Random Forest: Classification and Regression Tree (CART) and Improvement via Randomization

Joonjae Ryu

Maravelias Group

Dec. 13, 2019

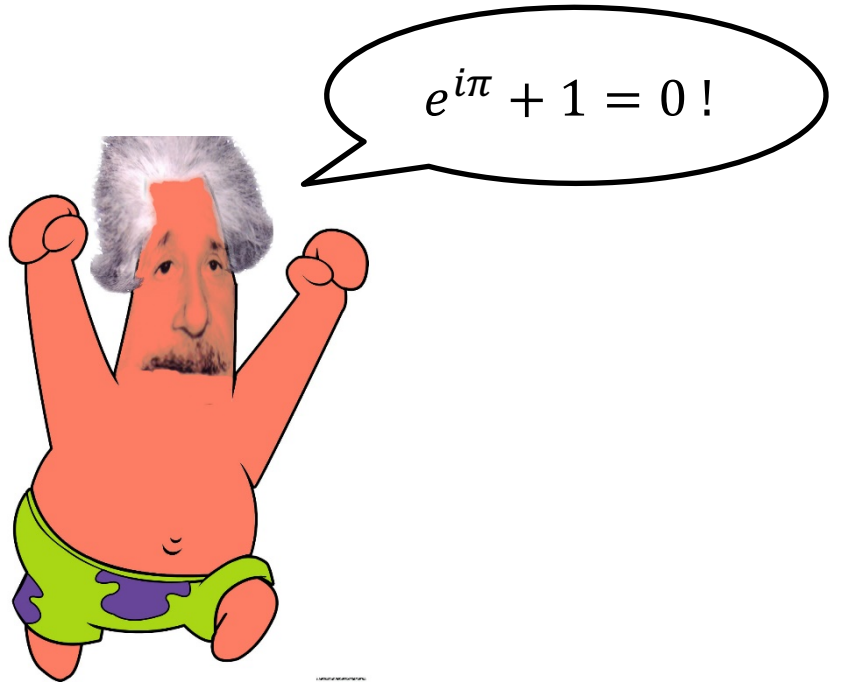
Chemical and Biological Engineering
University of Wisconsin – Madison

Machine Learning?

Book, note, etc.

How to learn

Learning



Dumb Patrick star

Einstein Patrick star!

Machine Learning?

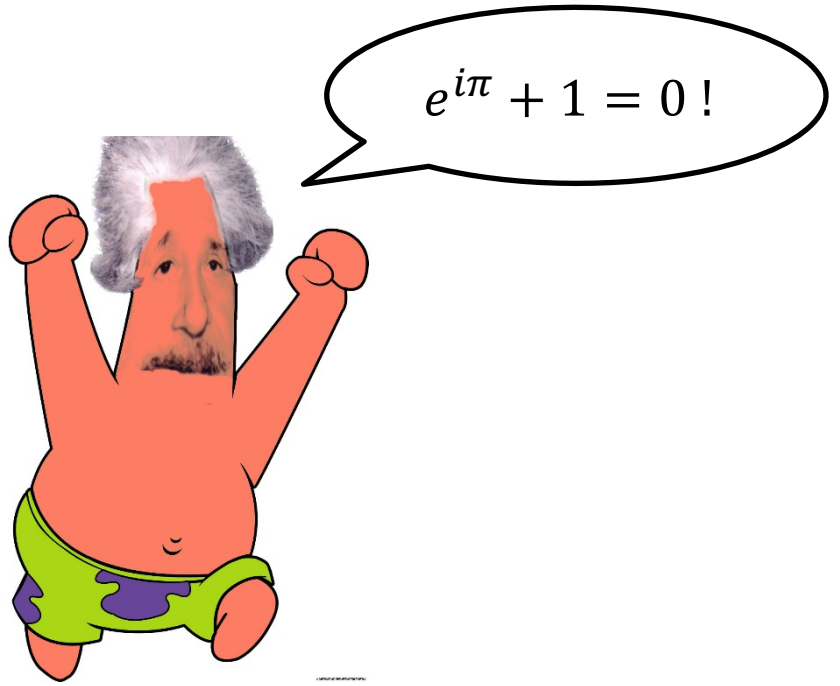
What is machine learning?

Teach machines how to learn!

Data ⇨

Algorithm ⇨

Training



Machine Learning?

Machine learning for prediction

Teach machines to predict answer with data

An instance

1+1=2

An example of answer

Training

Data

Data

Me

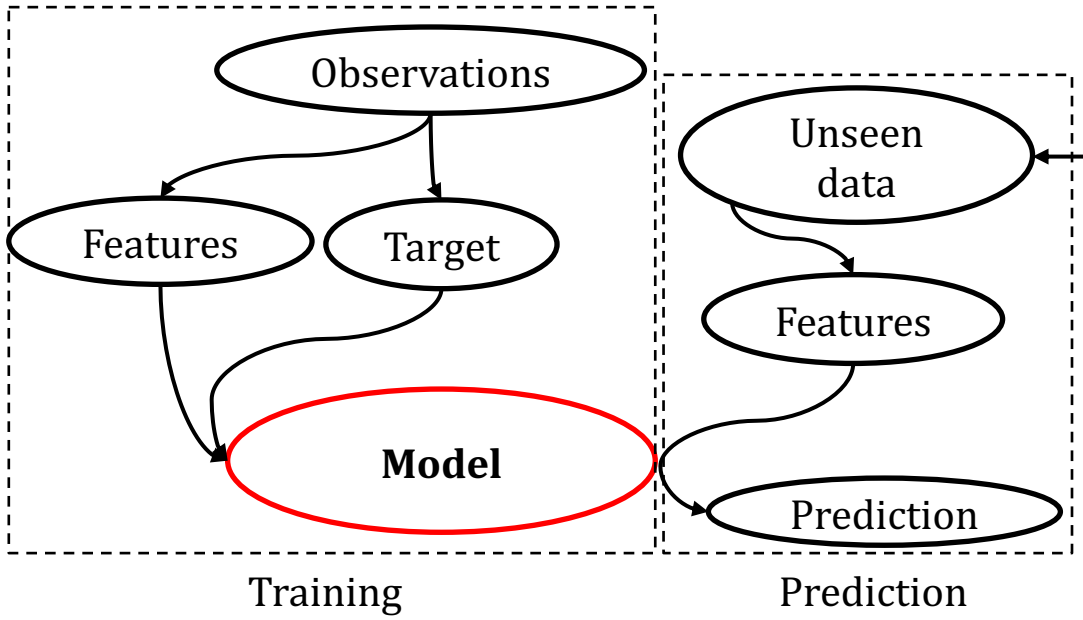
1+2=?

Prediction

3 !

Terminology

Prediction Problem



Target: Wine price (continuous)
Features: Age, Winter rain, Harvest rain, Seasonal temperature
Model: Regression

Target: Loan acceptance (discrete – yes or no)
Features: Gender, Married, Education, Income, Loan amount, etc.
Model: Classification

Observation (n)

- Data sample we collect (dataset)
  - Training set (Data for training of the model)
  - Test set (Data to test the model)

Target (y)

- The outcome we want to predict
  - Continuous (temperature, price, etc.)
  - Discrete (e.g., Yes or No)

Features (X)

- The information we use to predict the target
  - Collected raw data
  - Information made from raw data (e.g., $x_1 x_2$)

Model

- Regression: to predict continuous target
- Classification: to predict discrete target

