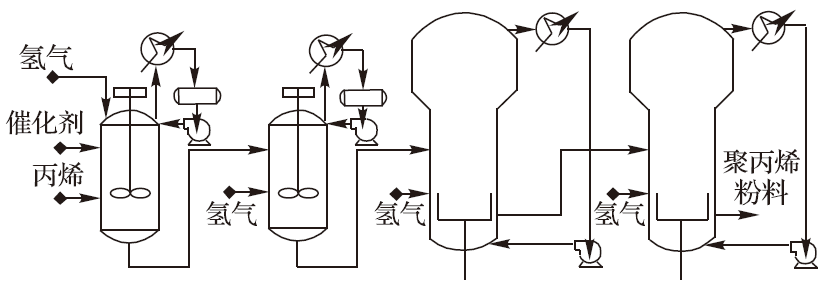
**聚丙烯熔融指数预报报告**

聚丙烯生产过程中的熔融指数是一个重要质量控制指标，决定了所生产产品的牌号与价格，但是难以在线测量，从而导致生产质量的控制品质大大降低。因此，采用生产中与该质量控制指标相关的可直接测量的操作变量，来在线预测该质量控制指标，从而提高生产质量控制品质。

1. **原理分析**

1.问题背景——聚丙烯熔融指数及其预报

聚丙烯是以丙烯为单体聚合而成的一种热塑性树脂，具有密度小、生产成本低、透明度高、化学稳定性好、无毒、易加工、抗冲击强度高、抗挠曲性及电绝缘性好等优点，是塑料中消费量最大的品种之一，其在塑料容器、办公用品、电子等方面已经得到了广泛的应用。熔融指数反映了聚丙烯的流动和加工性能．是最重要的质量指标，决定了聚丙烯的不同用途。对熔融指数准确可靠的预报在聚丙烯生产中具有至关重要的作用。它能有效的指导生产过程，进而提高聚丙烯生产的经济效益。

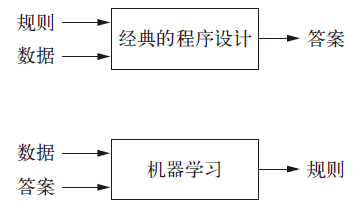
如图所示，图中是典型的聚丙烯生产流程。可以看到整个过程涉及到的反应关系错综复杂。目前聚丙烯生产中质量控制的一个主要困难就是缺乏熔融指数在线分析仪表。只能通过人工取样、离线化验的方法获得。但是这种方法有2～4 h的滞后，得到的结果无法指导实时生产控制。

因此，设计一个准确、可靠的熔融指数实时预报模型对熔融指数进行有效预测，对提高聚丙烯产品质量，减少经济损失，提升聚丙烯生产自控水平有着重要意义。此种软测量方法的实时性好，并且结果误差在工业应用的允许范围之内，能够满足工业测量的要求。因而成为当前国内外在聚丙烯熔融指数预报方面的主要研究方法。

本文中将结合机器学习的基本知识与模型，利用150组聚丙烯熔融指数数据进行建模。从而作为机器学习的初学者，达到利用机器学习解决实际问题的目的。

2.实践过程中的理论依据

①机器学习基本流程

利用机器学习可以针对数据进行自动挖掘，找出数据的内在规律，并应用这个规律来预测新数据。右图直观的表明了机器学习与经典程序设计的不同。根据软件工程学知识，可以将机器学习项目的流程简单的分为：定义问题、理解数据、数据准备、评估算法、优化模型、结果部署。本文在程序实现部分也将以此为顺序。

②环境配置（编程语言及类库）

编程语言：Python。

NumPy： Python的一种开源数值计算扩展。可用来存储和处理大型矩阵，提供了许多高级的数值编程工具，如矩阵数据类型、矢量处理、精密的运算库。

Matplotlib：Python中最著名的2D绘图库。

Pandas：纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据集所需的工具，也提供了大量能快速、便捷地处理数据的函数和方法。

scikit-learn：Python中开发和实践机器学习的著名类库之一。基本功能主要分为六大部分：分类、回归、聚类、数据降维、模型选择和数据预处理。

③数据理解与可视化

对数据的分析是机器学习中的重要领域，通过对数据的理解，可以选择有效的算法来建立模型。类库中提供了多种方法来观察和理解数据，这为发现数据的特征和选择合适的算法提供了思路。也可以通过Matplotlib提供的可视化图表来展示数据，以便于发现数据的特征。

④数据预处理与特征选择

数据预处理大致分为三个步骤：数据的准备、数据的转换、数据的输出。数据处理可以有效提高算法准确度。数据预处理根据数据本身的特性，有不同的格式和不同的要求，可能需要填补缺失值，剔除无效数据，或进行降维处理。本文利用scikit-learn来转换数据。

