# 实验七：数据处理综合应用1

**【实验目的】**

1. 使用所学的第三方库进行数据处理的综合应用；

**【实验内容】**

1. 完成缺省值、异常值和重复值的处理；
2. 完成数据的标准化和哑数据的处理；
3. 机器学习模型的使用。

**【预备知识】**

1. Pandas数据预处理相关函数
2. Sklearn库的应用

**【实验设备】**

1.硬件：PC机

2.软件：python3开发环境

**【实验步骤】**

1. 读取实验7数据csv文件，并将数据保存在data中。
2. 使用合适的方法将数据中的Temp列的缺省值用25替换，将Humi列中有缺省值的行删除；
3. 将Fog列中的缺省值使用前后两个相邻值的均值的方式完成填充；
4. 删除数据中的重复行；
5. 使用corr函数分析各列的相关性；
6. 将Temp列和Humi列进行标准差标准化，将Fog列进行离差标准化；
7. 将Target列中取值为正常的替换成数值0，取值为一般的替换成1，取值为严重的替换成2。
8. （选做）使用sklearn库创建基于SVC的分类模型，对标准化后的data进行分类分析；将数据集中的70%作为训练集，30%作为测试集，并分析模型的性能；

**参考代码：**

from sklearn.preprocessing import MinMaxScaler

from sklearn.model\_selection import train\_test\_split

from sklearn.svm import SVC

data = data.iloc[:,:-1]##提取数据集中的特征

target = data.iloc[:,-1] ## 提取数据集中的标签

scale = MinMaxScaler().fit(data)## 训练标准化规则

dataScale = scale.transform(data) ## 应用标准化规则

## 将数据划分为训练集测试集

data\_train,data\_test, target\_train,target\_test = train\_test\_split(dataScale,target,test\_size = 0.3,random\_state = 22)

##构建模型

svm = SVC()

#训练模型

svm.fit(data\_train,target\_train)

## 预测测试集的结果

target\_pred = svm.predict(data\_test)

#其中target\_pred为预测输出，查看该数组与target\_test是否相同