**《数据挖掘》**

**实验报告1**

**班 级 数212**

**学 号 21011171**

**姓 名 范鸿玥**

**批阅教师 钱夕元**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**评分标准：**

1. 成绩分为A/B/C/D四个等级（A为最高）
2. A：实验结果正确；过程进行了清晰的描述；作图清晰美观；对结果进行了详细的分析、讨论
3. B：实验结果基本正确；过程进行了部分描述；作图基本正确；对结果进行了简单的分析、讨论
4. C：实验结果部分正确；过程描述很少；作图不够正确；对结果的分析不够正确
5. D：实验结果出现明显的错误；过程混乱；没有结果呈现；对结果没有分析或分析完全错误
6. 没有按时提交作业、补交作业，成绩降一档
7. 抄袭作业成绩为零分。

**教师评语：**

|  |
| --- |
|  |

**实验成绩：**

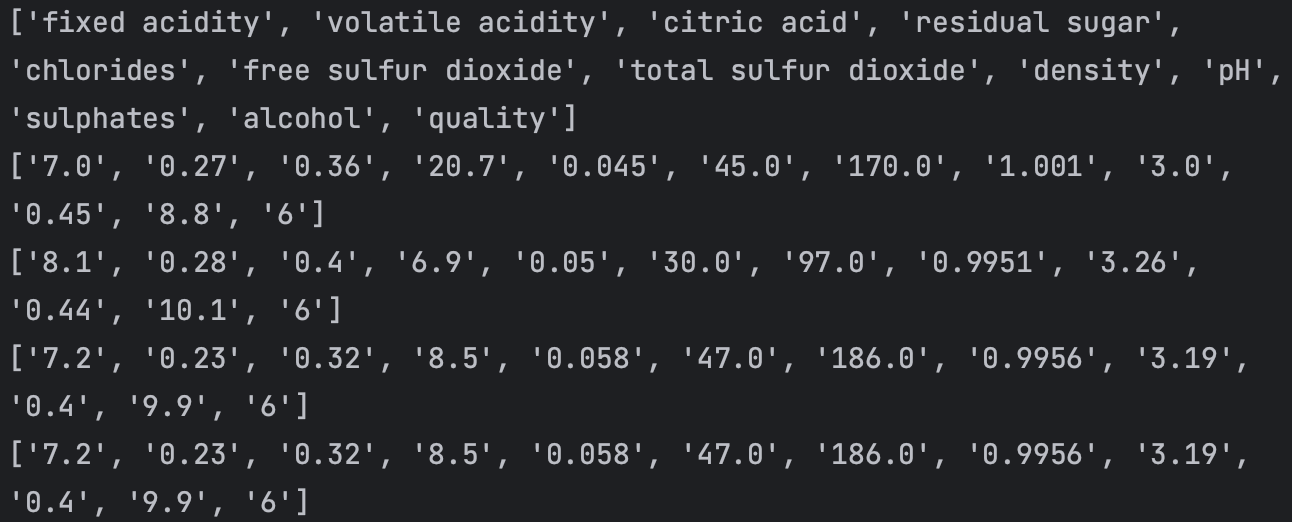
|  |  |
| --- | --- |
| 实验名称：Python及数据处理 | 实验地点：八教611 |
| 所使用的工具软件及环境：Anaconda | |
| 1. 实验目的：    1. 熟悉Python运行环境，安装anaconda，选定Python IDE工具    2. 熟练掌握Python数据分析基本操作    3. 熟悉Python数据分析常用的包    4. 完成2个实践案例的数据探索性分析； | |
| 1. 实验内容   （要求参照Python基础（上机操作）PPT内容及实践练习和实践案例要求，完成下述实验操作）   * 1. 完成实践案例：白葡萄酒品质探索   2. 完成实践案例：青年失业率探索 | |
| 1. 操作步骤、结果及讨论   **注：内容从下页开始。** | |

