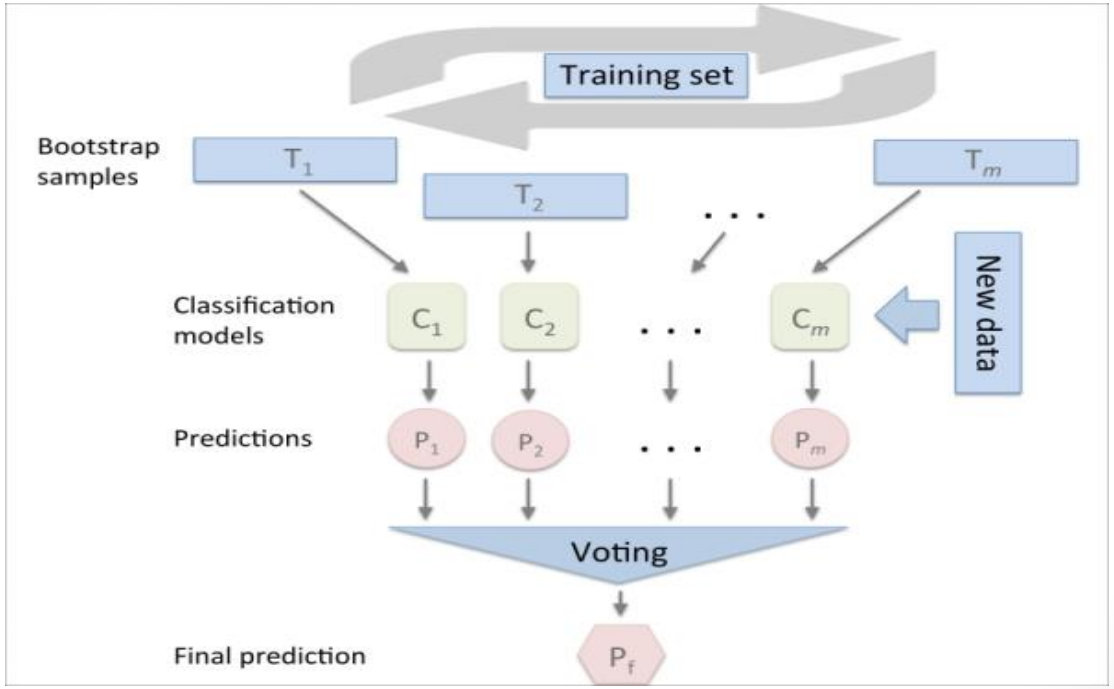
Ensemble

1. Bagging
2. Boosting

Ensemble

Machine learning paradigm which combine weak learners to become a strong learner

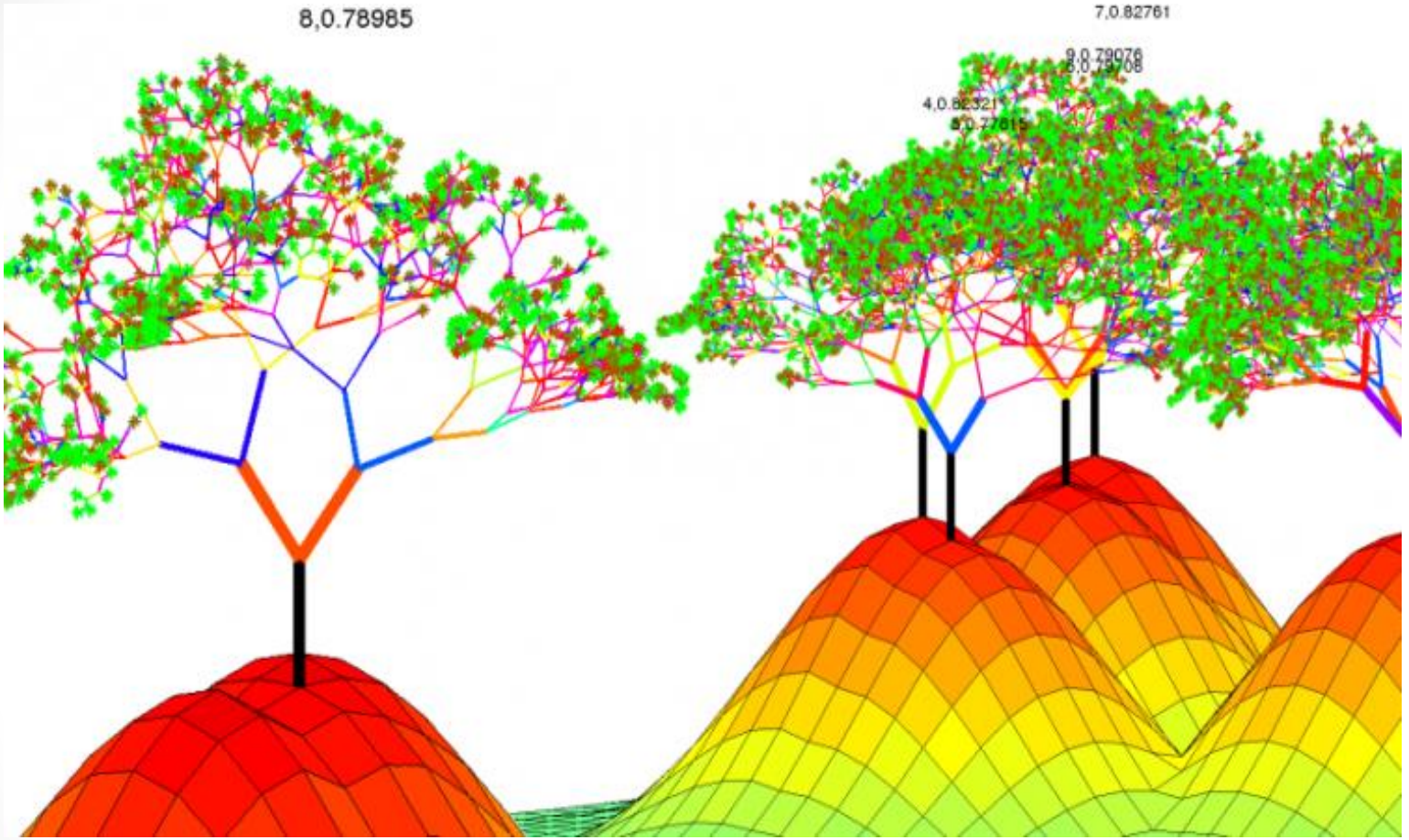
Model1	Model2	Model3	VotingPrediction
1	0	1	1



