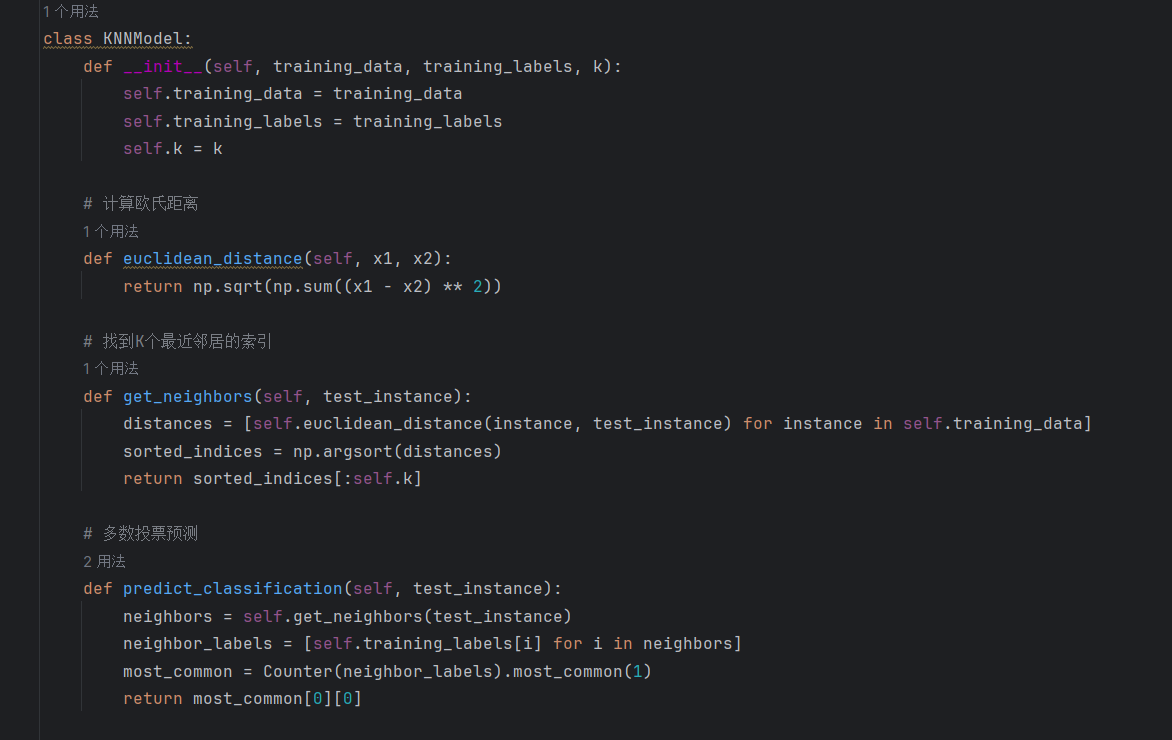
实验一：鸢尾花分类

1. 实验要求
2. 分类问题：鸢尾花分类
3. 根据萼片长、萼片宽度、花瓣宽度找到鸢尾植物的类别
4. 使用Accuracy作为评估指标
5. 实验思路
6. 多类特征分类，使用KNN（k近邻分类器）完成分类任务，因为特征数据的差别不大，且没有缺省值等，所以不做数据处理。
7. 使用KNN进行分类，我尝试了自定义函数和调用库函数，分类的结果差别不大。最后对结果进行Accuracy评估，输出结果。
8. 实验步骤
9. 加载数据集
10. 自定义KNN分类器，创建类，并定义相关函数



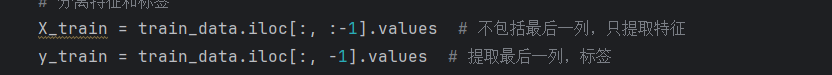
函数euclidean\_distance(self, x1, x2) 用来计算两个特征点之间的欧氏距离；

函数get\_neighbors(self, test\_instance) 用来寻找最近邻居的索引

函数predict\_classification(self, test\_instance) 用来进行预测，他会记录出现最多的标签并把这个标签作为预测结果。

1. 特征标签分离

以训练集为例：

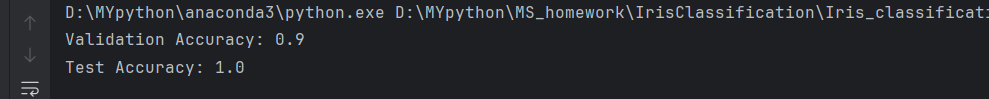


1. 模型分类并评估

因为我定义的KNN每次只能进行单个测试用例的预测，所以使用时要进行一个循环，依次对每个测试案例进行预测，并对预测正确的个体数进行递加操作，并以此结果进行Accuracy的评估。