* **案例1：白葡萄酒品质探索**

葡萄酒的价格与其品质相关，下面根据数据对白葡萄酒品质进行分析。whine.xlsm中共有3431组白葡萄酒数据，各项指标（共12个）分别为：固定酸度、挥发性酸度、柠檬酸、残余糖、氯化物、游离二氧化硫、总硫度、密度、PH值、硫酸盐、酒精度数、品质。

* + **读取白葡萄酒品质数据集**

我们将存储在本地的数据集white\_wine.csv读取入内存中。引入csv模块，打开文件将数据保存于列表content中，打印content前5行。



* **查看白葡萄酒中总共分为几个品质**

品质quality变量在数据中是一个离散变量，而不是连续的，所以它只会有固定的几个等级。那么我们用Python中自带的集合set来查看白葡萄酒中总共的品质等级。使用集合set查看白葡萄酒总共分为几个品质，并将所有品质等级保存在集合unity\_quality中。其中，品质等级数据在最后一列。



* **按白葡萄酒等级将数据集划分为7个子集**

将数据按白葡萄酒等级quality进行切分为7个子集，保存到一个字典中，字典的键为quality具体数值，值为归属于该quality的样本列表.按白葡萄酒等级将数据集划分为7个子集，用字典保存每个子集，字典变量名为content\_dict，变量的关键词key为品质，值value为每个品质子集的数据列表。



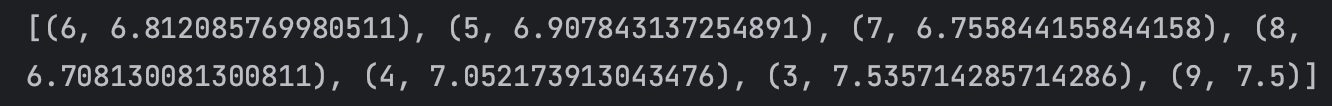
* **统计在每个品质的样本量**

统计每个品质下的样本数，保存为number\_tuple，该变量为一个列表，每个元素是一个二元元组，元组第一个元素是品质，第二个元素是样本数.



* **求每个数据集中fixed acidity的均值**

计算每个品质下变量fixed acidity的均值，并保存于列表mean\_tuple中。要求列表中每一个元组的第一个元素为quality，第二个元素为该品质下fix acidity的均值。

