Thank You

For more information please visit our website
www.edureka.co

Features	Python for Data Science	Data Science Masters	AIML PGD NIT Warangal
Instructor-led LIVE Session	✓	✓	✓
24/7 Doubt Clearing Support	✓	✓	✓
Capstone Project	✓	✓	✓
Cloud Lab	✓	✓	✓
Valid Certification	✓	✓	✓
Career Assistance			✓
Alumni Status			✓
Number of Projects/ Assignments /Case Studies/ Hands-on	15+	50+	120+
Hours of LIVE Class	42 hours	250+ hours	420+ hours
Price	₹21,995	₹ 89,999	₹2,22,450

edureka!

Courses	Python for Data Science	Data Science Masters	Data Science PGP by IIT Guwahati
Python Basics	✓	✓	✓
Machine Learning	✓	✓	✓
Statistics		✓	✓
Deep Learning		✓	✓
Natural Language Processing			✓
Sequence Learning			✓
Reinforcement Learning			✓
Apache Spark and Scala		✓	
Tableau		✓	
Data Science using R		✓	

DATA SCIENCE @edureka!

edureka!

Bootstrapped Dataset: 2 Variables – 3 Variable

Family History	High BP	Overweight	Weight (kg)	Diabetes
Yes	Yes	Yes	100	Yes
No	No	No	65	No
Yes	Yes	No	75	No
Yes	Yes	No	75	No

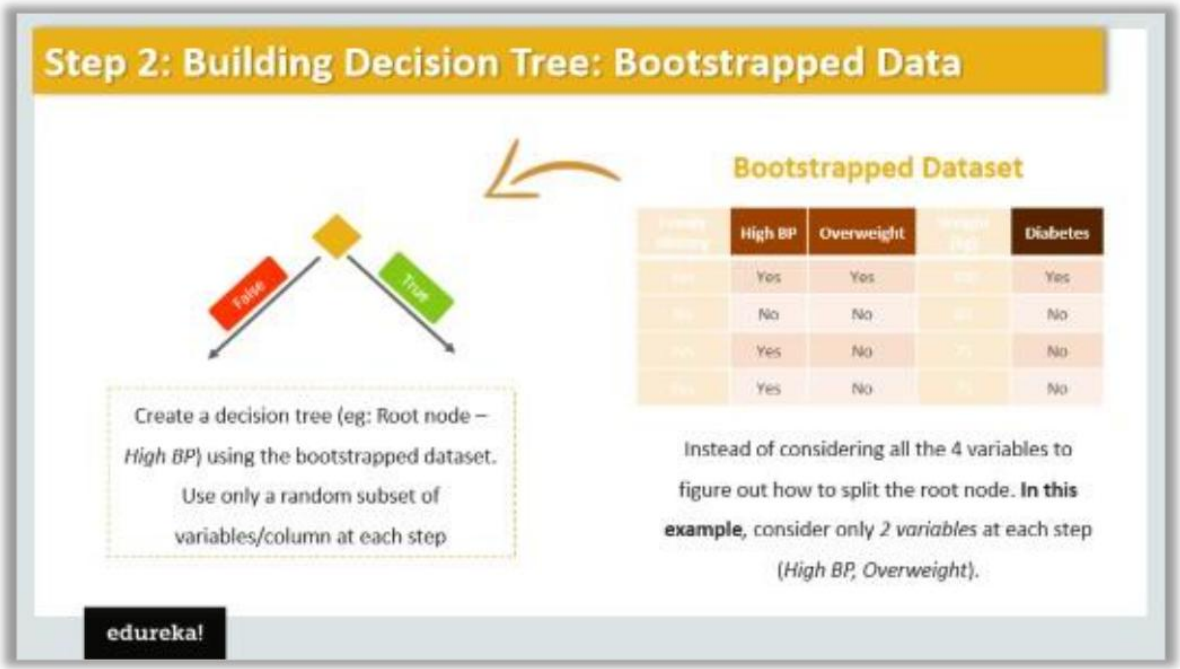
2 Variables

Compare the Out-of-Bag error for a random forest built using 2 variables vs 3 variables and select the most accurate random forest

Family History	High BP	Overweight	Weight (kg)	Diabetes
Yes	Yes	Yes	100	Yes
No	No	No	65	No
Yes	Yes	No	75	No
Yes	Yes	No	75	No

3 Variables

Bootstrapped Dataset: 2 Variables



Remember when we built our first tree, we only used 2 variables to make decision at each step?