1. 实验结果



1. 实验心得

多特征分类问题，这个之前在数据挖掘课上学过，做起来并不是很困难，最初我是调用的函数 KNeighborsClassifier() 直接进行分类。发现分类的结果不错，然后再转去搞自定义的函数来进行分类，一开始只是定义了各个函数，然后进行调用。但是因为要保存模型，所以去搜了一些资料，了解到要创建class类来进行函数包装，这里花费了比较多的时间。

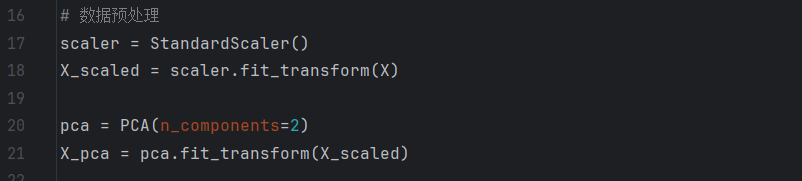
实验二：乳腺癌良恶性标签聚类评估

1. 实验要求
2. 聚类问题：乳腺癌良恶性标签聚类
3. 在这组数据中，只有病人的病理特征，而没有病人的乳腺癌是良性还是恶性的标签信息，请用聚类算法对数据进行聚类，用聚类结果对数据进行标签。
4. 使用标准互信息（NMI）作为评估指标。
5. 实验思路
6. 因为是对乳腺癌预测，病理特征的指标大有不同，所以需要对数据进行预处理，这是必要的。
7. 聚类考虑使用K-means聚类算法，因为只聚成两类，可能效果欠佳。
8. 模型训练完成后NMI进行评估。
9. 实验步骤
   1. 数据集读取和预处理
      1. 数据集读取

