Outline

1. Briefly explain random forest.
2. How to implement random forest in Weka.

Why use random forest? A reflection of the decision tree algorithm

1. **Decision tree is very sensitive to small variations in the training data, which makes the model very unstable.**
2. **Very likely to overfit the training data**

- "**Variance** is an error resulting from sensitivity to small fluctuations in the dataset used for training. High variance will cause an algorithm to model irrelevant data, or noise, in the dataset instead of the intended outputs, called signal. This problem is called **overfitting**. An overfitted model will perform well in training, but won't be able to distinguish the noise from the signal in an actual test."

https://towardsdatascience.com/decision-trees-6-important-things-to-always-remember-85636858da51
https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/

Random Forest

Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a "forest."

Steps of Random Forest Algorithm:

Step 1: Bootstrapping
Step 2: Feature selection
Step 3: Construction of trees
Step 4: Voting and aggregation

https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/
https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm

ID	x1	x2	x3	class
1	3	3	5	1
2	4	6	3	0
3	0	6	0	1
4	1	5	6	1
5	3	2	9	0

Sample 1
ID: 1, 2, 4, 3, 3

Sample 2
ID: 3, 2, 1, 5, 5

Sample 3
ID: 4, 2, 1, 3, 5

Sample 4
ID: 1, 2, 1, 3, 5

Step 1

Sample 1
Features: x1, x2

Sample 2
Features: x3, x2

Sample 3
Features: x1, x3

Sample 4
Features: x1, x2

Step 2

Step 4: Voting and aggregation