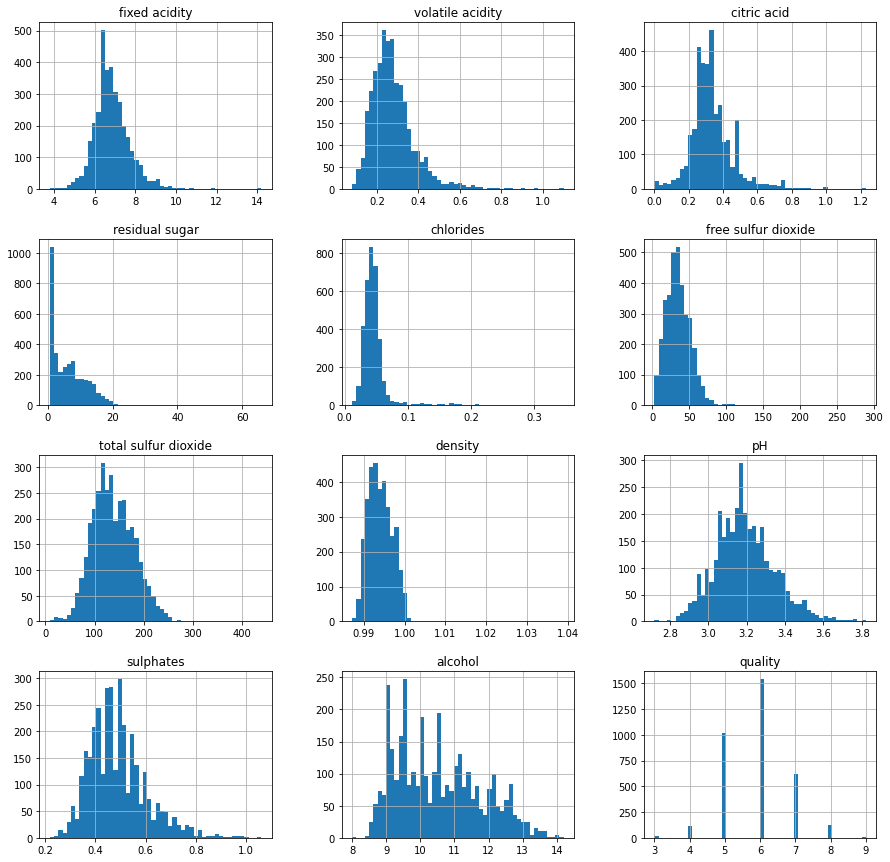


* + **补充：预测品质（随机森林）**

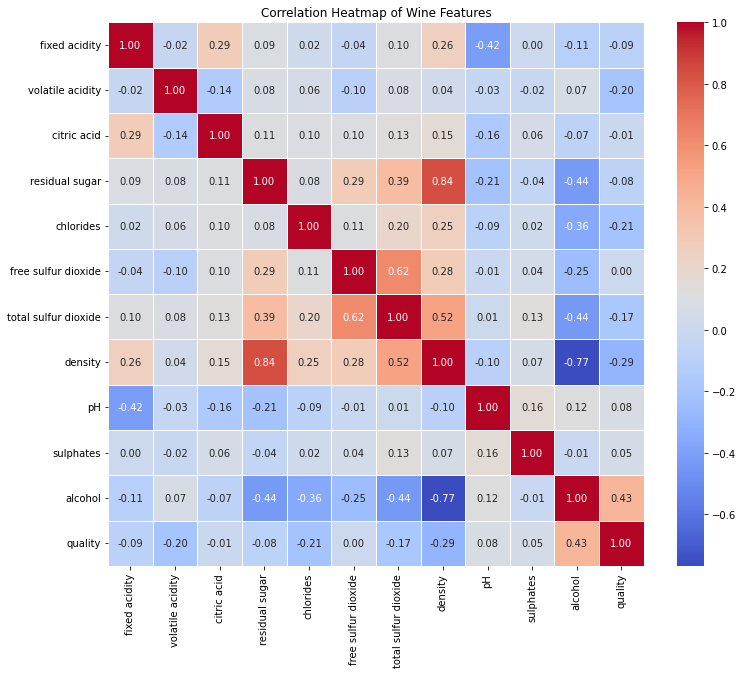
每种葡萄酒的品质得分从 0 到 10 不等，这里考虑一个二元分类问题，我们根据 11 个特征来预测葡萄酒的品质是高品质（葡萄酒品质6 分或以上）还是低品质（葡萄酒品质低于 6 分）。