How Good is Your Model: Out-of-Bag Dataset

Classification of Out-of-Bag Dataset

Diabetes	
YES	NO
1	4

Diabetes	
YES	NO
4	0

Diabetes	
YES	NO
3	1

Out-of-Bag Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
No	No	No	65	No

Next we do the same for all of the other Out-of-Bag samples for all the trees, and finally the accuracy of random forest can be measured by the proportion of Out-of-Bag samples that were correctly classified by the algorithm

How Good is Your Model: Out-of-Bag Dataset

Classification of Out-of-Bag Dataset

Diabetes	
YES	NO
1	4

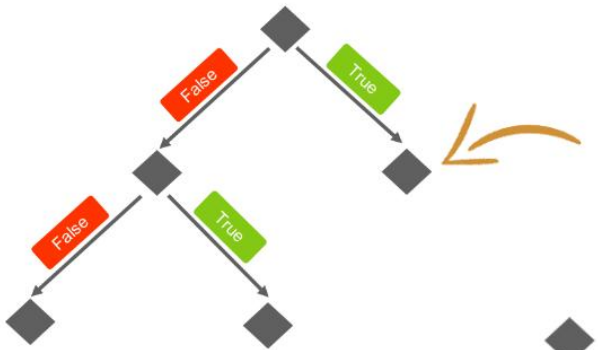
Out-of-Bag Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
Yes	No	Yes	110	Yes

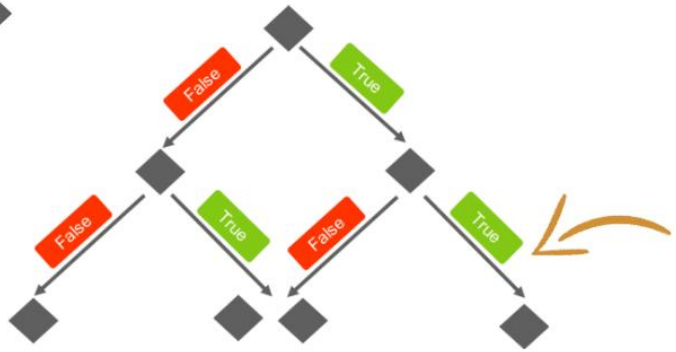
Incorrectly classified Out-of-Bag samples are known as '**Out-of-Bag Error**'

- Run this out of bag sample through other trees and keep a count the label
- Label with most votes wins and is assigned, RIGHT! In this case, Random Forest has correctly labelled the Out-of-Bag sample

How Good is Your Model: Out-of-Bag Dataset



Predicted Diabetes: **NO**



Predicted Diabetes: **YES**

. . .

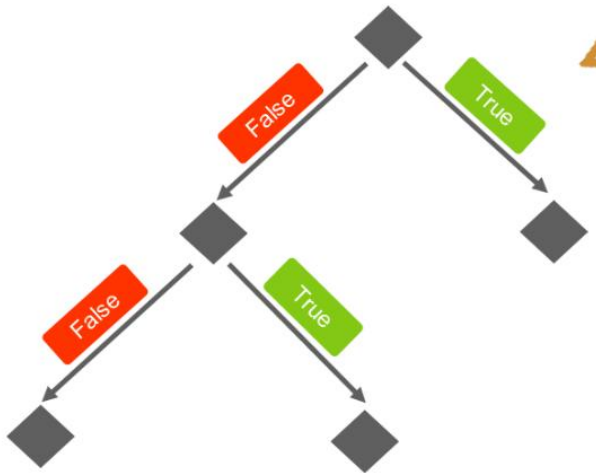
Predicted Diabetes: **YES**

Out-of-Bag Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
Yes	No	Yes	110	Yes

Run this Out-of-Bag sample data through all the trees that were built without it

How Good is Your Model: Out-of-Bag Dataset



Out-of-Bag Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
Yes	No	Yes	110	Yes

Since Out-of-Bag Dataset was not used to create this tree, we can run it through and check if it correctly classifies the sample as "**Yes**" in Diabetes

How Good is Your Model: Out-of-Bag Dataset

Original Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
No	No	No	65	No
Yes	Yes	Yes	100	Yes
Yes	Yes	No	75	No
Yes	No	Yes	110	Yes

Out-of-Bag Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
Yes	No	Yes	110	Yes

This is the entry that did not end up in the bootstrapped dataset and the collection of such entries as a dataset is known as **" Out of Bag Dataset"**

NOTE: Just in case, the original dataset were larger, we would have more than 1 entry over here

How Good is Your Model: Bootstrapped Dataset

Original Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
No	No	No	65	No
Yes	Yes	Yes	100	Yes
Yes	Yes	No	75	No
Yes	No	Yes	110	Yes

Bootstrapped Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
No	No	No	65	No
Yes	Yes	Yes	100	Yes
Yes	Yes	No	75	No
Yes	Yes	No	75	No

Remember! this entry (almost 1/3rd of the original dataset) was not included in the bootstrapped dataset because of the duplicate entry

Step 3 & 4: Counting the Votes for Predicting

Bootstrapped Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
Yes	Yes	No	75	?