特征选定能够选择有助于提高预测结果准确度的特征数据。如果数据中包含无关的特征属性，会降低算法的准确度，对预测新数据造成干扰，尤其是线性相关算法。执行特征选定有助于降低数据的拟合度、提高算法精度、减少训练时间。常见的数据特征选择方法包括递归特征消除（RFE），主要成分分析（PCA）。

⑤评估算法

最简单的评估算法方法就是将评估数据集和训练数据集完全分开，采用评估数据集来评估算法模型。将原始数据集分为两部分，第一部分用来训练算法生成模型，第二部分通过模型来预测结果，并与己知的结果进行比较，来评估算法模型的准确度。通常将80%的数据作为训练集，将20%的数据作为评估数据集。

交叉验证是用来验证分类器的性能的一种统计分析方法。基本思想是按照某种规则将原始数据进行分组，一部分作为训练数据集，另一部分作为评估数据集。首先用训练数据集对分类器进行训练，再利用评估数据集来测试训练得到的模型，以此作为评价分类器的性能指标。K折交叉验证是将原始数据分成K组（一般是均分），将每个子集数据分别做一次验证集，其余的K-1组子集数据作为训练集，这样会得到K个模型，再用这K个模型最终的验证集的分类准确率的平均数，作为此K折交叉验证分类器的性能指标。K折交叉验证可以有效地避免过学习及欠学习状态的发生，最后得到的结果也较有说服力。

⑥回归算法的准确度评估

本文中采取平均绝对误差（MAE）、均方误差（MSE）、均方根误差（RMSE）与决定系数（R2）来评估最后的模型精度。

平均绝对误差是所有单个观测值与算术平均值的偏差的绝对值的平均值。均方误差是衡量平均误差的方法，可以评价数据的变化程度。均方根误差是均方误差的算术平方根。均方误差的值越小，说明用该预测模型描述实验数据的准确度越高。

决定系数反映因变量的全部变异能通过回归关系被自变量解释的比例。拟合优度越大，决定系数越接近1。

⑦回归算法概述

本文中出现了六种基本的回归算法。线性算法有：线性回归算法、套索回归算法、弹性网络（Elastic Net）回归算法；非线性算法有：K近邻算法（KNN）、分类与回归树算法、支持向量机（SVM）。

线性回归：如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。

套索回归：在套索回归中会惩罚回归系数的绝对值大小。此外，它能够减少变化程度并提高线性回归模型的精度。

弹性网络回归：在模型训练时，弹性网络回归算法综合使用L1和L2两种正则化方法。当有多个相关的特征时，弹性网络回归算法很有用。

K近邻算法：K近邻算法是一种理论上较成熟的方法，也是最简单的机器学习算法之一。该方法的思路是：如果一个样本在特征空间中的k个最相似（特征空间中最邻近）的样本中的大多数属于某一个类别，则该样本也属于这个类别。KNN通过依据k个对象中占优的类别进行决策，而不是通过单一的对象类别决策。这是KNN算法的优势。

分类与回归树：分类与回归树的英文缩写是CART，算法由以下两步组成。树的生成：基于训练数据集生成决策树，生成的决策树要尽量大；树的剪枝：用验证数据集对已生成的树进行剪枝，并选择最优子树，这时以损失函数最小作为剪枝的标准。

支持向量机：支持向量机在解决小样本、非线性及高维模式识别问题上有很多优势，并能够推广应用到函数拟合等其他机器学习问题中。在机器学习中，支持向量机可以分析数据、识别模式，用于分类和回归分析。

⑧集成算法概述

装袋（Bagging）算法，如随机森林、极端随机树。装袋算法是一种提高分类准确率的算法，通过给定组合投票的方式获得最优解。

提升（Boosting）算法，如AdaBoost、随机梯度提升（Stochastic Gradient Boosting）。提升算法也是一种提高任意给定学习算法准确度的方法，主要通过对样本集的操作获得样本子集，然后用弱分类算法在样本子集上训练生成一系列的基分类器。提升算法可以用来提高其他弱分类算法的识别率。

⑨网格搜索优化参数

网格搜索优化参数是一种算法参数优化的方法。通过遍历已定义参数的列表，来评估算法的参数，从而找到最优参数。在scikit-learn中使用GridSearchCV来实现对参数的跟踪、调整与评估。

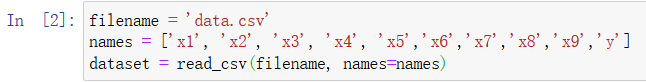
1. **程序实现方法**

**1.定义问题**

通过丙烯聚合机理和工艺流程分析可知,影响聚丙烯熔融指数的过程变量主要有9个：主、辅催化剂进料流率，3 股丙烯进料流率，釜内温度、压强、液位、氢气体积浓度，且这9个变量之间存在着不同程度的相关。通过对于这些特征属性的描述，可以发现输入的特征属性单位并不统一，或许应该对数据进行度量单位的调整。

**2.导入数据**

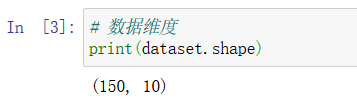
将项目所需要的类库导入，然后将数据集导入到Python中，在导入时设定数据属性特征的名字。



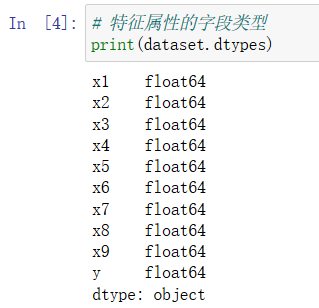
**3.理解数据**

对导入的数据进行分析，以便于构建合适的模型。