* 1. **读取数据**
  2. **描述特征分布及各特征相关性**

描述特征分布，输出结果为：



描述特征相关性，输出结果如下，即酒精度数、密度、氯化物与葡萄酒品质的相关性更强。



* 1. **模型选择之5折交叉验证**

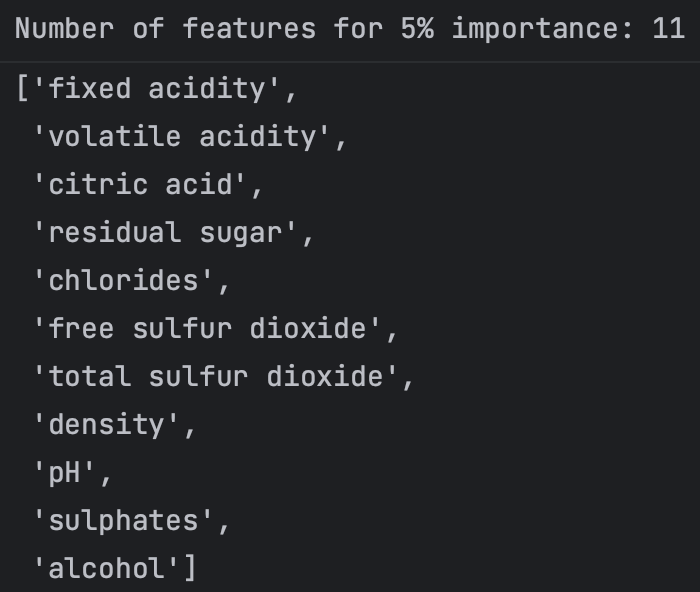
这里采用了Logistic regression、LinearSVC、Decision tree、Randomforest、XGBClassifier五种分类器去进行五折交叉验证。输出结果为：这5种模型的平均准确度分别为0.7534、0.7549、0.7566、0.8263、0.8196。因此确认采用随机森林去进行训练。

* 1. **随机森林**

这里用来评估随机森林模型中各个特征的重要性，使用平均减少不纯度和基于排列重要性两种不同的方法来可视化这些特征的重要性。

|  |  |
| --- | --- |
|  |  |

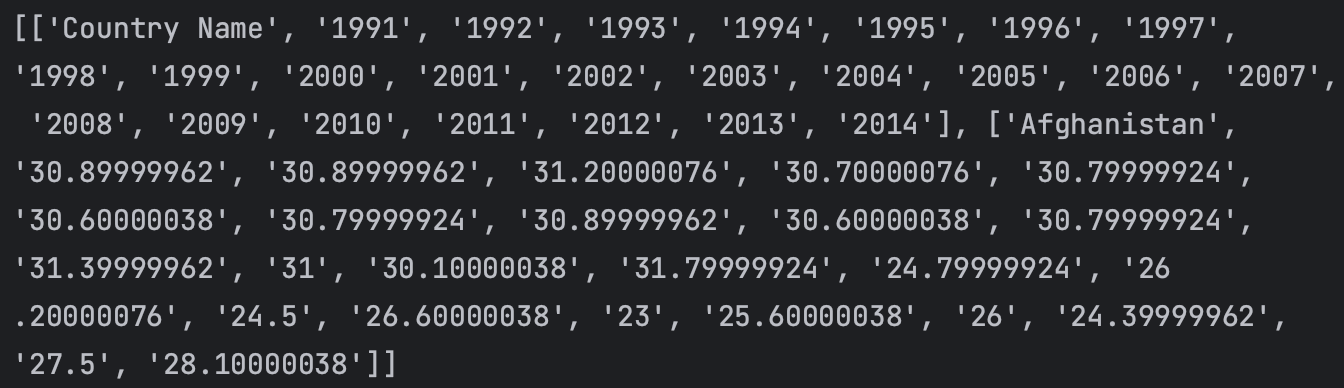
从特征重要性分析中筛选出重要性大于某个阈值的特征，并输出这些特征的数量和名称。



使用筛选出的重要特征重新训练随机森林模型，并评估模型的性能提升。结果表示：采用筛选出的重要特征训练出的随机森林模型精确度达到了0.8265，提升了0.02%。

* **案例2：青年失业率探索**
  + **将数据集读入内存**

引入模块 csv，创建函数 read\_csv() ，用于读取数据，其中函数的参数为 path，使用函数 read\_csv() ，将数据读入内存，分别将两个数据集保存到列表 unemploy\_male 及 unemploy\_female 中，打印前两行数据。

