# 人工智能Project2

## 要求：

1、回归（Air quality dataset）：线性回归；

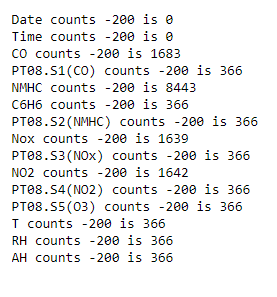
2、分类（BLE&RSSI dataset）：SVM、决策树、随机森林；

3、聚类（BLE&RSSI dataset）：DBScan、kmeans、GMM、层次聚类算法

其中聚类算法要求以t-SNE实现结果可视化

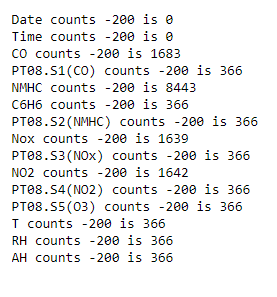
## 线性回归（Air quality dataset）：

### 数据预处理：

1. 数据了解：
   1. -200为缺失值，发现NMHC中缺失值非常的多
   2. 通过画图查看数据随时间的变化，发现数据随时间的变化趋势不明。
2. 数据处理：
   1. 数据归一化：
      1. 由于注意到某些项的数据特别大，导致它占的权重比较高，对数据影响过大，所以我把传感器的数值作了一个归一化的处理，将它化到与真实温度一样的数量级。
   2. 缺失值处理：将所有的带有缺失值的行删去
   3. 特征筛选：由于NMHC的缺失值太多所以将它从特征中删去
3. 调用sklearn 中的线性回归模型：
   1. 将数据分为训练集和测试集
   2. 训练集的准确率：0.99
   3. 测试集的准确率：0.90

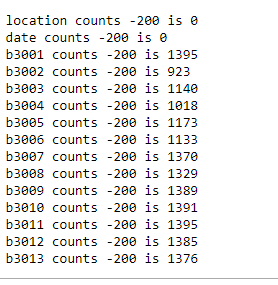
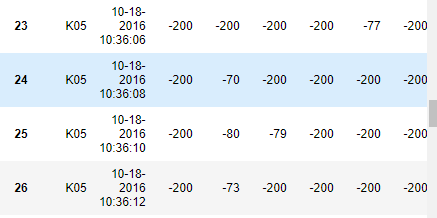
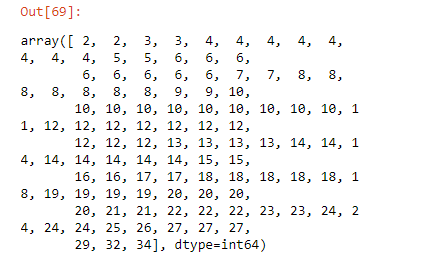
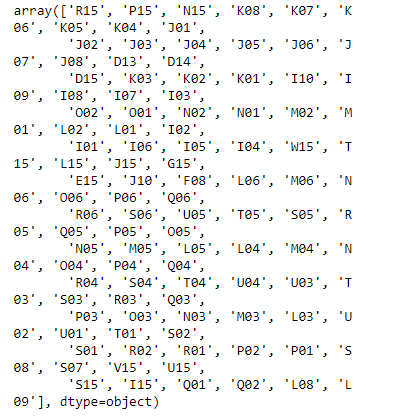
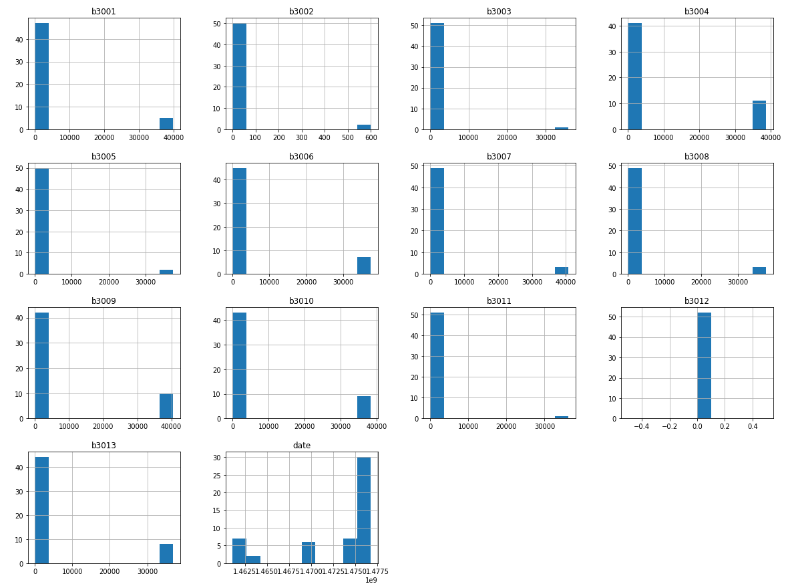
## 分类（BLE&RSSI dataset）

### 数据预处理：

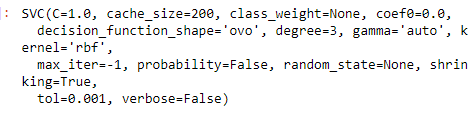
1. 数据了解：
   1. -200为缺失值，发现NMHC中缺失值非常的多
   2. 通过画图查看数据随时间的变化，发现数据随时间的变化趋势不明。
2. 数据处理：
   1. 数据归一化：
      1. 由于注意到某些项的数据特别大，导致它占的权重比较高，对数据影响过大，所以我把传感器的数值作了一个归一化的处理，将它化到与真实温度一样的数量级。
   2. 缺失值处理：将所有的带有缺失值的行删去
   3. 特征筛选：由于NMHC的缺失值太多所以将它从特征中删去

