Random Forest

Decision trees

Large decision trees have :

- Low Bias
- high variance
- tend to overfit.

Stumps have:

- High bias
- Low Variance
- Tend to underfit

Bagging : Reduce Variance

COMBINES STUMPS SO THAT THE VARIANCE IS REDUCED IN COMPARISON THE THE Dtrees.

Randomness :

- Subset of **Samples** is taken for training each tree.
- Subset of **variables** is taken for training each tree.
- For each **split** rules are not considered from all N variables of the tree but on a smaller random subset of N.
- The random subset of features selected for splits are different for each split.

Hyperparameters:

- **n_estimator**: sint, default=100

  The number of trees in the forest.

- **max_depth**: int, default=None

  The maximum depth of the tree.

- **max_samples:** int or float, default=None

  If bootstrap is True, the number of samples to draw from X to train each base estimator.

- **min_impurity_decrease**: float, default=0.0

  A node will be split if this split induces a decrease of the impurity greater than or equal to this value.

- **warm_start:** bool, default=False

  When set to True, reuse the solution of the previous call to fit and add more estimators to the ensemble, otherwise, just fit a whole new forest.

- **n_jobs:** int, default=None

  The number of jobs to run in parallel.