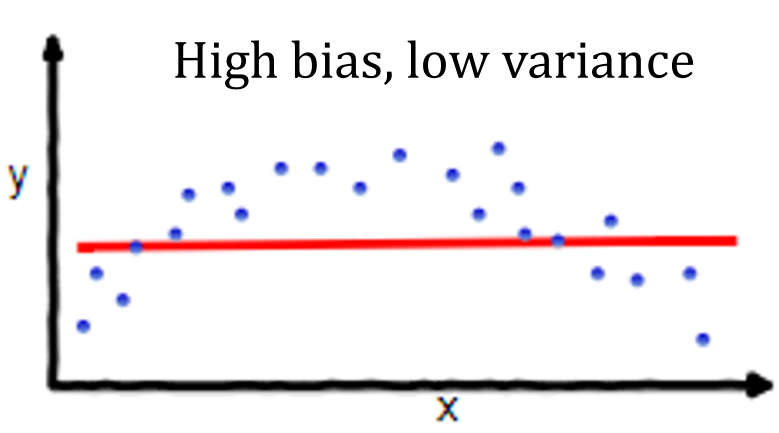
Errors in Prediction

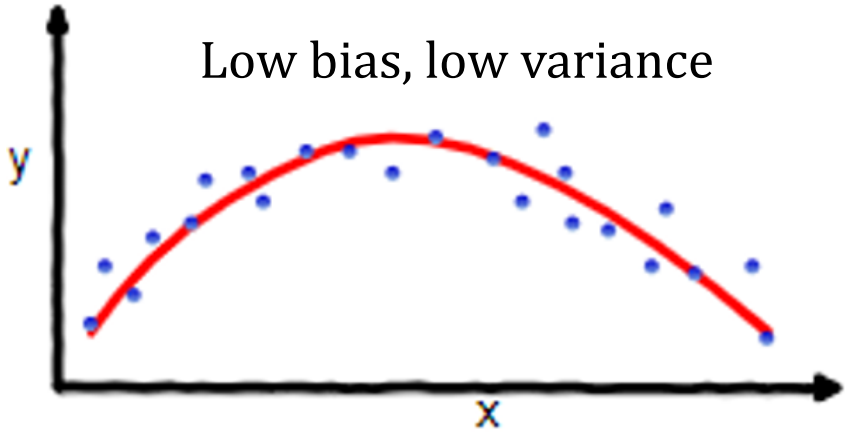
Bias

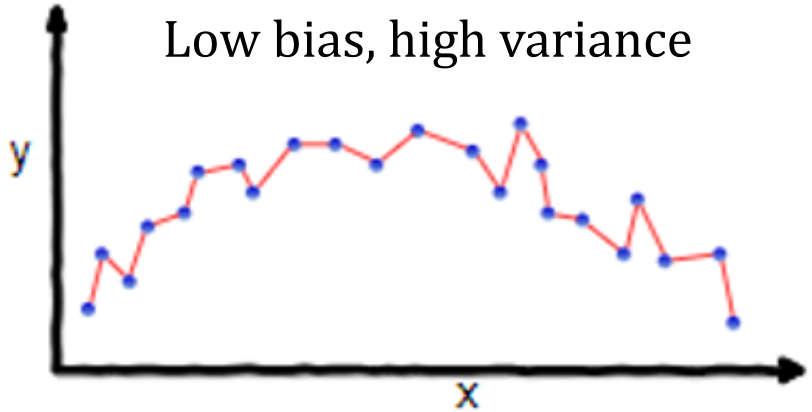
- Error from **simplifying** assumptions made by the model
- Underfitting

Variance

- Variation in prediction with **different training dataset**
- Overfitting
- Can be tested with unseen data







underfitting

Good balance

overfitting

Figure from https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

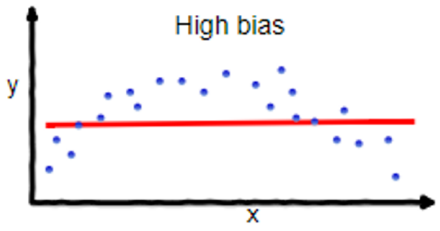
Bias vs Variance

Bias

- Error from **simplifying** assumptions made by the model
- Underfitting

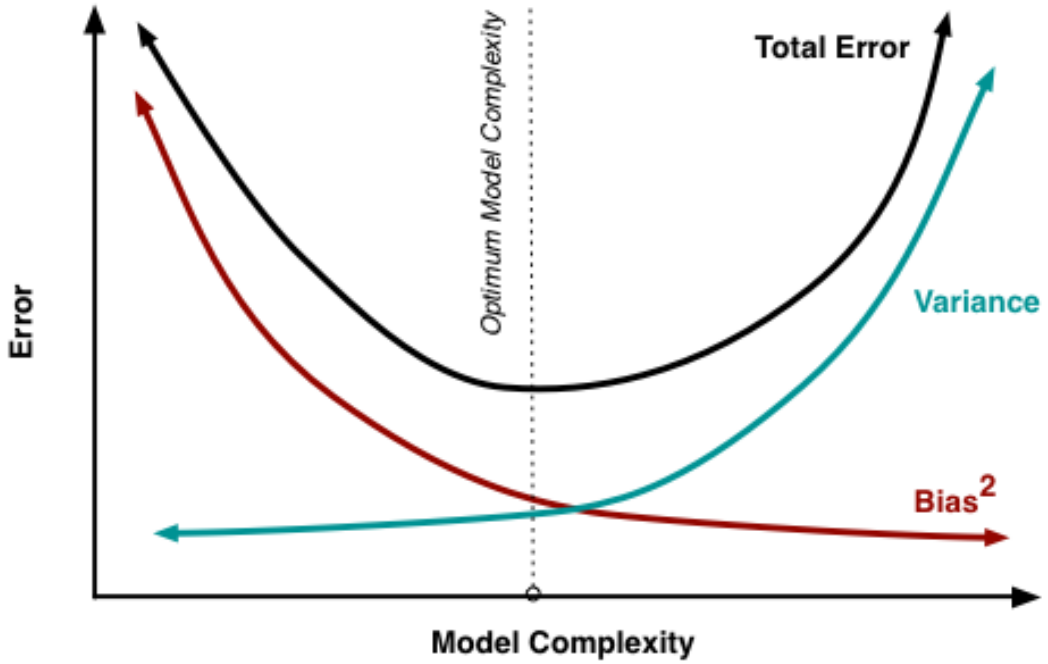
Variance

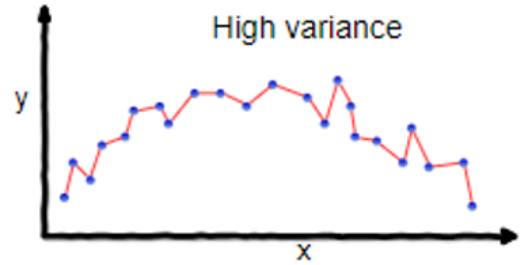
- Variation in prediction with **different training dataset**
- Overfitting
- Can be tested with unseen data



underfitting

Underfitting Overfitting





overfitting

It is very important to find this optimum model complexity!

Figure from http://scott.fortmann-roe.com/docs/BiasVariance.html

Classification and Regression Tree (CART)

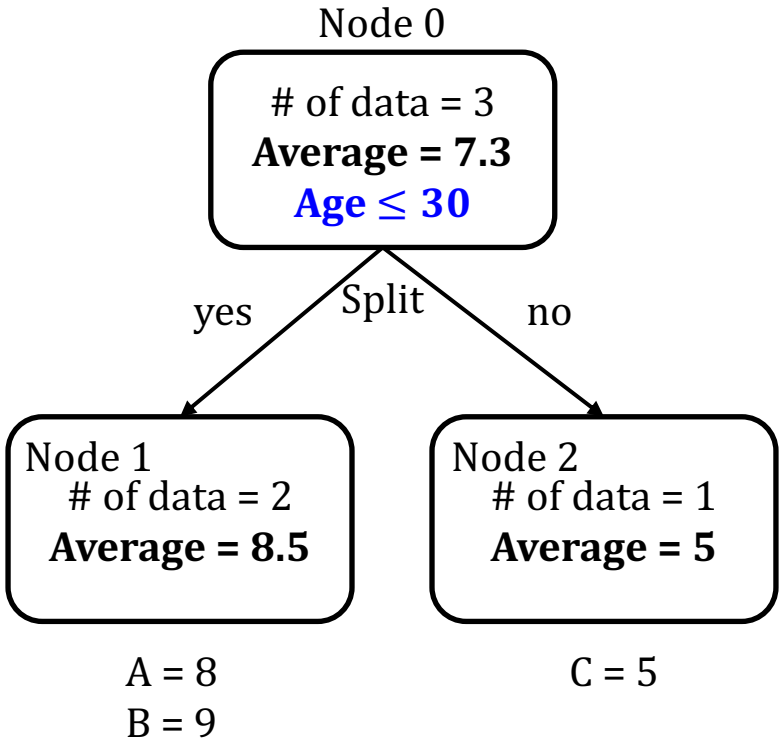
CART

- Tree-based learning algorithm
- Pros
  - Versatile (Discrete & continuous variables)
  - High-order interaction can be captured (More than linear)
  - Nice interpretability