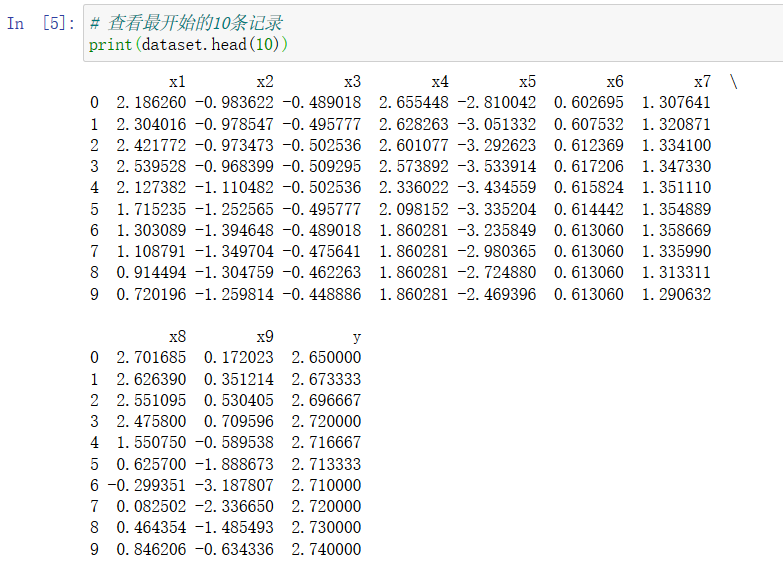
首先查看数据维度，共有150条记录，10个特征属性，与问题定义中的描述一致。

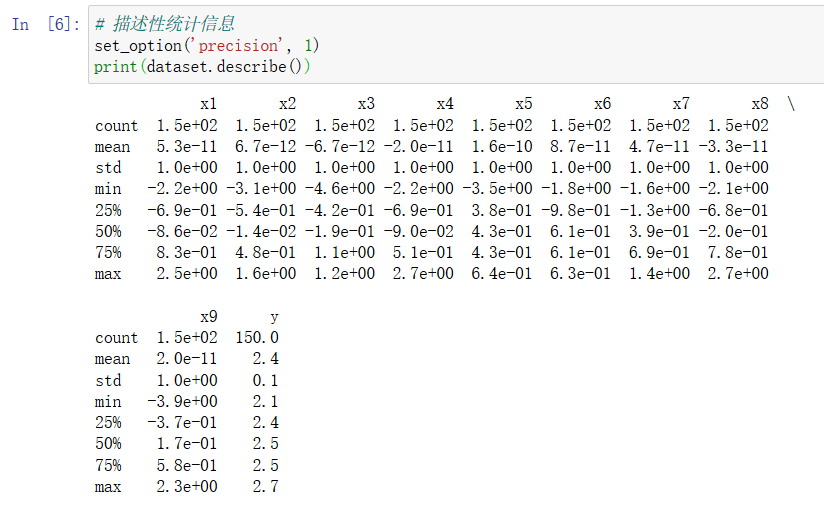


再查看各个特征属性的字段类型，看到所有的特征属性都是浮点数。

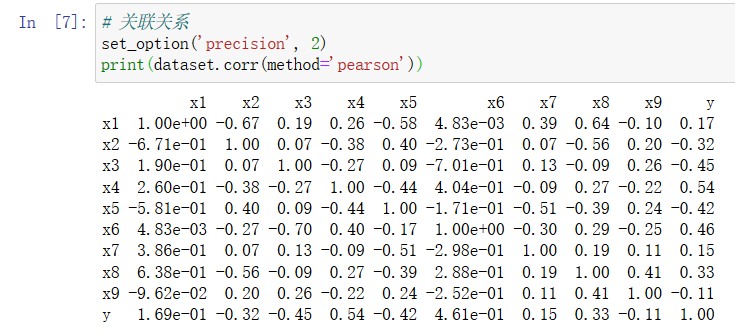


查看最开始的10条记录，并得到描述性统计信息，观察数据的分布情况。



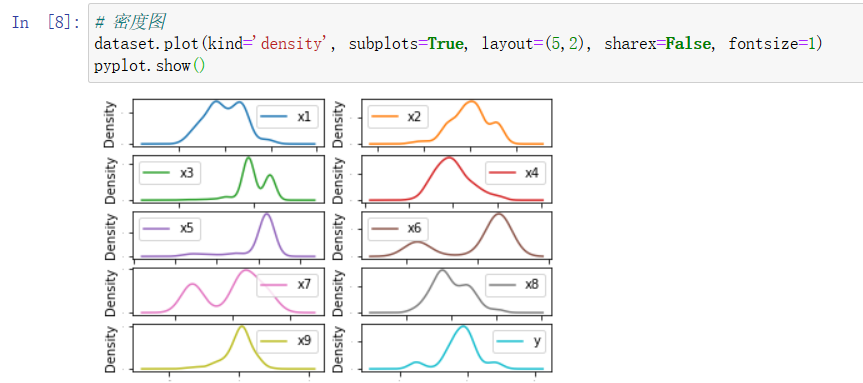


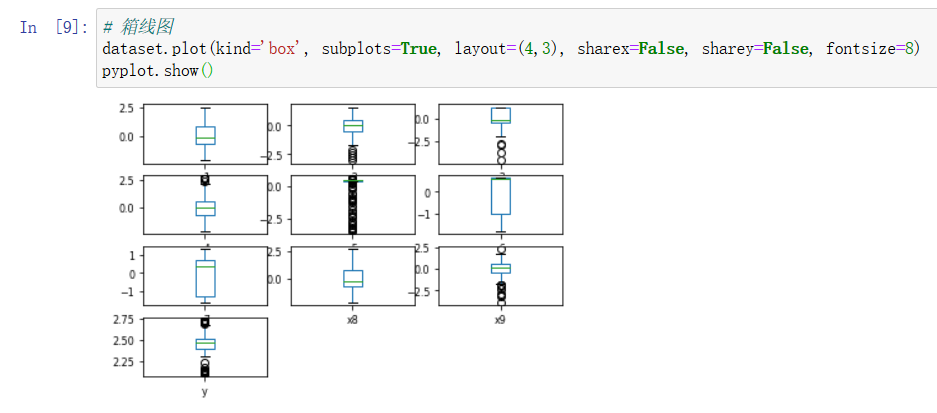
接下来分析数据特征之间的关联关系，查看数据的皮尔逊相关系数。发现部分特征属性之间具有强关联关系。



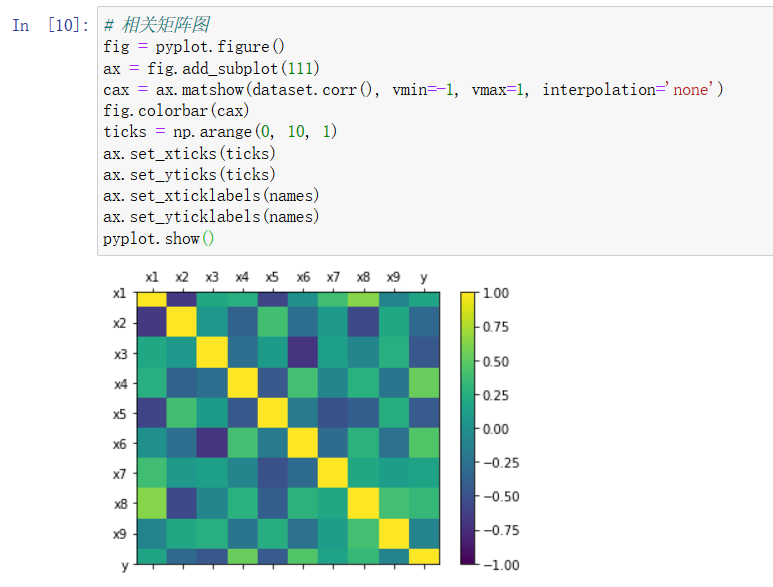
**4.数据可视化**

通过密度图直观的展示数据特征，通过箱线图查看数据分布的偏态程度。



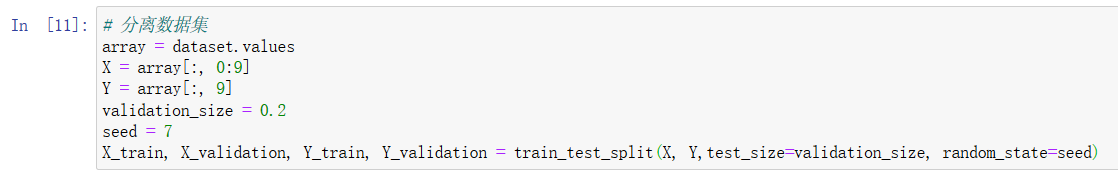


绘制数据相互影响的相关矩阵图，发现有些属性之间是强相关的，后续处理中可以移除这些特征属性，以提高算法准确度。



通过对数据分布的探索，发现数据结构较为复杂，需要考虑对数据进行转换，以提高模型的准确度。

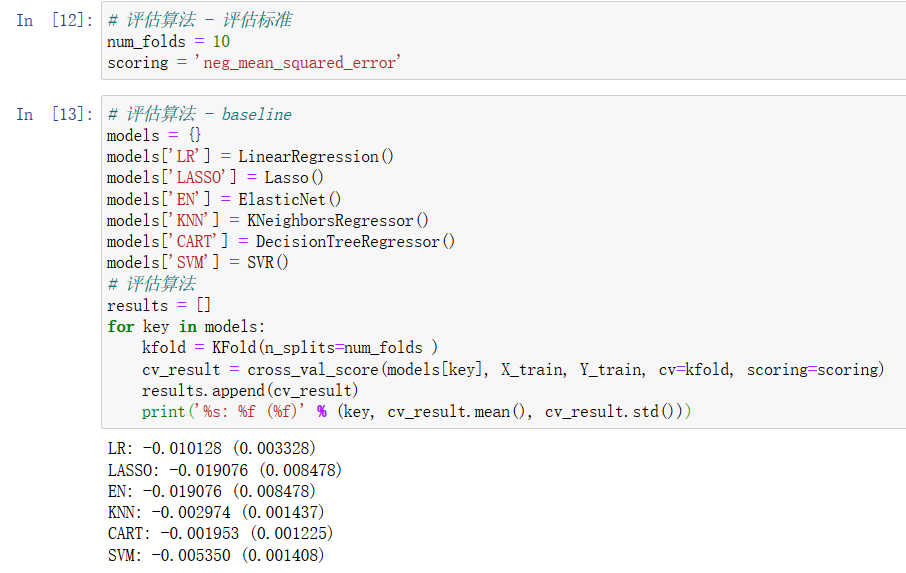
**5.分离评估数据集**

分离出部分数据集与训练集完全隔离，用来判断和报告最终模型的准确度。此处采用80%的数据进行训练，20%的数据来做测试。

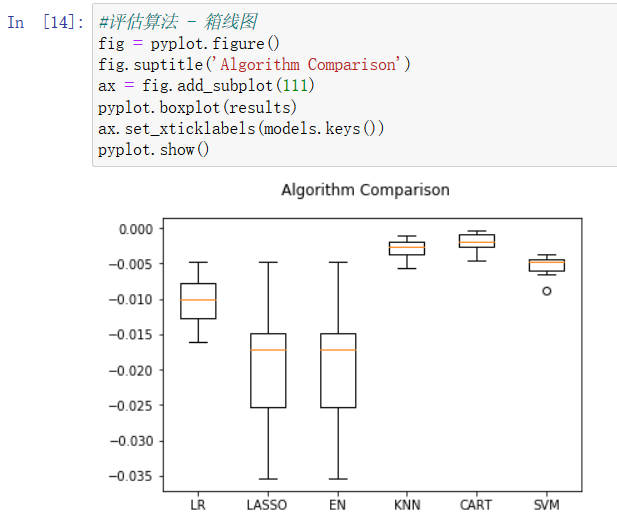
**6.评估算法**

①使用原始数据

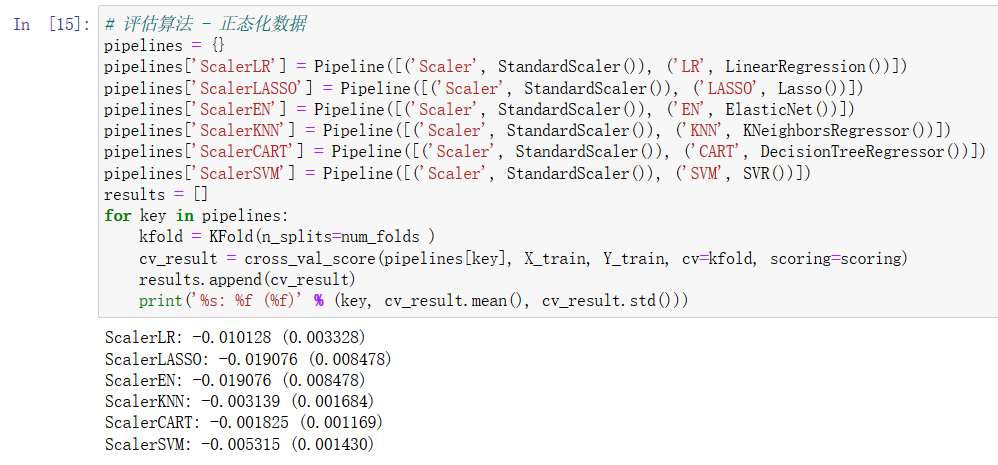
由于不能立刻选择出哪个算法对需要解决的问题最有效，因此需要设计一个评估框架来选择合适的算法。采用10折交叉验证来分离数据，通过均方误差来比较算法的准确度，均方误差越接近0，算法准确度越高。

对原始数据不进行处理，对算法进行一个评估，形成评估基准，此时选择三个线性算法和三个非线性算法，线性算法：线性回归（LR）、套索回归（LASSO）和弹性网络回归（EN）。非线性算法：分类与回归树（CART）、支持向量机（SVM）和K近邻算法（KNN）

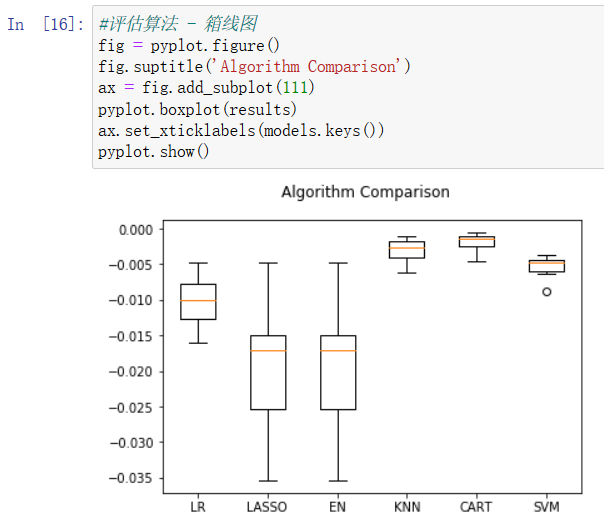
从执行结果来看，分类与回归树（CART）具有最优的MSE，接下来是K近邻算法（KNN）。

查看所有的10折交叉分离验证的结果，执行结果如图，从图中可以看到，非线性算法的分布比较类似，并且K近邻算法和CART算法的结果分布非常紧凑。

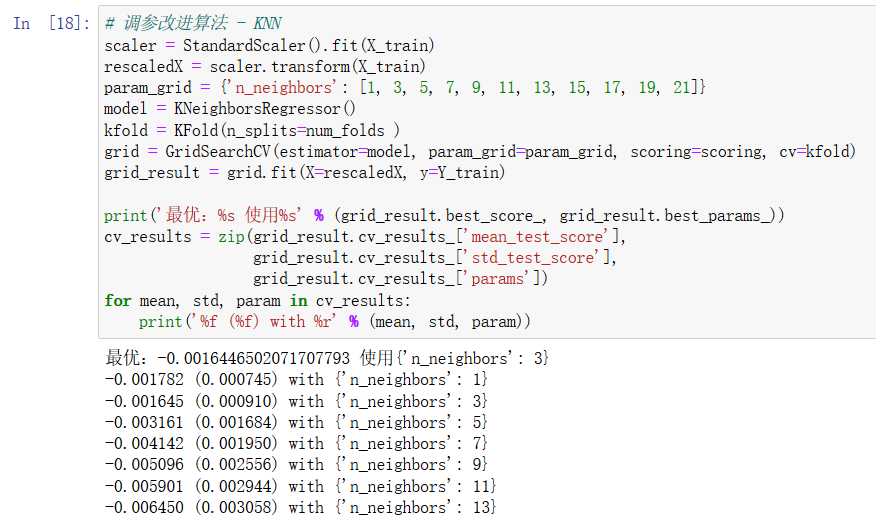
②正态化数据

猜测可能由于原始数据中不同数据特征的度量单位不同，导致某些算法结果不理想，于是将所有数据进行正态化，再次评估这些算法。对训练集数据进行数据转换处理，将所有数据特征值化为0为中位值，标准差为1的数据。对数据正态化时，为了防止数据泄露，采用Pipeline来正态化数据和对模型进行评估。为了与之前的结果比较，采用与之前相同的评估框架来评估算法模型。

