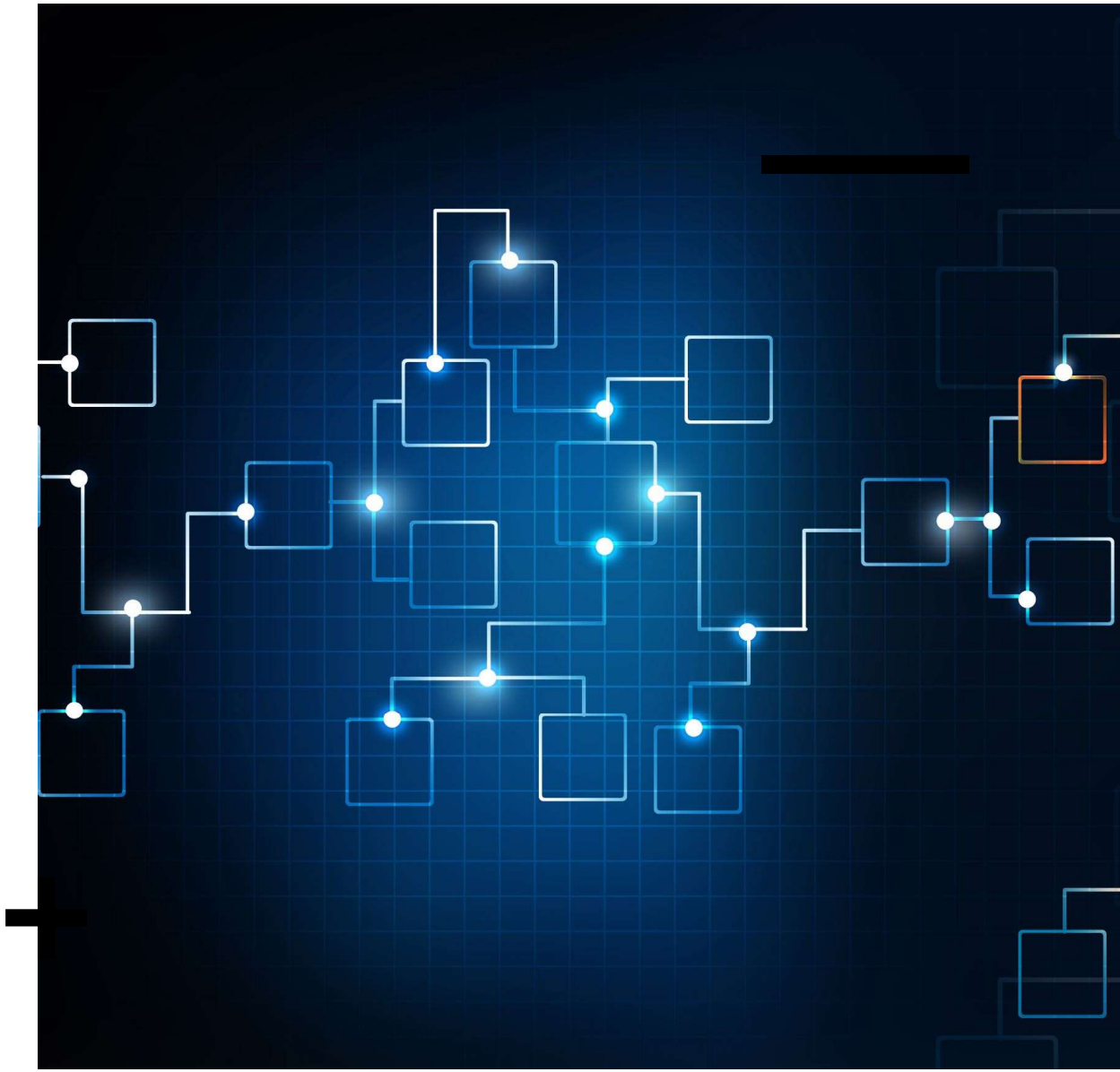
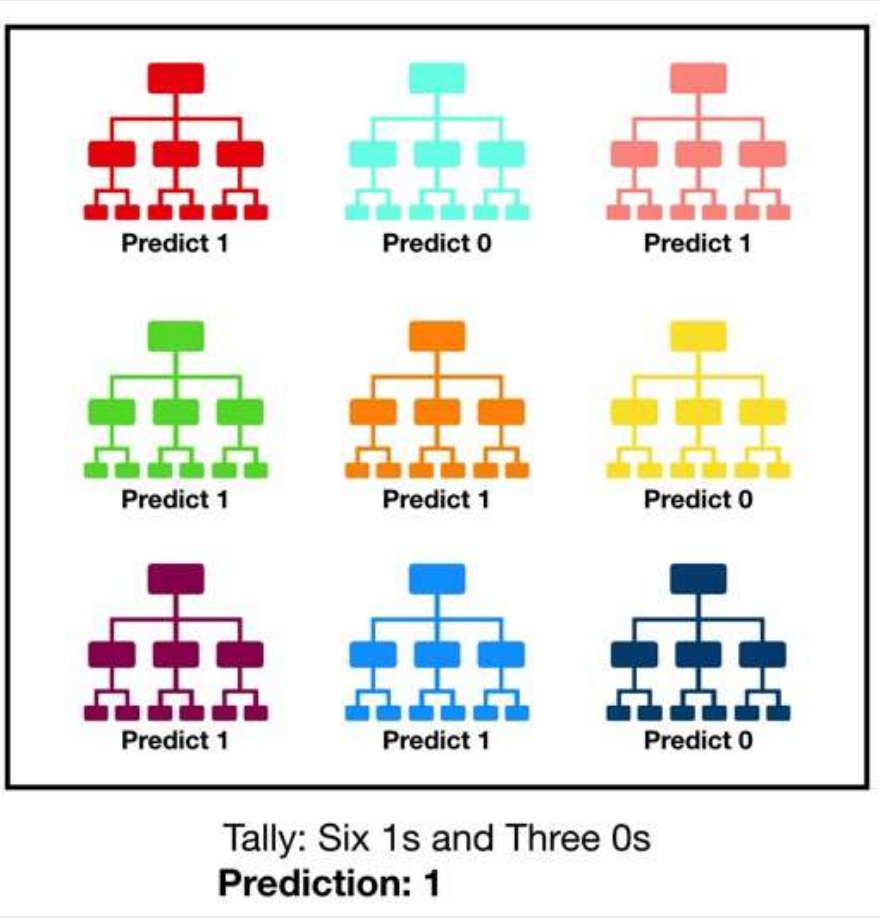
Random Forest