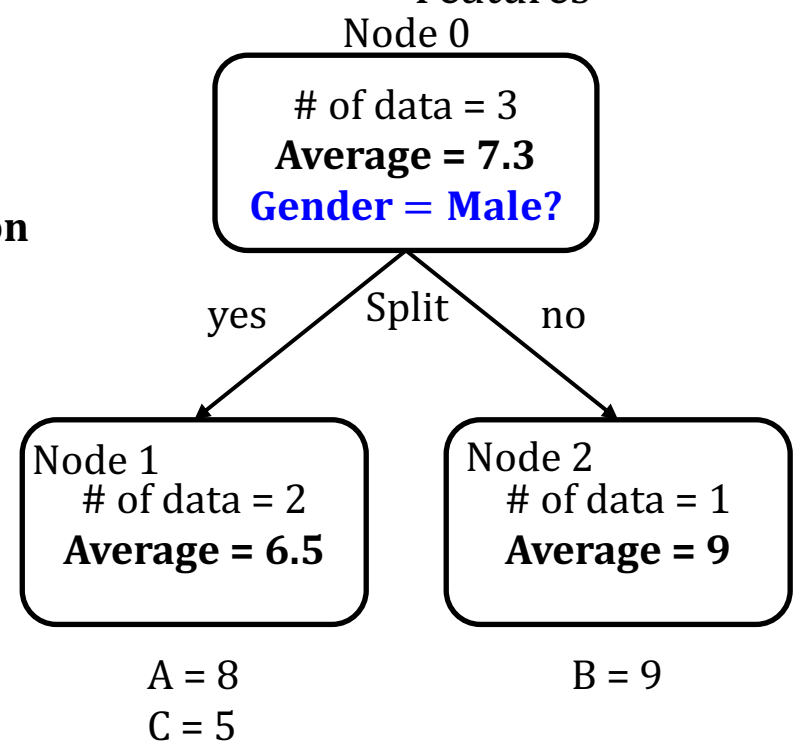
Data	Age	Gender	Exam Grade
A	18	Male	8
B	25	Female	9
C	40	Male	5

Features

Target



Prediction
Split



Which split is better?

Classification and Regression Tree (CART)

Regression criteria

Mean Squared Error (MSE)= $\frac{1}{N}\sum_i (y_i - \bar{y}_i)^2$

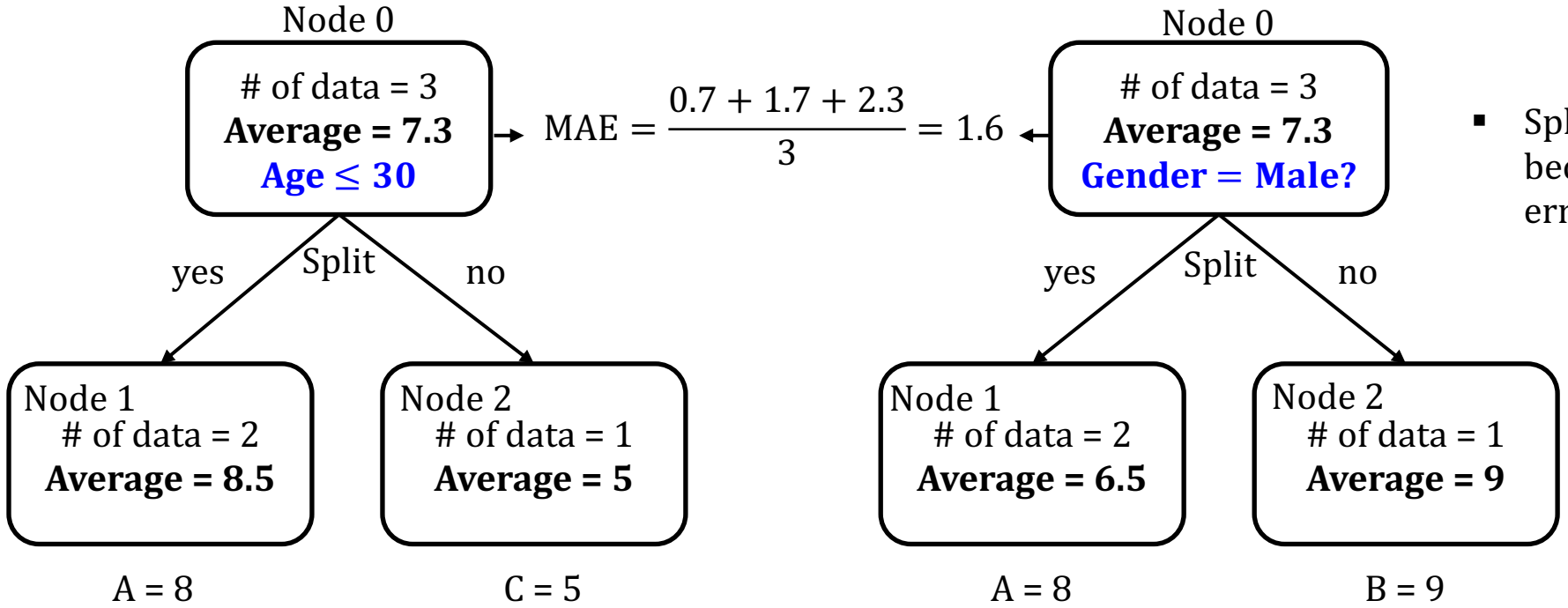
Mean Absolute Error (MAE) = $\frac{1}{N}\sum_i |y_i - \bar{y}_i|$

Classification criteria

Gini impurity = $\sum_k p_k(1 - p_k)$

Entropy = $\sum_k p_k \log(p_k)$

Data	Age	Gender	Exam Grade
A	18	Male	8
B	25	Female	9
C	40	Male	5



$\text{MAE} = \frac{1}{2}(|8.5 - 8| + |8.5 - 9|) = 0.5$ $\text{MAE} = 0.0$

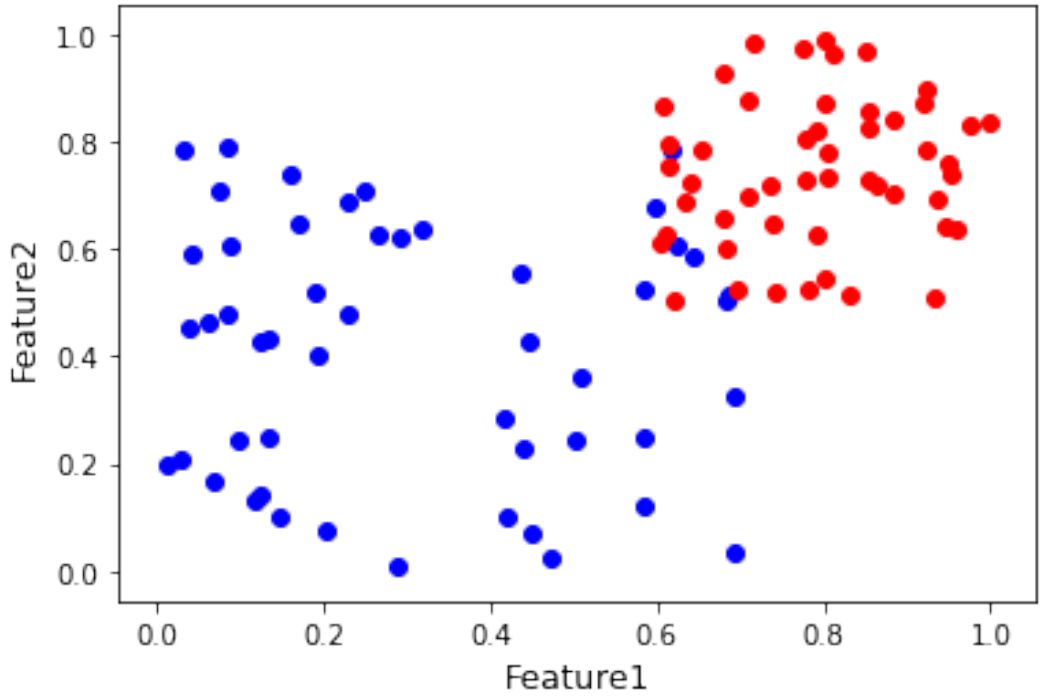
$\text{MAE} = \frac{1}{2}(|6.5 - 8| + |6.5 - 9|) = 1.5$ $\text{MAE} = 0.0$

- Split with **Age** is better because it can reduce more error (MAE) in prediction

Classification and Regression Tree (CART)

CART

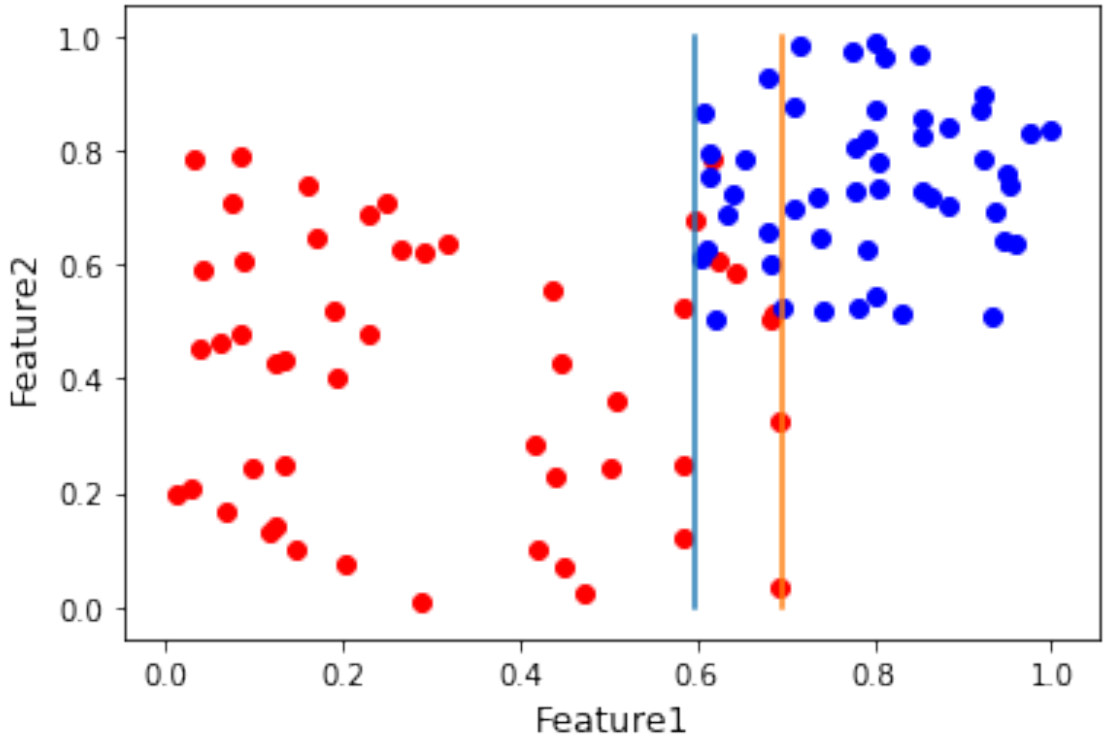
- Cons
  - Easy to overfit
- 100 data points for classification problem
  - 80 training data/20 test data
- Two features are used to classify into red and blue



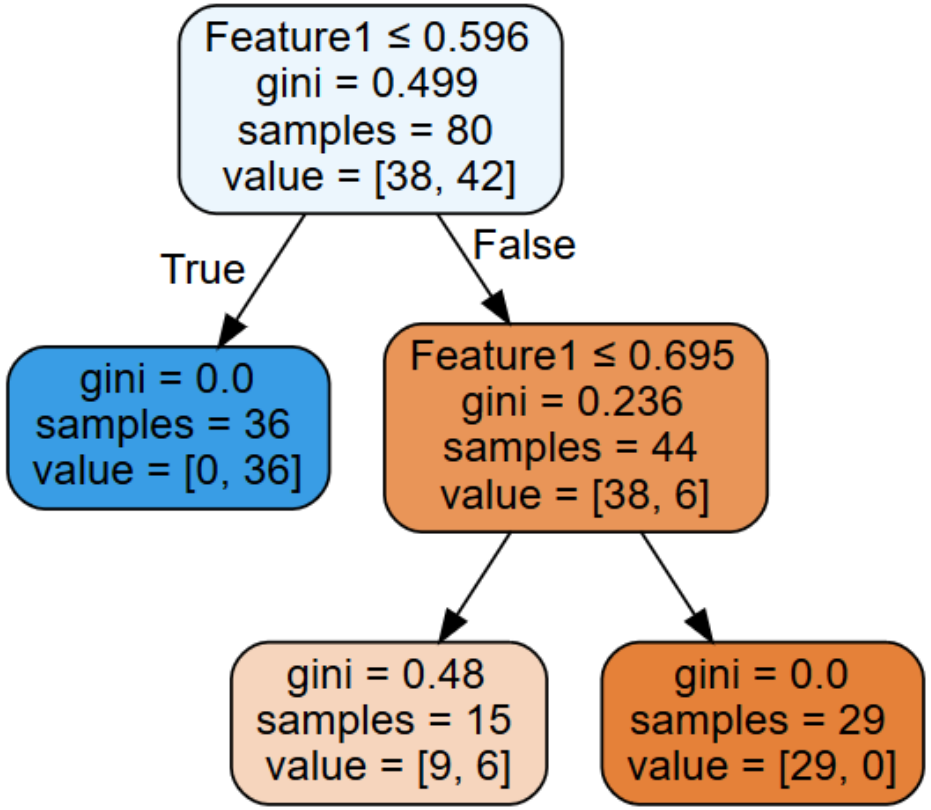
Classification and Regression Tree (CART)

CART

- Decent fitting
- Difference in training set vs test set is low (0.25)



Maximum depth = 3

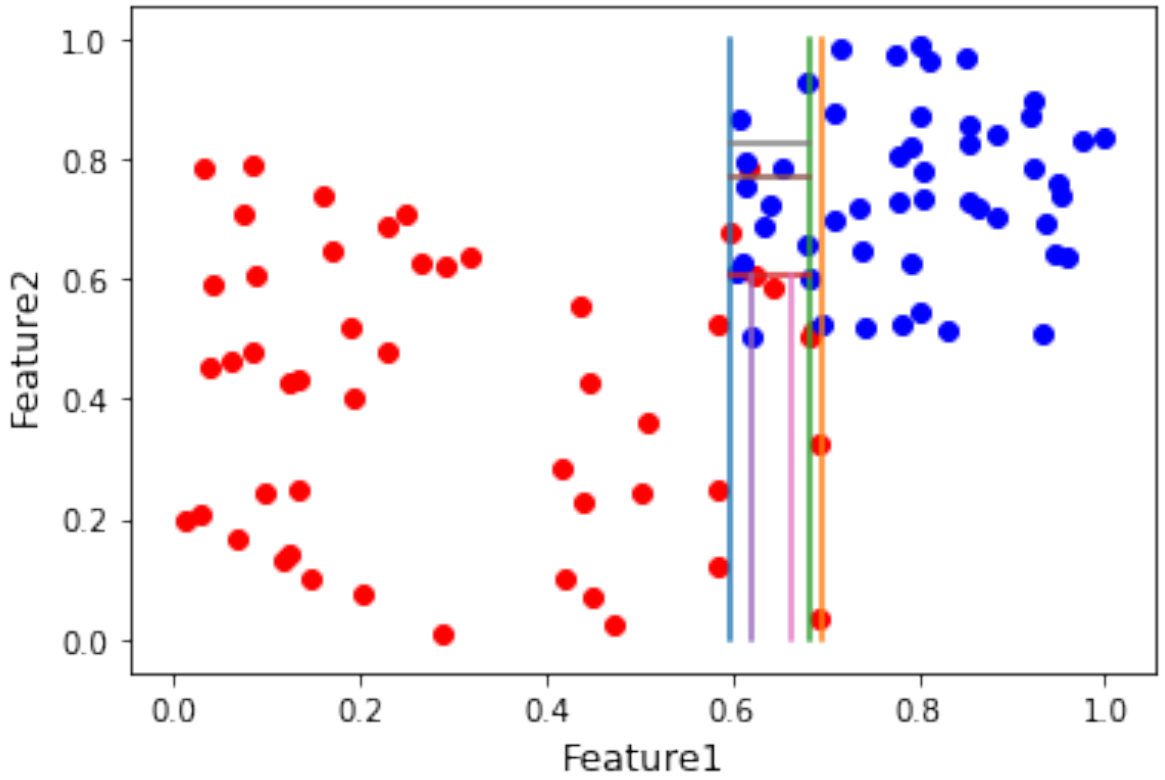


Training accuracy: 0.925
Test accuracy: 0.90

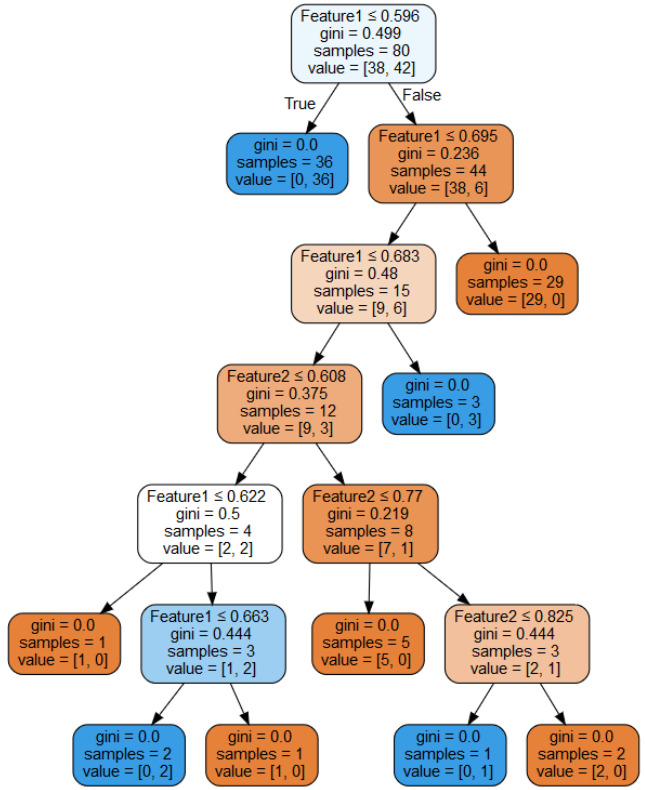
Classification and Regression Tree (CART)