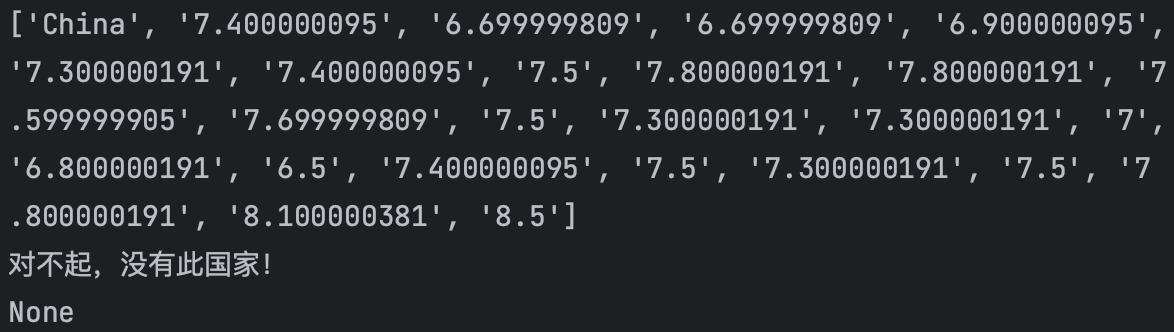
****

* **定义青年失业率类**

新建类，类名为 Youth。类内新建初始化函数 \_\_init\_\_，传入参数为 dataset ，指传入的数据集。添加类实例对象self.dataset，并赋值dataset。

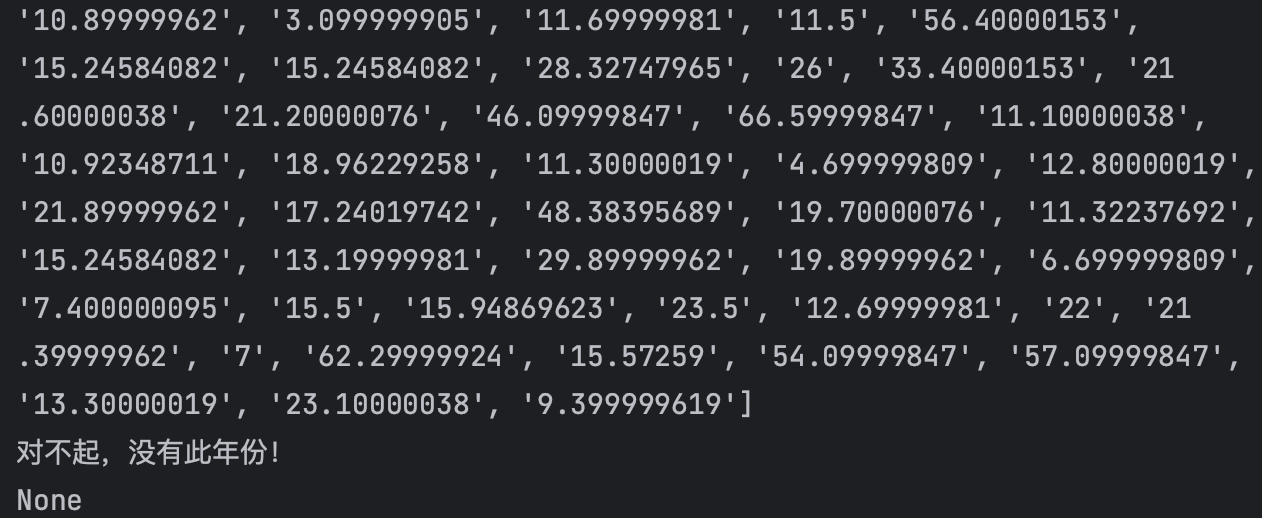
* **按名字选取数据**

在类Youth中添加函数，用于抽取该国家对应的数据。

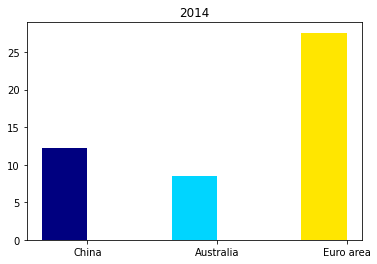


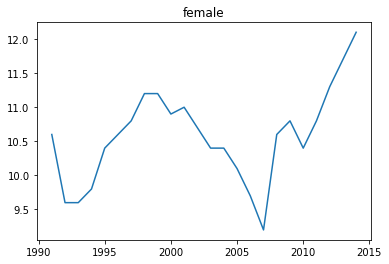
* **按时间选取数据**

在Youth类中添加按时间选择数据的函数。



* **比较不同国家不同性别的青年失业率**

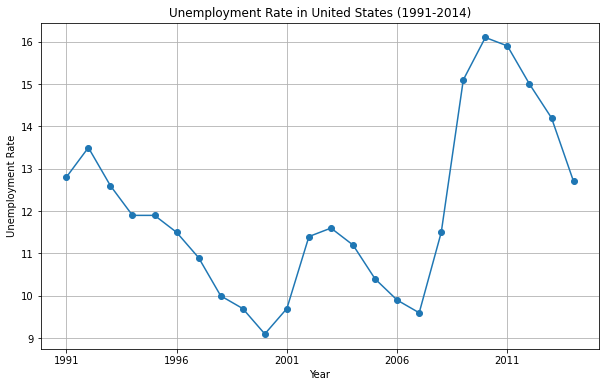




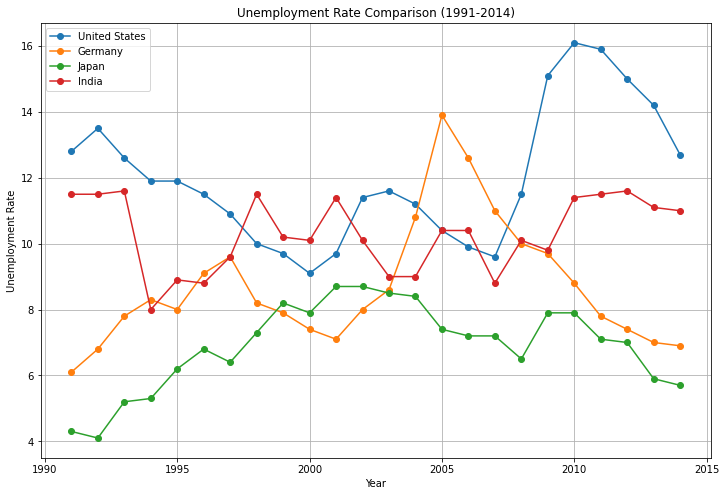
* **补充分析**

1. **数据分析**

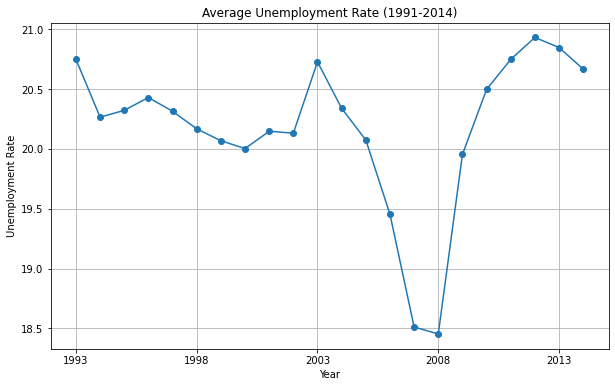
选择特定国家，查看其失业率数据，并描述其变化。这里以美国为例，如图所示，可以得知在1991-2014年，于2000年达到了最低的失业率。



也可比较不同国家的失业率差异，这里选取'United States', 'Germany', 'Japan', 'India'四个国家进行比较，并进行可视化。如图所示，这四个国家相比较而言，日本保持着一个较低的失业率，而美国的失业率相对较高。德国和印度位于中间位置。



再计算每年的平均失业率，并进行可视化：



由图可得，各国的平均失业率在2008年达到最低。

1. **时间序列分析和预测**

这里使用ARIMA模型预测下一年各国的失业率，这里以美国为例，使用其1991-2014年共24年的失业率数据来预测2015年的失业率。如图所示，预测出的2015年美国的失业率为11.14301024315765。