What is Random Forest?

– Random Forest is a machine learning method that is used to solve regression and classification problems

– The method creates multiple ***decision trees*** during training and makes a prediction from those decision trees

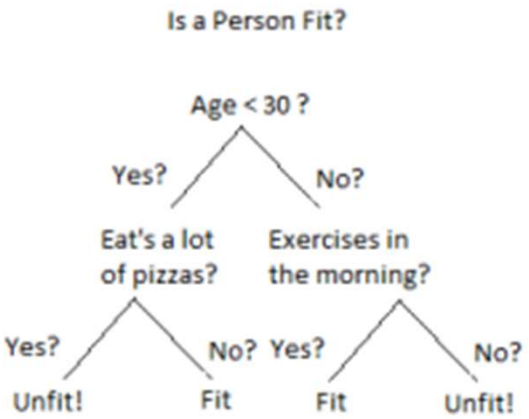


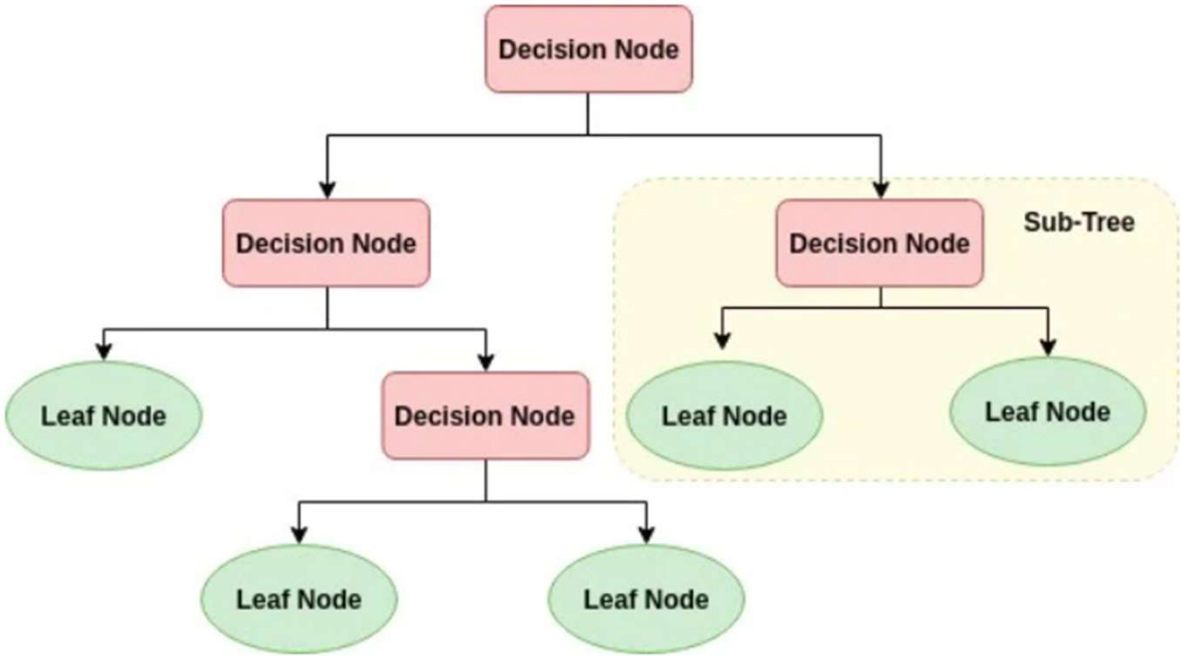
Decision Trees

A Decision Tree is an upside-down tree that makes decisions based on conditions of the data

