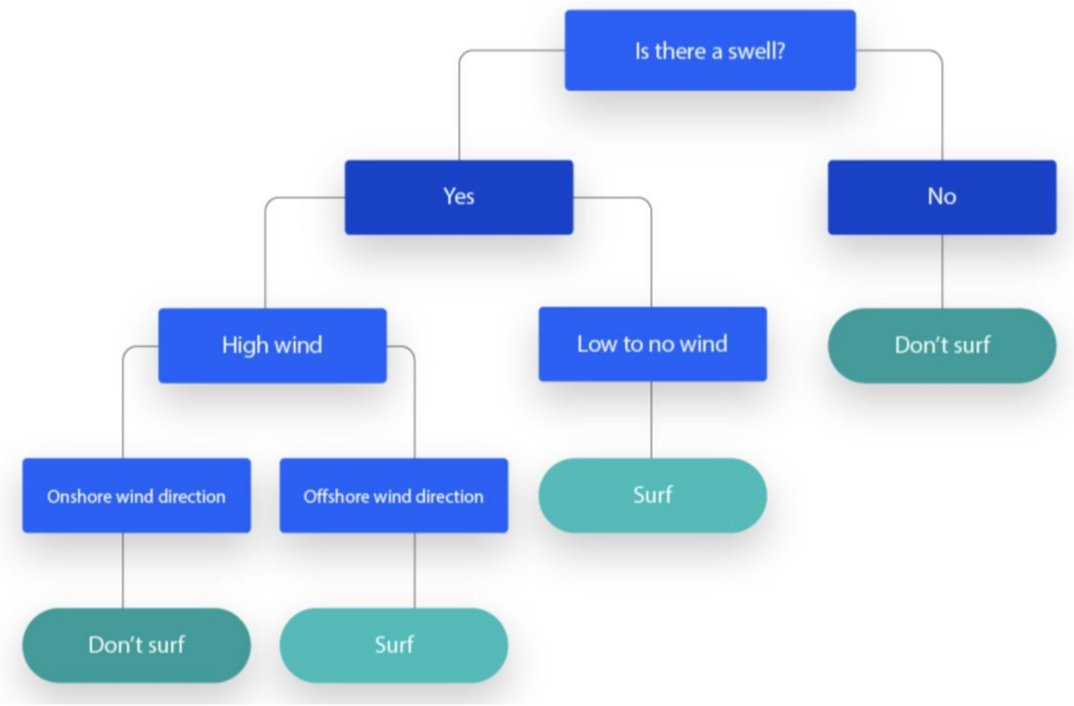
Decision Tree

- A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.



Source:https://www.ibm.com/topics/decision-trees

Decision Tree



Source:https://www.ibm.com/topics/decision-trees

It uses a "divide and conquer" approach, searching greedily to find the best points to split the data in the tree. This splitting process happens recursively from the top down until most or all of the records are grouped into specific class labels. The homogeneity of these sets depends on the tree's complexity. Smaller trees can easily have pure leaf nodes where all data points belong to a single class. However, as the tree grows, it gets harder to keep this purity, often resulting in too few data points in a subtree. This situation is called data fragmentation and can lead to overfitting.

To avoid overfitting, decision trees should not be multiplied beyond necessity. In simpler terms, decision trees should only get complex when needed, as the simplest explanation tends to be the best.

To manage complexity and prevent overfitting we can use pruning. Pruning removes branches that split on less important features. The model's accuracy is then checked using cross-validation. This way, we ensure our decision tree is both effective and not too complex for the data at hand.

Decision Tree

Advantages:

•**Interpretability**: Decision Trees are relatively simple to understand and interpret, making them desirable for collaborative decision-making and explaining results to non-technically oriented stakeholders.

•**Deals with Unbalanced Data**: This method is highly competent at handling diverse datasets and doesn't require balanced data to generate a robust model.

•**Variable Selection**: Decision Trees can identify the most significant variables and the relation between two or more variables, serving as a worthwhile tool for data exploration.

•**Handles Missing Values**: They have the ability to handle missing values in the dataset by looking at the probability of observing the various classes.

•**Non-parametric Nature**: They are a non-parametric method, meaning no assumptions about the space distribution and the classifier structure are made, which keeps the model simple and less prone to significant errors.

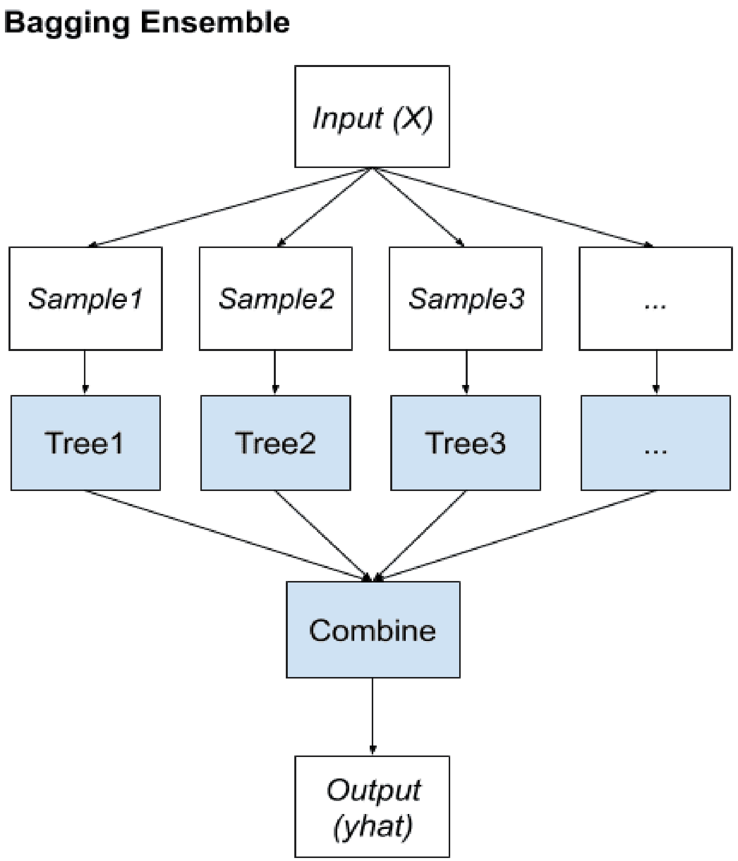
Decision Trees

Disadvantages:

- **Overfitting**: This refers to the creation of overly complex trees that fit the training data too closely and perform poorly on unseen data.
- **Sensitive to Small Variations**: Even slight changes in the input data can drastically alter the structure of the decision tree, impacting its stability.
- **Biased Learning**: Without proper parameter tuning, Decision Trees have a tendency to create biased trees if some classes dominate.

Ensemble Methods

- These methods consist of a collection of classifiers, such as decision trees, whose predictions are combined to determine the most popular outcome. Two widely recognized ensemble methods are bagging, also known as bootstrap aggregation, and boosting.



Random Forest

1. For $b = 1$ to B:

   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.

   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

      i. Select m variables at random from the p variables.

      ii. Pick the best variable/split-point among the m.

      iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$.

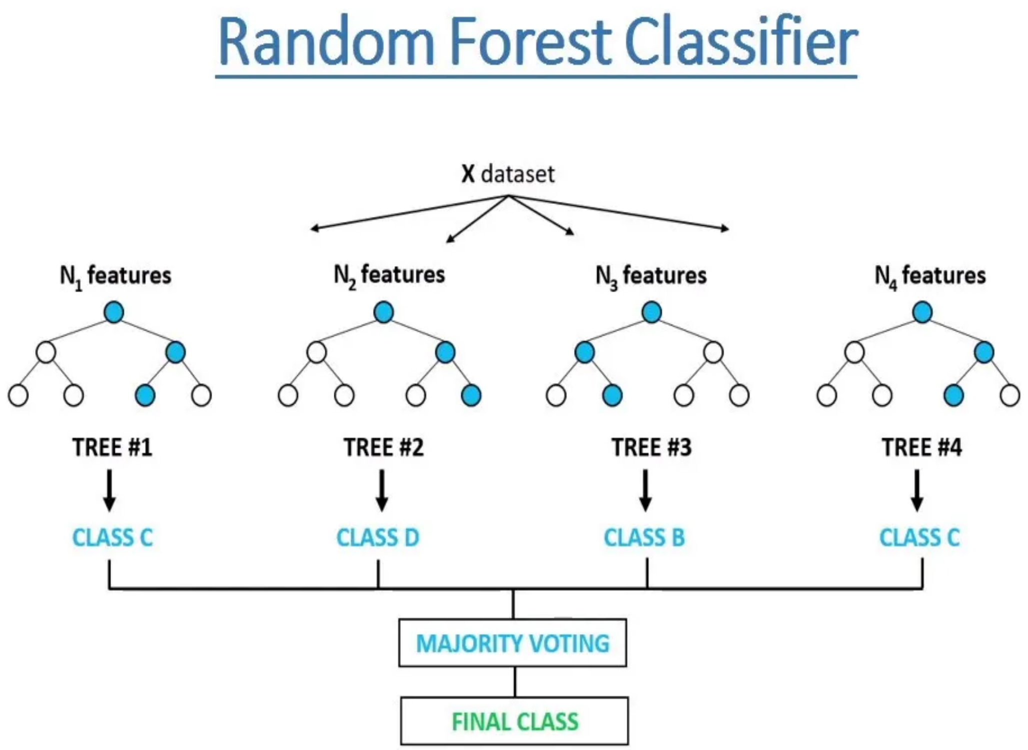
To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}^B(x) = \textit{majority vote } \{\hat{C}_b(x)\}_1^B$.

Source: Elements of statistical learning, chapter 15

Random Forest



we can see that random forest algorithm consists of a group of decision trees, where each tree is built using a data sample drawn from a training set with replacement, known as the bootstrap sample. Within this training sample, one-third is reserved as test data, called the out-of-bag (oob) sample. To add further randomness, feature bagging is applied, introducing more diversity and reducing correlation among the decision trees.

The prediction process varies based on the problem type. For regression tasks, the predictions from individual decision trees are averaged. In classification tasks, the prediction is determined by a majority vote—selecting the most frequent categorical variable as the predicted class.

Source: https://medium.com/@mrmaster907/introduction-random-forest-classification-by-example-6983d95c7b91

Random Forest

The random forest algorithm builds upon the bagging method by incorporating both bagging and feature randomness to construct an uncorrelated forest of decision trees.

Random forest main hyperparameters:

1. Number of estimators
2. Max depth
3. Minimum sample split
4. Max Featuess

Random Forest

Benefits:

- Reduced risk of overfitting
- Provides flexibility
- Easy to determine feature importance

Challenges:

- Time-consuming
- Memory usage
- Interpretability

Random Forest

Applications:

- **Classification**
- **Regression**
- **Anomaly Detection**
- **Recommendation Systems**
- **Customer Segmentation**
- **….**

Contact

mousas27@mcmaster.ca

Book an appointment with DASH: https://library.mcmaster.ca/services/dash

Contact DASH: Data Analysis Support Hub: libdash@mcmaster.ca