实验三　集成学习一

一．简单介绍 AdaBoost 和 Random Forest 算法的原理

二．Breast Cancer 数据实验

1.对 Breast Cancer 数据进行探索性分析

2.数据预处理

3.分别以决策树、逻辑回归、SVM 为基函数，利用网格搜索等方法寻找不同基函数下 AdaBoost 算法的最优参数。利用 Precision、Recall、F1 和 Auc 等指标评价模型，探究和对比不同基函数下的 AdaBoost 算法性能。

4.对比以决策树为基函数的 AdaBoost、Random Forest 以及 Lars 算法在 Breast Cancer 分类数据上的重要特征，得出影响 Breast Cancer 分类的关键因素。

三．Boston 数据实验

1.对 Boston 房价数据进行探索性分析

2.数据预处理

3.以 R^2、MSE、MAE 等指标为评价标准，探究 Random Forest 算法的参数对模型性能的影响

4.对单棵决策树以及以决策树为基函数的集成算法(AdaBoost，Random Forest)进行性能对比，探索相较于单模型而言，集成学习的特点。

5.分别对以决策树为基函数的 AdaBoost 算法、Random Forest 以及 Lars 算法得出的特征重要性进行对比分析，得出影响 Boston 房价的关键因素。