It is comprised of ***decision nodes*** and ***decision leaves***

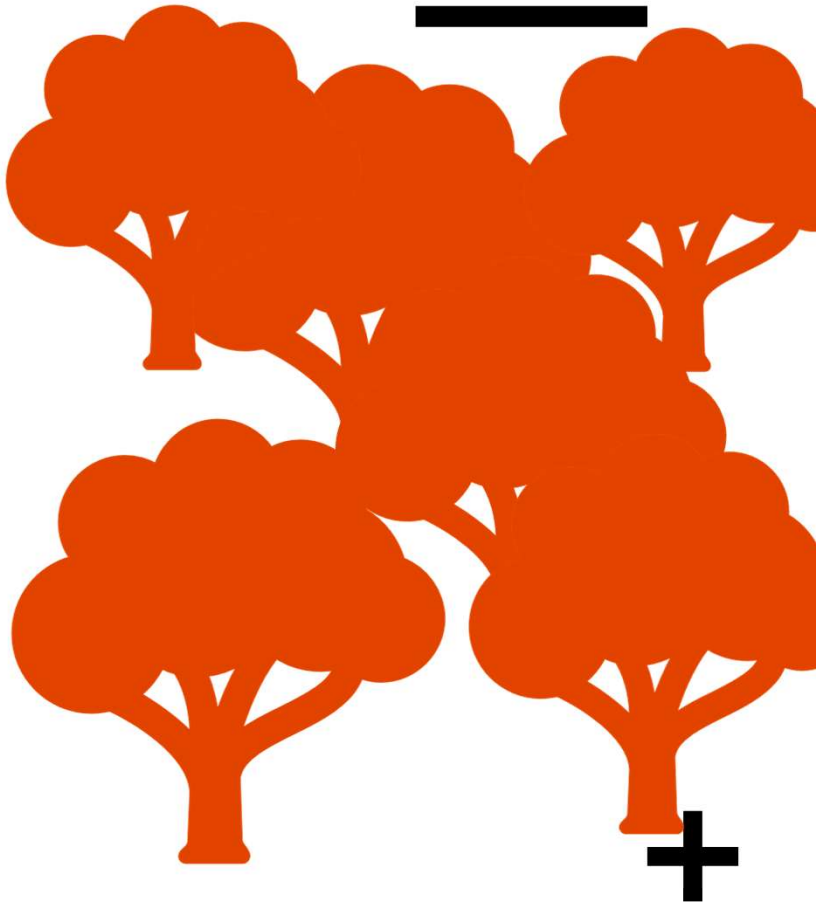
The nodes are where the data is split off and the decision leaves are the final prediction/outcomes

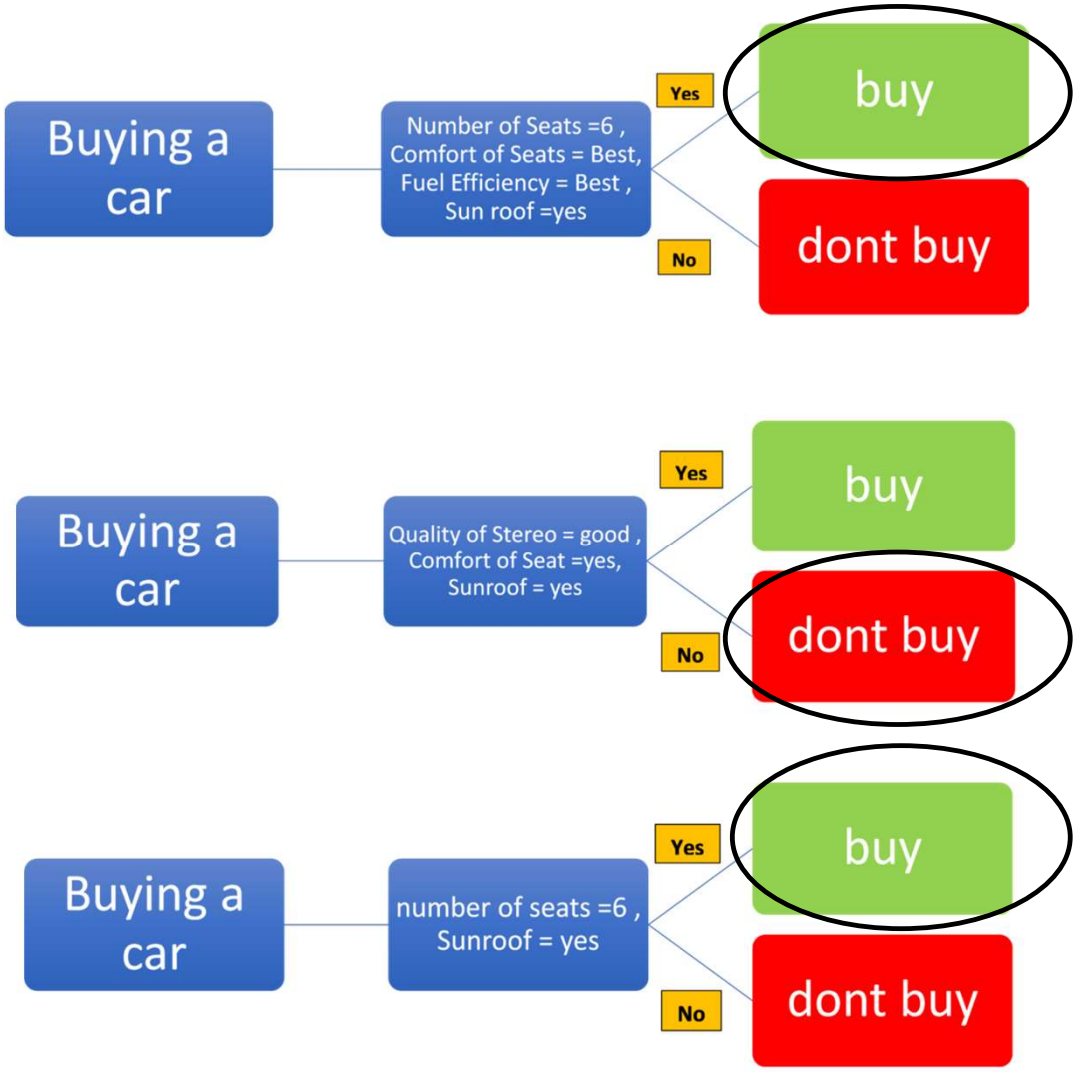

Is a Person Fit?
Age < 30 ?
Yes?
No?
Eat's a lot of pizzas?
Exercises in the morning?
Yes?
No?
Yes?
No?
Unfit!
Fit
Fit
Unfit!


