

**2024年《机器学习》**

**工程报告**

**（个人版）**



**课 程：** 机器学习

**姓 名：** 党存远

**学 号：** 2022217587

**完成时间：** 2024.5.19

|  |  |  |  |
| --- | --- | --- | --- |
| **“机器学习-大作业”评分细则** | | | |
| **成绩等级** | **具体表现** | **教师评分** | |
| 优秀（100-90] | 报告撰写优秀，题目本身难度大，工作量饱满；Notebook代码清晰准确，实验与可视化优秀 | □ |  |
| 良好（90-80] | 报告撰写良好，题目本身难度较大，工作量较多；Notebook代码较好，实验与可视化良好 | □ |  |
| 中等（80-70] | 报告撰写中等，题目本身难度一般，工作量达标；Notebook代码一般，实验与可视化达标 | □ |  |
| 及格（70-60] | 能完成基本要求，工作量较少，非Notebook代码 | □ |  |
| 不及格（<60） | 未达最低要求或抄袭线上线下资源 | □ |  |

教师签名：

一． 自己所作工作的简介

摘要：介绍自己在本项目中的具体工作，比如分工的具体内容，使用方法，如何设计，实验结果与分析等。

在本次项目我担任小组组长，负责整个项目的选题、规划与成员间的任务分配。在项目中，我主要负责数据的分析、处理，以及神经网络模型和集成学习分类模型的构建与测试。具体而言，我利用pandas和numpy对数据进行预处理，使用matplotlib和seaborn进行数据可视化，确保数据特征与预测结果的相关性，确保数据集中无缺省同时对部分数据的标签进行更改保证数据的质量以及数据特征选择的准确性。同时我设计并实现了几种不同的机器学习模型，包括神经网络实现分类，通过多次实验迭代优化模型参数。实验结果表明，我们的模型在测试集上达到了较高的准确率，对于项目主题的预测具有较好的效果。

（说明：以下分内容，根据自己做的划分内容，写在下面的对应分项）

二、研究背景与意义

说明本工程所涉及的研究背景研究的意义，包括主要涉及领域，主要研究方法，主要存在问题，现有解决方案等等，可以自行补充其他内容。

我们选择的数据集是苹果品质的数据集。该数据集包含有关一组水果的各种属性的信息，提供对其特征的深入了解。该数据集包括水果 ID、大小、重量、甜度、松脆度、多汁性、成熟度、酸度和质量等详细信息。在现代农业和食品工业中，苹果质量的监测和控制是至关重要的。随着技术的发展，传统的人工检查方法逐渐被更加高效、准确的自动化检测技术所取代。苹果质量数据集的研究背景涉及农业、食品科学、计算机视觉和机器学习等多个领域。通过自动化检测技术，可以大幅提高苹果质量检测的效率，减少人工成本和时间，同时高质量的苹果质量检测可以保证市场上的苹果符合安全标准，减少食品安全隐患。

**主要研究方法:**

数据采集和整理：收集苹果的多维度数据，如重量、颜色、大小、甜度、酸度等。

统计分析:使用描述性统计和推论统计方法分析数据特征。

模型构建:通过使用机器学习常见模型对数据进行分析，最终模型用于预测出苹果的质量。

主要存在问题:

数据质量:数据的准确性和完整性直接影响分析结果。

数据多样性:不同品种和不同生长环境的苹果数据具有较大差异性。

模型鲁棒性:模型在不同条件下的稳定性和适用性需要进一步验证。

现有的解决方法:数据清洗：通过数据清洗和预处理，提升数据的质量和可用性。

多变量分析:利用多变量统计分析方法，深入挖掘数据之间的关系。

模型优化:通过交叉验证和参数调优，提升模型的预测能力和稳定性。

三、模型方法

**详细说明**本工程所使用的模型方法和理论等

在本项目中matplotlib.pyplot matplotlib.pyplotmatplotlib.pyplot利用上述数据处理以及机器学习库进行后续编程内容的实现。在本项目的数据预处理中使用了数据清理和数据类型转化将原始数据更换为更为适合模型训练的格式。之后在特征工程部分进行特征缩放提高模型收敛速度以及主成成分分析来减少降维技术，同时利用 imblearn.over\_sampling 进行不平衡数据的处理操作。

我训练时所用的模型有:

1. MLP多层感知机：一种经典的前馈神经网络，用于处理各种类型的数据，包括图像、文本和结构化数据。MLP通过多个隐藏层和非线性激活函数，能够学习和表示复杂的非线性关系。该模型构建的重点在于：在MLP中，数据从输入层开始，通过各隐藏层逐层传播到输出层。在每一层，神经元对输入信号进行加权求和，并通过激活函数引入非线性变换。激活函数是MLP的核心之一，负责引入非线性，增强网络表达复杂模式的能力。常用的激活函数包括：Relu,sigmoid以及softmax。
2. 卷积神经网络：一种深度学习模型，广泛应用于图像和视频处理等领域。CNN通过引入卷积层、池化层和全连接层，能够高效提取和处理数据的空间特征。卷积神经网络主要包括：卷积层(卷积层是CNN的核心组件，主要作用是提取输入数据的局部特征。卷积操作通过一个或多个卷积核在输入数据上滑动，计算局部区域的加权和。每个卷积核可以提取不同的特征，如边缘、纹理等。)，激活函数(激活函数引入非线性，使网络能够表达更复杂的特征。卷积层输出通常经过Relu激活函数),池化层(池化层用于对特征图进行下采样，减少特征图的尺寸，从而降低计算复杂度并提高模型的抗噪性。常用的池化方法包括最大池化和平均池化。最大池化选取局部区域的最大值，平均池化则计算局部区域的平均值。),全连接层(用于综合前面提取的特征并进行分类或回归任务。全连接层中的每个神经元与前一层的所有神经元相连，类似于传统的多层感知机),损失函数(衡量模型预测值与真实值之间的差异，是模型优化的目标。对于分类任务，常用的损失函数是交叉熵损失),以及反向传播(用于计算损失函数对每个参数的梯度，并使用优化算法更新权重和偏置。)
3. 残差网络: 通过引入残差块和跳跃连接，ResNet能够成功训练超过百层的深度网络，极大地提升了模型的性能。
4. 监督自编码器:由编码器和解码器两部分构成，编码器将高维输入数据压缩到低维潜在空间表示，解码器则将这个潜在表示重建回原始数据。自编码器的目标是最小化输入数据与重建数据之间的差异，通常使用均方误差作为损失函数,由于自编码器是一种无监督的学习方法因此我在编码器中加入了分类任务。
5. 集成学习: 种通过结合多个基学习器的预测结果来提高整体模型性能的方法。集成学习方法通过将多个弱学习器组合成一个强学习器，从而在准确性和稳健性上优于单一模型。在我构建的集成学习模型中我使用基于投票分类器的基模型结合了SVM，决策树以及K临近，对于本次的二分类我使用了逻辑回归作为元模型进行训练训练。

四、系统设计

系统的**详细**设计，系统流程，系统的每一步的具体流程，例如，如何处理语料，如何训练模型，如何测试模型，最后对模型进行评估等。

对于该数据集，我们的目的旨在机器学习模型准确地分类苹果的质量。该系统整合了数据预处理、特征工程、模型训练、测试和评估等步骤，使用多种模型来实现高精度的分类效果。以下是这个系统的构建:

1、首先进行数据预处理：

a.数据的加载与检查:首先从CSV文件中加载苹果质量数据，使用Pandas进行初步的数据查看，包括基本信息（如列名称、数据类型和统计描述）。通过df.info()和df.describe()方法，我们能够获取关于数据集的初步统计概览，例如各特征的计数、均值、标准差等。

b.之后进行数据的清洗操作:我们的系统数据清洗的步骤包括处理缺失值、删除重复项以及转换数据类型。例如，将'Acidity'列从字符串转换为浮点数，将'Quality'列的分类标签（如"good"、"bad"）转换为数值型标签。这一步骤是通过自定义的clean\_data()函数实现的，确保数据的一致性和适用性。

c.最后在数据处理阶段进行了数据的可视化:使用Seaborn和Matplotlib库对清洗后的数据进行可视化，通过直方图以及散点图展示各特征的分布情况。

2、之后进行数据的重采样以及缩放

a.使用SMOTE在少数类的样本之间插值生成新的样本。这有助于平衡类别，改善模型在少数类上的预测性能。

b.特征缩放:使用归一化，确保所有特征在相同的尺度上，使得模型训练更加稳定和快速。

3、进行数据的降维处理

a.通过PCA对数据进行降维处理以适合模型

4、选择合适的模型进行训练

a.使用交叉验证等技术来训练模型，确保模型在未见数据上的泛化能力,选择合适的损失函数和优化算法，调整模型参数。

b.通过网格搜索、随机搜索或贝叶斯优化方法进行超参数的调优，以找到最佳的模型配置。

c.通过常见的有监督训练模型对模型进行训练和测试，我选用了神经网络模型中的MLP,CNN,残差网络以及使用集成学习进行训练和测试，经验证发现其在测试集上的准确率较高。

5、结果的可视化和解释:

a.通过ROC曲线，混淆矩阵以及学习曲线进行结果的可视化以及解释。

五．实验结果分析、对比和讨论

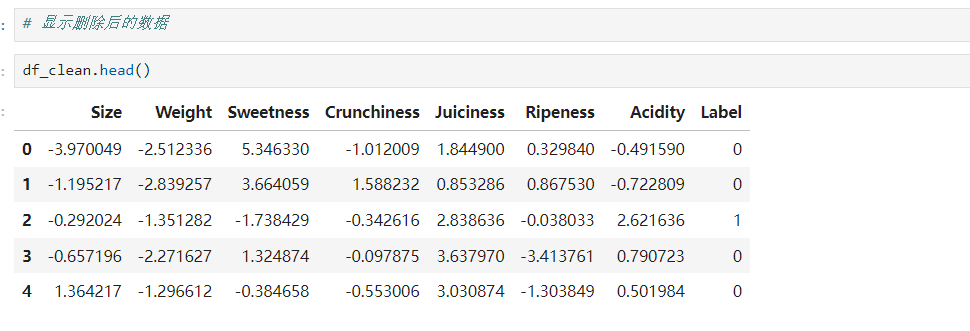
对实验结果进行分析说明，测试一些课程中演示的样例，根据结果说明为什么对或者为什么错等具体分析；对不同模型和参数的对比进行分析；最后对系统提出改进方案等。

特征工程部分

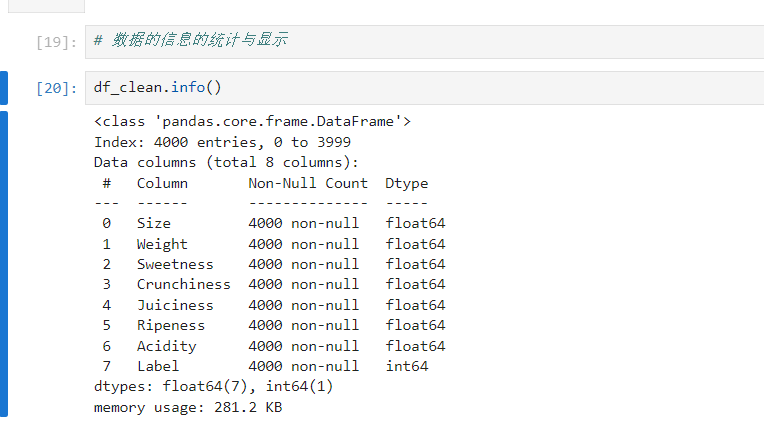
数据清洗



删除特征后的数据显示



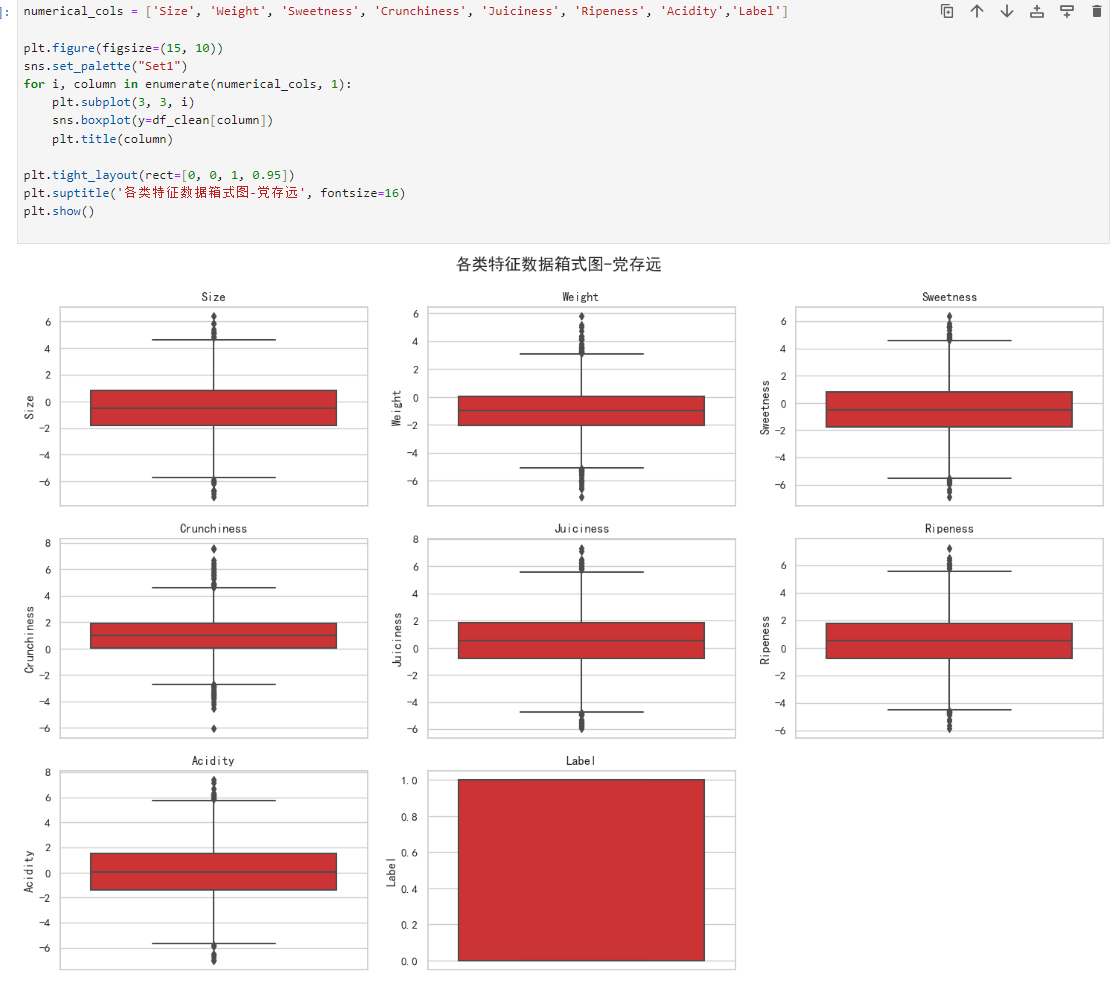
数据特征的显示



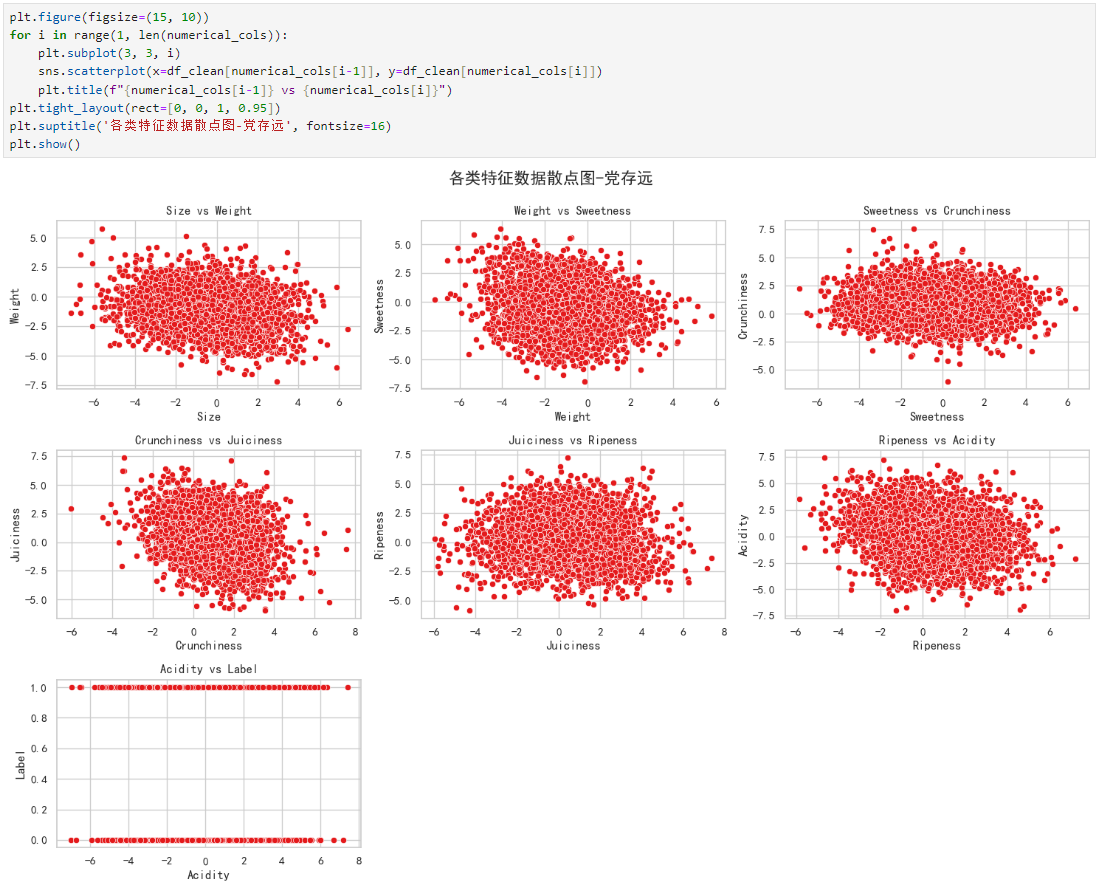
各类数据特征



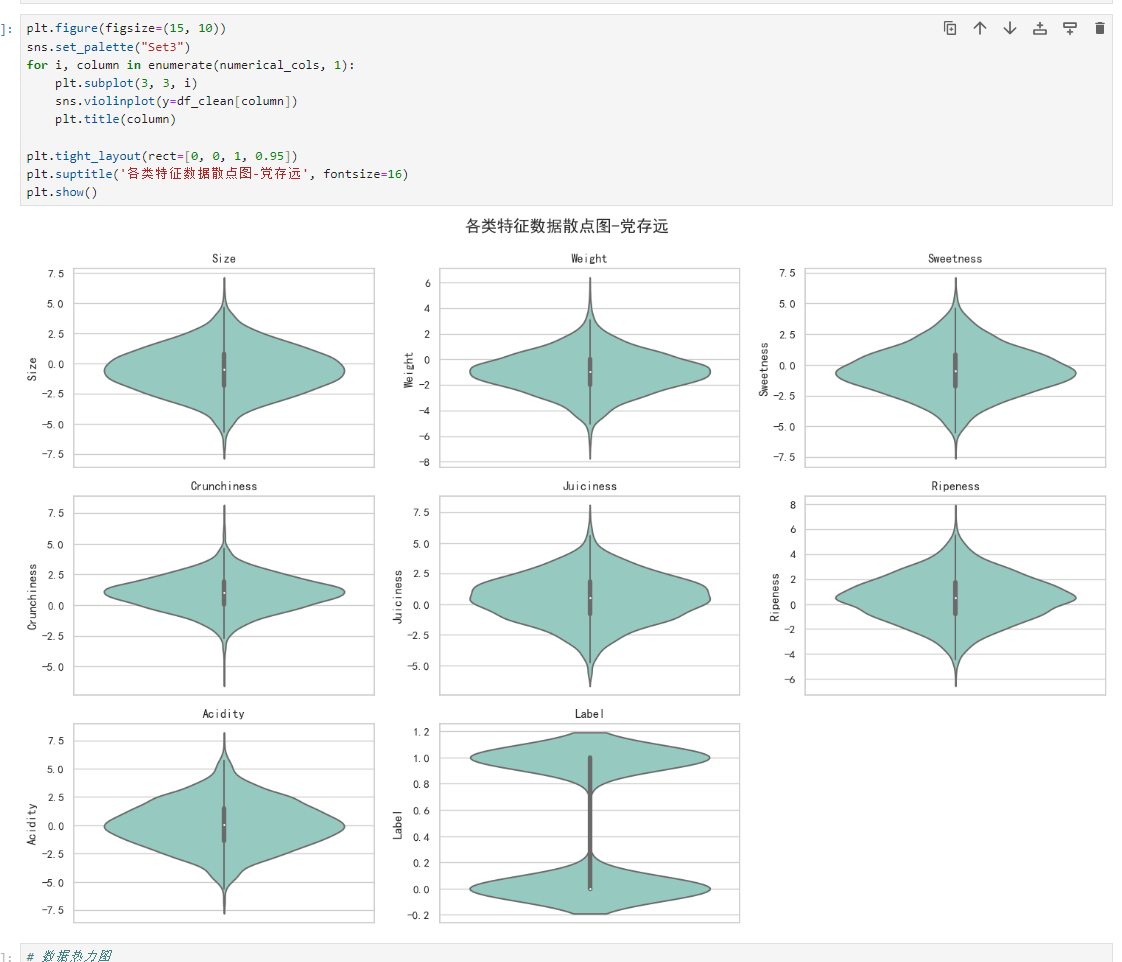
各类数据箱式



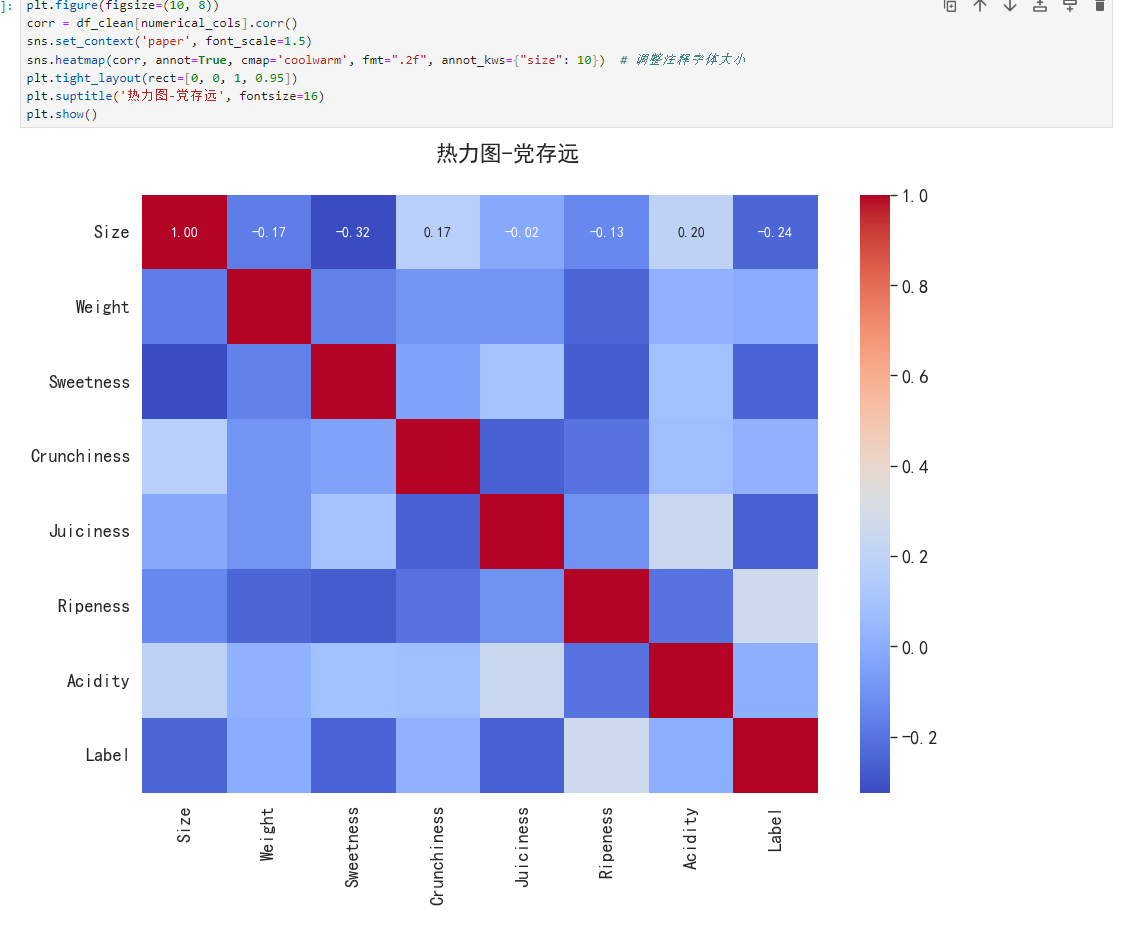
各类数据散点



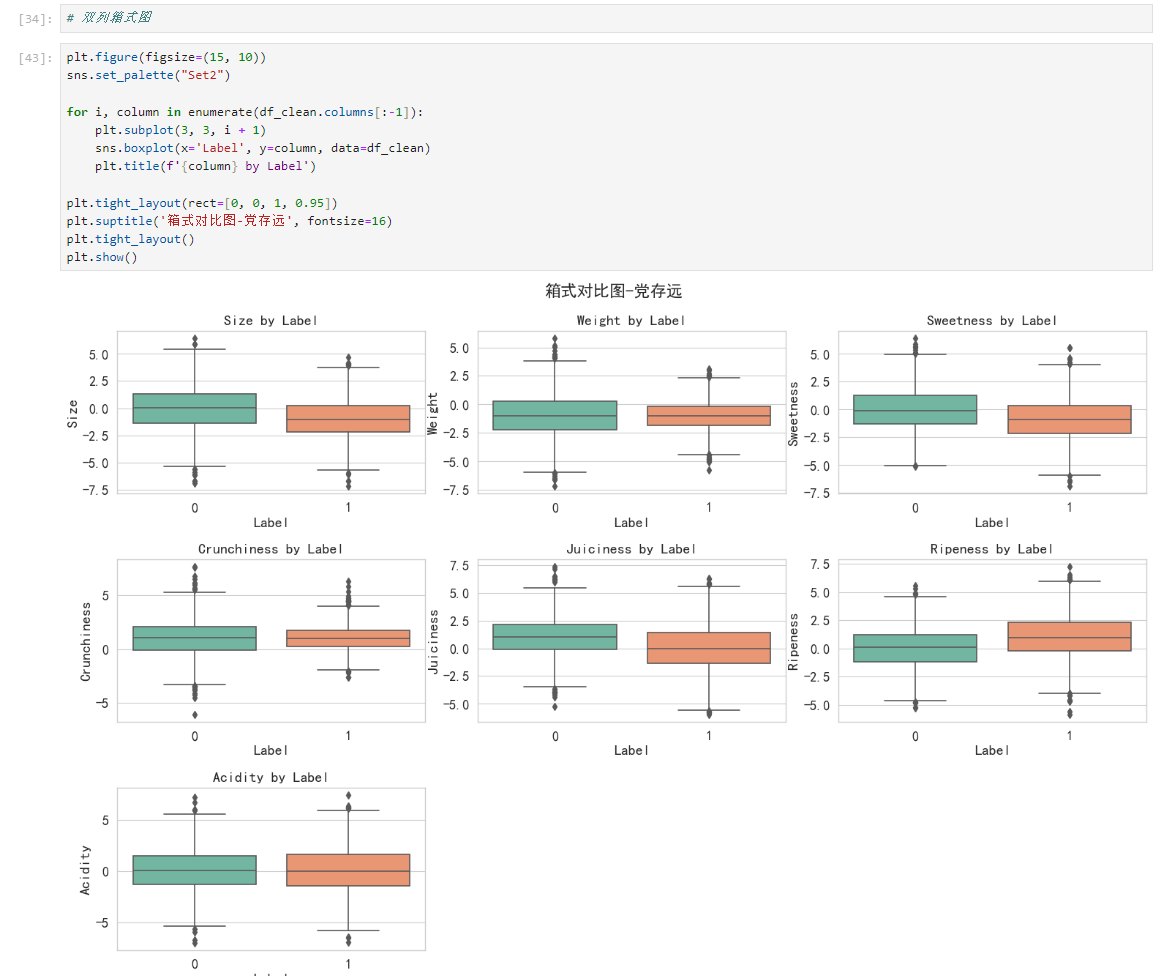
各类数据小提琴



数据热力



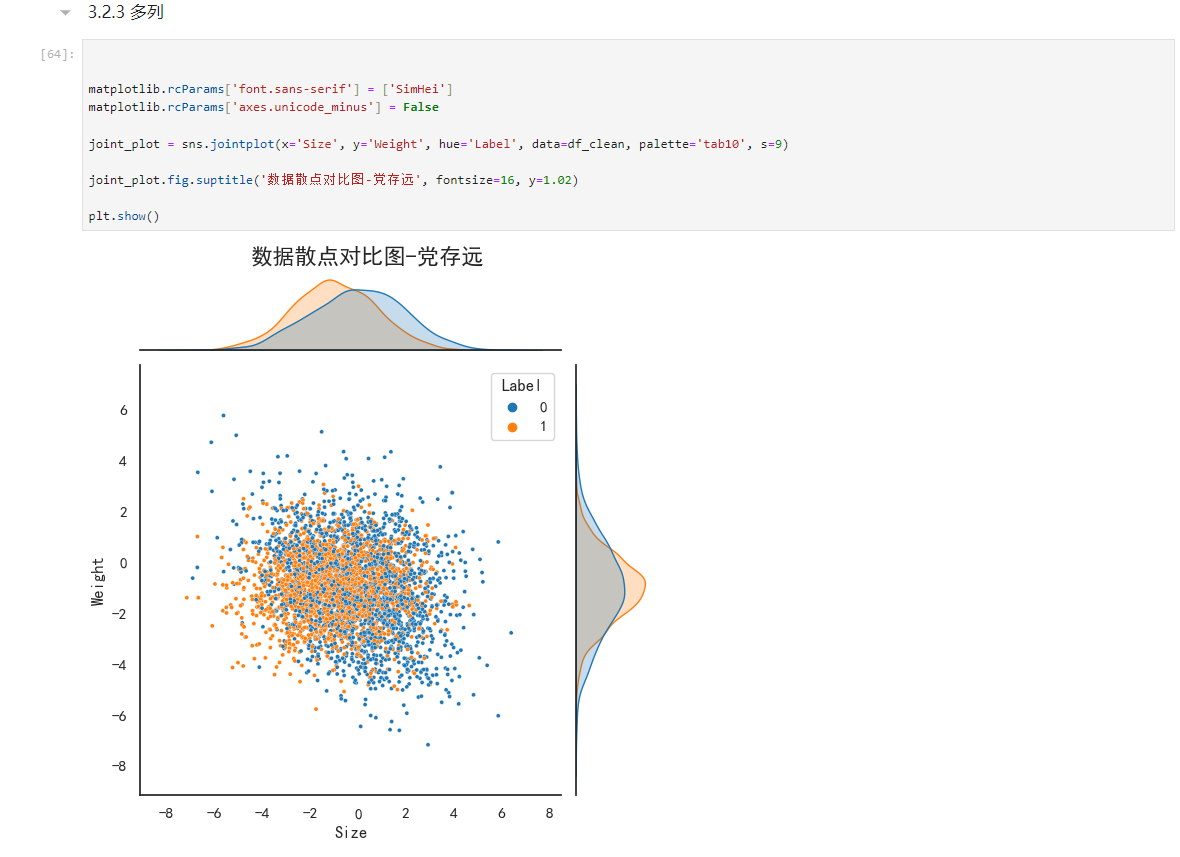
数据特征箱式对比



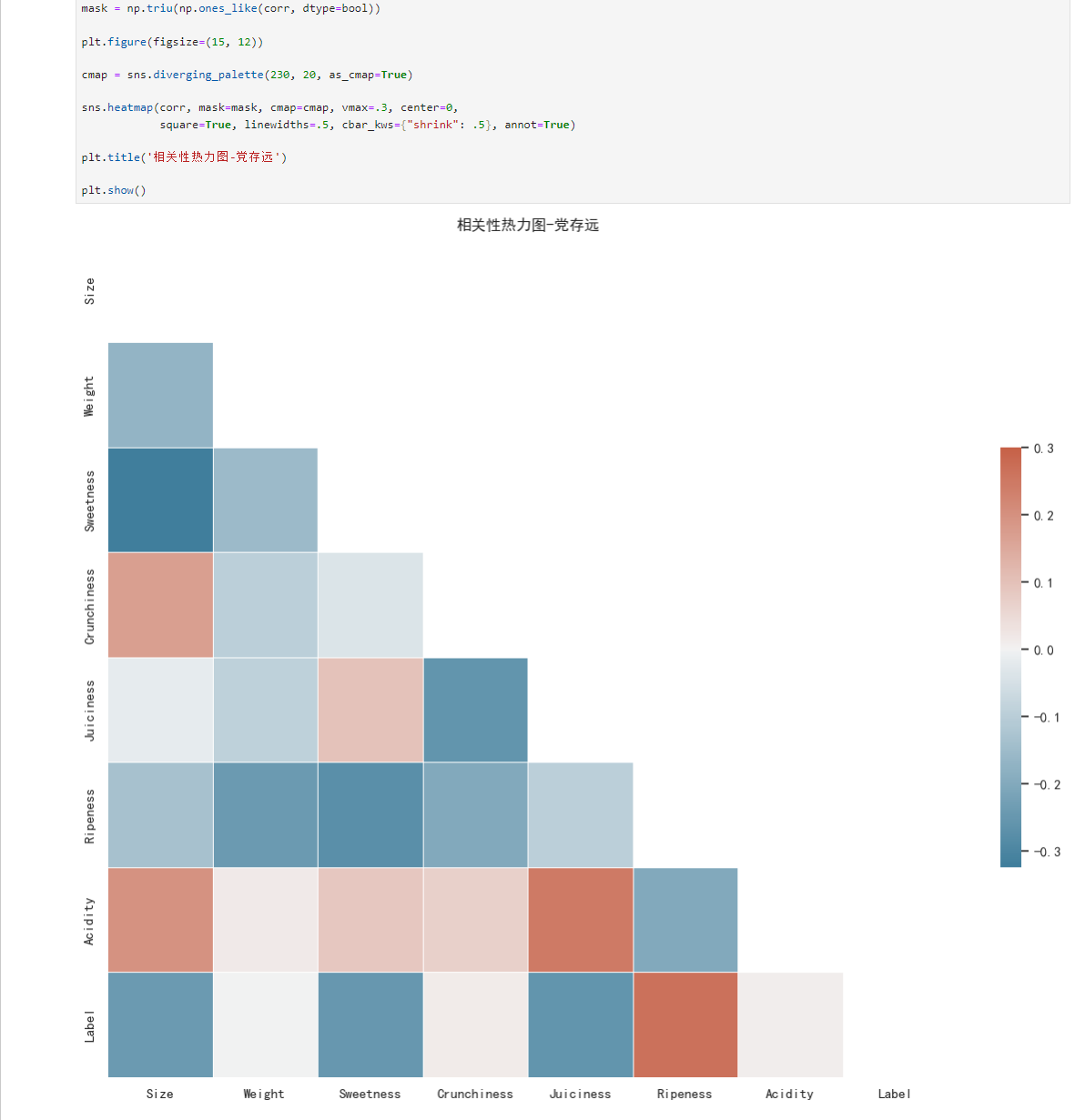
数据散点



数据对比散点



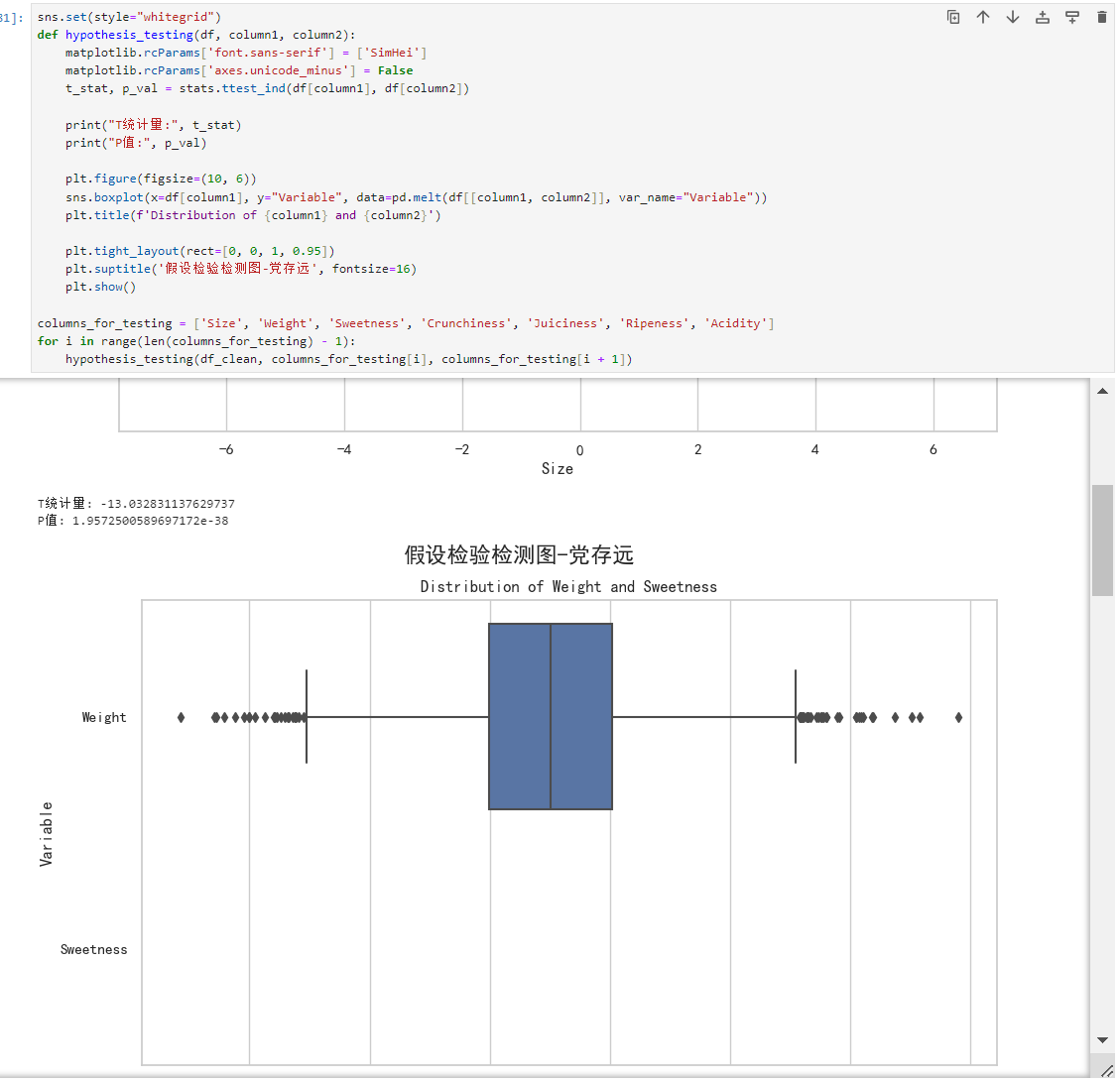
相关性热力



异常处理



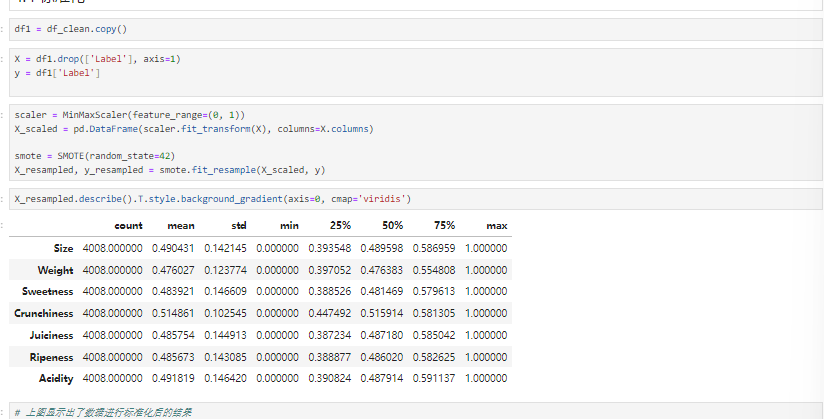
假设检验



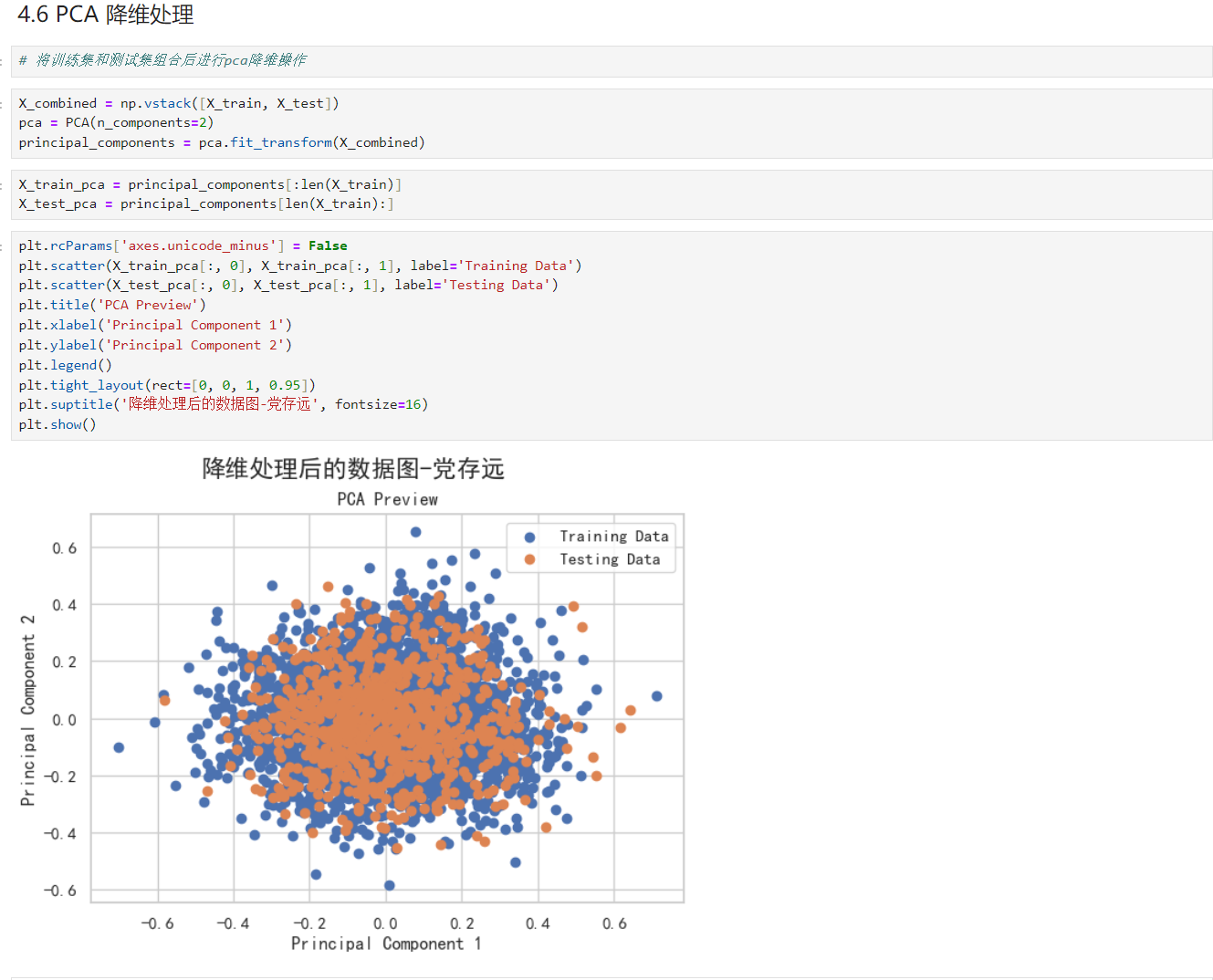
不平衡数据处理



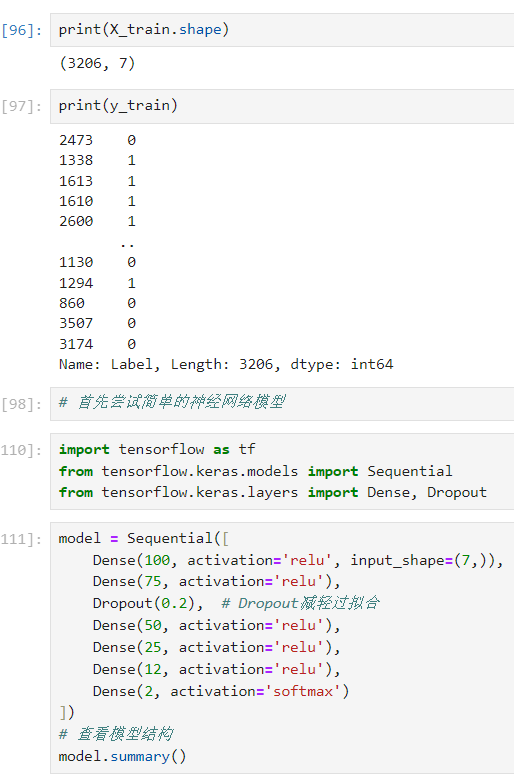
标准化

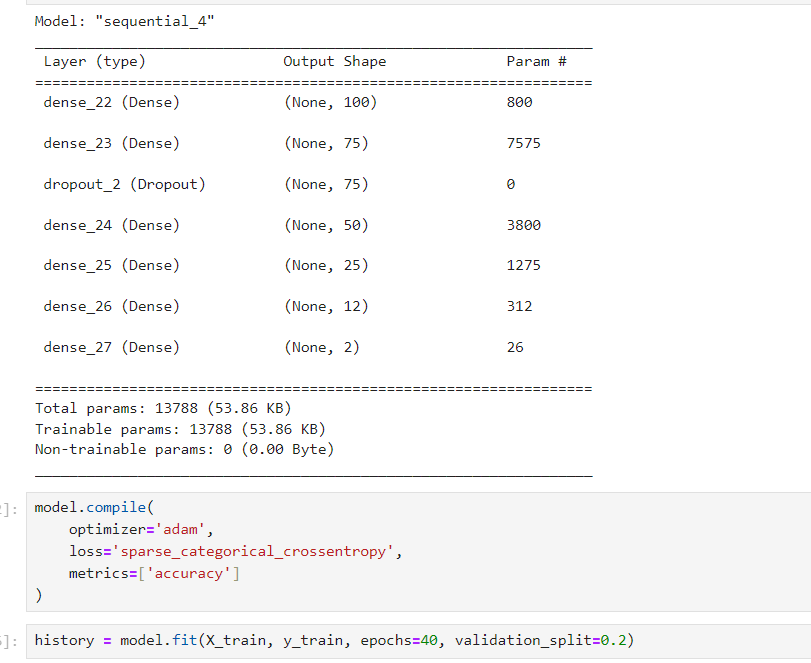


降维处理后的数据图

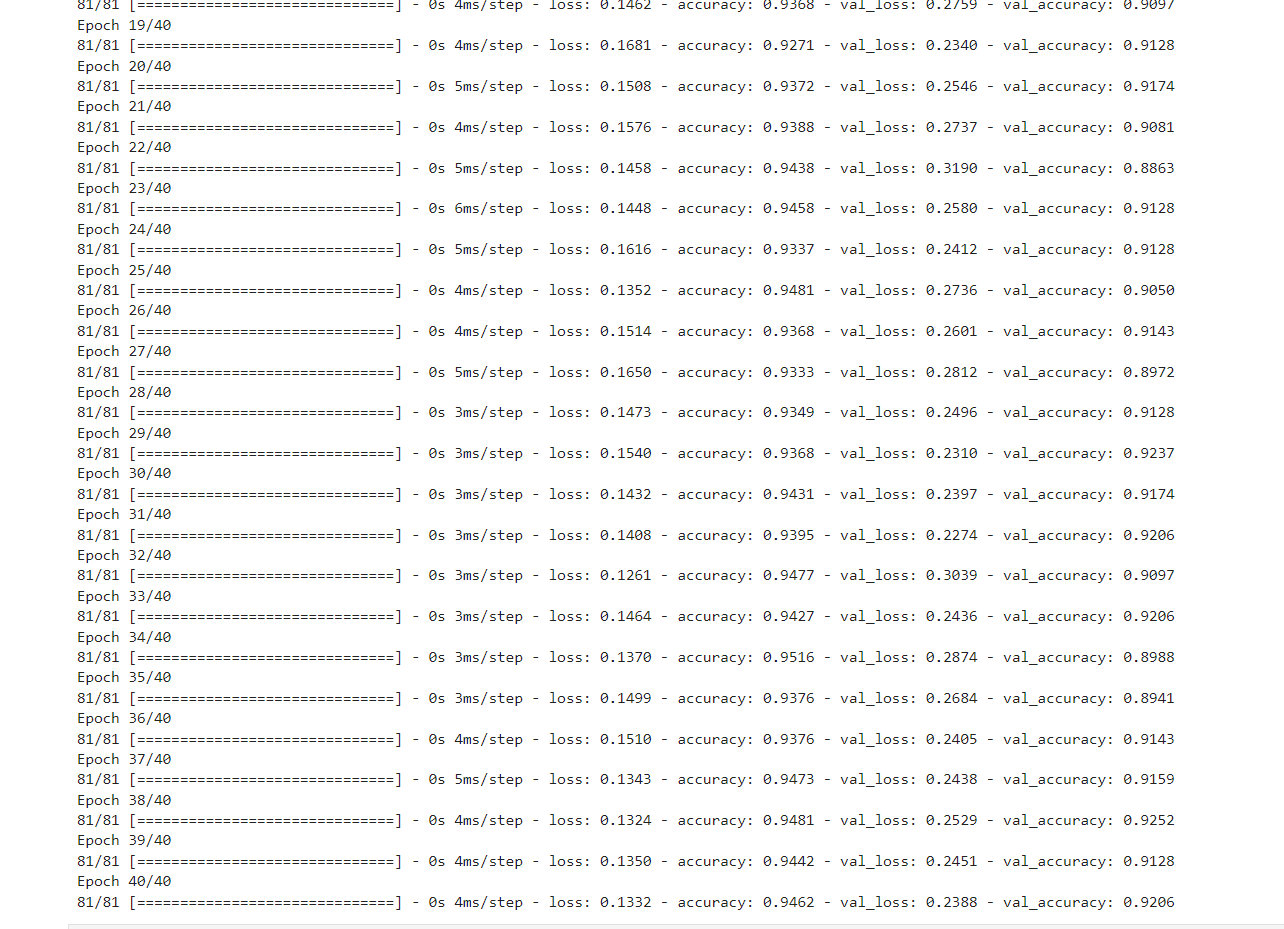


MLP的搭建

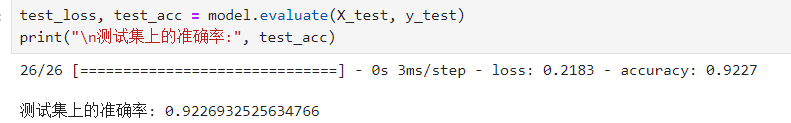




训练过程



准确率

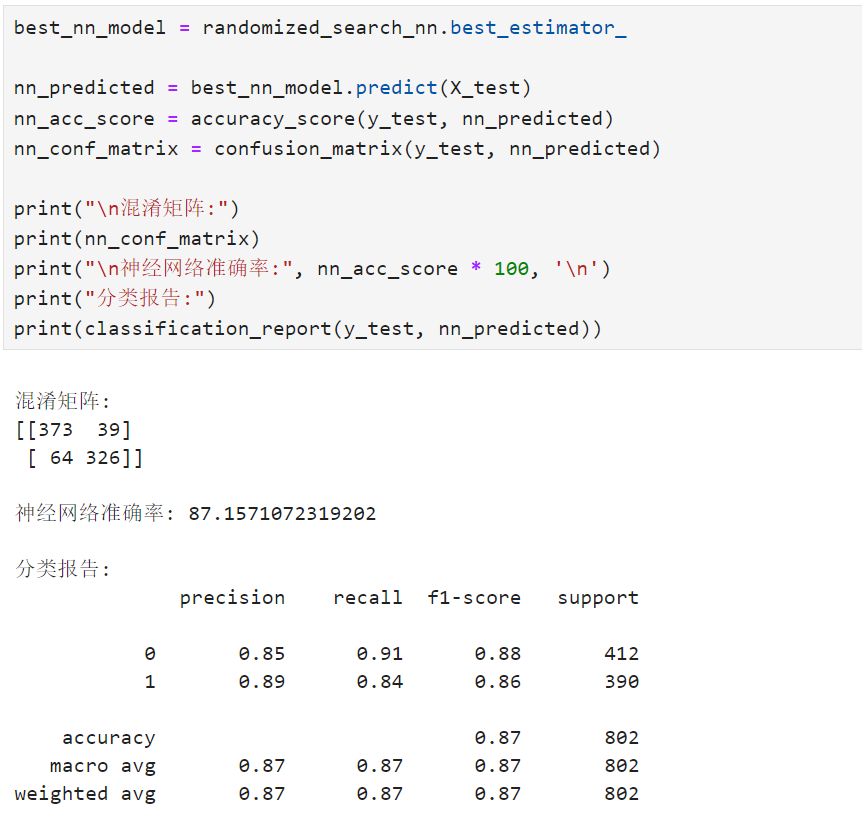


如图所示普通前馈神经网络在测试集上的准确率达到了0.92 可以看出该神经网络的泛化能力较强，由于该种较为简单的神经网络层数和结点数量多，进行递归的次数多，同时使用了Dropout层减少过拟合保证了该模型较优，该类模型适用于像本项目种所用的数据集的特征数量较少但是数据量较大的情况。

MLPClassifier

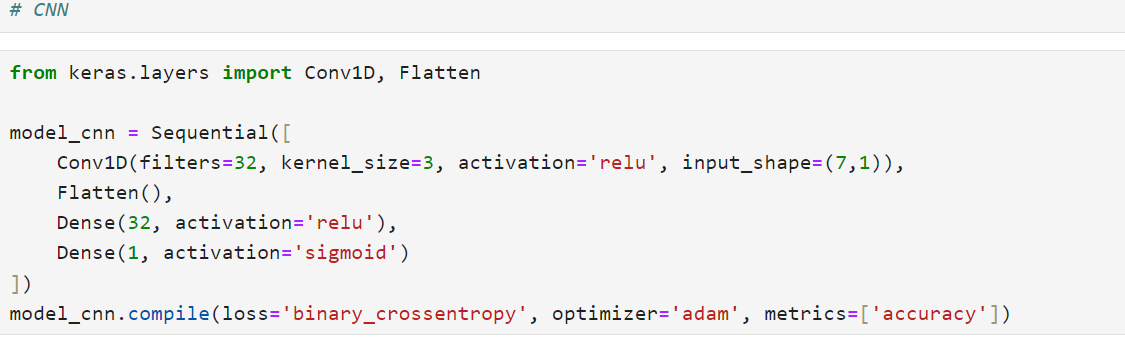


准确率



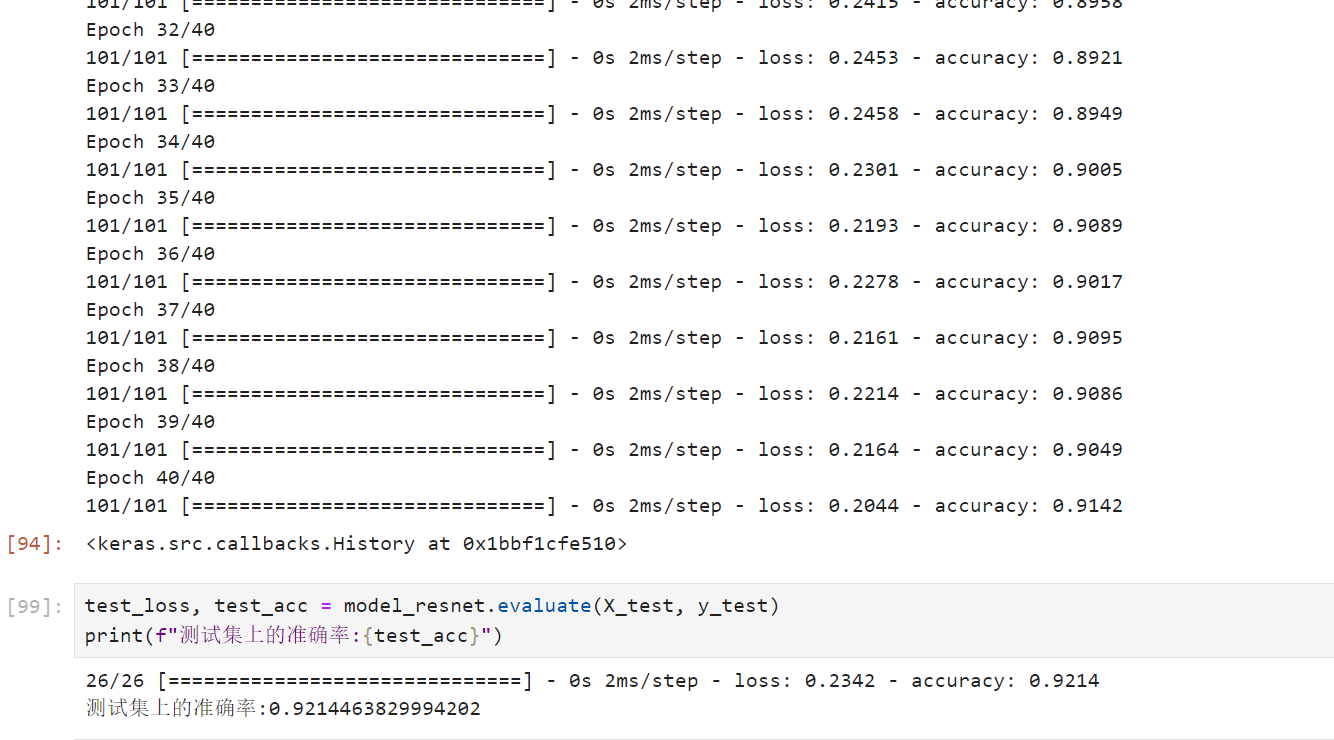
通过测试集的准确率可以发现，该种模型相较于MLP前馈神经网络，模型的效果变差，这是由于我们小组所选择的数据集类型是表格形数据类型，一般该种类型下，最为适合的模型是MLP前馈神经网络。

卷积神经网络



可以看到我创建了一个非常简单的卷积神经网络，通过设立卷积层，展平层以及全连接层和输出层来实现模型的搭建，并且在全连接层中利用relu作为激活函数。

训练过程



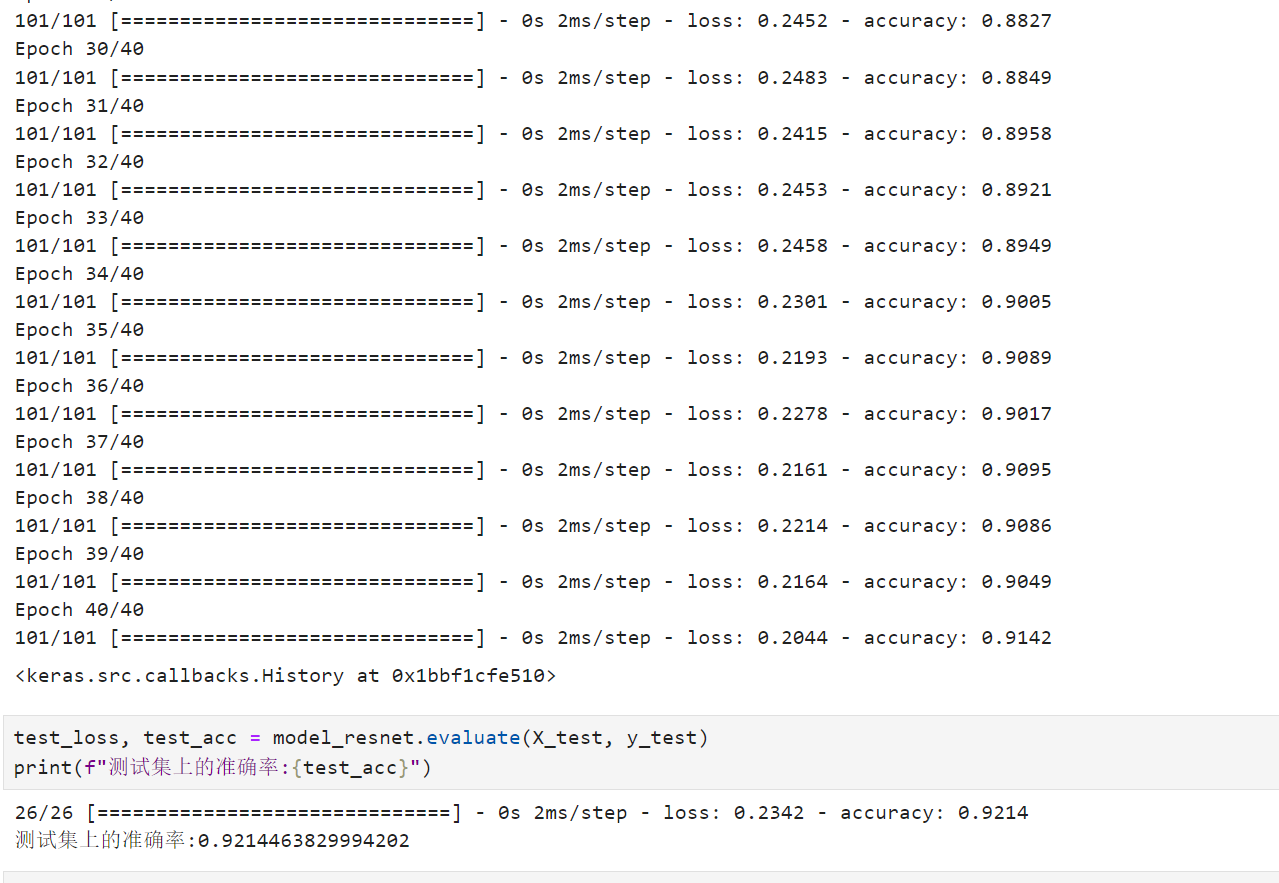
可以看到经过多次训练，数据集在卷积神经网络中的准确率较高。

残差网络



在构建残差网络的过程中，我首先创建了输入层并且设置了网络的输入层，之后我设计了全连接层和残差块部分，使用Relu作为激活函数。

模型的训练



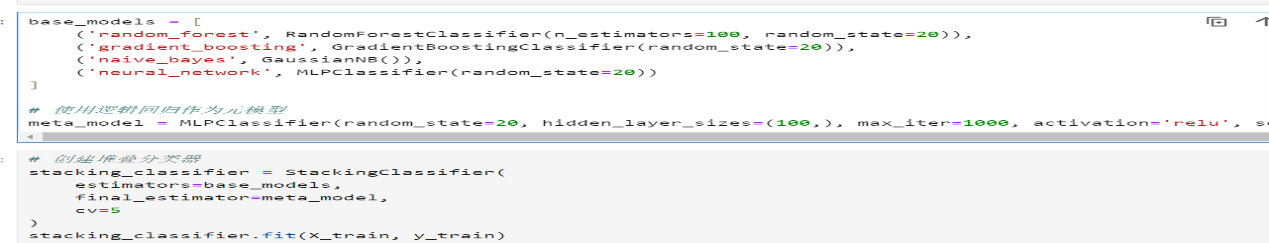
观察模型在数据集上的运行结果可以看到，本模型在测试集上的准确率较好，说明该模型的表现较优，分类能力较强。

集成学习模型构建

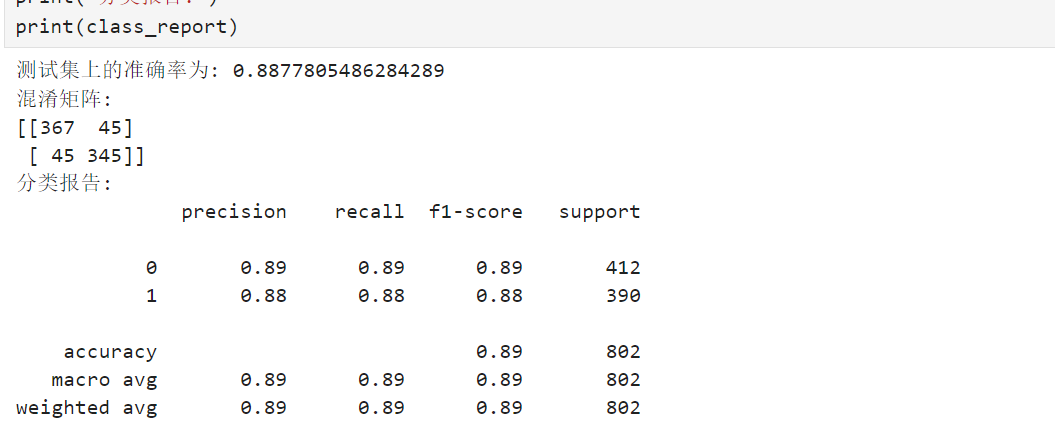


通过堆叠分类器来构建模型，使用逻辑回归,朴素贝叶斯以及adaboost作为堆叠分类器中的元模型，对于该二分类问题，选择逻辑回归作为基模型进行训练。

训练过程:



准确率

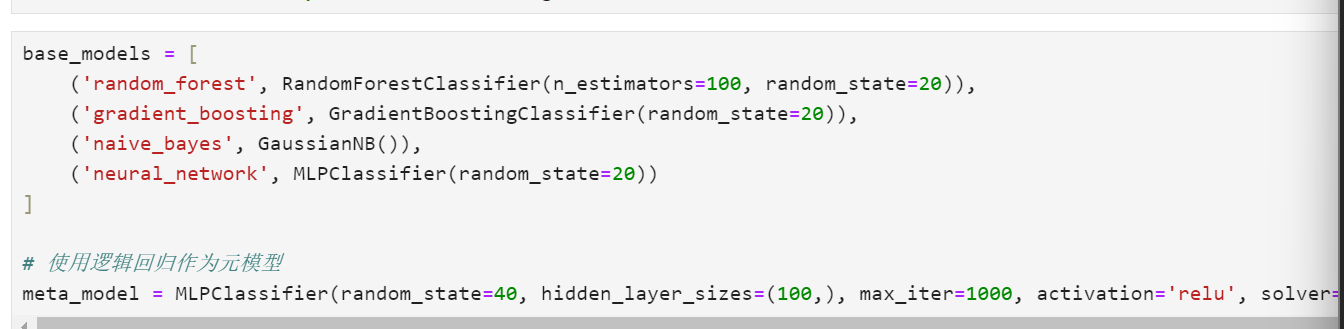


分析该模型导致模型的准确率发生变化的原因是基模型个和元模型。

由于我使用的逻辑回归，朴素贝叶斯不适合该类数据集。

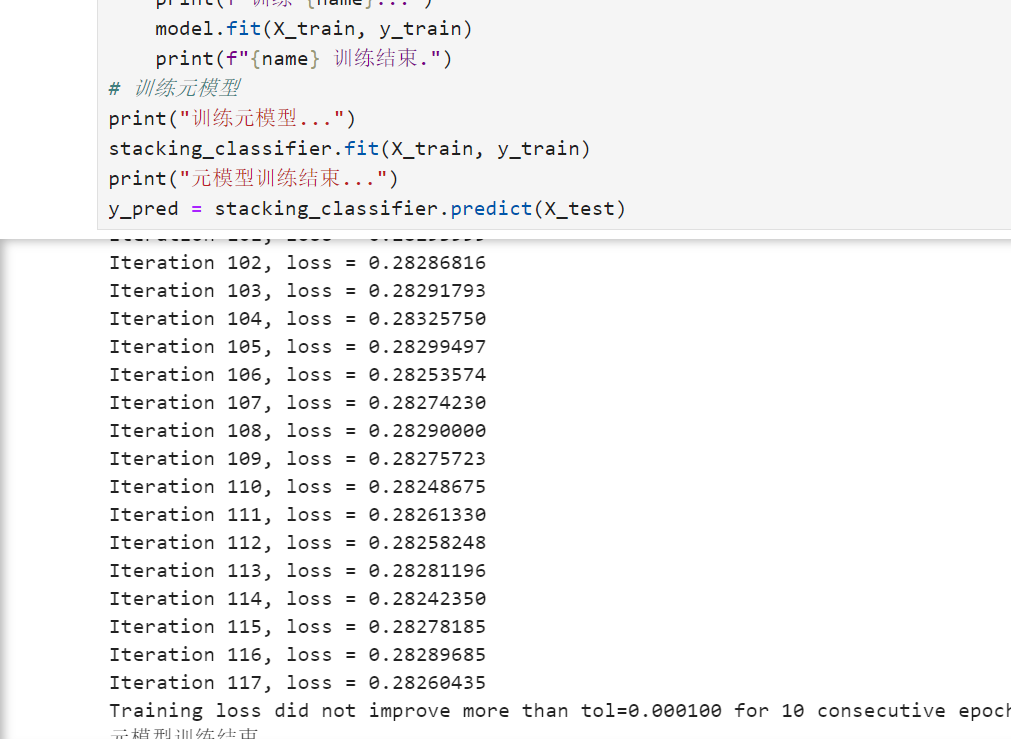
尝试使用不同的元模型和基模型进行修改和测试。

集成学习模型构建



该模型使用随机森林，朴素贝叶斯以及神经网络作为基模型，对于二分类使用用逻辑回归作为元模型。

集成学习训练过程及准确率





如图所示，通过更改基模型提升了模型的准确率，原因是神经网络，逻辑回归更适合作为训练二分类的基模型类型。

六．其他

上面五部分之外的内容，写在这里。

通过这次项目，我对机器学习有了更加深刻的理解。我了解到了数据处理对于项目的后续进程的重要性，数据的质量直接影响到模型训练的结果，合适的数据清洗和特征处理是成功应用机器学习算法的前提。同时数据处理时需要通过可视化对数据的特征进行更深入的了解。在数据清洗过程中，我学到了处理缺失值和异常值的重要性。在本项目中，我选择移除了包含缺失值的行，因为数据集足够大，这样处理不会导致信息的严重丢失。此外，移除不必要的列也是清洗的一部分，因为这些列对于建立模型并没有帮助，反而可能引入噪声。在实际工作中，正确的处理缺失值和无关特征对提升数据质量至关重要，可以防止模型学到错误的关联，提高模型的泛化能力。在进行数据处理时，我尝试了多种方法并对比了它们的效果。这个过程是一个不断试错和学习的过程。例如，在标准化数据时，我比较了MinMaxScaler和RobustScaler的效果，选择了最适合当前数据集的方法，之后我了解到不同的模型有着不同的假设前提和适用场景，并不是越复杂的模型就一定会在测试集上表现更为优异，了解并选择合适的模型对于解决特定问题至关重要，对于不同场景一定要选择不同的模型进行训练否则会适得其反导致模型效果变差。最后，通过本次项目，我不仅提升了自己在数据处理、模型构建和评估的技能，也加深了对机器学习理论和实践应用的理解，特别是在模型评估和选择上，我学到了如何根据实际问题灵活调整策略和方法，比如在神经网络的搭建中，如何通过修改神经元数量以及神经网络的层数和激活函数的种类，迭代的次数来修改模型以得到表现更加优秀的模型。

七．对本门课的感想、意见和建议

希望听到你的心声。

对于本课程我认为课程内容丰富，难度也较大，同时讲课的速度也较快，在监督学习的部分花费了大量的篇幅，但是在无监督学习的部分讲解较少，希望可以增加这一部分的讲解，同时也希望能够降低讲解速度使学生有更深入的了解。