Random Forest (*Most used algorithm)*

Random Forest (*Most used algorithm)*

- Bagging Technique (**B**ootstrap **agg**regat**ing** - **Bagging**)


8,0.78985
7,0.82761

Why Random Forest?

No overfitting

Use of multiple trees reduce the risk of overfitting

Training time is less

High accuracy

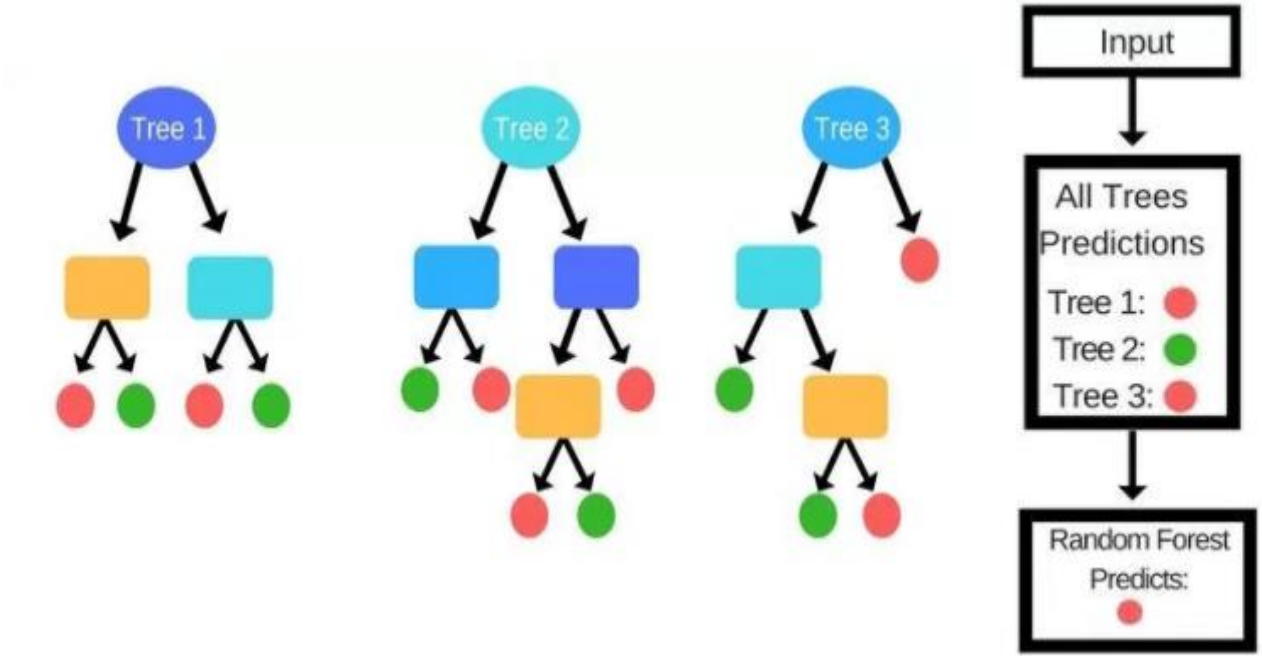
Runs efficiently on large database

For large data, it produces highly accurate predictions

Estimates missing data

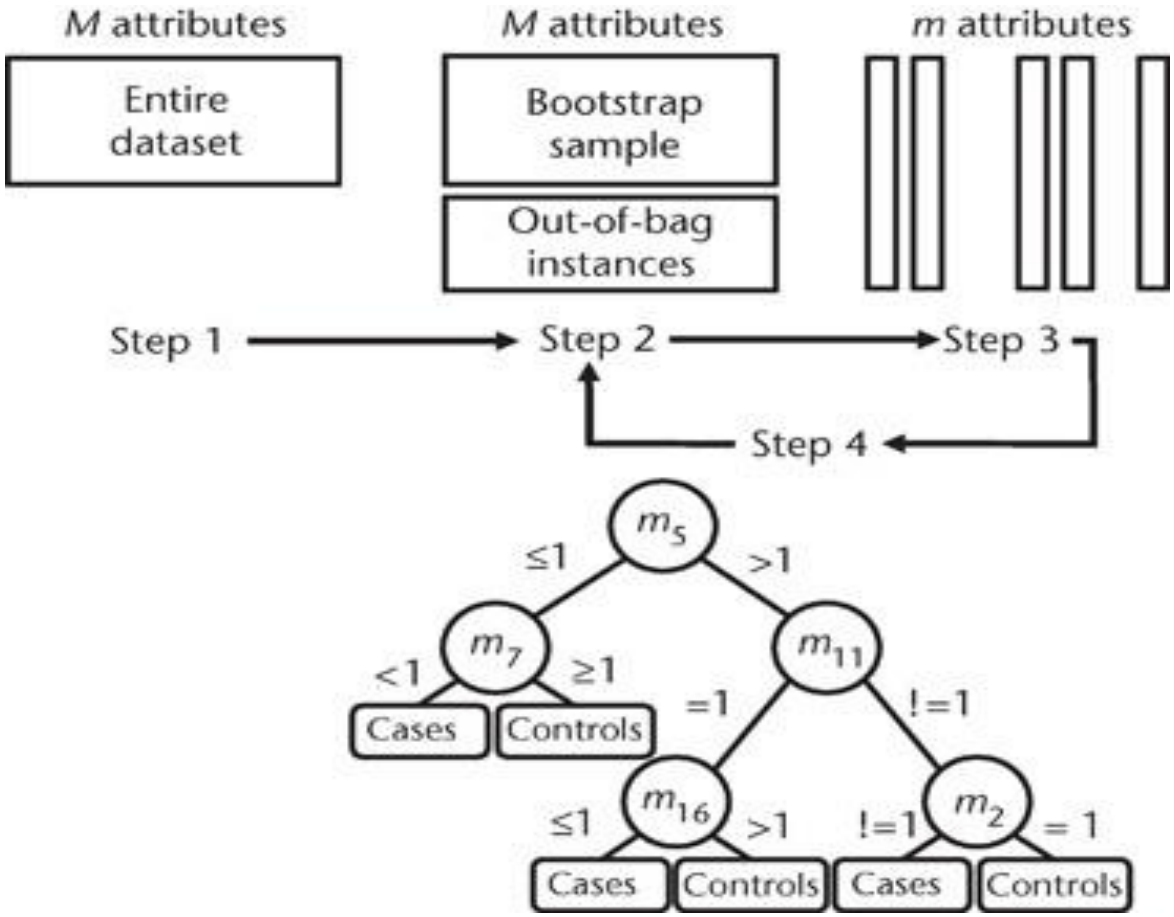
Random Forest can maintain accuracy when a large proportion of data is missing

HOW THE RANDOM FOREST ALGORITHM WORKS IN MACHINE LEARNING



- Supervised learning algorithm
- **Regression and classification problems**

Bagging


M attributes
Entire dataset
M attributes
Bootstrap sample
Out-of-bag instances
m attributes
Step 1
Step 2
Step 3
Step 4
m_5
≤1
>1
m_7
m_{11}
<1
≥1
=1
!=1
Cases
Controls
m_{16}
m_2
≤1
>1
!=1
= 1
Cases
Controls
Cases
Controls

Random Forest pseudocode

- Randomly select **“k”** features from total **“m”** features.