Decision Node
Decision Node
Decision Node
Sub-Tree
Leaf Node
Decision Node
Leaf Node
Leaf Node
Leaf Node
Leaf Node

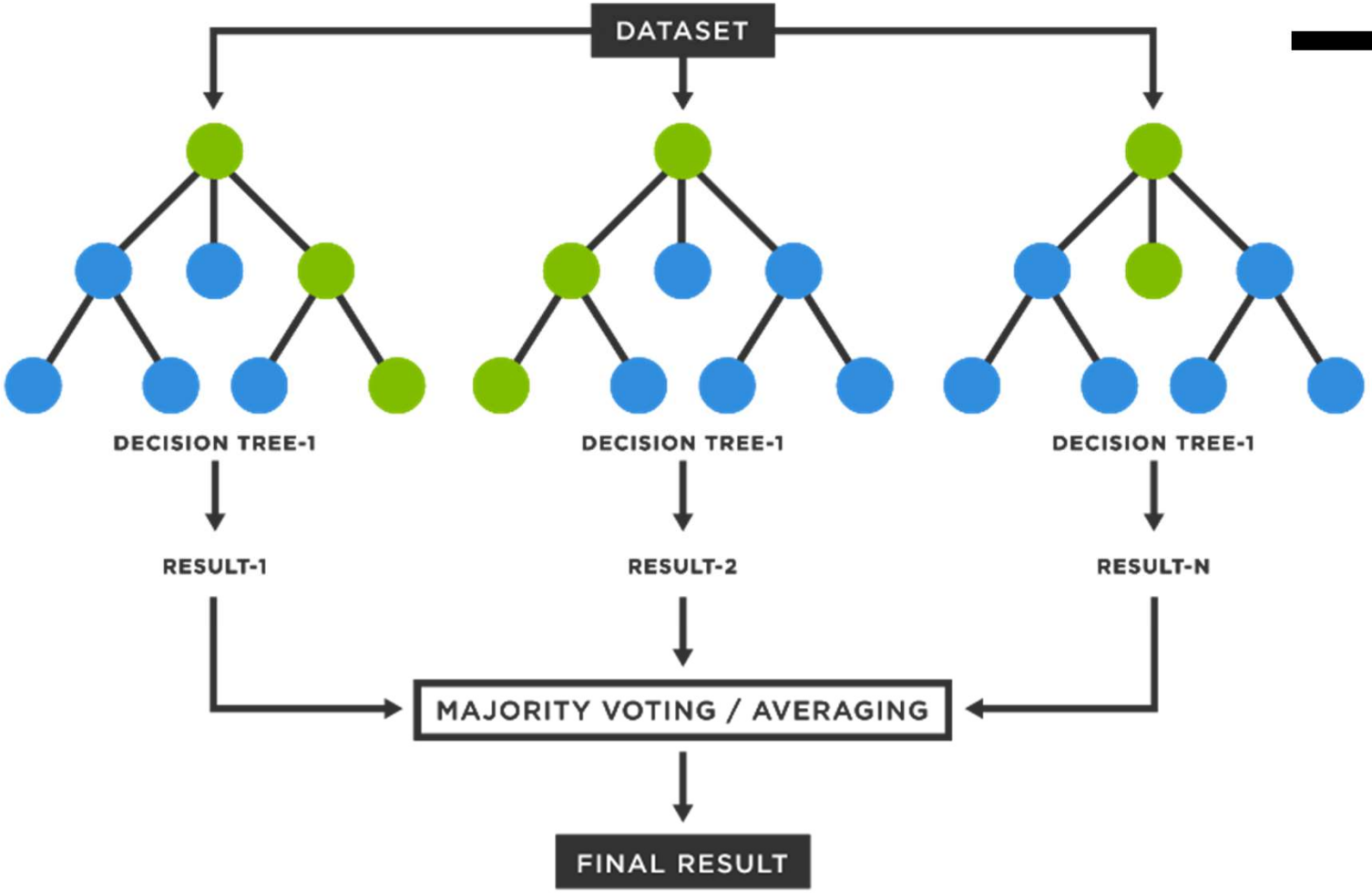
Random Forest and Decision Trees

- Random Forest's algorithm creates a multitude of decision trees
- Each decision tree will take a random set of rows from the data and will be checking a random set of inputs and map those inputs to an output/prediction
- If it is determining a regression, Random Forest will take the average from every decision tree to determine the prediction
- For classification, the prediction is selected by the mode of the decision tree outputs.





What would the output be?


DATASET
DECISION TREE-1
DECISION TREE-1
DECISION TREE-1
RESULT-1
RESULT-2
RESULT-N
MAJORITY VOTING / AVERAGING
FINAL RESULT

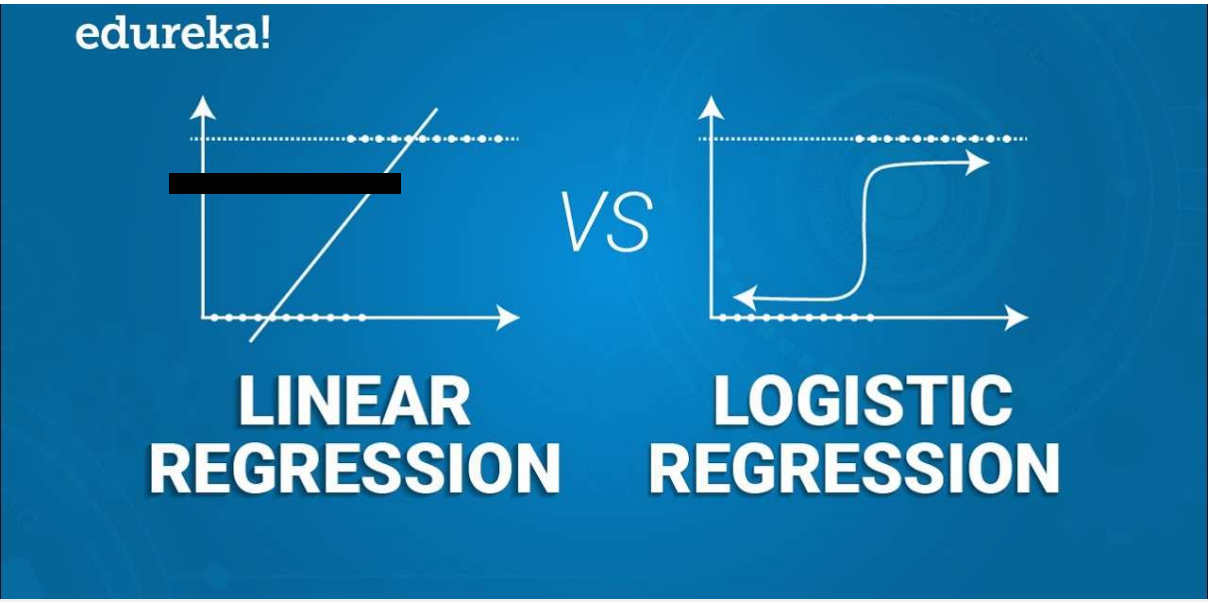
Pros & Cons

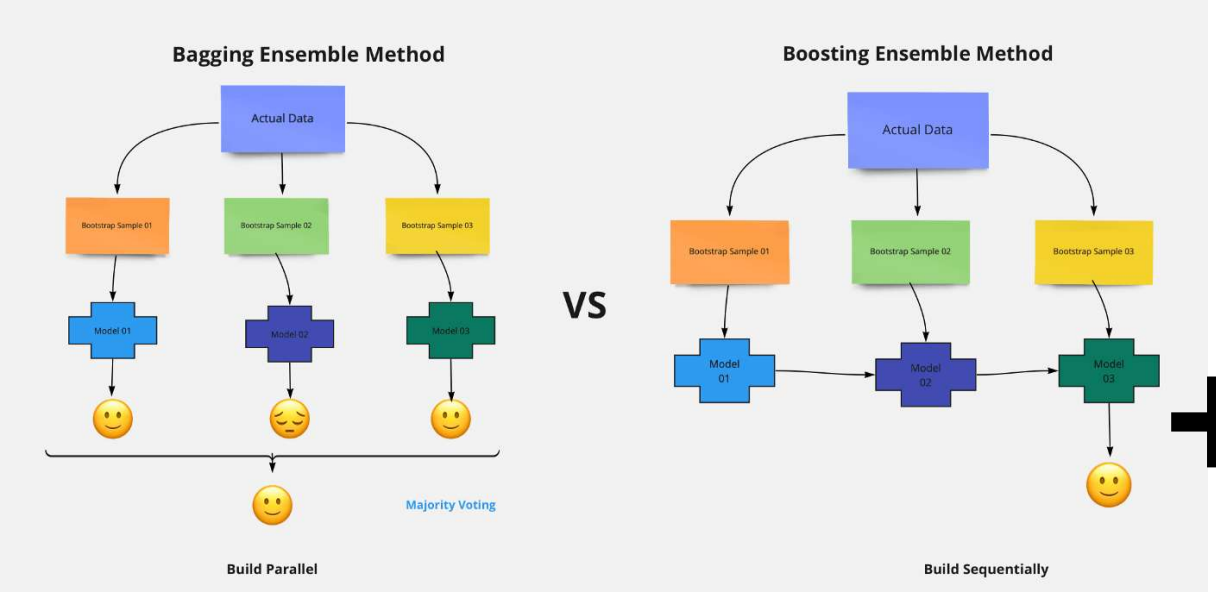
Advantages

- Versatile (regression & classification problems, categorical and continuous values)
- Robust against outliers
- Addresses overfitting, decreases variance
- Not necessary to remove irrelevant features, handles missing values
- Highly accurate

Disadvantages

- Due to number of trees created, requires high power and resources
- Training for and combining each tree adds time
- Tough to interpret, complex
- Additional work necessary to determine feature importance





Comparisons

Logistic Regression

- Similarity: classification problems, discrete target
- Difference: addresses overfitting, features' numerical value

Linear Regression

- Similarity: regression problems, continuous target
- Difference: sensitivity to outliers, affected by feature scale

XGBoost

- Similarity: ensemble decision tree methods
- Difference: gradient boosting, use of hyperparameters

Data Preprocessing

Generally less data transformation than basic ML models

– Outliers - insensitive to outliers
– Missing data – remove or impute preprocessing or impute on the fly
– Standardization and normalization – not necessary
– Categorical – get dummies or one hot encoder

Hyperparameters

- Random Forest has numerous hyperparameters
- The majority of parameters concern either sampling methods or the dimensions of the Random Forest

Arguably the most important Hyperparameters

– n_estimators (int)

– Bootstrap (Boolean)

– max_depth (int)

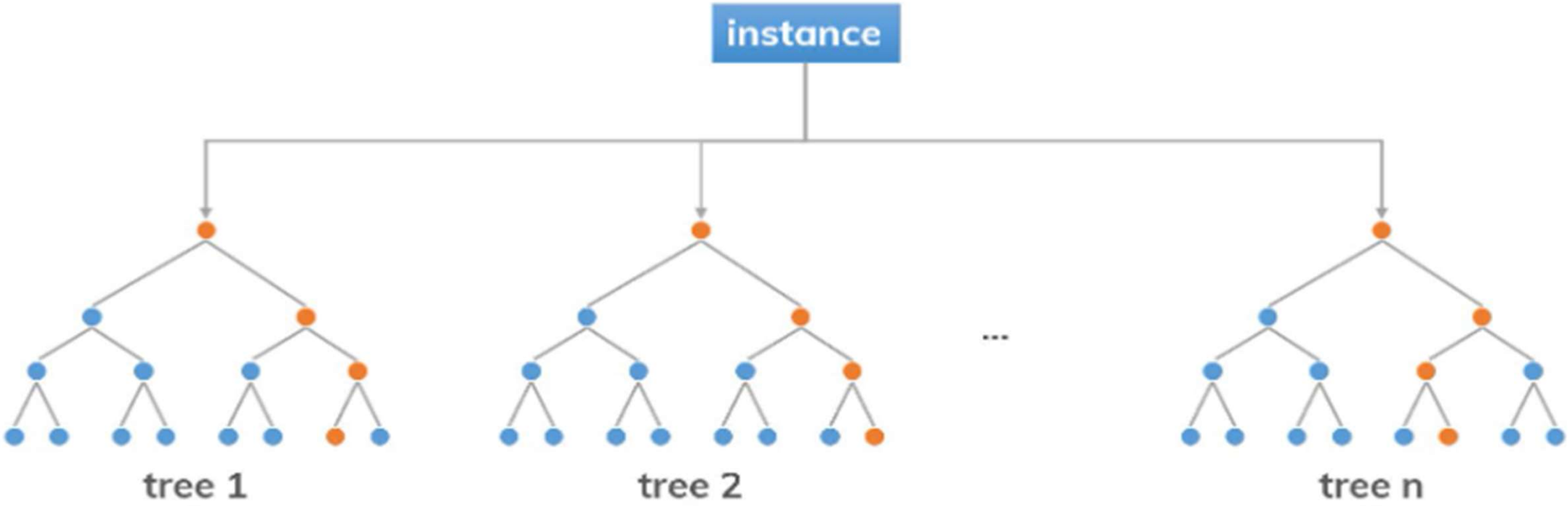
– min_samples_leaf (int)

– min_samples_split (int)

– max_features (int)

n_estimators (int)

–Controls the number of trees that are created

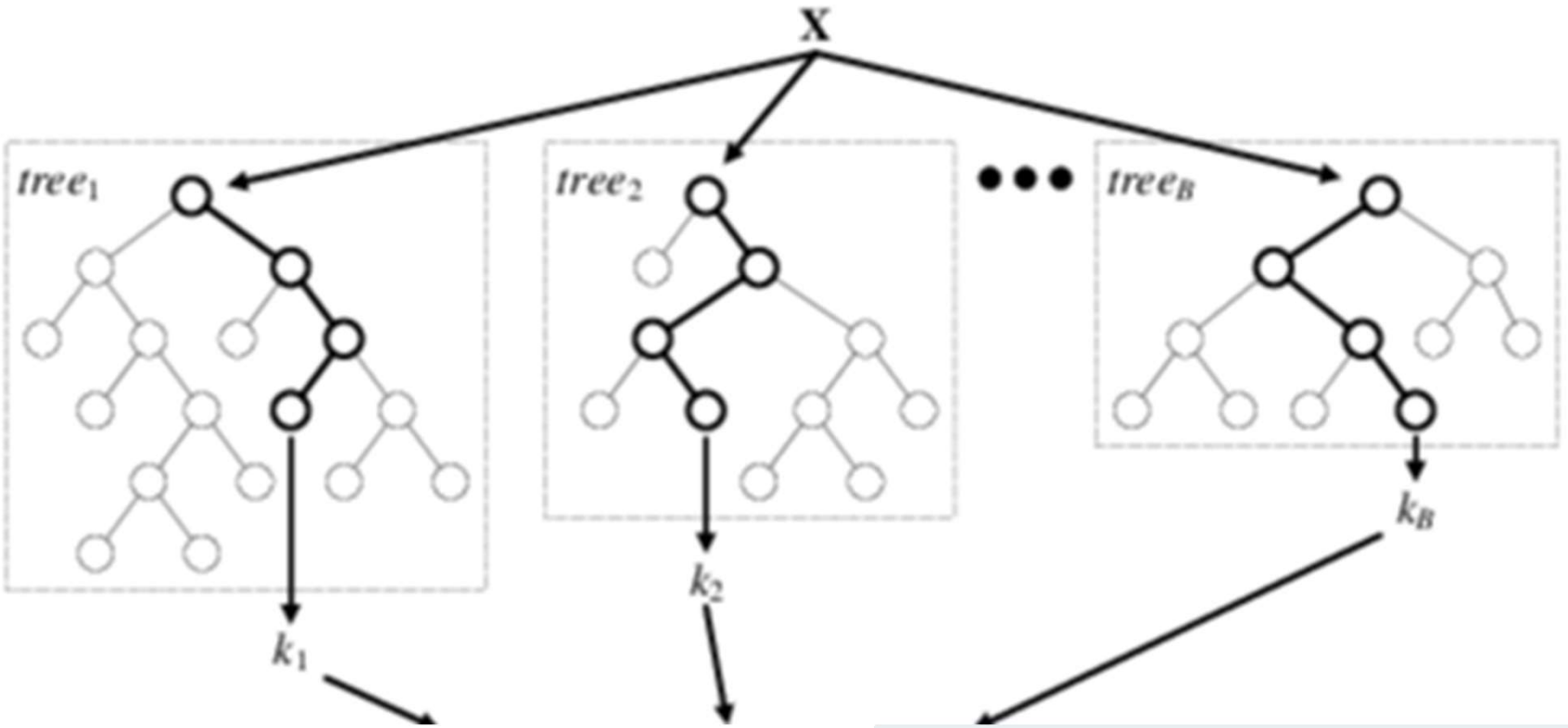


bootstrap

A boolean value that controls if the data sampled for the forest is replaced when sampled