执行后发现分类与回归树算法具有最优的MSE。

再次查看所有10折交叉分离验证的结果，结果如图，可以看到CART算法和KNN算法的MSE和数据分布情况差别不大。

**7.调参改善算法**

下面通过网格搜索算法来优化CART和KNN的参数

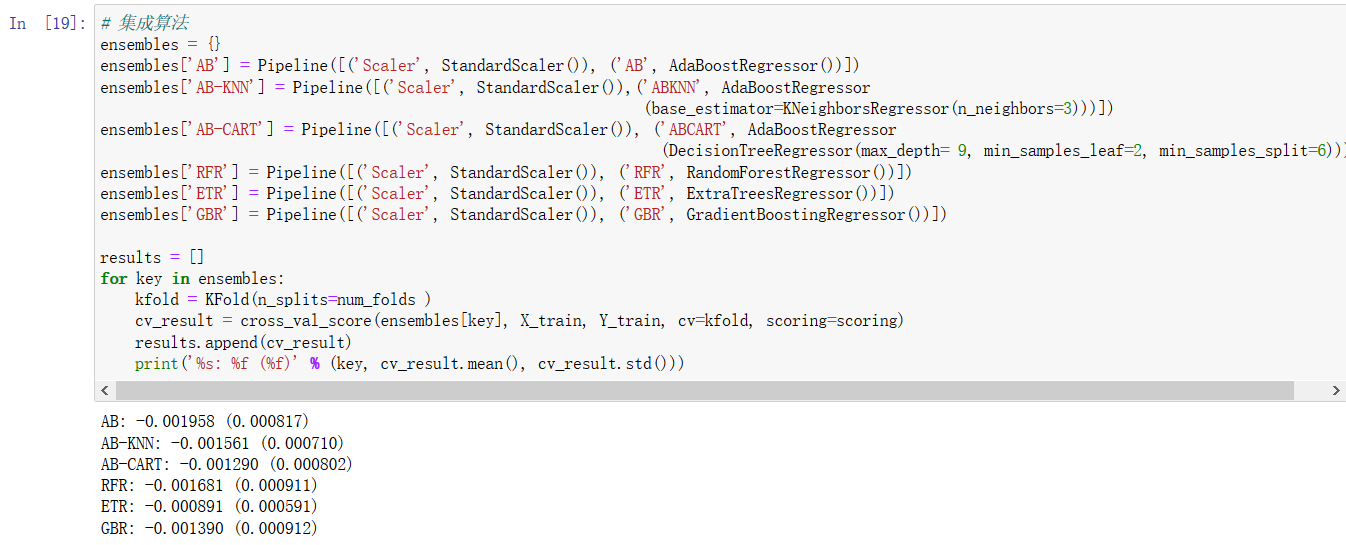
得到在这几种情况中的最优结果——KNN算法：参数近邻个数为3。

**8.集成算法**

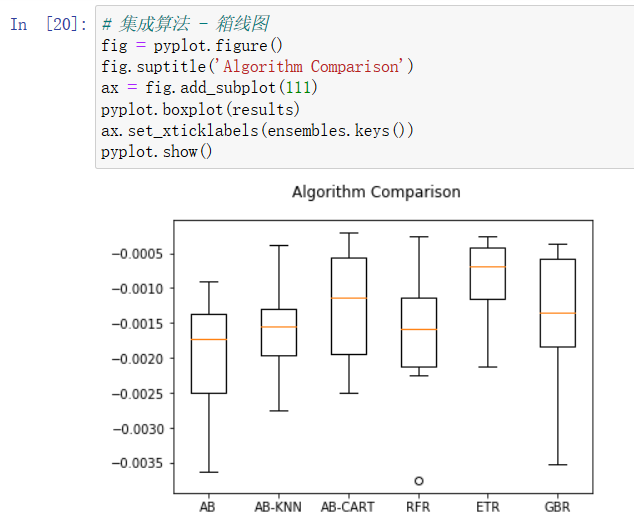
除了调参，提高模型准确度的方法还有使用集成算法，下面对表现较好的分类与回归树（CART）和K近邻算法（KNN）进行集成，看算法能否提高。

装袋算法：随机森林（RF）和极端随机树（ET）

提升算法：AdaBoost（AB）和随机梯度上升（GBM）

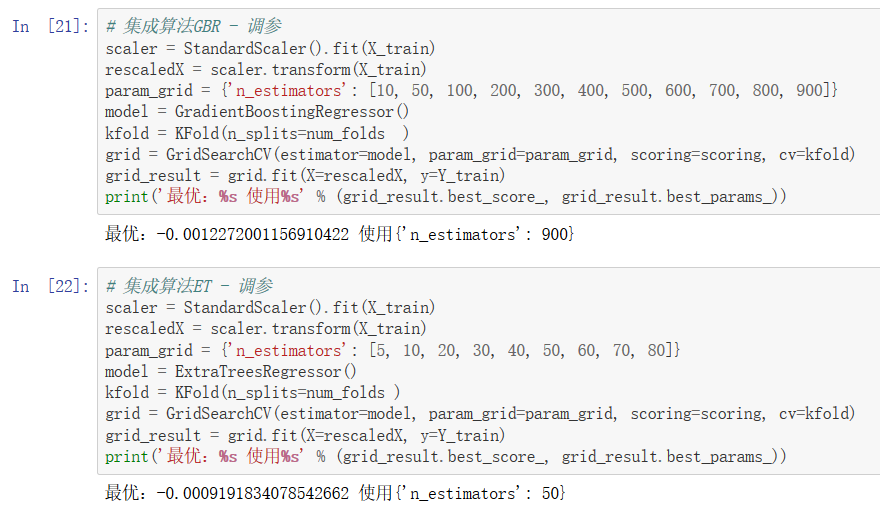
依旧采用与之前相同的评估框架和正态化之后的数据对相关算法进行分析

与之前的算法相比，这一次的准确度都有了较大的提高。通过箱线图查看集成算法在10折交叉验证中均方误差的分布状况。



**9.集成算法调参**

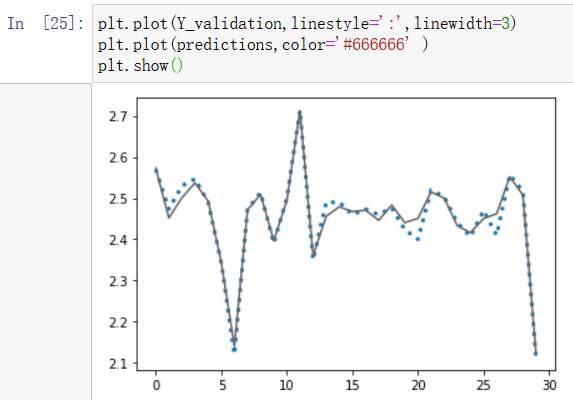
下面对集成算法中表现较好的随机梯度上升（GBM）和极端随机树（ET）算法进行调参，再次比较这两个算法模型的准确度，来确定最终的算法模型。

对于随机梯度上升（GBM）算法来说，最优的n\_estimators是900；对于极端随机树（ET）算法来说，最优的n\_estimators是50。执行结果，极端随机树（ET）算法优于随机梯度上升（GBM）算法，因此采用极端随机树（ET）算法来训练最终的模型。

**10.确定最终模型**

已经确定使用极端随机树（ET）算法来生成模型，对该算法进行训练和生成模型，并计算模型的准确度。





得到最后的结果，RMSE在0.018左右，由图像也可看出数据预测的最终结果。

1. **结果分析**

聚丙烯是工业生产与日常生活都不可缺少的化工合成产品，因此，丙烯聚合生产过程的有效控制和高质量产出保证就极为重要。其中最重要的质量指标就是聚丙烯熔融指数。在现阶段实验室离线分析的模式下，熔融指数的测量存在一定的技术难题，因此，建立准确、可靠的熔融指数实时预报模型对熔融指数进行有效预测是非常有意义的。当前在丙烯聚合过程中熔融指数软测量方面已经有了大量的研究，但是熔融指数预报的准确性和预报模型的泛化能力仍一直是学术界和工业界需要不断改善的目标。

本文在六种基本的回归算法 线性回归（LR）、套索回归（LASSO）、弹性网络回归（EN）、分类与回归树（CART）、支持向量机（SVM）和K近邻算法（KNN） 的基础上，根据得到的结果进行针对性的改进，采用调参的方式，提高模型的精度。对表现较好的CART和KNN进行集成。并且对随机森林（RF）、极端随机树（ET）、AdaBoost（AB）和随机梯度上升（GBM）算法进行了实现。并且选取其中表现较好的ET与GBM算法进行调参，得到最后的模型。

由得出的结果（MSE: 0.00033 MAE: 0.01267 RMSE: 0.01824 R2: 0.97447）可以看出模型的精度尚可，但是在一些情况下的预测值与实际值仍有较大偏差。可能需要进一步的进行优化。根据程序实现部分的流程，可以采用PCA降维的方法降低模型复杂度并运用新的模型结构以及新的智能算法进行优化。

作为从零开始的机器学习初学者，在这次项目经历中，我对于机器学习的各个流程都有了一定程度的了解，也产生了较强的兴趣，在今后打算继续学习相关知识，对这方面进行更深的了解。