A1 Resource

https://seaborn.pydata.org/examples/scatterplot\_matrix.html

<https://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_vs_lda.html#sphx-glr-auto-examples-decomposition-plot-pca-vs-lda-py>

<https://plot.ly/python/splom/>

https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.train\_test\_split.html

<https://www.jianshu.com/p/1da36bc3663b?utm_source=oschina-app>

<https://datascience.stackexchange.com/questions/15135/train-test-validation-set-splitting-in-sklearn>

<https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>

Similar to SVC with parameter kernel=’linear’, but implemented in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples.

<https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

<http://sofasofa.io/forum_main_post.php?postid=1004110>

<https://www.cnblogs.com/gczr/p/7098592.html>

好

5.1

<https://blog.csdn.net/bqw18744018044/article/details/80034394>

<https://zhuanlan.zhihu.com/p/48976706>

5.2

https://blog.csdn.net/zpf123456789zpf/article/details/88303833

5.3

https://blog.csdn.net/u013270326/article/details/81229738

<https://blog.csdn.net/u013270326/article/details/81229738>

5.4

<https://www.jianshu.com/p/503a56095297>

<https://www.jianshu.com/p/4e27ef04ace3>

5.5

<https://blog.csdn.net/HHTNAN/article/details/80237226>

<https://blog.csdn.net/lynn_001/article/details/85339034>

https://blog.csdn.net/Barry\_J/article/details/81938402

较多的子树可以让模型有更好的性能，但同时让你的代码变慢。开始的时候树的数量较小时会使分类器的分类误差大，性能也较差。增加树的数量会提高模型的精准性，但当数的数量达到一定数字时，模型会过于拟合导致精准度下降 从而丧失了在预测时的准确性。

随机森林中通过引入随机抽样和随机抽列，使模型对异常点有更好的鲁棒性，模型的泛化能力更强。如果是无限颗树，那其实会抵消随机性的引入，最后的模型会是一个过拟合的模型，其泛化性能也会降低。此外，噪音较大时，模型也会学习到更多噪音相关的信息，发生过拟合，降低泛化性能。

不仅仅是简单地运用集成思想，而且它是基于对残差的学习的。

随机森林是基于决策树的算法，只不过是利用集成的思想来提升单颗决策树的分类性能。主要特点是由于随机选择样本和特征，所以不容易陷入过拟合。随机森林算法的主要步骤是：从样本集中用Bootstrap随机选取n个样本，并从所有属性中随机选取K个属性，选择最佳分割属性作为节点建立分类器（CART，SVM等. 重复以上m次，即建立了m个分类器，并通过投票表决结果，决定数据属于哪一类。梯度提升决策树GBDT也是一种基于集成思想的决策树模型。算法简述是每次迭代生成一颗新的决策树，计算损失函数在每个训练样本点的一阶导数gi和二阶导数hi ，然后通过贪心策略生成新的决策树。计算每个叶节点对应的预测值 ，并将新生成的决策树添加到模型中。

GBDT和随机森林最本质的区别是GBDT中的每一棵树学的是之前所有树结论和的残差，残差即指真实值减去预测值。