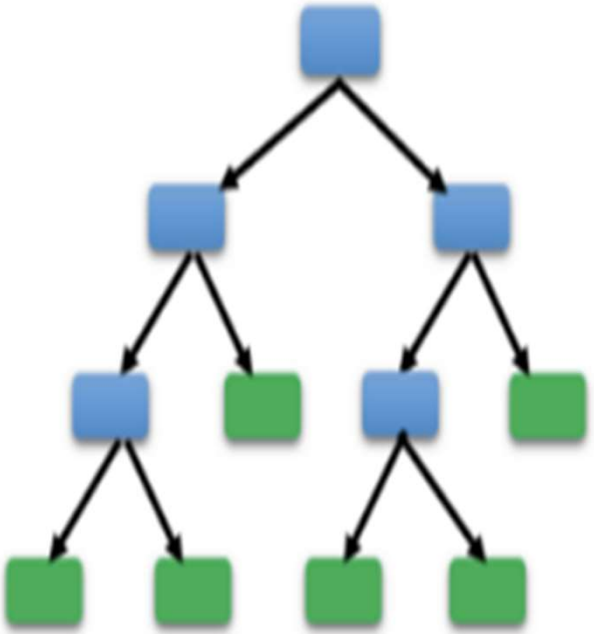
max_depth

– Controls the number of layers present in each tree.



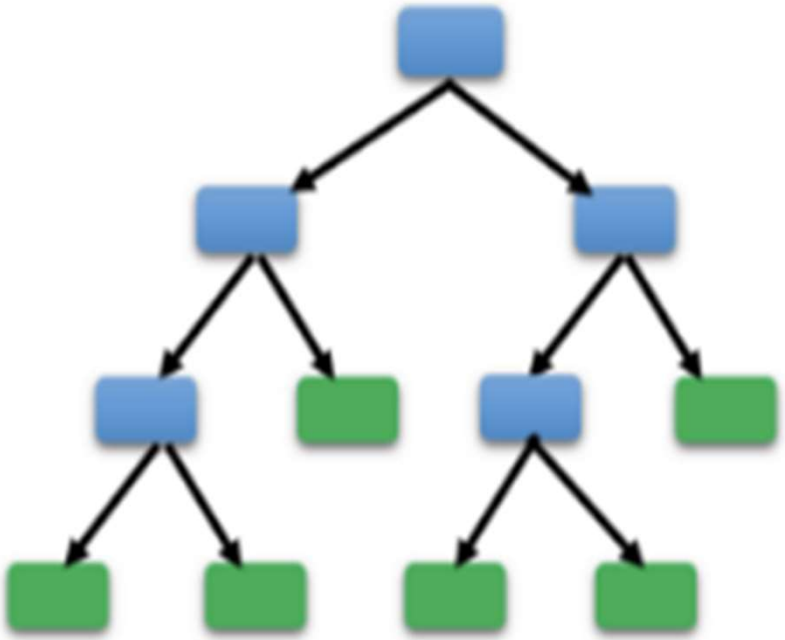
min_samples_leaf

–Controls the minimum number of sample points contained in the final nodes of a tree



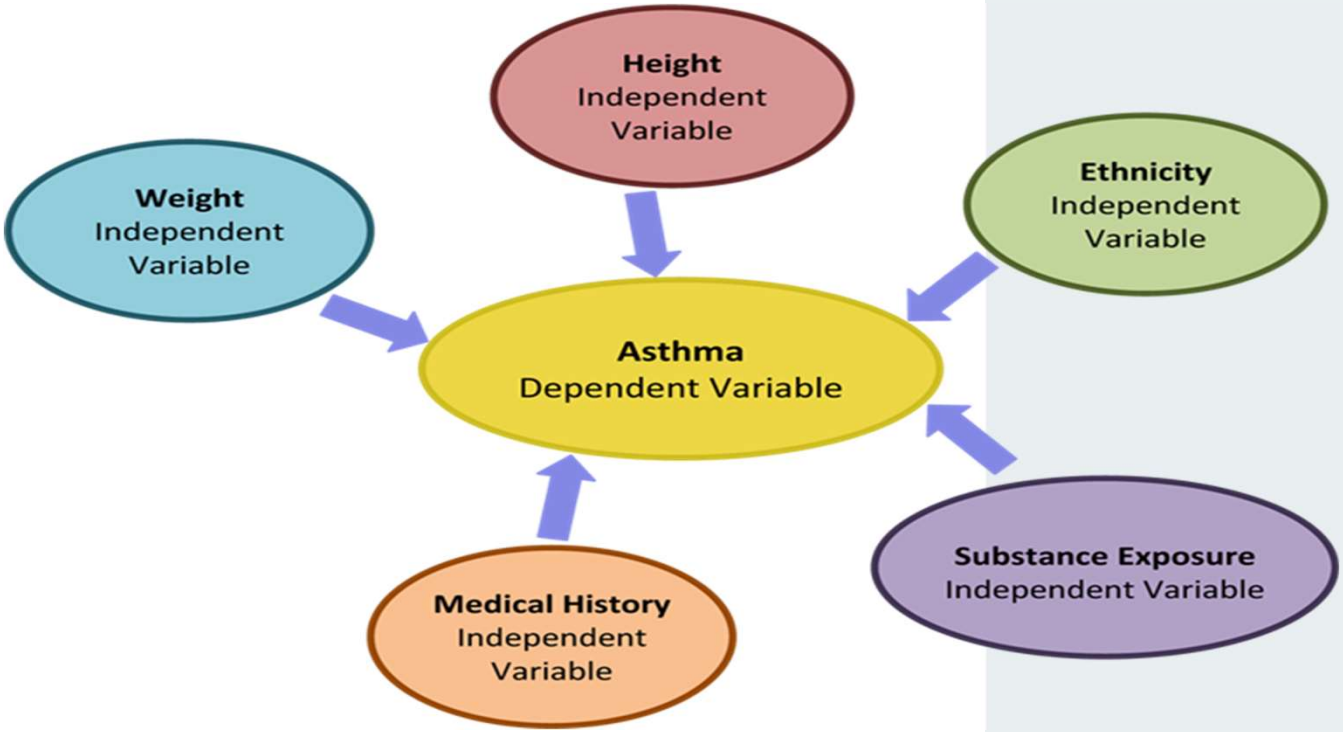
min_samples_split

– Controls the minimum number of sample points that are used before a split in the nodes.



Max_features

– The maximum number of columns/variables that are considered before a split in the nodes is made.



Appendix 1 - Documentation

– Random Forest Classifier: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

– *Random Forest Model (2018):* This documentation, while pertaining to the SAS language, contains useful conceptual explanations for Random Forest. Most notably, this document describes methods and concepts for the optimization of Random Forest models. https://documentation.sas.com/doc/en/fcmrcdc/15.1/fcmrug/n1tzufyosz12bhn12ls3imd63q94.htm?msclkid=125cdab1c72c11ec90e145055fbacb35

– Random Forest Regressor: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Appendix 2 - Videos

– *Random Forest Algorithm Clearly Explained (2021):* This YouTube video by Normalized Nerd gives a short overview on the difference between decision trees and random forest algorithms, outlines the intuition behind random forest, and walks through an example of how random forest functions: https://youtu.be/v6VJ2RO66Ag

– *Random Forests: Data Science Concepts (2020)*: This YouTube video by ritvikmath walks through a real-world example (predicting if a HS student will drop-out or not) using random forest as a prediction model. He shares a bit about decision trees and goes into depth on both bagging and random subspaces in the random forest algorithm as well as addresses advantages and disadvantages of random forest. He also highlights how to calculate feature importance: https://youtu.be/w-eWTxbRQcU

Appendix 3 – Research Articles

– *A random forest guided tour (2016)*: This article goes in-depth on the functionality, the math behind it, and the origins, updates, and use cases of random forest: https://link.springer.com/article/10.1007/s11749-016-0481-7

– *How many trees in a random forest? (2012):* This study focuses on the number of trees to optimize random forest models: https://www.researchgate.net/profile/Jose-Baranauskas/publication/230766603_How_Many_Trees_in_a_Random_Forest/links/0912f5040fb35357a1000000/How-Many-Trees-in-a-Random-Forest.pdf

Appendix 4-1 – Blog/Website Sources

– *Advantages and Disadvantages of Random Forest Algorithm in Machine Learning (2019):* Blog post covering bullet point advantages and disadvantages of using Random Forest. http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-random.html

– *Best Practices with Data Wrangling before running Random Forest Predictions:* This stack exchange presents the cases for dealing with data processing in random forest: https://stats.stackexchange.com/questions/172842/best-practices-with-data-wrangling-before-running-random-forest-predictions

– *Example of a Random Forest in Python* (2020): In this blog post the author demonstrates Random Forest using a fictional dataset, in Python. https://datatofish.com/random-forest-python/?msclkid=6697e1bac75211ec91f7cf0d18743ad2

– *Preprocessing: OneHotEncoder() vs pandas.get_dummies:* This article goes over the differences between one hot encoder and get dummies for transforming categorical data. https://albertum.medium.com/preprocessing-onehotencoder-vs-pandas-get-dummies-3de1f3d77dcc

Appendix 4-2 – Blog/Website Sources

– *Random Forest Algorithm - Random Forest Explained | Random Forest in Machine Learning | Simplilearn:* This explanation of the random forest includes a code along to and notebook with dataset. https://www.youtube.com/watch?v=eM4uJ6XGnSM Code along - https://drive.google.com/drive/folders/1MQ5Nnhj3gs6TcIl0fcKP_AfOh8EMDB42

– *Random Forest Algorithm: A Complete Guide (2022)*: This blog goes in-depth on how random forest works, the real-life use cases and analogies related to the algorithm, feature importance, difference between decision trees and random forests, important hyperparameters, and advantages and disadvantages: https://builtin.com/data-science/random-forest-algorithm

– *Random forest Algorithm in Machine learning: An Overview (2020):* All-encompassing blog post giving an overview of the method, examples, and application in Python and R. https://www.mygreatlearning.com/blog/random-forest-algorithm/#ApplyingRandomForestwithPythonandR

– *Tuning the parameters of your Random Forest model (2015)*: This blog details random forest models and the hyperparameters available. https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/

– *How data normalization affects your Random Forest algorithm:* This article discusses the potential of normalization within Random Forest. https://towardsdatascience.com/how-data-normalization-affects-your-random-forest-algorithm-fbc6753b4ddf