Predict if the patient has diabetes or not

Vote Count

Diabetes	
YES	NO
95	5

In this case, '**YES**' received most of the votes,
❼ **patient has diabetes**

Let's say we created 100 decision tree and this is the overall vote count of the predictions from all the decision tree. Next thing is to find out the option which received most votes

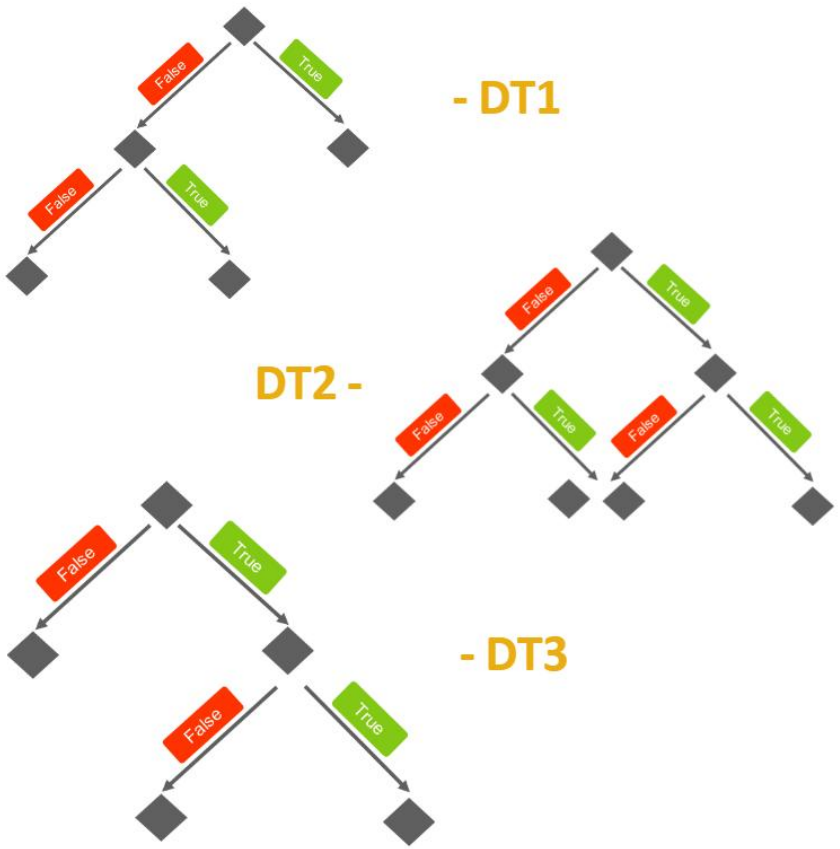
NOTE: Bootstrapping the data and using the aggregate to make a decision is called **Bagging**

Step 2: Building Decision Tree: Bootstrapped Data

Original Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
No	No	No	65	No
Yes	Yes	Yes	100	Yes
Yes	Yes	No	75	No
Yes	No	Yes	110	Yes

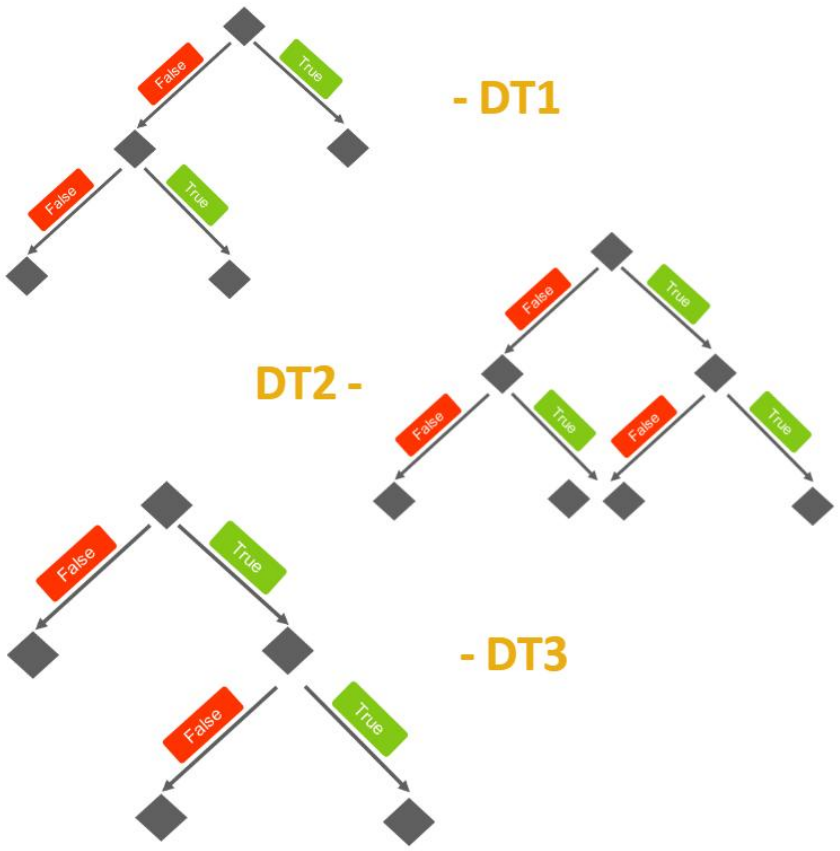
Using a bootstrapped sample and considering only a subset of variables at each step results in a wide variety of trees. This variety makes random forest **more effective** than individual decision trees



Step 2: Building Decision Tree: Bootstrapped Data

Original Dataset

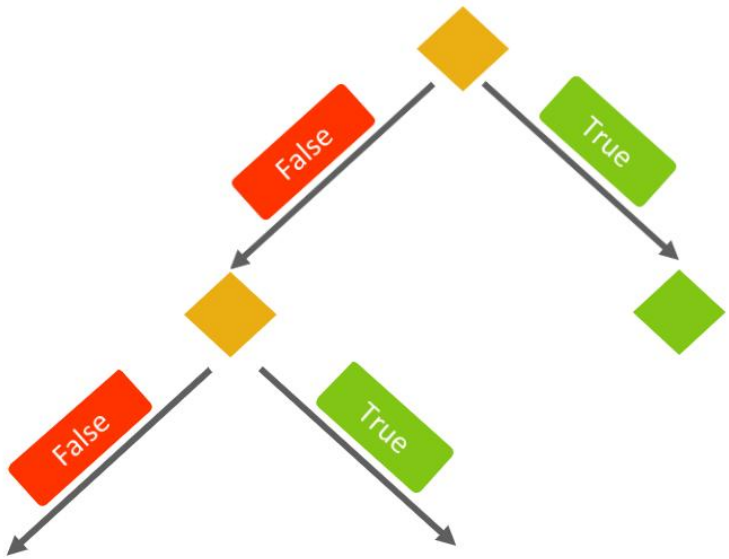
Family History	High BP	Overweight	Weight (kg)	Diabetes
No	No	No	65	No
Yes	Yes	Yes	100	Yes
Yes	Yes	No	75	No
Yes	No	Yes	110	Yes



Recursively splitting sample at other nodes

Get back to Step 1 and repeat: Build new bootstrapped dataset and rebuild decision trees considering subset of variables at each step (ideally you have to repeat this step 100's of time)

Step 2: Building Decision Tree: Bootstrapped Data



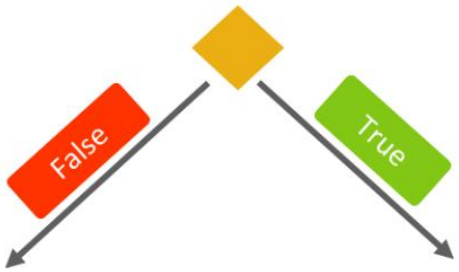
Now we need to figure out, how to split samples at this node

Bootstrapped Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
Yes	Yes	Yes	100	Yes
No	No	No	65	No
Yes	Yes	No	75	No
Yes	Yes	No	75	No

The decision tree is ready using the random subset of variables at each step

Step 2: Building Decision Tree: Bootstrapped Data



Create a decision tree (eg: Root node – *High BP*) using the bootstrapped dataset. Use only a random subset of variables/column at each step

Bootstrapped Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
Yes	Yes	Yes	100	Yes
No	No	No	65	No
Yes	Yes	No	75	No
Yes	Yes	No	75	No

Instead of considering all the 4 variables to figure out how to split the root node. **In this example**, consider only *2 variables* at each step (*High BP, Overweight*).

Step 1: Create a Bootstrap Dataset contd...

Original Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
No	No	No	65	No
Yes	Yes	Yes	100	Yes
Yes	Yes	No	75	No
Yes	No	Yes	110	Yes

Bootstrapped Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
Yes	Yes	Yes	100	Yes
No	No	No	65	No
Yes	Yes	No	75	No
Yes	Yes	No	75	No

Creating a bootstrapped dataset with randomly selected sample from the original dataset

Note: 3rd and 4th randomly selected sample in the bootstrapped dataset are same

Step 1: Create a Bootstrap Dataset

Bootstrapped Dataset

The **bootstrap** method is a resampling technique used to estimate statistics on a population by sampling a **dataset** with replacement. It can be used to estimate summary statistics such as the mean or standard deviation.

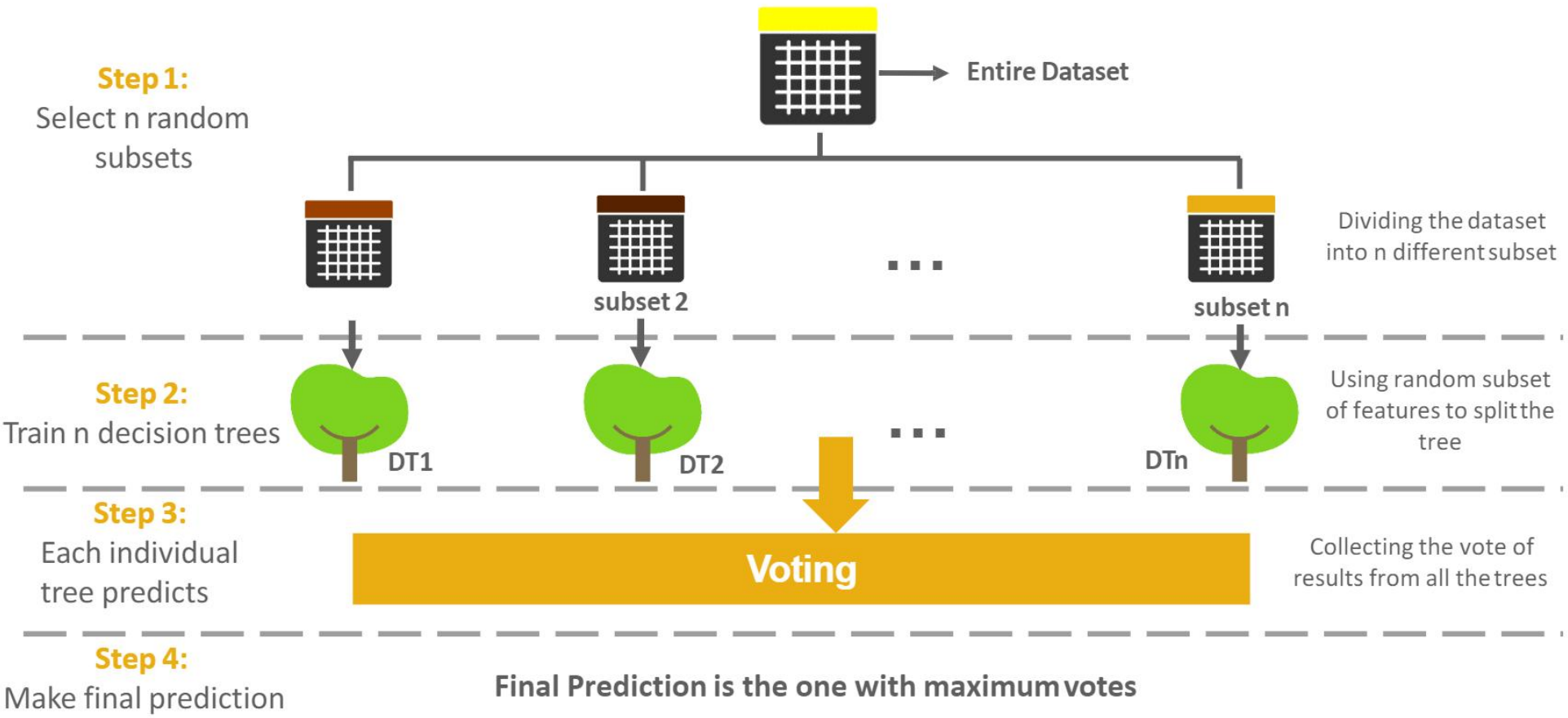
The **bootstrap** dataset (same size as original) is created by randomly selecting samples from the original dataset.

Family History	High BP	Overweight	Weight (kg)	Diabetes
No	No	No	65	No
Yes	Yes	Yes	100	Yes
Yes	Yes	No	75	No
Yes	No	Yes	110	Yes

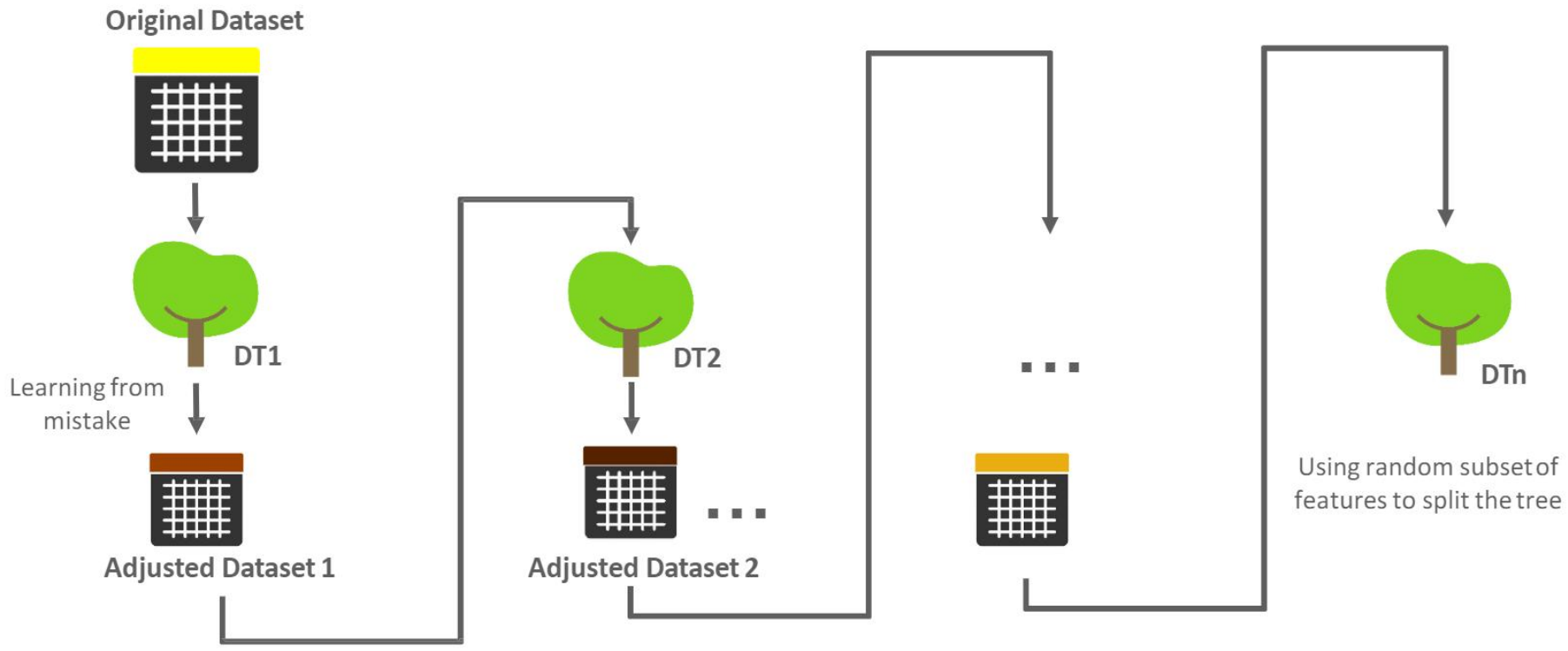
This is our sample dataset...

NOTE: You can pick the same sample more than once

Building a Random Forest



Ensemble Learning: Types - Boosting



Ensemble Learning: Types - Boosting

Boosting is training a bunch of individual models in a **sequential** way. Each individual model learns from mistakes made by the previous model.