因为原数据的行数是打乱的，不方便观察出时间关系，所以在读取文件前我将数据进行了重新排序并保存在了新文件中，除此外未做任何处理。随后读取排序后的文件。

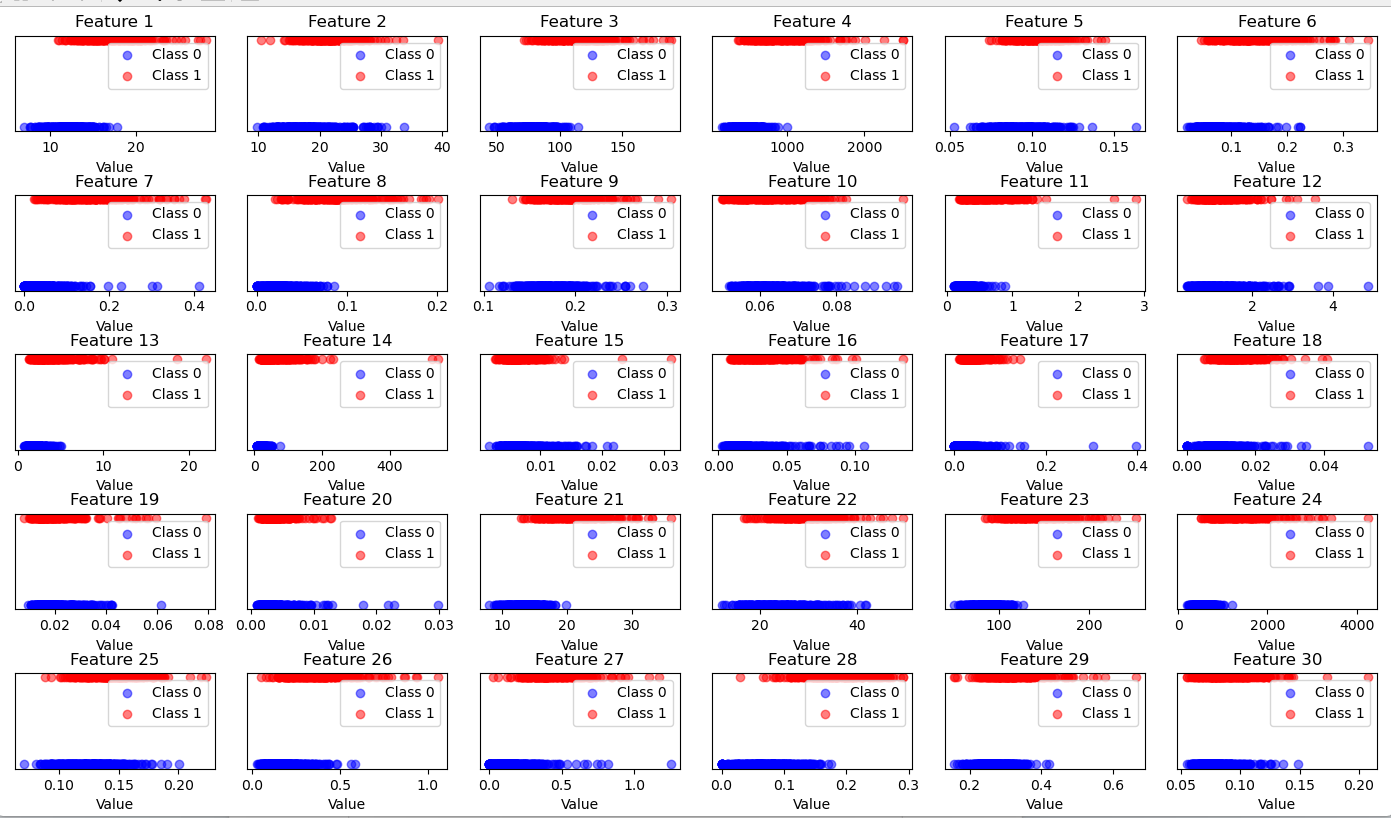
* + 1. 数据预处理

数据涉及多个特征和指标，需要进行PCA和标准化。

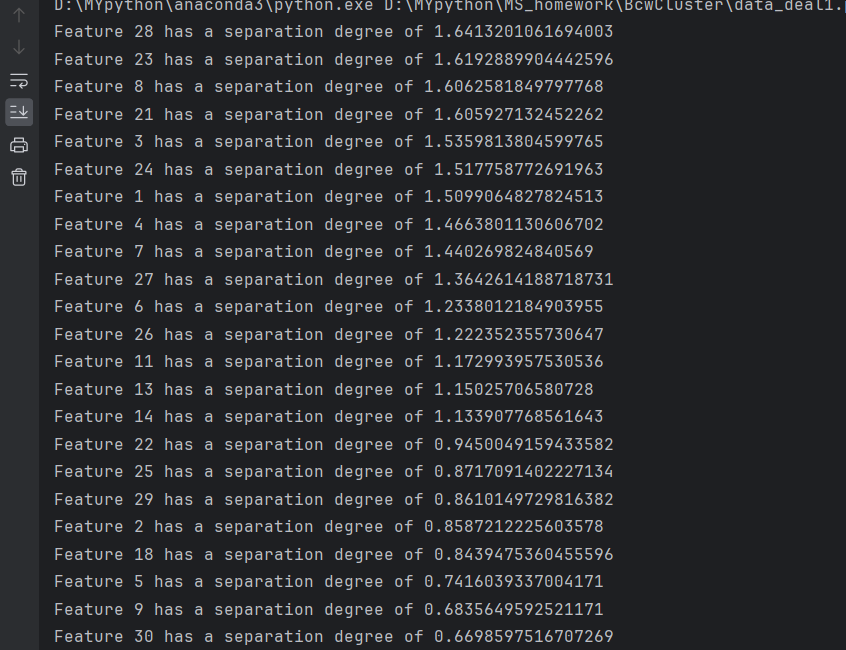


* 1. 特征目标分离和提取

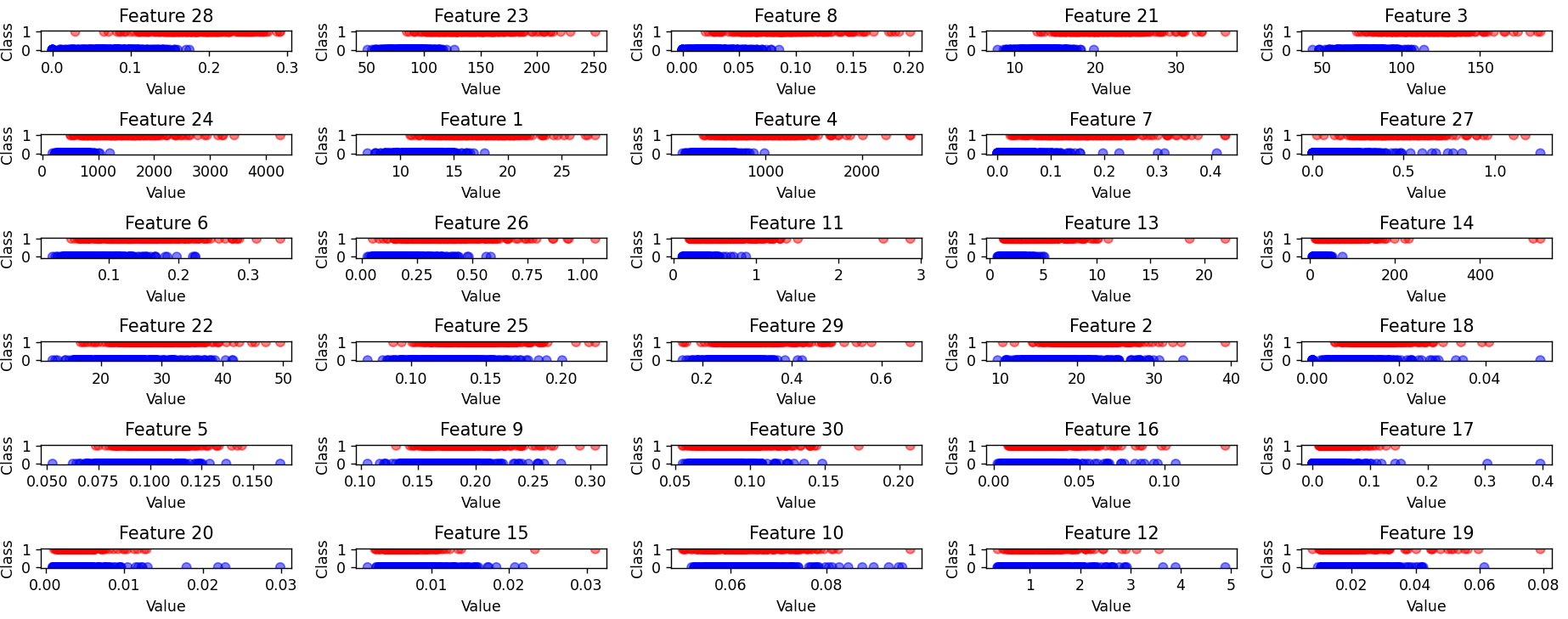
这里我对数据集的特征值进行了一定的处理，因为特征值取自病人的病理特征，但是并非所有的病理特征都属于乳腺癌的范畴，所以需要找出更为合理的特征列。我先对30个特征进行了可视化，也就是表达出两个类别在每个特征上的分布，可视化结果如图：



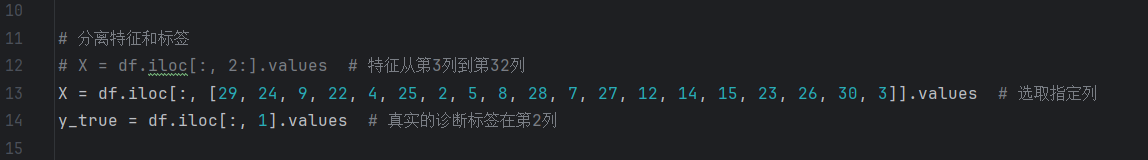
其中，class0指的是B类型，class1指的是M类型。通过观察我们可以看出在一些特征上，俩各个类别有很大的重合部分，这些特征我们就可以不分析。为了更方便地观察，我计算了每个特征上两个类别的分离程度，计算的是欧式距离，然后进行排序输出，以下是部分输出结果：



