Random Forest is a combination of Decision Trees.

One of the main**drawbacks of Decision Trees** is that they are very prone to over-fitting: they do well on training data, but are not so flexible for making predictions on unseen samples.

(insert bias, variance image)

(insert image with overlapping circles)

We use random sampling to overcome over fitting.

1. Randomly select partial rows (data sample)
2. Randomly select partial columns (features)
3. Combine and Average

In conclusion, the whole process goes as follows:

1. Create a bootstrapped data set for each tree.
   * Randomly select partial rows
2. Create a decision tree using its corresponding data set, but at each node use a random sub sample of variables or features to split on.
   * Randomly select partial columns (features)
3. Repeat all these three steps hundreds of times to build a massive forest with a wide variety of trees. This variety is what makes a Random Forest way better than a single decision tree.
   * Each tree has high variance but low bias (centered)
   * Many overlapping tree variances average to one predictor

Once we have built our forest, we are ready to use it to make our predictions.

1. Categorical Model – discrete levels predictor
2. Regression Model – continuous numerical predictor

Reference: Random Forest Explained, <https://towardsdatascience.com/random-forest-explained-7eae084f3ebe>