Random Forest

Ensemble Methods – **Random Forest:**

1. Each tree in the ensemble is built from a sample drawn with replacement (bootstrap) from the training set

1. In addition, when splitting a node during the construction of a tree, the split that is chosen is no longer the best split among all the features

1. Instead, the split is picked is the best split among a random subset of the features

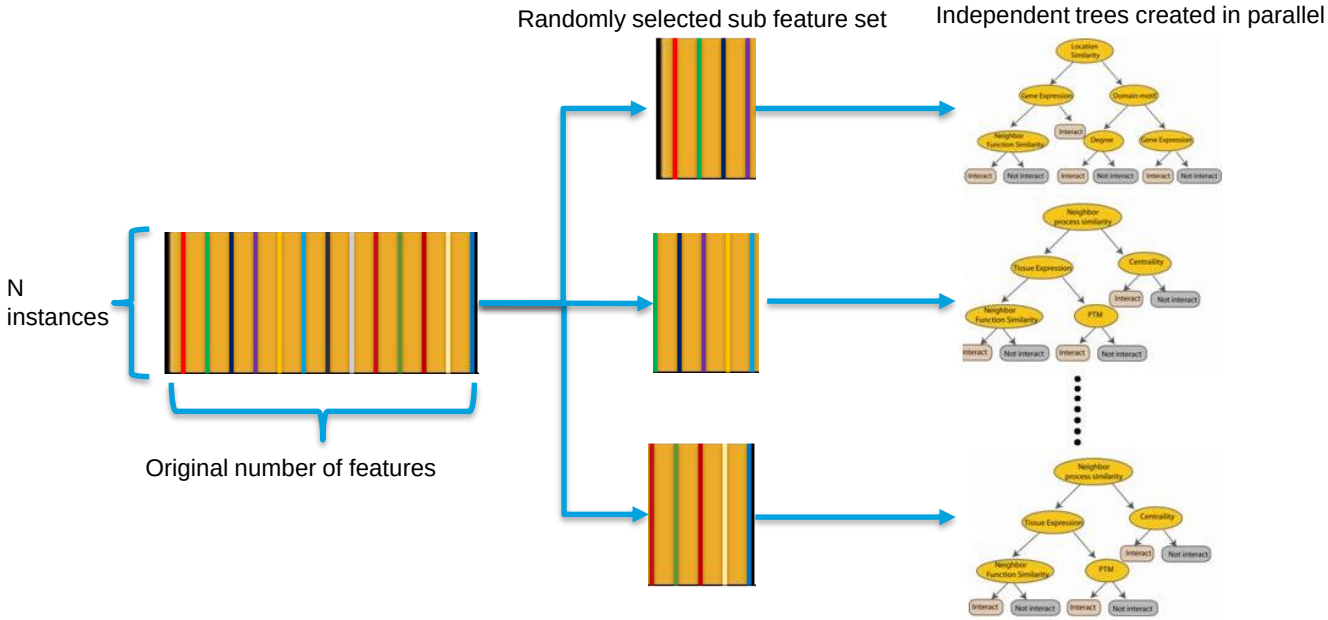
1. As a result of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree)

1. Due to averaging, its variance decreases, usually more than compensating the increase in bias, hence yielding overall a better result

Source: scikit-learn user guide , chapter 3 , page 231

Ensemble Methods - **Random Forest:**

1. Used with Decision Trees. Create different trees by providing different sub-features from the feature set to the tree creating algorithm. The optimization function is Entropy or Gini index



Ensemble Learning – **Random Forest**:

Lab- 9 Improve defaulter prediction of the decision tree using Random Forest

Description – Sample data is available at local file system as credit.csv

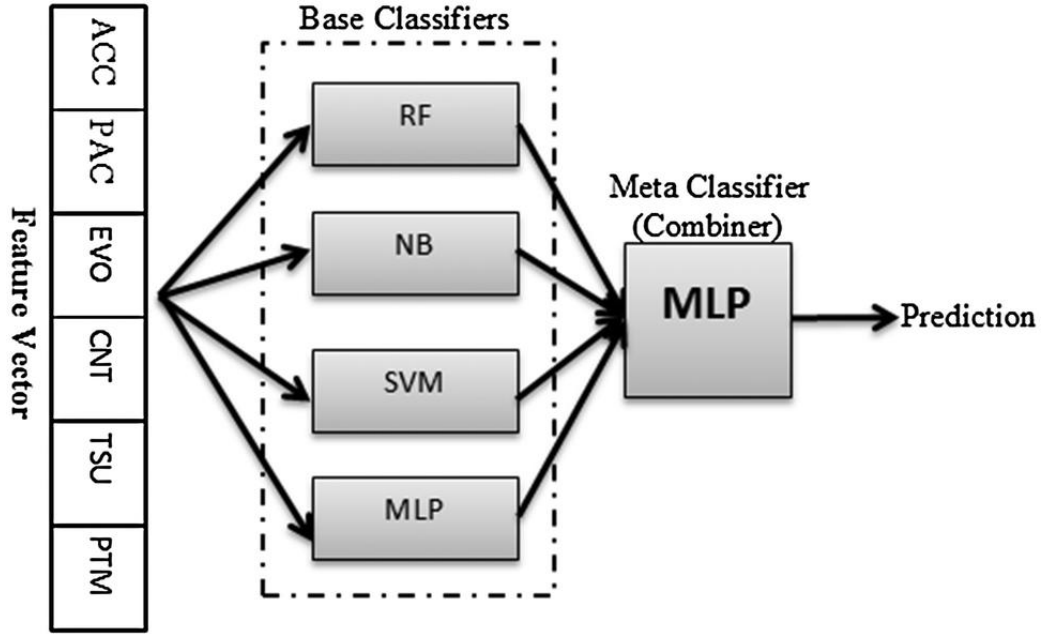
The dataset has 16 attributes described at https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
or in the notes page of this slide

Sol: RF+Credit+Decision+Tree.ipynb

Ensemble Methods – **Stacking:**

1. Similar to bagging, but apply several different models to original data
2. The weights for each model is determined based on how well they perform on the given input data
3. Similar classifiers usually make similar errors (bagging), so forming an ensemble with similar classifiers may not improve the classification rate
4. Presence of a poorly performing classifier may cause performance deterioration in the overall performance
5. Similarly, even on presence of a classifier that performs much better than all of the other available base classifiers, may cause degradation in the overall performance
6. Another important factor is the amount of correlation among the incorrect classifications made by each classifier
7. If the consistent classifiers tend to misclassify the same instances, then combining their results will have no benefit
8. In contrast, a greater amount of independence among the classifiers can result in errors by individual classifiers being overlooked when the results of the ensemble are combined.

Ensemble Methods – **Stacking:**



Source: http://pubs.rsc.org/-/content/articlelanding/2014/mb/c4mb00410h/unauth#!divAbstract

Ensemble Learning – **Stacking**:

Lab- 10 Improve defaulter prediction of the decision tree using Stacking

Description – Sample data is available at local file system as credit.csv

The dataset has 16 attributes described at
https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
or in the notes page of this slide

Sol: Stacking+Credit+Decision+Tree.ipynb