这是排序后的可视化散点图：



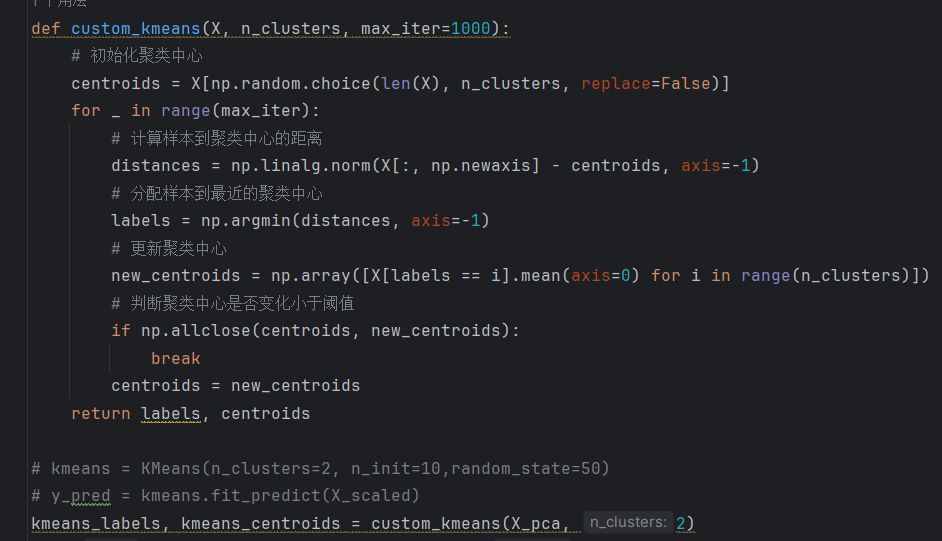
然后我们选出分离度较高特征列参与聚类



这些列数是特征列在原数据集的列数。把这些特征提取出来进行聚类。

* 1. 模型训练

自定义函数



1. 实验结果



1. 实验心得

感觉这个实验最主要的还是对数据的预处理和聚类方法的选用，一开始我直接拿原始数据和全部的特征列进行聚类，最终的NMI只有0.5左右，效果很差。在对数据进行处理后，现在的NMI接近0.68，较之前的结果有了很大的进步。另外，我也尝试过使用层次聚类和DBSCAN聚类，效果要差得多，仅有0.35和0.23。所以选对模型和数据处理还是很重要的。我的实验中选择的K-means聚类的效果也并不是最佳的。

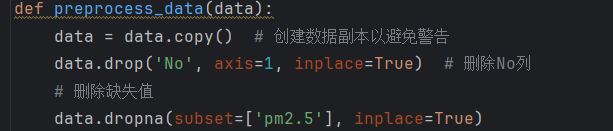
实验三：预测北京PM2.5浓度

1. 实验要求
2. 回归问题：北京PM2.5评估。
3. 目标是构建北京地区的PM2.5的预测模型，通过该地区的温度、湿度等气候特征对该地区的PM2.5进行准确预测。
4. 使用Mean Squared Error作为评估指标。
5. 实验思路
6. 对数据进行预处理。
7. 考虑到通过多种气候特征进行预测，特征值的种类较多，使用随机森林模型预测，可以避免对多特征数据的归一化和标准化处理，同时选择最有效的特征进行训练。
8. 模型训练完成后分别在测试集和验证集上进行对比测试，因为使用MSE得到的值比较大，所以我使用RMSE作为评估指标，并将预测结果和实际值进行可视化比较。
9. 实验步骤
   1. 数据集读取和预处理
      1. 数据集读取

因为原数据的行数是打乱的，不方便观察出时间关系，所以在读取文件前我将数据进行了重新排序并保存在了新文件中，除此外未做任何处理。随后读取排序后的文件。

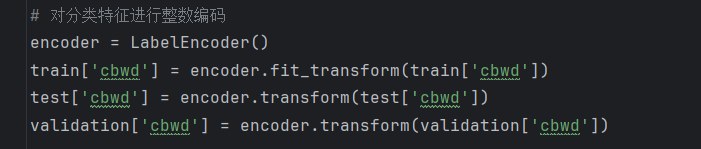
* + 1. 数据预处理
       1. 处理缺省值

把第一列行数列删除。另外，因为数据中有较多的缺省值，针对预测目标PM2.5列的缺省值我进行了删除处理（在实验过程中我尝试过全补为0，但是对预测结果产生了不良的影响，导致模型的性能下降）。