CART

- Overfitting case
- CART can be trained until 100% accuracy for given data (Low bias)
- Less prediction accuracy for test data (High variance!)



Maximum depth = 6

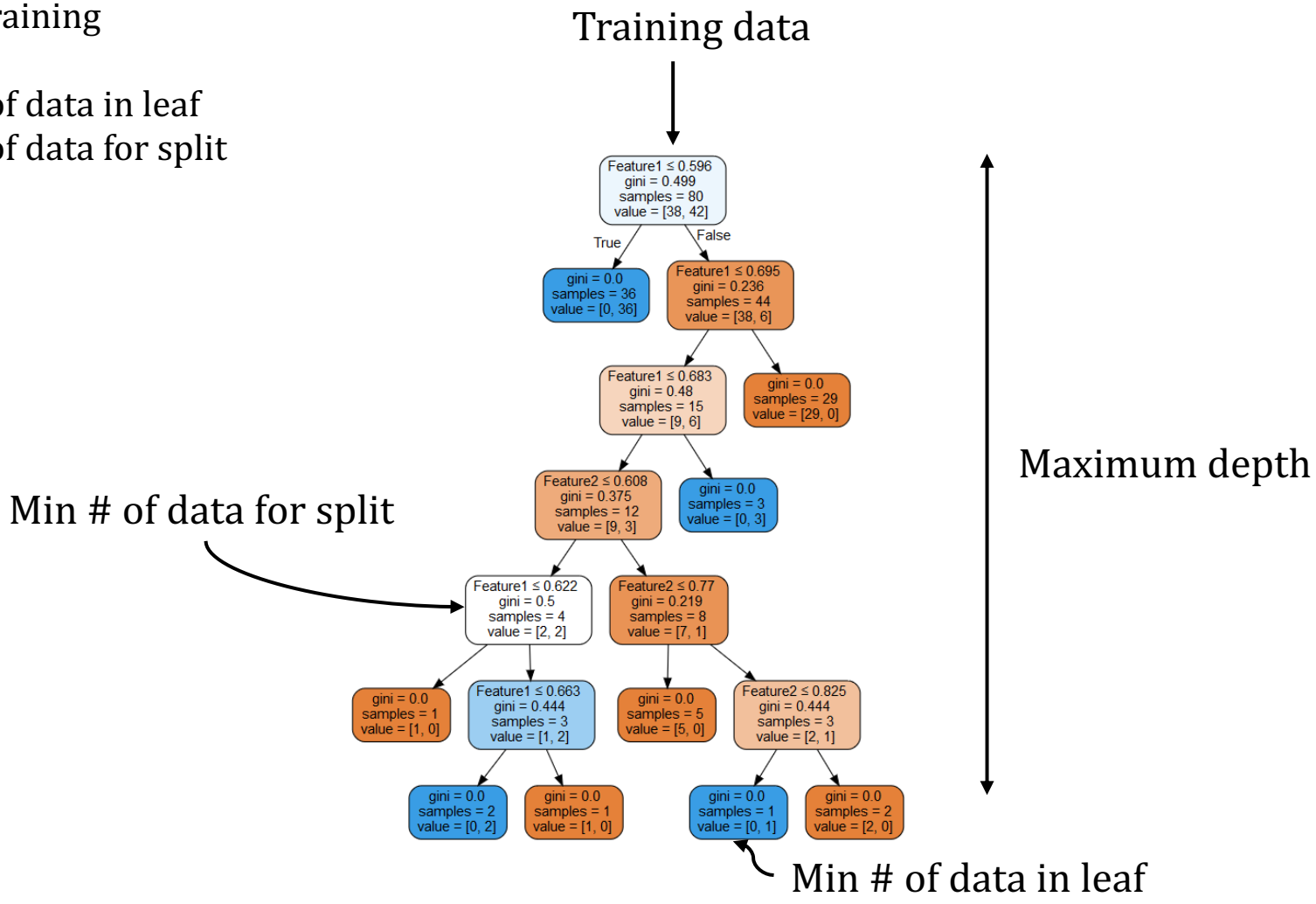


Training accuracy: 1.0
Test accuracy: 0.85

Classification and Regression Tree (CART)

CART

- Parameter tuning is critical
  - Data selection for training
  - Maximum depth
  - Minimum number of data in leaf
  - Minimum number of data for split



Characteristics of CART

CART

- Nice prediction for training data
  - Can learn very nicely for the given data **(low bias)**
- But…
  - Easy to overfit (very sensitive to training data) **(high variance)**
  - Maybe not very powerful for prediction with unseen data

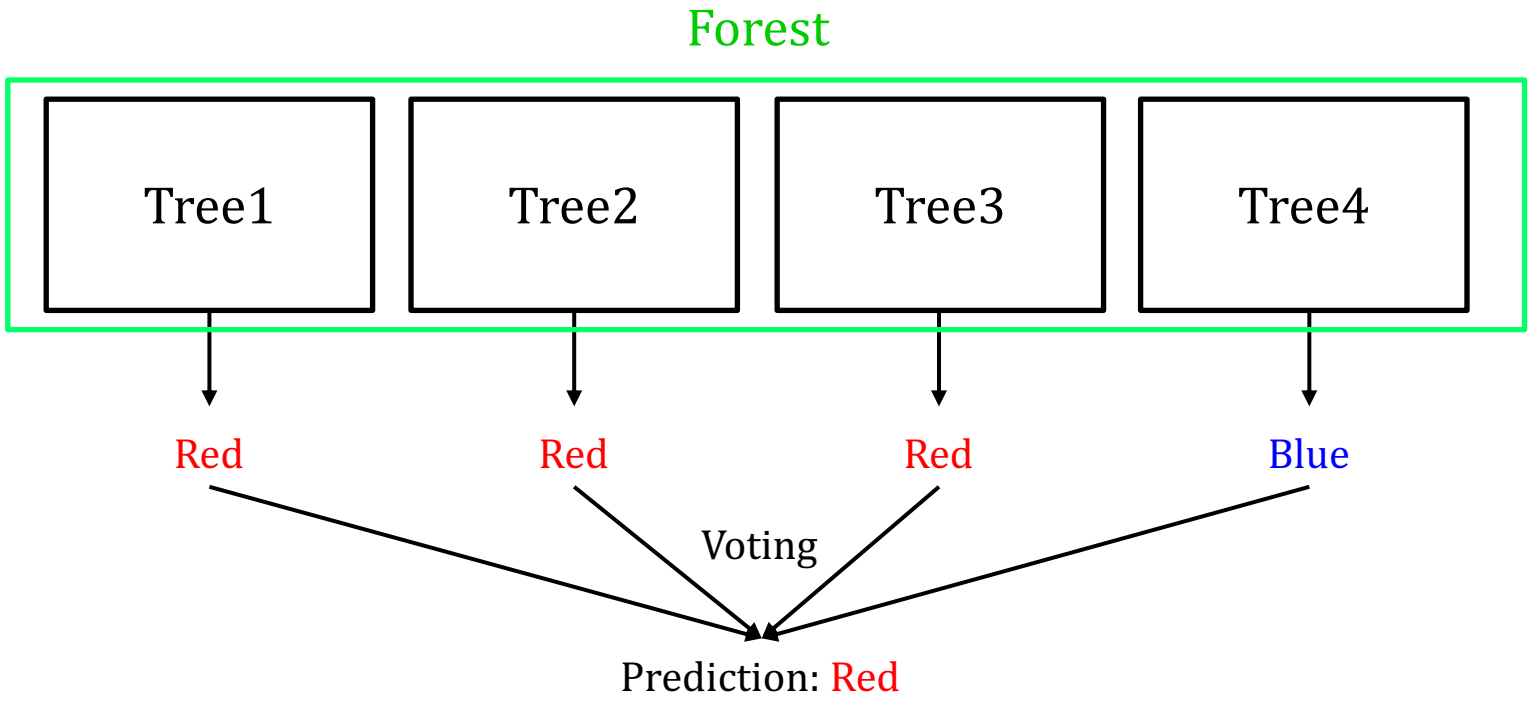
Can we **reduce variance** while preserving **low bias**?

Yes!

Random Forest (RF)

Random Forest

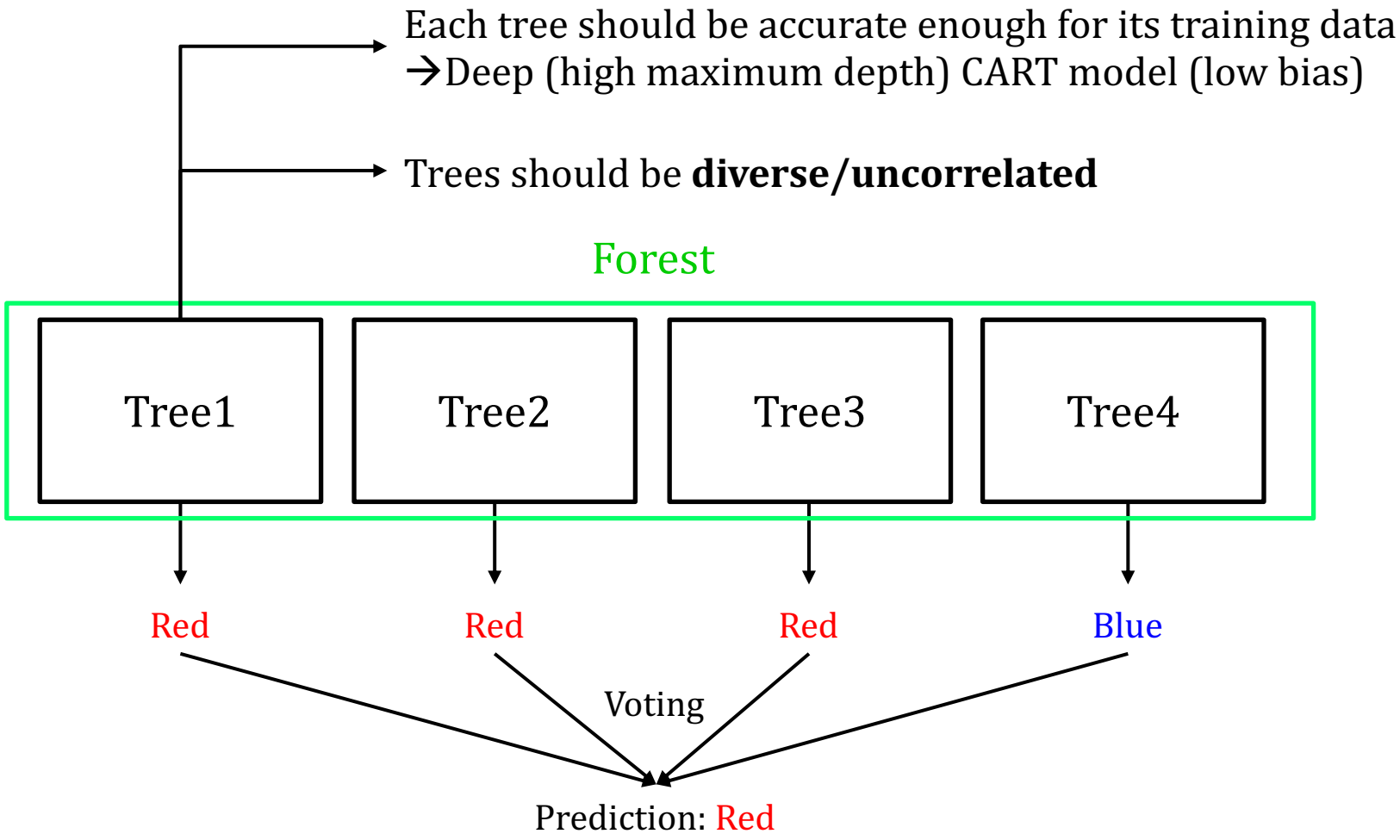
- It is a collection of decision trees to predict the target
- Final prediction is determined by “voting” from all the trees



Random Forest (RF)

Random Forest

- How to generate the trees?



Random Forest (RF)

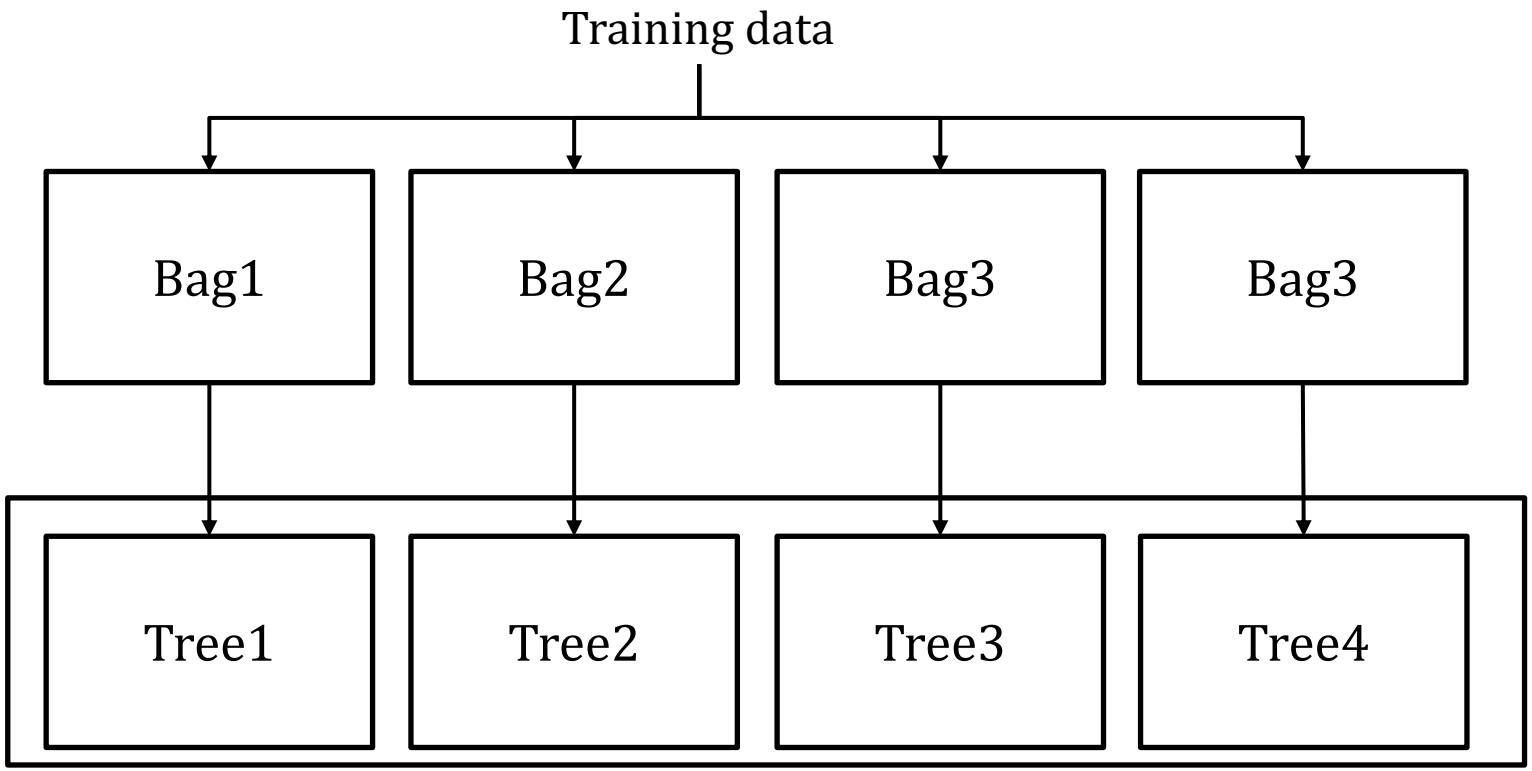
Random Forest

- How to generate diverse trees?
  - **Different training data** for each tree!
    - Bootstrap aggregation (Bagging)
      - Data with **N** observations
      - Randomly select **N** data **with replacement** for each bag
      - ~37% of the observations are duplicated

Training data: (A,B,C)

→ bag1: (A,A,B)
→ bag2: (A,B,B)
→ bag3: (B,B,C)
→ bag4: (A,B,C)

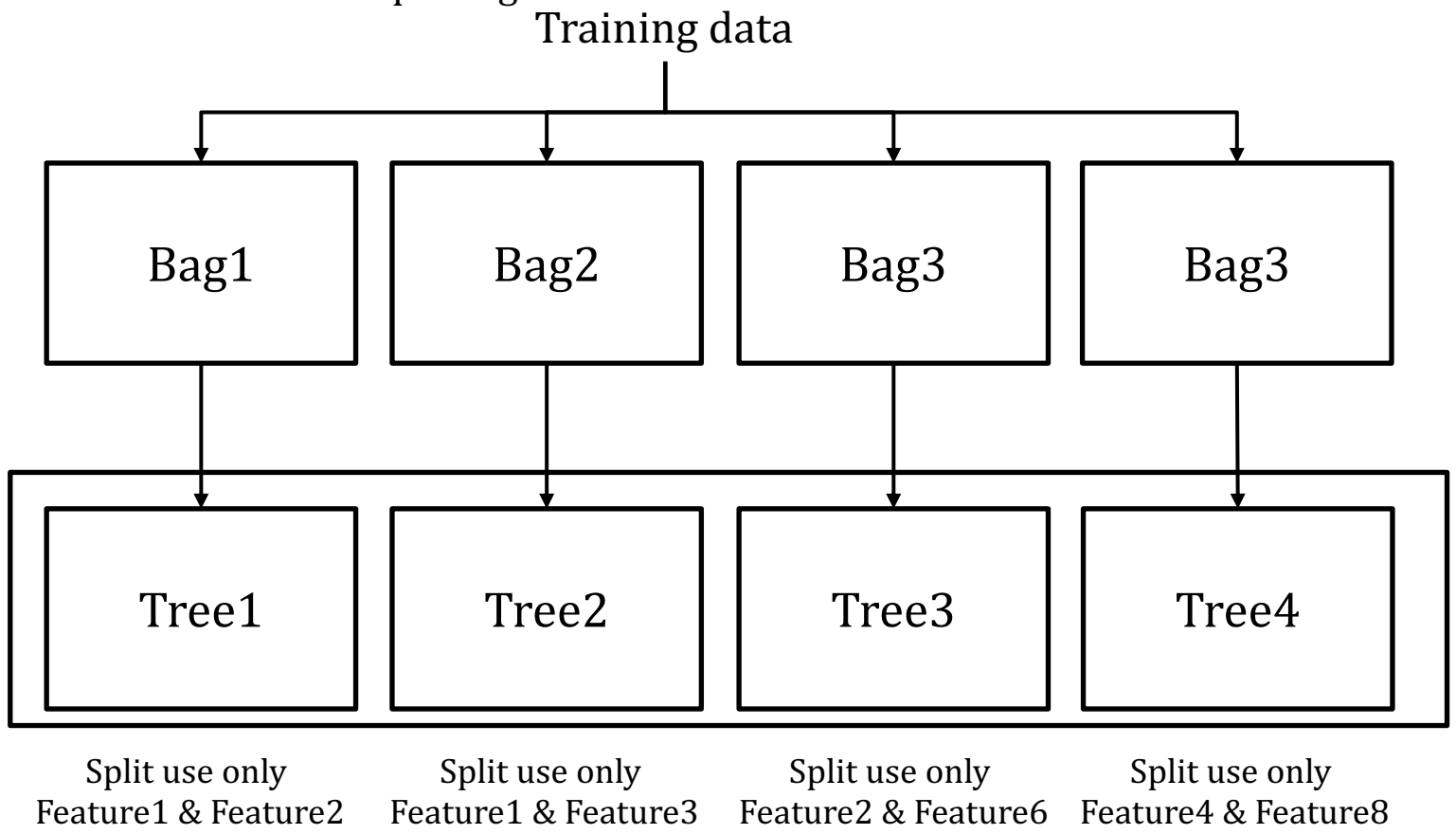
Bagging



Random Forest (RF)

Random Forest

- How to generate diverse trees?
  - **Different features** to split for each tree!
    - Use only a subset of features for splitting



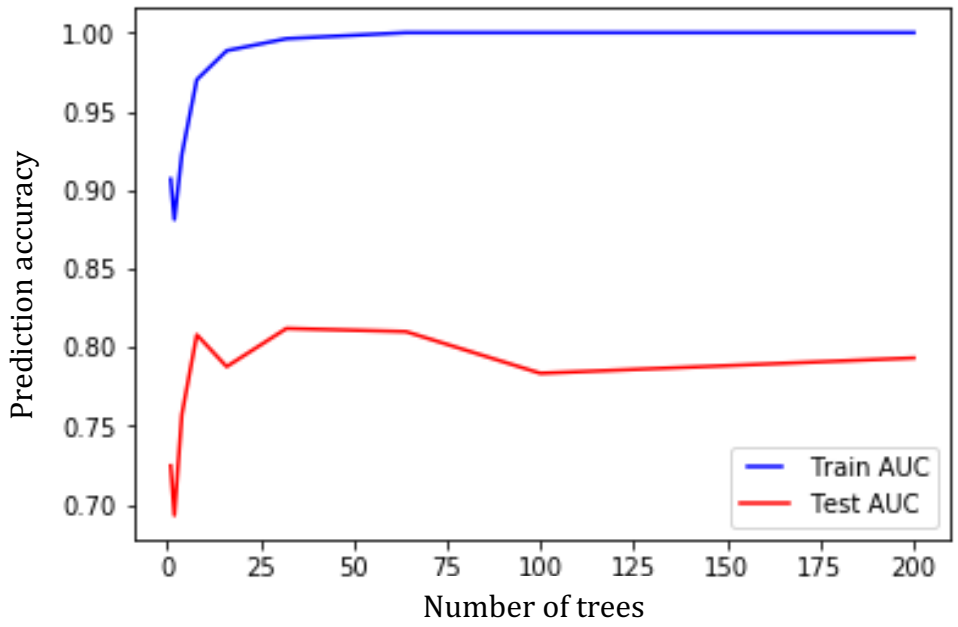
Randomly choose features for split in each tree!

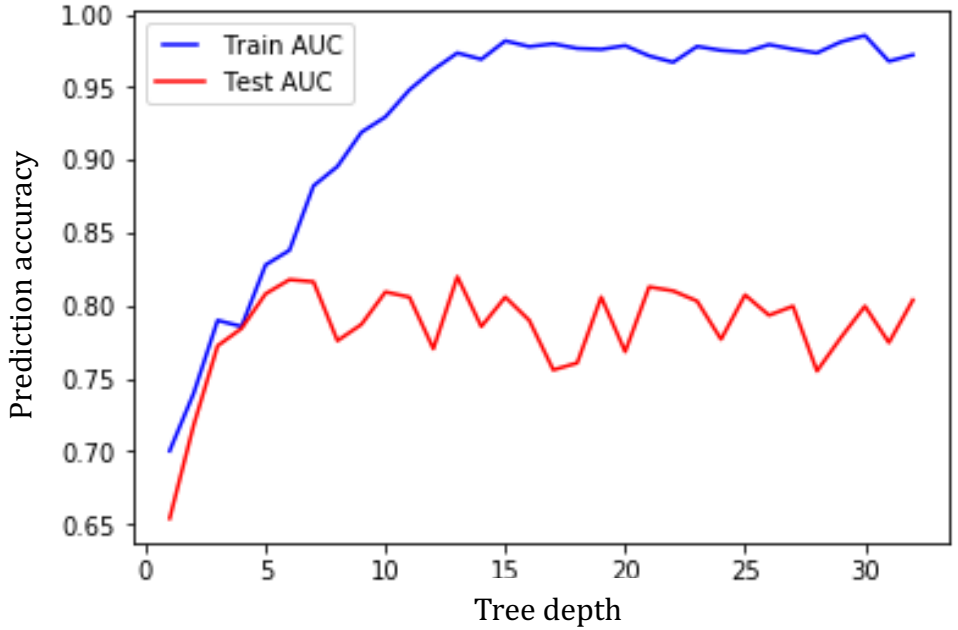
More diverse trees with less correlation