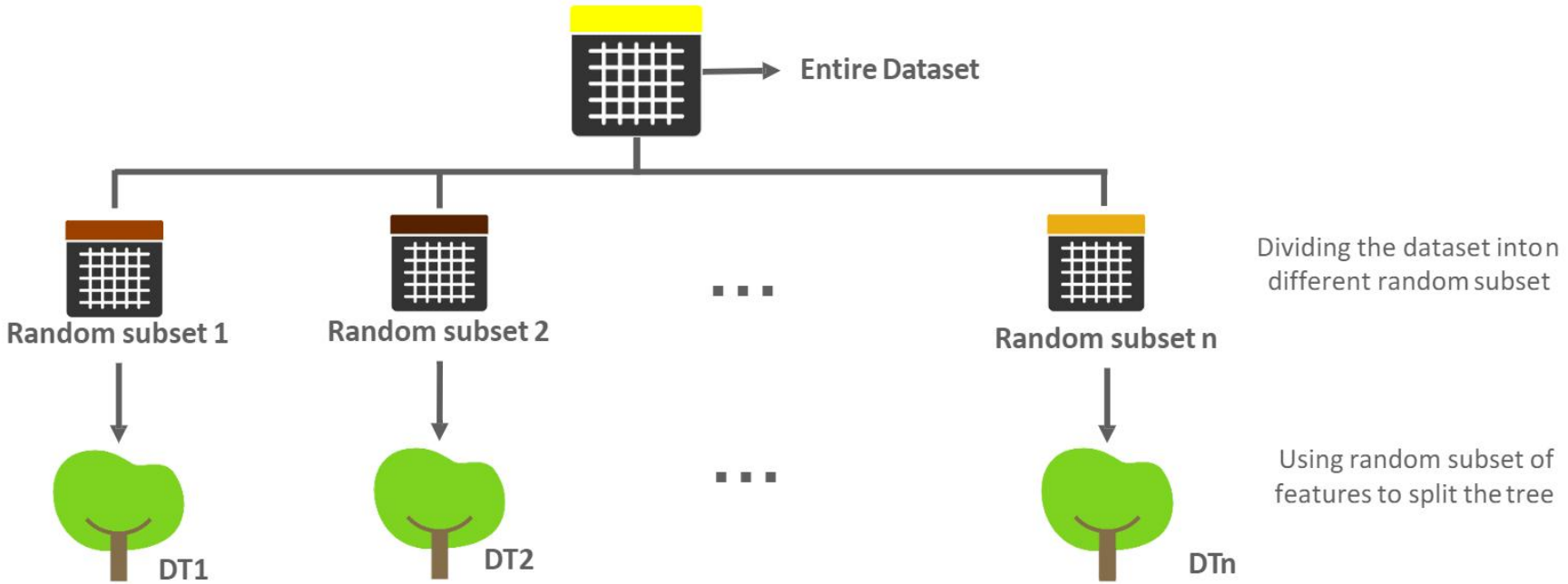
Bootstrapped Dataset

Family History	High BP	Overweight	Weight (kg)	Diabetes
Yes	Yes	Yes	110	YES

Vote Count

Diabetes	
YES	NO
95	5

Ensemble Learning: Types - Bagging



Bagging

Ensemble Learning: Types - Bagging

Bootstrapping the data and using its aggregate to make a decision is known as **Bagging**. *In other words*, **Bagging is** training a bunch of individual models **parallelly**, and each model is trained by a random subset of the data.

Bootstrapped Dataset

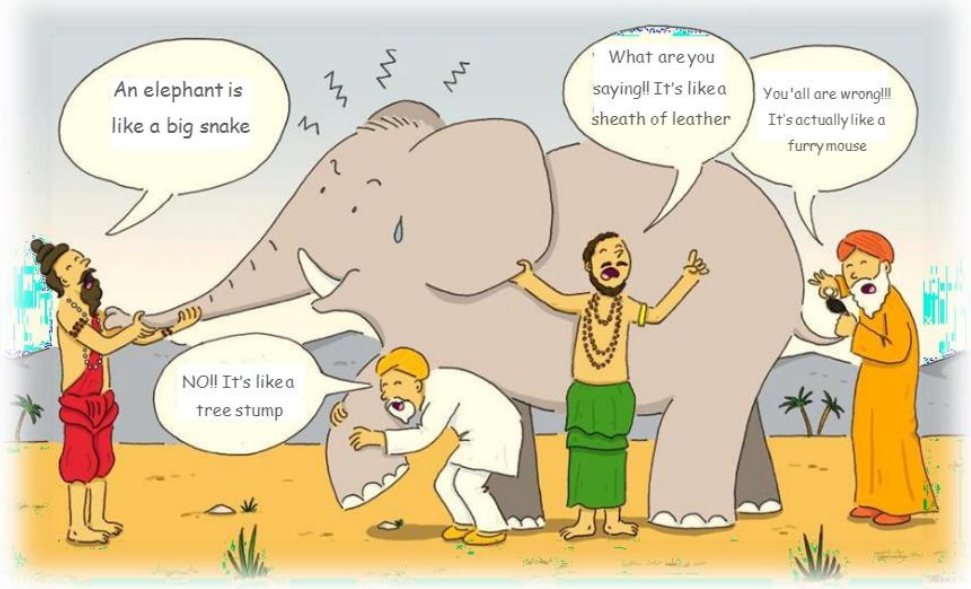
Family History	High BP	Overweight	Weight (kg)	Diabetes
Yes	Yes	Yes	110	YES

Vote Count

Diabetes	
YES	NO
95	5

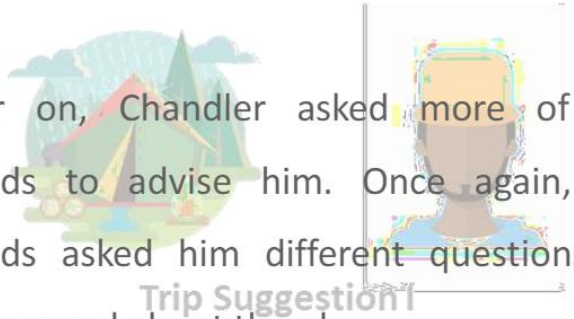
What is Ensemble Learning?

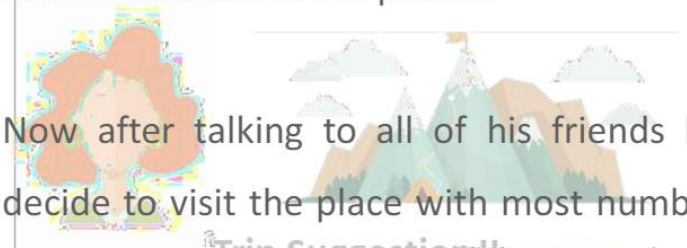
Random forest uses Ensemble learning method in which the predictions are based on the **combined results** *of various individual models*



Fable of blind men and an elephant

Random Forest Analogy





Later on, Chandler asked more of his friends to advise him. Once again, his friends asked him different questions to recommend about the places.

Now after talking to all of his friends he decide to visit the place with most number of votes. the above scenario is a typical example of **Random Forest Algorithm**.