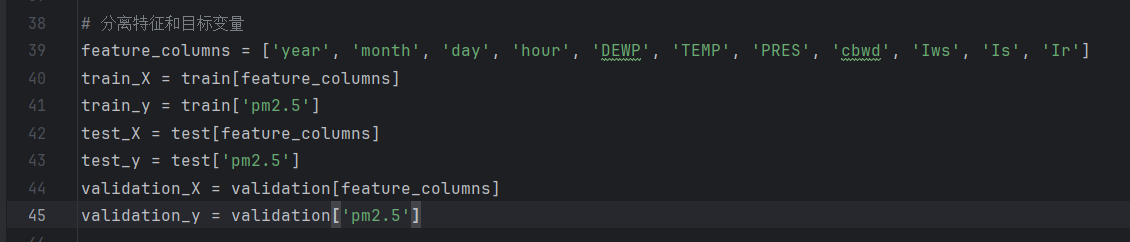


* + - 1. 分类特征转换

在特征中，cbwd的格式特殊，对他进行整数编码。

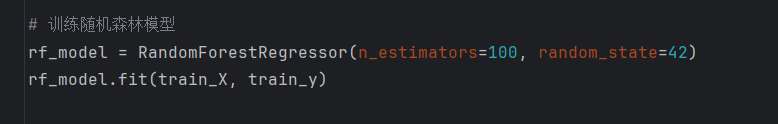


* 1. 特征目标分离和提取



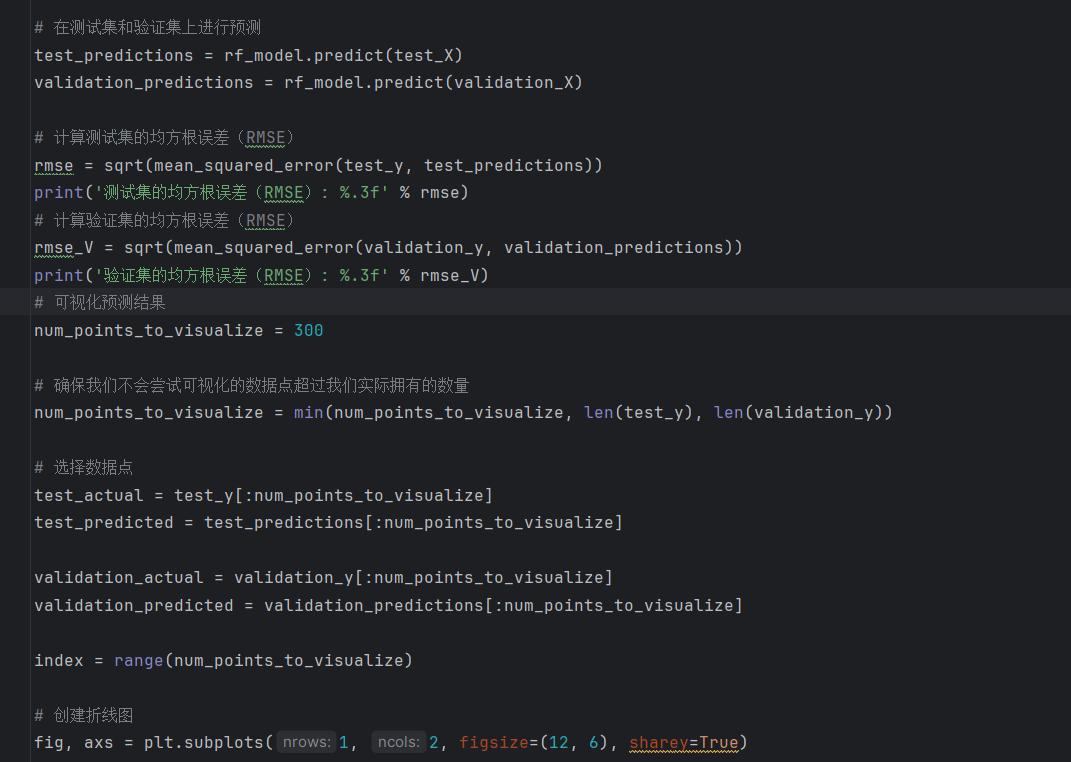