Random Forest (RF)

Random Forest

- Performance of Random Forest
  - Performance is affected by
    - Number of trees (variance)
    - Accuracy of each tree (bias)





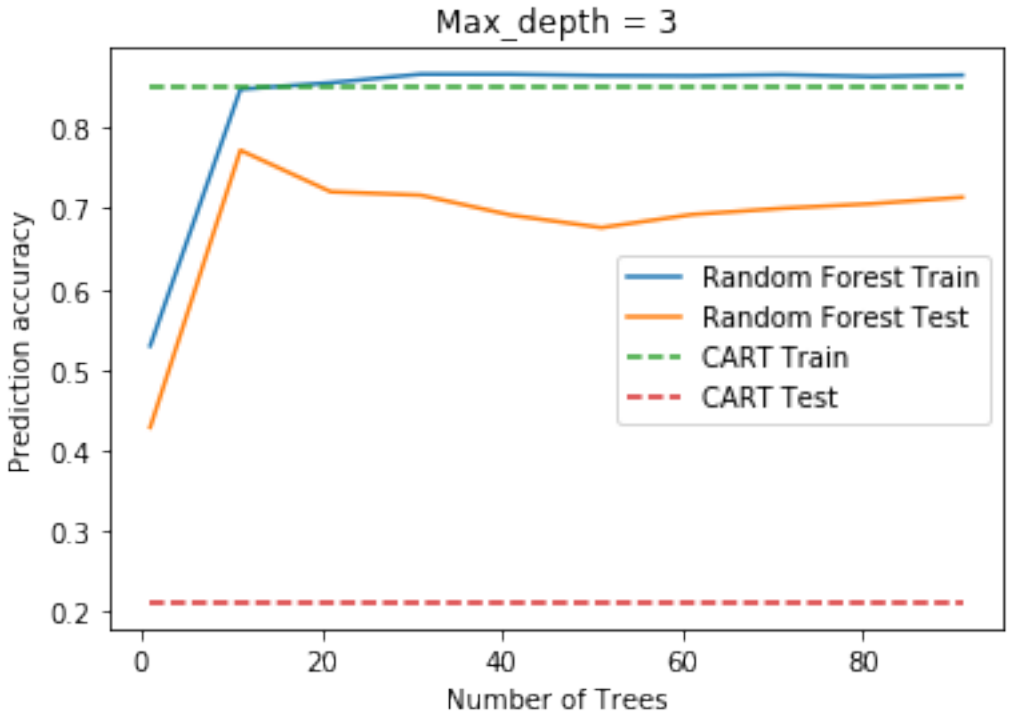
General trends of the performance of RF with different number of tress/tree depth

Comparison between CART and RF

Example

- Target: Life Expectancy of the state
- 7 Features: Population, Income, Illiteracy%, Murder, High School Graduate%, Frost days, Area
- 50 data (small data set) → 40 training data/10 test data

- By using RF, the prediction accuracy on the test set is way better than that of CART
  - Using RF, 0.21 → 0.7



Training set accuracy ~ 0.9

Test set accuracy (RF) ~ 0.7

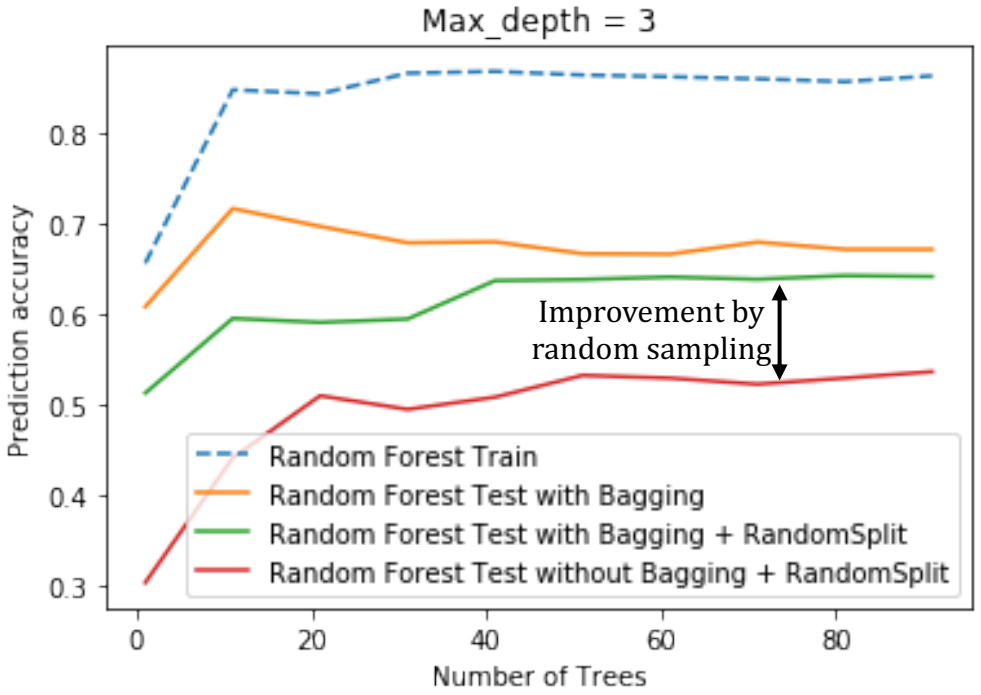
Test set accuracy (CART) ~ 0.21

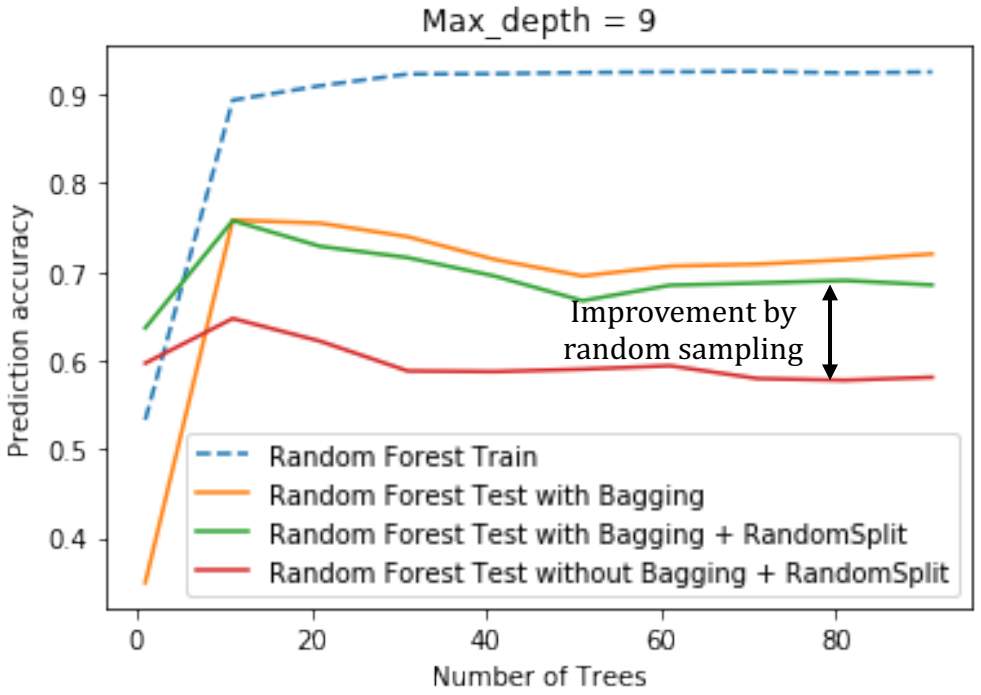
Impact of Bagging in RF

Example

- To see the impact of random sampling of data,
  - Entire data (instead of randomly sampled with replacement) is used for each tree
  - For split, only a random subset (3 out of 7) of the features is selected → To make each tree different

- Sampling with replacement is clearly impactful to increase the prediction accuracy
- Split with less features has almost the same result with a sufficient number of trees





Summary

- Classification and Regression Tree (CART) model is useful
  - Can do regression & classification
  - Can handle continuous & discrete variables easily
  - Explanation of the result is possible
- However, due to overfitting, it is important to tune hyperparameter correctly
  - Maximum depth/minimum number of data in leaf, etc.
  - Data selection for training is critical
- Random Forest (RF) can **reduce variance** while preserving low bias of CART
  - Use Bagging (Random sampling with replacement)
  - Use different random splits in different trees
- By doing so, better prediction accuracy can be obtained by reducing variance!
- Bagging can be used with different building block algorithms that are easy to overfit to reduce variance