调用sklearn 中的线性回归模型

### 数据预处理：

1. 数据了解：
   1. 通过官网我们了解到：-200 是为检测到信号的点，横坐标为型号源，纵坐标为location。
   2. 查看各个信号源的缺失信号的个数：
      1. 发现 -200 的数据非常多
   3. 观察到由于信号不稳定或者是其它原因，某些location 某些时段的数据存在缺失值：
      1. 如图，明显第23 行的k05 存在缺失值
   4. 观察location的值的分布：
      1. 
      2. 可以看到location的个数分布不均匀：
         1. 
   5. 观察信号源的数据分布：
      1. 
      2. 发现大多数的信号源的差别在于高频率的检测次数
2. 数据处理：
   1. 数据化为正值：
      1. 发现 -200 的数据非常多，为了降低无效数据的影响，所以我将所有数据+200,变成了正值
   2. 缺失值处理：
      1. 求均值补充数据：
         1. 将每一个位置对应的每一个信号源的数据求非零行的均值，如果没有非零行那么设置均值为0
   3. 将时间数据变成时间戳
   4. 筛选分类目标：
      1. 由于只有十三个特征（即信号源），所以目标location 的个数不能超过13个，不然会出现过拟合，所以我从其中筛选了13个出现次数最高的location的值，来作为分类目标的种类。
   5. 分割数据集，将数据集分割为测试集和训练集

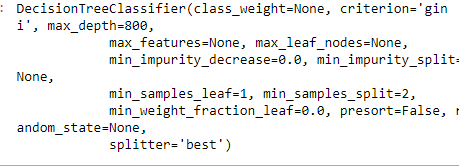
### 建模：

* + - 1. 大体步骤：
         1. 是通过grid 网格搜索最佳参数
         2. 将参数输入得到最后训练结果
         3. 评估模型
      2. Svm:
         1. 
         2. 准确率：

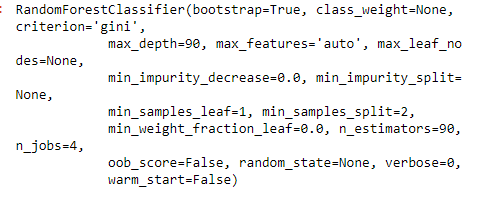
测试集：0.125

训练集：1

* + - * 1. 混淆矩阵：TP: 0 ,TN: 0, FP:0 ,TN :0
      1. 决策树
         1. 模型参数：



* + - * 1. 准确率：
        2. 混淆矩阵：TP: 1 ,TN: 2, FP:0 ,TN :1
      1. 随机森林
         1. 模型参数:



* + - * 1. 准确率：

测试集： 0.8125

训练集： 1.0

* + - * 1. 混淆矩阵: TP: 1 ,TN: 3, FP:0 ,TN :0

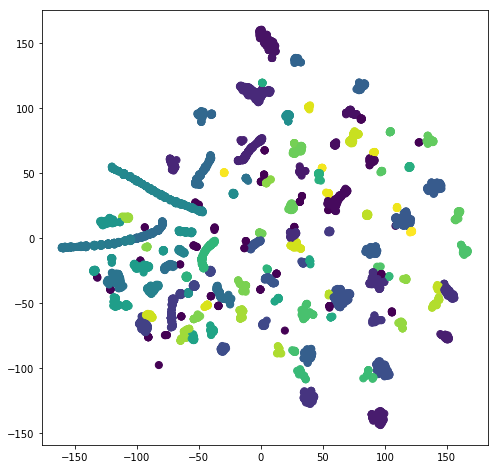
## 聚类（BLE&RSSI dataset）

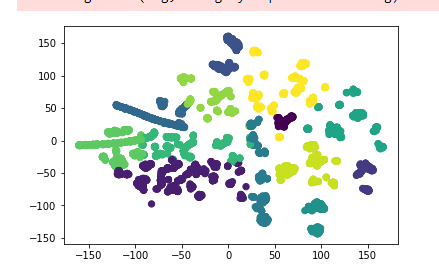
### 数据预处理：

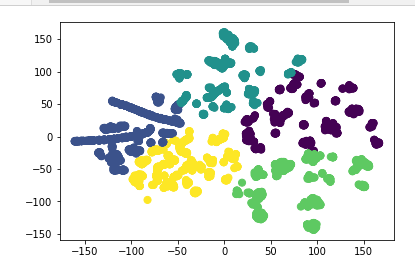
1. 直接处理：
2. 数据化为正值：
   * 1. 发现 -200 的数据非常多，为了降低无效数据的影响，所以我将所有数据+200,变成了正值
3. 将时间数据变成时间戳
4. 分割数据集，将数据集分割为测试集和训练集
5. 将测试集用tsne 降维到两维

### 建模：

* + - 1. DBSCAN：
         1. 发现分成五个簇时准确率最高
         2. 用轮廓系数来评估准确率：0.4418
         3. 可视化：



* + - 1. K-means：
         1. 发现分成五个簇时准确率最高
         2. 用轮廓系数来评估准确率：0.4418
         3. 可视化：
* ·
  + - 1. 层次聚类：
         1. 发现分成五个簇时准确率最高
         2. 用轮廓系数来评估准确率：0.452
         3. 可视化：



* + - 1. GMM：
         1. 用轮廓系数来评估准确率：0.42
         2. 可视化：