* 1. 模型训练

这里使用库自带的函数进行训练

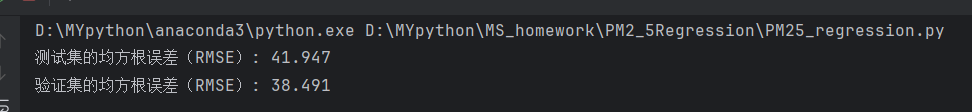


* 1. 预测和可视化

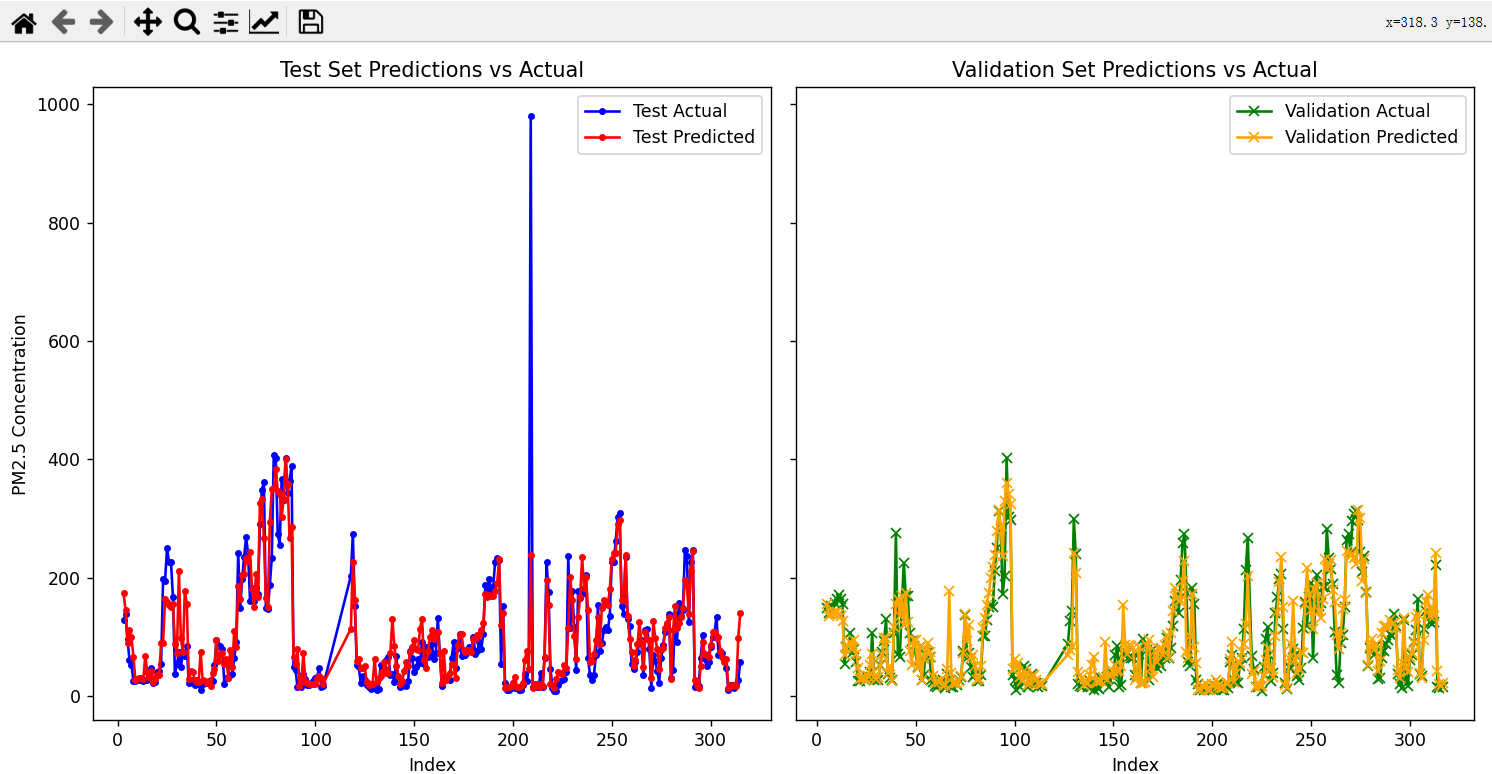
调用函数进行预测和可视化。可视化的时候，因为数据过多，所以只可视化训练集和验证集前300个数据的预测和实际值的比较折线图。