Random Forest Analogy

Every friends gave suggestion by asking him few questions

Trip Suggestion I

Trip Suggestion II

Trip Suggestion III

Random Forest Analogy

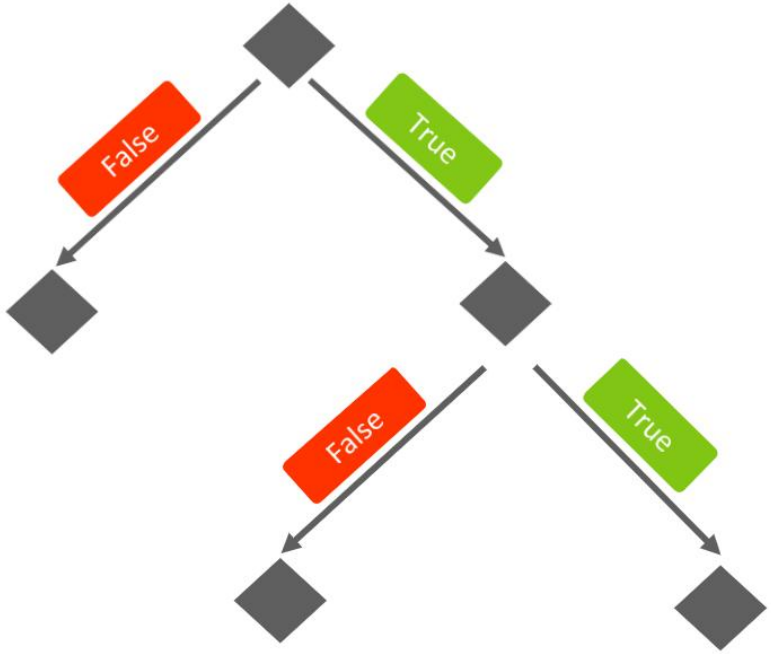
Chandler is planning for a one-year vacation trip. So in order to decide which places should he travel to, he asks his friends for their advice.

Why Random Forest?

Decision Trees are easy to build and interpret.

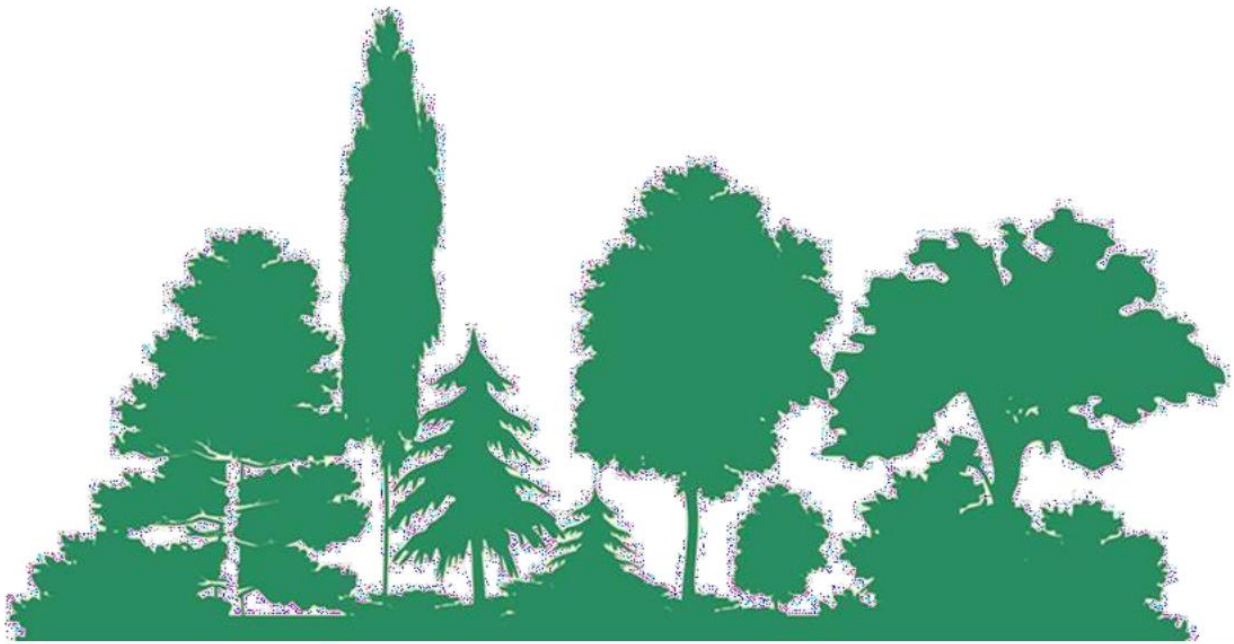
Then **WHY** Random Forest?

Decision Trees have only **one aspect,** therefore are **less accurate and inflexible** when it comes to **classifying new samples.**



What is Random Forest?

Random Forest is the most used supervised machine learning algorithm for classification and regression



MODEL BU
RANDOM

edureka!