1. <https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume>

从以上网址下载数据，并采用回归分析的方法全面分析交通流数据，找到影响交通流数据的主要因素，并对所得结果进行详细解释（15分）

2．https://archive.ics.uci.edu/ml/datasets/Forest+Fires

从以上网址中提供的数据，用前400个数据建立合理的统计模型，并对后118个数据进行预测分析，最后与实际Y 进行比较，并对自己所建模型进行解释分析（15分）

1. <https://archive.ics.uci.edu/ml/datasets/Adult>(15分)

采用决策树和贝叶斯方法对以上链接中的数据[adult.data](https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data)进行分类，并在数据集[adult.test](https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test)上验证分类的精度，并比较各种方法的优劣。

注：在分析的过程中，每个人可以选择[adult.data](https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data)数据中的1/10加以分析，测试集都采用[adult.test](https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test)即可。

1. <https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand> (15分)

对上述数据采用三种聚类分析的方法进行聚类分析，并评价每种聚类方法的效果。

1. 每人选择上证股市中的一个行业(证券，电力，煤炭等)，使用每只股票2018年年报中股票的基本面数据(财务状况，盈利情况，市场占有率，公司高管受教育程度，公司董事长的性别，董事长目前薪水，分红策略，市盈率，净资产收益率，销售增长率，资产负债率，流动比率，收入规模，利润规模，行业中所处的地位（按中位数分为前，后），公司总资产，公司总部所处的地理位置（华东，华中，华北，东北，西北，西南）等信息)，

（25分）

(1). 建立合适的统计模型，分析年报公布后第二天股票价格的影响因素。

(2). 若重新设置新的响应变量为Y\*=1（年报公布后第二天股票上涨），Y\*=0(年报公布后第二天股票下跌),采用logistic模型对数据进行分析，并与第一问所得结果进行比较。

(3). 对所选行业中所有股票按照4：1分成training data 和 test data, 并在training data上采用神经网络，支持向量机方法进行分类，并求解预测效果最好的解，最后将所得的分类准则用在test data上，验证分类的效果。

6．从UCI网站中（https://archive.ics.uci.edu/ml）任选一个数据集，分别利用两种集成的学习方法进行分析，并比较优劣。(15分)

要求：

1.写出每一题数据分析的思路和步骤。

1. 写清每一题分析的结果，并详细解释。
2. 详细写出所用函数，并将分析的主要代码附到报告后面。
3. 第三题，第四题中必选一道从UCI网站中上找其他数据进行分析，在报告中写清数据的来源和变量的具体含义。
4. 提交时间：12月25号