  Where **k << m**

  For classification a good default is: k = sqrt(m)
  For regression a good default is: k = m/3

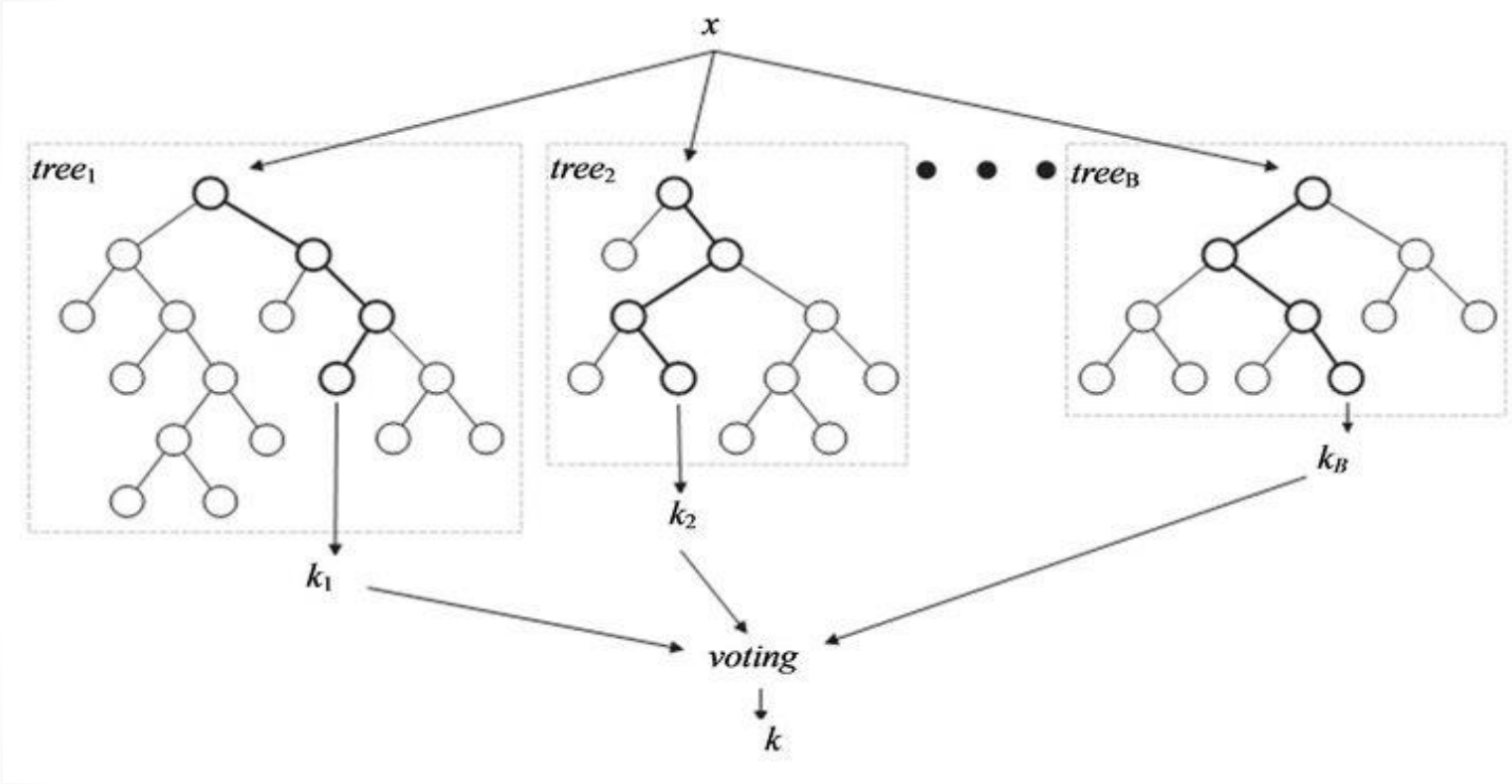
- Among the **“k”** features, calculate the node **“d”**.
- Split the node into **daughter nodes**.
- Repeat **1 to 3** steps
- Build forest by repeating steps **1 to 4** for “n” number times to create **“n” number of trees**.

Key Points

- **Majority voting**.
- **Higher the number** of trees in the forest = **High accuracy**.

- When we have more trees in the forest, random forest classifier won't **overfit** the model.
- For each bootstrap sample taken from the training data, there will be samples left behind that were not included. These samples are called **Out-Of-Bag samples** or OOB.
- The performance of each model on its left out samples when averaged can provide an estimated accuracy of the bagged models. This estimated performance is often called the **OOB estimate of performance**.