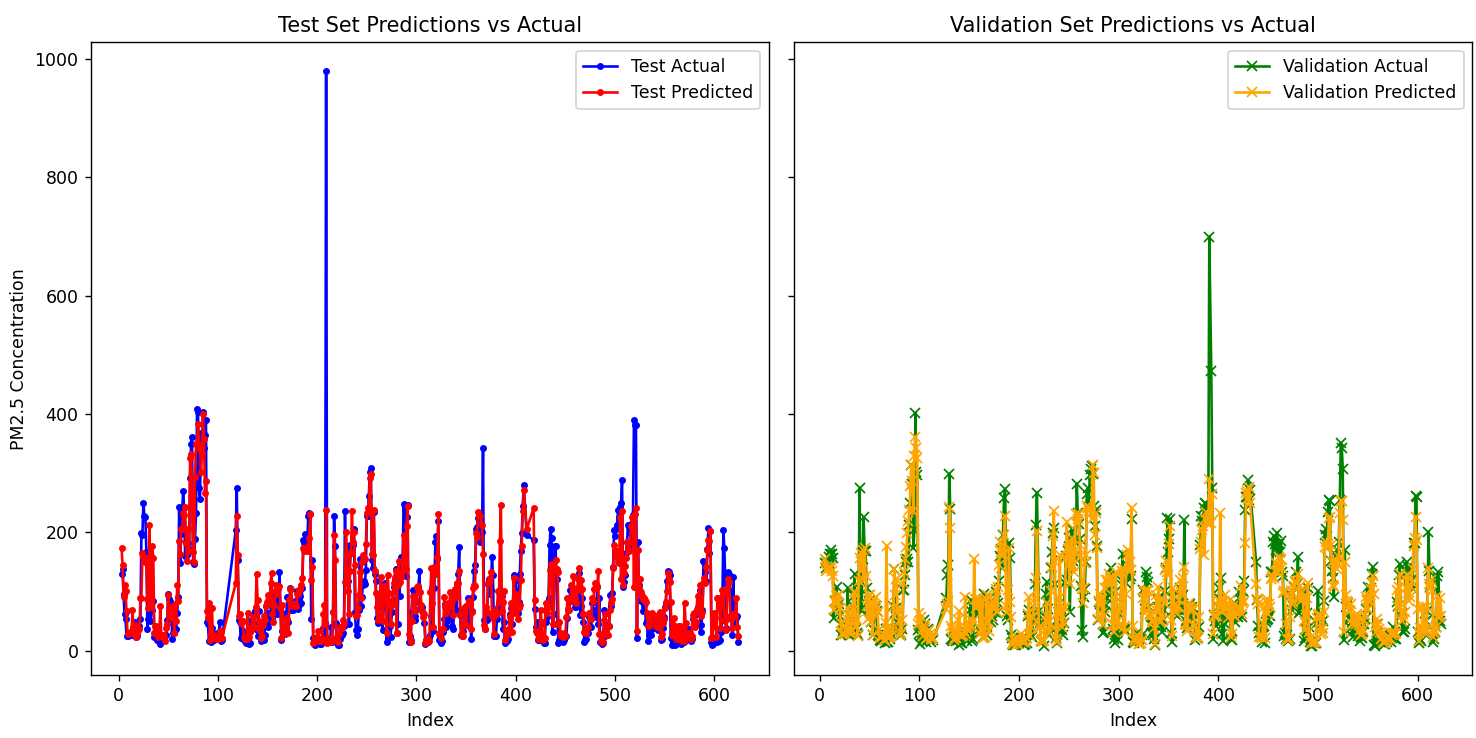
1. 实验结果及分析



通过对数据的整理可以得到PM2.5数值的范围大致为 ，因为数据有删除过，但是大部分数据的数值都保留在，在这个区间内考虑，RMSE的数值分别为41.95和38.50左右，是可以接受的。另外，从预测结果在两个数据集上的对比折线图（这里只取了数据集前300个数据）来看，图像的走势如下：



两个数据集相对应的折线图内，预测值和实际值的走势和形状基本是相同的，说明模型的性能还是很好的。但是观察测试集的对比图，可以观察到有一处预测的结果相差较大，这也是在测试集上的RMSE比验证集高的主要因素之一。下图是取前600个数据的对比折线图：



整体上还是可观的，但是也可以观察到验证集上出现了个别偏差。

1. 实验心得

在最初做实验时，我没有仔细地看过数据集，只是大致的略了一遍，后面在对实验数据进行处理时，我考虑的是把PM2.5的缺省值全部补为0，再去进行线性回归训练，但是这样简单的数据处理后，最后得到的RMSE高达70-80，虽然相对于区间来看并不是很大，但是在折线图上反映时，走势虽然一样，但是数值差距还是很大的。随后我仔细看了数据，发现行数打乱和时间标签的问题，我在排序完之后，考虑时间的循环性和可能的相关性，我把四个时间标签合并成了一个时间戳，再去进行预测，但是效果也并不是很好，RMSE也在70左右，说明可能四个时间组件各自作为特征效果会更好。同时观察数据，我发现相对于数据的总量（指训练集）共26294条，其中PM2.5为空值的行数仅有1254条，所以删去这些数据对整个数据集的影响并不会很大，所以我在对所有数据集进行预处理时，会删去这些无效行，再进行训练，这样得到的RMSE在40左右，说明效果更好。

另外，在选择模型时，最初使用的是线性模型，但是由于特征的维数较高需要大量的归一化和标准化处理，所以最后改为随机森林模型，这个模型的效果在实验时表现得很好（从折线图可以看出来）。