Random Forest algorithm

Decision tree

* Decision tree is a non-linear model built using linear boundaries.
* Completely learning data can be a downside because it can lead to overfitting.
  + Overfitting occurs when the model is very flexible (high capacity).
  + A flexible model has high variance because the learned parameters vary considerably from the training data.
* Gini Impurity represent the probability that a randomly selected node will be incorrectly classified.

Random forest

* Two key concepts:
  + Random sampling of training data points when building trees.
  + Random subset of features considered when splitting nodes.
* Each tree goes through testing and could use some samples more than once.
* Overall, the entire forest will have lower variance (good) without increasing bias (bad).
* Bias-variance trade-off is a common issue with machine learning, it is the balance between giving a model high flexibility (high variance) that can learn training data but cannot generalize new data vs method which cannot learn the training data (high bias). A random forest reduces variance by use multiple trees.