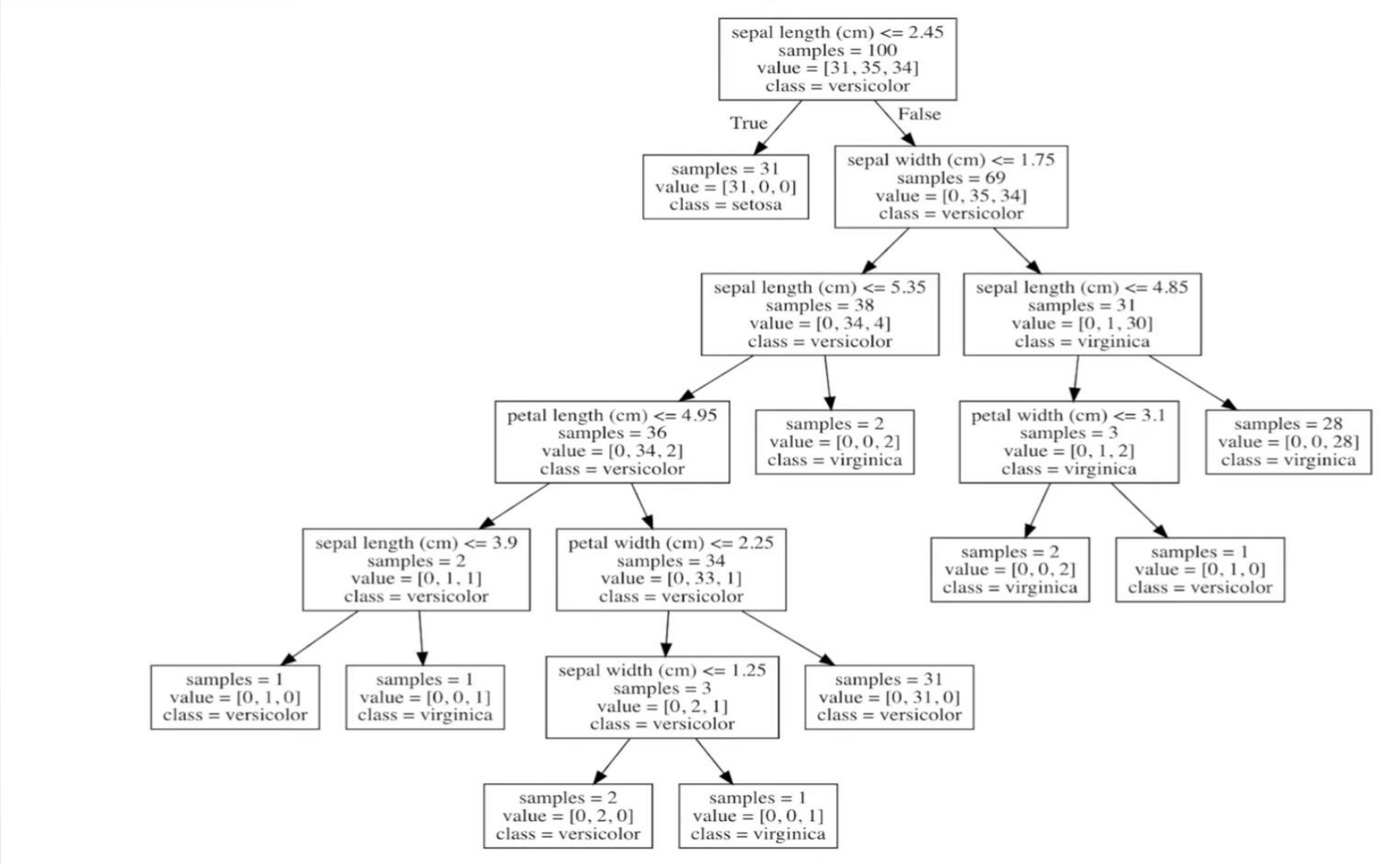
Random Forest - Skeleton



Important hyper-parameters for RF tuning:

- n_estimators - *# of trees in a forest*
- max_features - *# of features for each tree*
- max_depth – *# of levels in a tree*
- min_sample_split – *Min no. of samples before internal node split*
- min_sample_leaf – *Min no. of samples in a leaf node*
- n_jobs – *For parallel processing across multiple processors (if any)*
- oob_score – *Score of Out Of Bag (OOB) sample*
- Bootstrap – *Random selection of samples with replacement*
- random_state – *Fixed random state of samples in a tree*

Before max_depth tuning:



After max_depth tuning: