基于多元数据分析和机器学习的生物质和煤共热解问题研究

摘　要

针对问题一，本文首先将附件一数据进行预处理，将空白项填充为 0 并拆分合并项以便后续分析。然后利用 SPSS 软件对正己烷不溶物(INS)和热解产率的三项指标做**描述性分析**，计算出 INS 均值约为 0.13，标准差约为 0.183，反映数据较为分散。再通过 pearson **相关系数**进行相关性分析构建线性回归模型进一步验证正己烷不溶物(INS)对焦油产率、水产率、焦渣产率是否产生显著影响，并利用数据可视化量化分析绘制散点图、箱线图以及趋势线，计算决定系数（R^2）和 p 值。最后得出随着 INS 的增加，二者产率均增加，水产率减少；焦油产率和焦渣产率的 p 值统计显著，说明 INS 对二者产率有显著影响，可能对减少焦渣产率具有一定的作用的结论。

针对问题二，本文采用**线性回归模型**对热解实验数据进行了交互效应分析，构建的模型成功量化了 INS 质量(m)、混合比例(p)及其交互项(mp)对产物产率的贡献。分析结果显示：混合比例对产率有显著正向影响，而 INS 与混合比例的交互效应不显著，意味着它们对焦油产率的影响相对独立。观察到 INS 与混合比例的交互效应显著正向影响水产率，表明在特定 INS 浓度下增加混合比例可提升水产率。最为关键的发现是，INS 与混合比例的交互效应对焦渣产率产生了极其显著的负向影响，伴随混合比例升高，尤其是在高 INS 浓度下，焦渣产率显著降低。此外，混合比例本身也对焦渣产率有显著的负向作用。

针对问题三，我们建立了线性回归模型，设置出由焦油产率、混合比例和正己烷不溶物量三者数据驱动的回归模型，再通过使用**边界优化**中的 Brent **算法**用一个抛物线逼近目标函数，利用计算抛物线的极值来逼近目标极值，进而在选定边界上寻找到最优值，即确定了最大化焦油产率的混合比例，最终根据优化结果得到的最优配比与预测产率分析出最大化焦油产率与煤所占混合物比例的关系。

针对问题四，首先本文通过数据预处理将配比数据转化为浮点数形式，并利用配对样本 t **检验**探索了实验值与理论计算值间的差异。结果显示，诸如 CS/HN 组合的焦渣产率在不同配比下存在显著差异，以及其他多个组合和产物也显示出显著差异，如 CS/HN 组合在 50/100 配比下焦渣产率差异最大；深入到子组层面，对每个显著差异组合按混合比例评估了实际与理论值的差异，提取出差异最大的前两个混合比例，通过可视化展示，为优化热解过程提供了直观指导，指出了需要优先考虑调整的具体条件。

针对问题五，我们首先利用 LabelEncoder 函数和 lambda 函数将定类变量转化，然后，本文采用**随机森林**与**支持向量回归**算法构建模型进行预测，比较精度后发现随机森林的预测效果更好，我们设置决策树的数量为 100，确立随机种子作为训练集，使用训练好的模型对测试集进行预测，利用均方误差（MSE）和决定系数（R2）评估模型的性能，最后作出预测结果图。

关键词　相关性分析　Brent 算法　t 检验　随机森林　支持向量回归

目 录

一、问题重述

1.1 问题背景

随着全球能源需求的快速增长以及传统化石燃料资源的日益枯竭，对可再生能源的利用和高效能源转化技术的研究愈发重要。生物质，作为一种可再生且环境友好的能源来源，与煤等传统化石燃料共热解的技术逐渐受到广泛关注。生物质与煤共热解技术不仅能够有效提高能源利用效率，还能促进资源的综合利用，对于保障能源安全、减少环境污染具有重要作用。

1.2 问题提出

化工实验室引入微晶纤维素作为模型化合物，分析比较棉杆(CS)热解、神木煤(SM)热解、棉杆/神木煤(CS/SM)共热解和微晶纤维素/神木煤共热解产生的正己烷可溶物(HEX)组分变化。建立不同混合比例进行固定热解实验。实验结果如附件 1 和附件 2 所示。基于以上条件，建立数学模型研究生物质热解产物的特性和化学反应机理，解决以下问题：

问题一：分析附件一中提供的实验数据，确定正己烷不溶物(INS)对热解产率（焦油产率、水产率、焦渣产率）的影响是否显著。利用图表或图像展示 INS 含量与各种热解产物产率之间的关系，并解释其影响机制。

问题二：在热解实验中，研究正己烷不溶物(INS)与混合比例之间是否存在交互效应，这种交互效应如何影响热解产物的产量。如果交互效应显著，请识别出在哪些具体的热解产物（如焦油、水、焦渣等）上，样品重量和混合比例的交互效应最为明显。

问题三：基于附件一中的共热解产物特性和组成数据，构建一个数学模型来优化共解热混合比例。模型的目标应是提高产物的整体利用率和能源的转化效率。通过模型分析，找出最佳的混合比例组合。

问题四：根据附件二中的实验数据，分析每种共热解组合的产物收率实验值与理论计算值之间是否存在显著差异。如果存在显著差异，请利用子组分析的方法，进一步确定这些差异主要体现在哪些混合比例的组合上。

问题五：基于实验数据，构建一个数学模型来预测热解产物的产率。模型应能够考虑各种影响因素（如正己烷不溶物含量、混合比例等），并给出在不同条件下热解产物产率的准确预测值，这个模型将为实验条件的优化提供有力的理论支持。

二、问题分析

2.1 问题一的分析

问题一要求研究正己烷不溶物(INS)与焦油产率、水产率、焦渣产率的影响，需要对正己烷不溶物(INS)和热解产率的三项指标做描述性分析，考虑到正己烷不溶物(INS)和热解产率的三项指标均为定量变量，故使用描述性统计，计算出均值、最大值、最小值、标准差等统计量反映数据的集中趋势与离散趋势，进行数据可视化绘制散点图。最后通过相关性分析进一步分析正己烷不溶物(INS)对焦油产率、水产率、焦渣产率是否产生显著影响。

2.2 问题二的分析

问题二要求探讨热解过程中正己烷不溶物（INS）的量以及生物质和煤的混合比例对热解产物（焦油、水、焦渣）产率的交互效应。通过建立线性回归模型并运用 OLS（普通最小二乘法）进行拟合，我们获得了各产物产率与两个关键因子及其交互项之间的关系，并对其统计显著性进行了评估。

2.3 问题三的分析

问题三要求通过建立生物质和煤的混合比例的优化模型，使共热解产物中的焦油产率最大化。其中，为了能够实现煤的较高纯度和较低的处理成本，我们希望焦油的质量尽可能接近纯煤热解所产生的焦油。我们选择继续沿用问题二中的线性回归模型，通过最小化观测值和模型预测之间的平方误差来估计参数。同时建立边界优化模型处理混合比例的边界条件从而确保解不会超出指定范围。

2.4 问题四的分析

问题四要求我们对附件二提供的热解实验数据进行深入研究，特别是针对不同生物质与煤炭混合比例（如 5/100 至 50/100 表示的生物质占比）下热解产物（焦油、己烷可溶物（HEX）、水、焦渣）的实际产量与理论计算产量之间的差异。通过执行配对样本 t 检验，我们旨在验证实际测量值与理论预测值之间是否存在显著差异，并找出这些差异在哪些特定条件下最为显著。

2.5 问题五的分析

问题五要求基于实验数据，建立相应的模型，对热解产物产率进行预测，首先进行数据预处理，将定类变量转化为定量变量，再利用随机森林和支持向量回归算法建立机器学习模型对原有数据进行训练，选择精度较高的模型用于预测和优化热解过程，最后作出预测结果图。

三、模型假设

1、假设热解过程史不受外界环境(如温度波动、湿度等)的影响,只考虑内部的化学反应和物理变化。
2、在进行热解反应过程模拟时,假设系统处于稳态条件,即进料速率、温度、压力等在整个反应过程中保持不变。
3、假设热解反应在一定条件下稳定进行，产率仅受原料组成和操作条件的影响。

四、定义与符号说明

符号定义	符号说明
T	INS 的质量比例
Y_t	焦油产率
Y_w	水产率
Y_c	焦渣产率
X	INS 的量
p	配比
ε	误差项
δ	拉格朗日乘数

五、模型的建立与求解

5.1 研究正己烷不溶物(INS)对三项热解产率的影响

为研究正己烷不溶物(INS)与焦油产率、水产率、焦渣产率的影响，需要对正己烷不溶物(INS)和热解产率的三项指标做描述性分析，考虑到正己烷不溶物(INS)和热解产率的三项指标均为定量变量，故使用描述性统计，计算出均值、最大值、最小值、标准差等统计量反映数据的集中趋势与离散趋势，进行数据可视化绘制散点图。最后通过相关性分析进一步分析正己烷不溶物(INS)对焦油产率、水产率、焦渣产率是否产生显著影响。

5.1.1 数据清洗与数据预处理

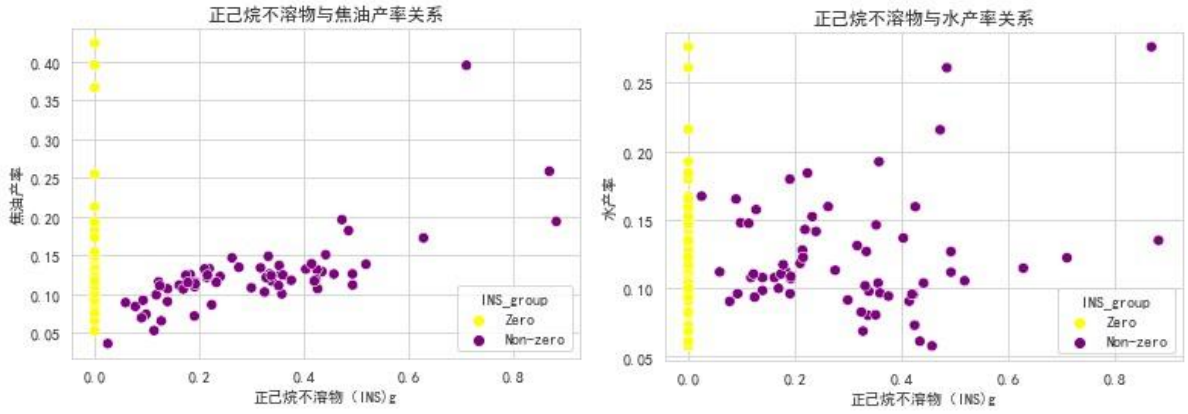
对于附件一，我们将正己烷不溶物、正己烷可溶物产率数据空白项填充为 0，并把合并项拆分以便后续计算:

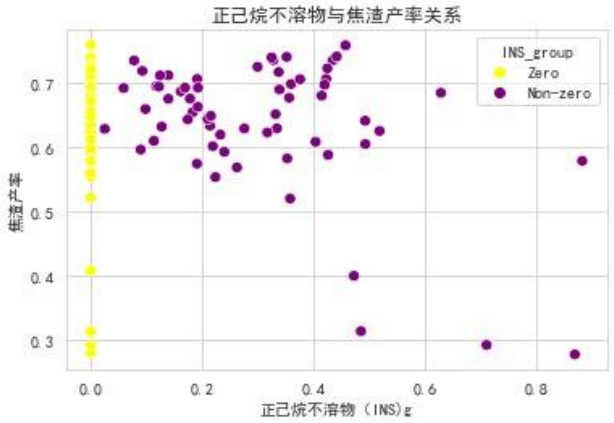
5.1.2 描述性分析

利用 SPSSPRO 软件对整理后的部分附件一数据进行描述性统计分析结果如下：

变量名	样本量	最大值	最小值	平均值	标准差	中位数	方差	峰度	偏度	变异系数（CV）
时间	100	20151123	20130224	20142278.64	8557.838	20140328	73236594.48	-1.633	-0.194	0
样品 g	100	12.055	5.072	8.478	1.62	8.316	2.623	-0.696	-0.219	0.191
焦油(Char) g	100	2.139	0.361	0.995	0.311	0.982	0.097	2.096	0.803	0.312
水(Water) mL	100	1.9	0.58	1.019	0.283	0.947	0.08	1.02	1.151	0.278
正己烷不溶物（INS) g	100	0.868	0	0.13	0.183	0	0.034	2.196	1.519	1.412
焦油产率	100	0.424	0.036	0.13	0.064	0.122	0.004	10.46	2.99	0.495
水产率	100	0.276	0.058	0.125	0.043	0.112	0.002	3.428	1.571	0.344
焦渣产率	100	0.76	0.278	0.633	0.105	0.646	0.011	4.585	-2.095	0.166
正己烷可溶物产率	100	0.889	0	0.049	0.098	0	0.01	55.163	6.572	2.018

从表中得出正己烷不溶物(INS)平均值约为 0.13，标准差约为 0.183，表明此项数据点更远离均值，分布较为离散，具有一定的波动性。且基于焦油产率、水产率、焦渣产率变异系数（CV）分别为 0.495，0.344，0.166 均大于 0.15，显示出不同程度的差异性。据此将数据可视化，绘制散点图与如下：





整个散点图的分布表明，正己烷不溶物与焦油产率、水产率、焦渣产率之间可能存在某种关联，但这种关联也不直接表现为简单的线性关系。由于数据点分布较为分散，要更准确地评估 INS 与热解产物之间的关系，需要进行统计分析，如计算皮尔逊相关系数或者绘制散点图的趋势线来探索潜在的相关性。

5.1.3 相关性分析及线性回归模型的建立

我们定义如下变量：

1、T：INS 的质量比例

2、Y_t：焦油产率

3、Y_w：水产率

4、Y_c：焦渣产率

5、X:INS 的量

一、首先，我们利用皮尔逊相关系数来衡量正已烷与焦油产率、水产率、焦渣产率的相关性，皮尔逊相关系数衡量的是线性相关关系，度量两个变量 X 和 Y 之间的相关程度，其值介于-1 与 1 之间。若系数等于 0，则说明两个变量之间无线性相关关系，不存在无相关关系。相关系数的绝对值越大，相关性越强；相关系数越接近于 1 或-1，相关度越强；相关系数越接近于 0，相关度越弱。公式如下：

$$r_{xy}=\frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum\left(x_i-\bar{x}\right)^2\sum\left(y_i-\bar{y}\right)^2}} \quad (1)$$

将表格数据结果代入公式，利用 Python 代码得出正已烷与三项热解产物产率相关系数如下表：

	正已烷不溶物（INS)g	焦油产率	水产率	焦渣产率
正已烷不溶物 (INS)g	1	0.209114	0.08446	-0.199367
焦油产率	0.209114	1	0.257735	-0.541579
水产率	0.084460	0.257735	1	-0.845404
焦渣产率	-0.199367	-0.541579	-0.845404	1

分析结果：
1、正已烷不溶物（INS）与焦油产率的相关系数为 0.209，存在较弱的正相关关系；
2、正已烷不溶物（INS）与水产率的相关系数为 0.084，关系轻微；
3、正已烷不溶物（INS）与焦渣产率的相关系数为-0.199，存在较弱的负相关关系。
二、为了进一步确认正已烷对油产率、水产率、焦渣产率是否产生显著影响，我们建立线性回归模型进行回归分析：

$$\begin{aligned} Y_t &= \alpha_0+\alpha_1 X+\epsilon \\ Y_w &= \beta_0+\beta_1 X+\epsilon \\ Y_c &= \gamma_0+\gamma_1 X+\epsilon \end{aligned} \quad (2)$$

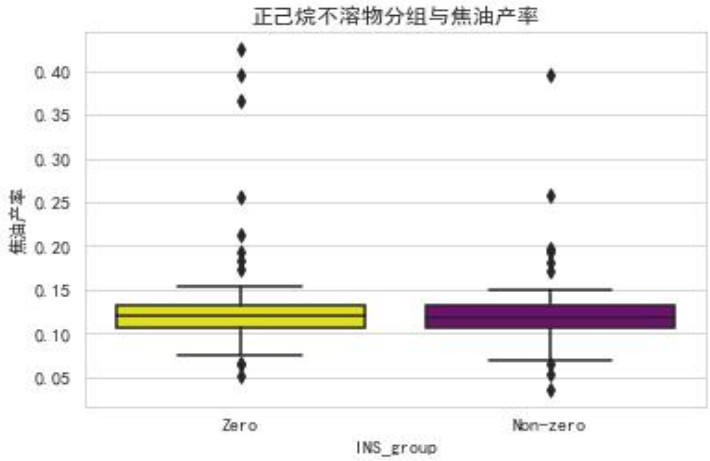
5.1.4 线性回归模型的求解

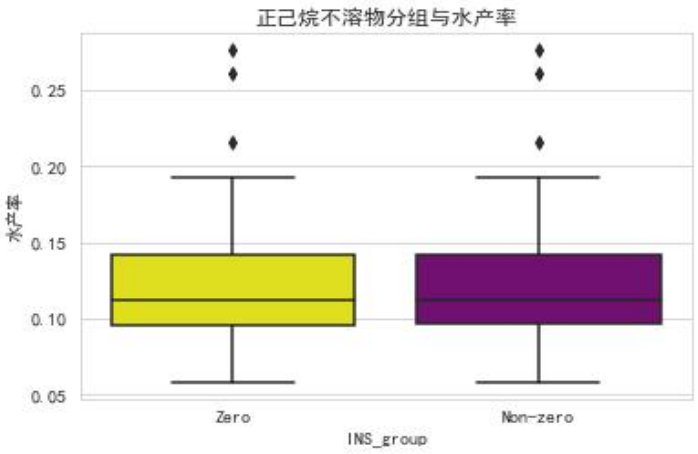
我们使用最小二乘法来估计上述模型的参数，进行 t 检验来评估每个模型参数的显著性，利用 Python 代码编写程序求解上述模型。部分关键值如下表：

	R-squared	系数	p 值
焦油产率	0.044	0.0611	0.015
水产率	0.007	0.0243	0.412
焦渣产率	0.04	-0.0989	0.02

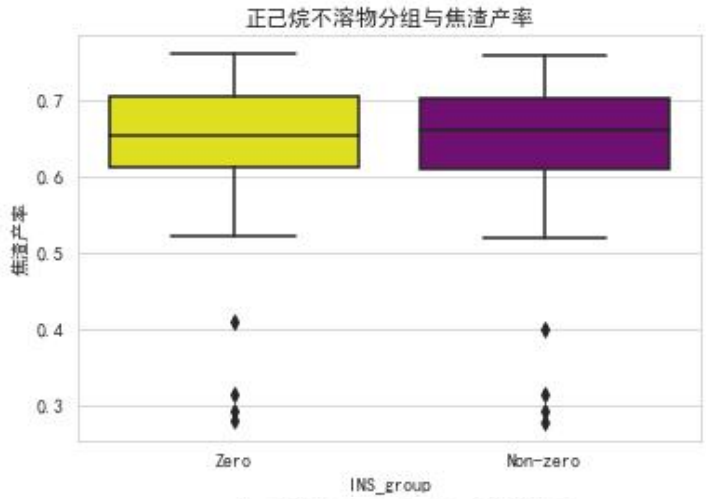
三者的R^2（决定系数）均较小，说明正已烷不溶物对三者的影响力较小；而焦油产率和水产率的系数均为正值，表明随着 INS 的增加，二者产率均增加，水产率减少；焦油产率和焦渣产率的 p 值统计显著，说明 INS 对二者产率有显著影响，可能对减少焦渣产率具有一定的作用。

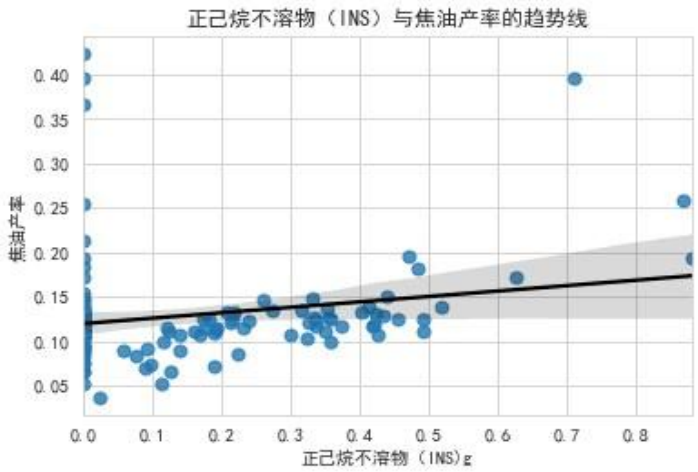
绘制箱线图和趋势图如下：

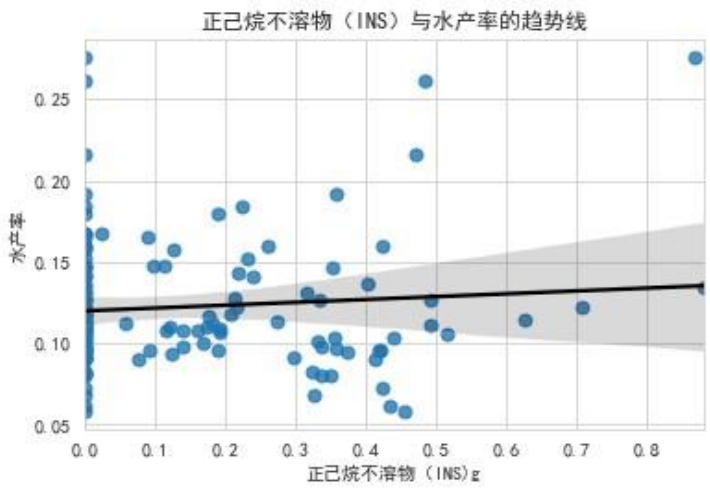


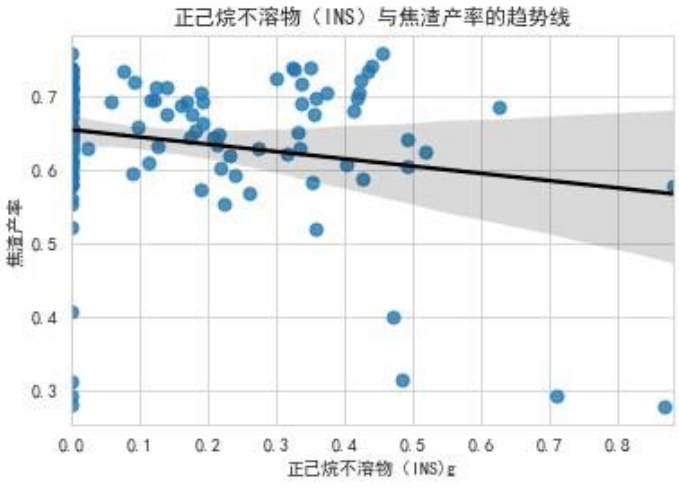


箱线图和趋势图进一步验证了描述性统计中的发现，且与回归分析结果相符合。









5.2 问题二模型建立与求解

5.2.1 交互效应分析模型

1.数据处理

在热解实验的数据集中，我们将处理正己烷不溶物(INS)的质量（表示为 m）以及生物质和煤的混合比例（表示为 p）。混合比例需要从原始的“配比”形式（如 5/100）转换为实际的比例值（即小数形式，如 0.05）。

2. 交互效应分析模型建立

为了量化 INS 和混合比例对热解产物产量的交互效应，我们将采用线性回归模型。该模型将考虑 INS 的量（m）、混合比例（p）以及它们的交互项（mp）。模型可以表示为：

$$Y = \alpha_0 + \alpha_1 m + \alpha_2 p + \alpha_3 mp + \varepsilon \quad （3$$

3. 变量定义

Y：热解产物的产率

X：正己烷不溶物的量。

配比：生物质和煤的混合比例（已转换为小数形式，如 0.05 表示生物质占 5%，煤占 95%）

$\alpha_0, \alpha_1, \alpha_2, \alpha_3$：模型参数。

ϵ：误差项。

4. 交互效应分析模型评估

通过拟合上述模型，我们检查α_1、α_2和α_3的估计值和它们的统计显著性。特别地，如果α_3显著不为零，这表明正己烷不溶物的量（X）和混合比例（配比）之间存在显著的交互效应。

5. 结果解释

根据模型的输出，我们可以确定哪些热解产物的产率受到 X 和配比交互效应的显著影响。如果α_3显著，则表明在改变 X 和配比时，热解产物的产率会以一种非线性的方式变化。

6. 数据可视化

绘制三维曲面图或等高线图来直观展示不同 X 和配比水平下热解产物产率的变化，这些图形将帮助我们更好地理解交互效应的性质。

5.2.2 实际计算与结果

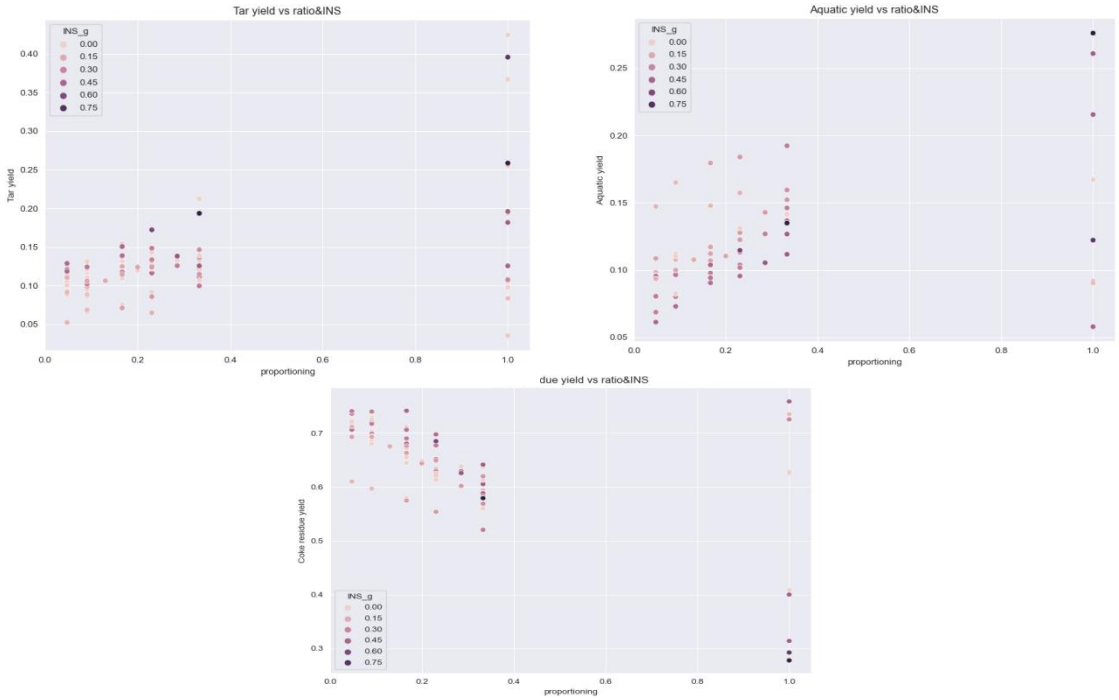
采用 Python 进行实际计算后，我们得到以下关于焦油产率模型的部分结果：

焦油产率模型结果：

Coefficients	Standard Errors	t-values	P-values	Conf. Interval Lower	Conf. Interval Upper
0.0981	0.007	13.383	0		0.113
0.0174	0.034	0.51	0.611		0.085
0.0835	0.017	4.965	0		0.117
0.0611	0.06	1.018	0.311		0.18

我们做如下可视化：

散点图可视化



通过散点图可视化，我们可以看到：
焦油产率和水产率随着混合比例的增加而增加，是在低正己烷不溶物浓度下更为明显焦渣产率随着混合比例的增加而降低，尤其在高正己烷不溶物浓度下，这种降低趋势更加显著。

5.2.3 交互效应对热解产物产率的影响分析

在评估热解过程参数对产物产率的影响时，我们特别关注了正己烷不溶物（INS_g）与混合比例之间的交互效应。以下是基于模型结果的主要发现：

焦油产率：

正己烷不溶物与混合比例的交互效应对焦油产率的影响并不显著（p 值=0.311），意味着两者在调控焦油产率时可能较为独立。然而，混合比例本身对焦油产率具有显著的正向影响，即随着混合比例的增加，焦油产率也相应增加。

焦渣产率：

在焦渣产率方面，正己烷不溶物与混合比例的交互效应呈现出高度显著性（p 值<0.001），并且该交互效应表现为负向影响。这意味着，在含有正己烷不溶物的条件下，混合比例的增加会导致焦渣产率更为显著地降低。
同时，混合比例对焦渣产率也表现出显著的负向影响，即混合比例的增加会直接导致焦渣产率降低。

水产率：

对于水产率而言，正己烷不溶物与混合比例的交互效应同样显著（p 值=0.019），且为正向影响。这提示我们，在含有正己烷不溶物的环境中，混合比例的增加会促进水产率的提升。

交互效应最为显著的热解产物：

焦渣产率受到正己烷不溶物与混合比例交互效应的影响最为显著，这一发现对于控制热解过程中的焦渣产率至关重要。
水产率虽也表现出显著的交互效应，但其影响程度相对焦渣产率较低。

结论：

我们的分析表明，焦渣产率和水产率是受正己烷不溶物与混合比例交互效应影响最为显著的热解产物。其中，焦渣产率的交互效应尤为显著，为优化热解过程参数提供了重要依据。通过调整这两个因素，我们可以实现更高效的能源转化和更好的产物利用，从而为热解技术的优化提供有价值的指导。

5.3 问题三模型的建立与求解

5.3.1 优化模型的确立

根据题目的要求，我们的主要目的是通过建立生物质和煤的混合比例的优化模型，使共热解产物中的焦油产率最大化。其中，为了能够实现煤的较高纯度和较低的处理成本，我们希望焦油的质量尽可能接近纯煤热解所产生的焦油。

焦油产率Y：取决于混合比例X和正己烷不溶物量I。

形式化为:

$$Y = f(I,X) \tag{4}$$

其中f是由数据驱动的回归模型。

2 数据处理：

将配比数据由字符串转换为数值,清洗和准备数据为模型的建立打好基础，利用线性回归模型来拟合焦油产率与正己烷不溶物量和混合比例的关系。

焦油产率:

$$Y_t = \alpha_0 + \alpha_1 m + \alpha_2 p + \alpha_3 mp + \varepsilon \tag{5}$$

其中：α_0、α_1、α_2和α_3是模型的参数，ε是误差项。

3 参数估计：

通过最小化观测值和模型预测之间的平方误差来估计参数。

4 优化问题的定义：

确定目标函数，我们选择将焦油差率最大化：

$$\max X_{s,t} Y = f(I,X) \tag{6}$$

其中I是固定的正己烷不溶物量，X是可变混合比例，约束在 0 到 1 之间。

5.3.2 边界优化模型的建立与求解

5.3.2.1 优化算法介绍

通过调整优化算法的搜索范围在优化过程中处理自变量的边界条件从而确保解不会超出指定范围，并以此作为单个变量的取值范围来进行优化。其中我们主要利用了 minimize 函数在 scipy.optimize 包中的 Brfent’s Method 算法来解决这类问题：

1.Brent’s Method 方法是一种数值优化的迭代算法，用于寻找连续函数的局部最小值，它是一种解决无约束单变量优化非常有效的方法。

2.算法的内容主要是初始化已选中的初始点，模拟出对应函数的函数值，然后构造抛物线的插值，以估计最优解的位置，进而更新试探点，其中使用二分法确定是否接收试探点，最终确定更新区间。

3.Brent’s Method 方法在寻找单变量函数的最小值通常具有很高的效率和可靠性。

5.3.2.2 模型的建立

根据公式$\max X_{s,t} Y = f(I,x)$我们选择将在边界$X\epsilon[0,1]$内最大化Y转化为最小化$-Y$即使用 minimize_scalar，来寻找X的最优值：

$$\min mize - Y = -f(I,X) subject\ to\ X\varepsilon[0,1] \quad (7)$$

该方法直接应用 Brent 方法，高效地在指定区间内找到了最大化焦油产率的混合比例。

5.3.2.3 优化结果

最优配比：0.99999403913548
预测的最大焦油产率：0.19278573992795

而最优配比接近 1，说明了最大化焦油产率几乎全部是因为煤的热解，而这一结果也正与题目中关于更加倾向煤焦油的质量相呼应。预测的最大焦油产率数据表明了在最优条件下，焦油的产率约为 19.28%。

5.3.2.4 结论

根据 Brent 优化算法运算得出的结果，产生最高的焦油产率的条件是混合比例接近于完全煤的条件（即X接近于 1），并且焦油质量与纯煤热解产生的焦油相近。最终达到了利用煤的高纯度优势，同时减少由生物物质引入杂质和相关处理成本的目的。并且总结得出应尽量增增加煤在混合物中的比例来达到最大化焦油产率和质量，而这一结论为共热解过程的操作条件奠定了基础。

5.4 问题四模型建立与求解

5.4.1 数据预处理

为了方便计算，我们将附件二数据：
整理为如下形式：

GA/NM	5/100	10/100	20/100	30/100	50/100
Tar	6.9	8.86	9.56	11.52	13.26
Calculated tar	6.08	8.98	12.14	14.82	19.1
HEX	5.46	7.69	7.07	8.46	8.79
Calculated HEX	6.08	7.23	9.25	10.95	13.67
Water	9.77	10.11	12	12.58	13.62
Calculated water	10.12	9.88	9.47	9.12	8.56
Char	60.28	58.65	56.17	53.35	49.56
Calculated char	60.32	58.69	55.84	53.42	49.56
GA/HN	5/100	10/100	20/100	30/100	50/100
Tar	16.9	16.98	18.85	20.14	14.8
Calculated tar	17.48	19.05	21.73	23.94	27.37
HEX	12.81	12.3	13.26	15.61	8.17
Calculated HEX	12.46	13.46	15.18	16.6	17.92
Water	6.07	7.24	7.95	8.87	9.22
Calculated water	4.83	4.84	4.85	4.86	4.88
Char	66.54	64.41	60.56	57.45	52.85
Calculated char	68.56	66.56	63.05	60.08	55.33

5.4.2 配对样本 t 检验

理论知识

配对样本 t 检验（Paired Samples t-Test）通常用于比较同一组对象在不同条件、不同时间或不同处理下的两个观测值之间的差异。这种检验假设两个样本之间的差异服从正态分布，并且两个样本的观测值是配对的（即来自同一组对象）。

理论基础

配对样本 t 检验（Paired Samples t-Test）的基本假设是，对于每一对观测值（即来自同一对象的两个不同条件下的测量值），其差异（d）构成了一个样本，这个样本应该来自一个近似正态分布的总体。这个假设是进行配对样本 t 检验的前提，它确保了我们可以使用 t 分布来推断两个配对样本均值之间是否存在显著差异。

该检验的主要目标是检验两个配对样本的均值之间的差异是否显著不等于零。换句话说，我们想要确定这两个样本所代表的总体均值之间是否存在统计上的显著差异。

设有两个样本，每个样本都由 n 个观测值组成：

$$\text{样本 1：}\ X_1, X_2 \ldots\ldots X_n$$
$$\text{样本 2：}\ Y_1, Y_2 \ldots\ldots Y_n$$

对于每一个观测值对（即同一对象的两个不同条件下的测量值），我们计算它们的差异：

$$\backslash d_i = X_i - Y_i \tag{8}$$

整个样的差异的均值和标准偏差分别计算如下：

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{9}$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}{n-1}} \tag{10}$$

t 统计量的计算
t 值是用来衡量样均值之间差异的大小相对于样本中这些差异的离散程度。t 统计量计算公式如下：

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{11}$$

自由度和 P 值
在配对样本 t 检验中，自由度（degrees of freedom, df）通常是 n-1，其中 n 是配对的数量。这是因为我们在计算 t 统计量时使用了 n 个差异值，但同时也估计了一个参数（即差异的总体均值），因此减去了 1 个自由度。
t 统计量的值与 t 分布进行比较，以确定 P 值（也称为显著性水平或观察到的显著性）。P 值表示在零假设（即两个样本之间的差异均值为零）为真的情况下，观察到这样或更极端结果的概率。

应用场景：

配对样本 t 检验适用于多种场景，特别是在需要对同一组实验对象在两种不同条件下进行比较时。以下是几个典型的应用场景：

1. 同一组个体在两种不同处理或条件下的测量比较。
2. 个体在接受某种治疗或干预前后的状态或指标比较。
3. 在教育或心理学研究中，同一组学生在不同时间点的表现或技能水平比较。
4. 任何需要评估“前后”变化或相同对象在不同条件下的表现差异的情况。

检验结果：

在进行配对样本 t 检验后，通常会得到一系列结果，包括 t 值、自由度、P 值以及差异均值的估计值等。以下是一个简化的检验结果展示（假设只展示前 9 条关键信息）：
这些结果提供了关于两个配对样本之间是否存在显著差异的详细信息，并帮助研究人员和统计分析师做出基于数据的决策。

Combination	Product	T-Statistic	P-Value
CS/HN	Char	-4.411591	0.011586
CS/SM	HEX	4.508776	0.01075
CS/SM	Water	-3.187304	0.033302
CS/HS	Water	3.873166	0.017945
CS/HS	Char	-5.024556	0.007362
SD/SM	HEX	8.511092	0.001045
SD/SM	Char	-6.231656	0.003378
SD/HS	n-hexane soluble(HEX)	3.661904	0.021544
GA/HN	Water	5.385132	0.005749
GA/HN	Char	-20.499814	0.000033
RH/HN	Tar	-8.983966	0.00085
RH/HN	HEX	-11.196172	0.000362
RH/HN	Water	6.913362	0.002297
RH/SM	HEX	5.563402	0.005112
RH/SM	Char	-3.374359	0.027931

显著组合筛选：

我们将从修正后的结果中筛选出存在显著差异的组合和产物(即 P 值小于 0.05)：

Combination	Product	T-Statistic	P-Value
CS/HN	Char	-4.411591	0.011586
CS/SM	HEX	4.508776	0.01075
CS/SM	Water	-3.187304	0.033302
CS/HS	Water	3.873166	0.017945
CS/HS	Char	-5.024556	0.007362
SD/SM	HEX	8.511092	0.001045
SD/SM	Char	-6.231656	0.003378
SD/HS	n-hexane soluble(HEX)	3.661904	0.021544
GA/HN	Water	5.385132	0.005749
GA/HN	Char	-20.499814	0.000033
RH/HN	Tar	-8.983966	0.00085
RH/HN	HEX	-11.196172	0.000362
RH/HN	Water	6.913362	0.002297
RH/SM	HEX	5.563402	0.005112
RH/SM	Char	-3.374359	0.027931

5.4.3 子组分析

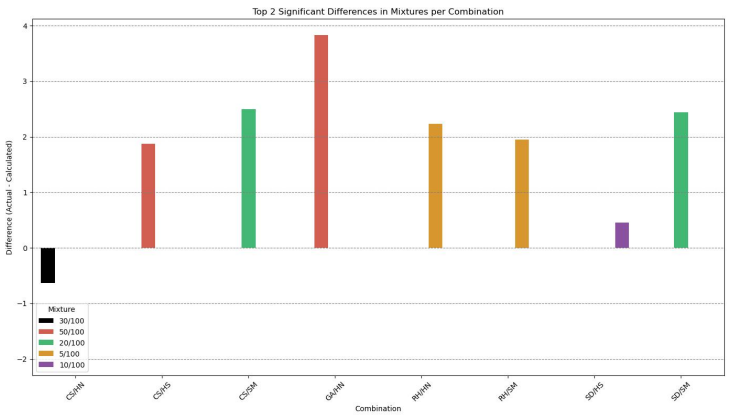
在观察到总体样本中存在显著差异后，为了进一步深入探索这些差异的具体来源和表现形式，我们将进行子组分析。子组分析的目的在于确定实验值与理论计算值之间的差异在哪些特定的混合比例或条件下表现得最为显著。

具体而言，子组分析将涉及以下几个步骤：

混合比例划分：首先，我们将总样本按照不同的混合比例或条件进行划分，形成多个子组。这些子组可以基于实验设计中的不同变量水平、处理条件或时间点等因素进行定义。
差异计算：对于每个子组，我们将计算实验值与理论值之间的差异。这可以通过计算每个子组内观测值的平均值或中位数，并与相应的理论值进行比较来实现。

差异评估：接下来，我们将对这些差异进行定量评估。这可以通过计算差异的绝对值、平方差、百分比误差等指标来衡量差异的大小和显著性。同时，我们也可以使用统计测试（如 t 检验、方差分析等）来评估这些差异是否具有统计学意义。
结果展示：最后，我们将展示子组分析的结果。这可以通过表格、图表或文字描述等形式来呈现。在结果中，我们将重点关注那些表现出显著差异的子组，并探讨这些差异的可能原因和解释。

在深入分析热解过程数据时，我们识别出在特定混合比例下，实验值与理论计算值之间存在显著的差异。例如，在 CS/HN 组合的焦渣(Char)产品中，我们观察到在 50/100 的比例下差异最大，达到了-3.46。为了更加直观地呈现这些关键信息，我们决定采用一种更为聚焦的可视化方法。
考虑到实际应用中优化和调整的需求，我们决定直接展示每个组合中差异最大的前两个混合比例。这种方法不仅实用，而且能够迅速识别出需要重点关注和调整的具体条件。我们将为每个组合提取出差异最大的两个混合比例，并通过可视化的方式展现出来。
通过这种方式，我们能够更直观地看到哪些混合比例和条件下实验值与理论值的差异最为显著，从而为热解过程的优化提供有力的数据支持。这种可视化的方法将有助于研究人员和工程师更快速、更准确地识别出需要调整的参数，从而进一步提高热解过程的效率和产物质量。下图是可视化结果：



5.2 问题 5 的模型建立与求解

5.5.1 数据预处理

对于附件一，我们使用 LabelEncoder 函数将试样列中的分类数据转换为数字标签，并将结果存储在新的试样编码列中，使用 lambda 函数将配比列字符串形式的分数转换为浮点数形式如下，以便后续建模：

试样	配比	试样编码
淮南煤(HN)	100	12
淮南煤(HN)	100	12
神木煤(SM)	100	13
神木煤(SM)	100	13
神木煤(SM)	100	13
神木煤(SM)	100	13
内蒙褐煤(NM)	100	0
内蒙褐煤(NM)	100	0
内蒙褐煤(NM)	100	0
黑山煤(HS)	100	17
黑山煤(HS)	100	17
棉杆(CS)	100	8
棉杆(CS)	100	8
木屑(SD)	100	5
木屑(SD)	100	5
小球藻(GA)	100	1
小球藻(GA)	100	1
稻壳(RH)	100	14
稻壳(RH)	100	14
棉杆/淮南煤(CS/HN)	0.05	9
棉杆/淮南煤(CS/HN)	0.05	9
棉杆/淮南煤(CS/HN)	0.1	9
棉杆/淮南煤(CS/HN)	0.1	9

5.5.2 随机森林预测模型的建立

基于热解反应的复杂性，假设热解产物产率 y 是自变量 x 的非线性函数：

$$y = f(x) + \epsilon \tag{12}$$

其中，ϵ是产生的随机误差，我们通过随机森林算法进行对 f(x)的拟合。
随机森林是从原始训练样本集N中有放回地重复随机抽取k个样本生成新的训练样本集合，然后根据自助样本集生成 k 个分类树组成随机森林，新数据的分类结果按分类树投票多少形成的分数而定。其实质是对决策树算法的一种改进，最终结果通过投票或取均值，使得整体模型的结果具有较高的精确度和泛化性能。随机森林的预测函数的公式如下：

$$f(x) = \frac{1}{A}\sum_{a=1}^{A} w_a(x;\theta_a) \tag{13}$$

5.5.3 支持向量回归（SVR）预测模型的建立

支持向量机回归是支持向量机算法的回归扩展,SVR 采用非线性方式建模,用来处理非线性回归问题。原理是用非线性映射将数据映射到高维数据特征空间中,使得在高维数据特征空间中自变量与因变量具有很好的线性回归特征,在该特征空间进行拟合后再返回到原始空间。其基本模型如下：

$$f(x) = < v, x > + b \tag{14}$$

$$f(x) = \sum_{i=1}^{n} (\delta_i - \delta_i^*) S(x, x_i) + b \tag{15}$$

其中 V 是权重向量，b 是截距项，δ 和δ_i^* 是拉格朗日乘数，$s(x, x_i)$是核函数。

5.5.4 模型求解

我们设置决策树的数量为 100，确立随机种子作为训练集，使用训练好的模型对测试集进行预测，利用均方误差（MSE）和决定系数（R2）评估模型的性能，使用 Python 代码编写程序得到两个模型预测精度结果如下：

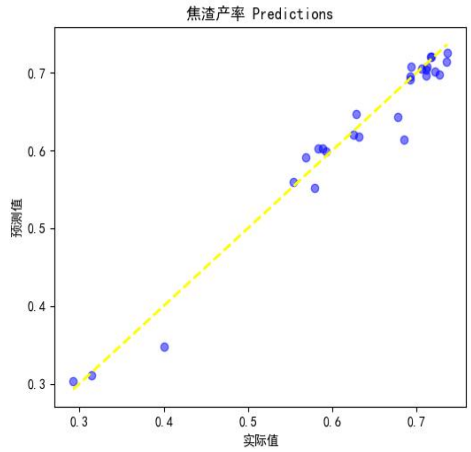
Random Forest Results:	MSE	R^2
焦油产率	0.0005	0.87
水产率	0	0.98
焦渣产率	0.0005	0.96
正己烷可溶物产率	0.003	0.19

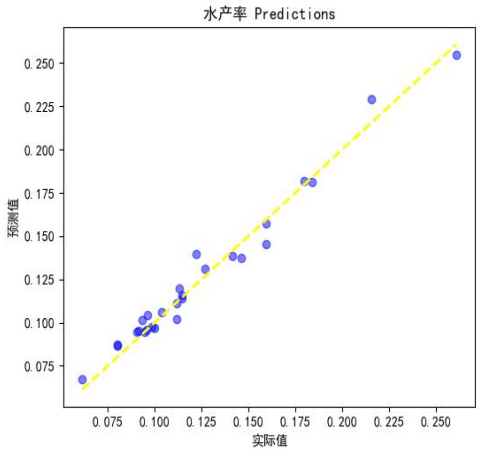
SVR Results:	MSE	R^2
焦油产率	0.0028	0.22
水产率	0.0033	-0.69
焦渣产率	0.0047	0.67
正己烷可溶物产率	0.0059	-0.57

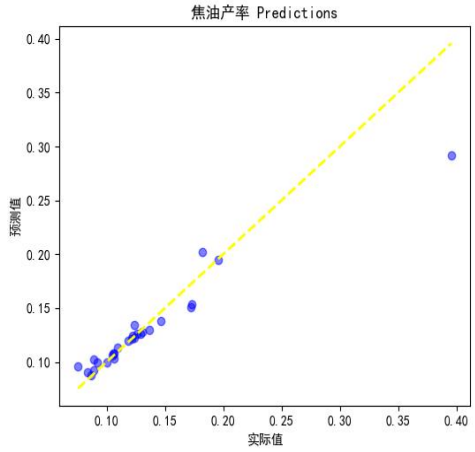
从表中看出，随机森林整体预测精度较好，故使用随机森林算法进行预测，展示随机三个预测结果如下：

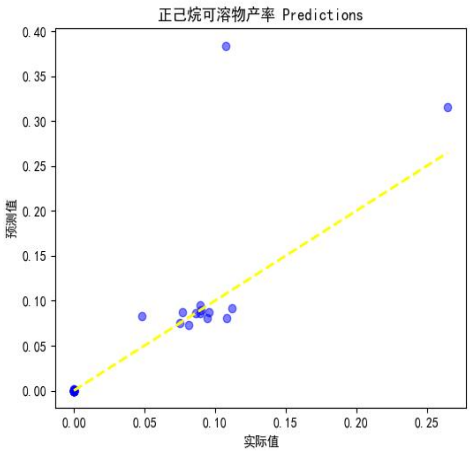
样本 85:	真实值	预测值
焦油产率	0.1236	0.1224
水产率	0.0802	0.0874
焦渣产率	0.7171	0.7206
正己烷可溶物产率	0.0895	0.0861
样本 27:		
焦油产率	0.1292	0.126
水产率	0.1268	0.1314
焦渣产率	0.6317	0.6179
正己烷可溶物产率	0	0
样本 4:		
焦油产率	0.1054	0.1079
水产率	0.0914	0.0955
焦渣产率	0.7269	0.6981
正己烷可溶物产率	0	0

预测结果图如下：









预测结果显示焦油产率、水产率、焦渣产率拟合结果较好，而正己烷可溶物产率的模型表现较差,这可能是因为该产物的动态范围和复杂性较高。

六、模型的评价与推广

（一）模型优点

1、本文采用的线性回归模型具有直观的结构，提供了清晰的参数估计，相比于复杂的机器学习模型，线性回归在计算上更加高效，能够快速地给出结果，便于实际应用和结果更新。

2、本文采用的随机森林模型通过整合多棵决策树的预测结果，降低了过拟合的风险，提高了模型的泛化能力，使其在不同数据集上都有良好的表现，具有强鲁棒性。

3、随机森林能够出色地处理数据中的非线性关系，特别适用于复杂的分类和回归任务。

（二）模型的缺点

1、线性假设限制:线性回归假设变量间关系是线性的,这限制了其在处理非线性关系时的有效性。
2、在大数据集上,随机森林的训练时间可能较长，需要较多的计算资源

（三）模型的推广

这些模型可能够为各行各业提供精准的决策支持和优化的建议。举例来说，在能源行业中，这些模型通过精确预测和优化热解过程，可以显著提升能源转化效率，从而有效降低成本。而在环保领域，它们同样具有强大的应用潜力，帮助企业实现废物减量化、资源化和无害化，降低污染物排放，为可持续发展贡献力量。

七、参考文献

●[1]张明昊,卓翔芝.基于模糊信息粒化和支持向量机的Brent原油期货价格预测[J].青岛大学学报(自然科学版),2021,34(04):127-132.

●[2]孙德山.支持向量机分类与回归方法研究[D]. 中南大学,2004.

● [3]刘琳岚;高声荣;舒坚.基于随机森林的链路质量预测[J]. 通信学报,2019(04).22

●[4]王彦顺,肖瑞瑞,丛兴顺,等.生物质与煤共热解动力学特性研究[J].广东化工,2023,50(05):20-21+16.

●[5]马萌,白永辉,卫俊涛,等.生物质与煤(共)热解/气化过程中挥发分-半焦交互作用研究与进展[J].化工学报,2022,73(11):5186-5200.

八、附录

附录 1
介绍：支撑材料的文件列表
附录 2 对模型进行检验 附录 3 ....

附录 1
介绍：该代码是 Matlab 语言编写的，作用是对模型的检验

```python
import pickle

# 载入焦油产率模型
with open('model_tar.pkl', 'rb') as f:
    model_t = pickle.load(f)

# 载入水产率模型
with open('model_w.pkl', 'rb') as f:
    model_w = pickle.load(f)

# 载入焦渣产率模型
with open('model_c.pkl', 'rb') as f:
    model_c = pickle.load(f)

# 现在可以使用这些模型进行优化或其他分析

#%%
import pandas as pd
from scipy.optimize import minimize_scalar
from statsmodels.formula.api import ols

# 加载数据
data_path = '附件一填充后.xlsx'
data = pd.read_excel(data_path)

# 配比数据处理，将配比转换为数值形式
def convert_ratio(r):
    if isinstance(r, str) and '/' in ratio:
        num, denom = map(int, ratio.split('/'))
        return num / (num + denom)
    elif isinstance(ratio, str):
        return int(ratio) / 100
    else:
        return ratio / 100

data['配比'] = data['配比'].apply(convert_ratio)

# 重命名列，避免编码问题
data = data.rename(columns={'正己烷不溶物（INS)g': 'INS_g'})
```

```

# 假定正己烷不溶物量为一个典型值，例如平均值
fixed_ins = data['INS_g'].mean()

# 使用已构建的模型进行优化
model_t = ols('焦油产率 ~ INS_g * 配比', data=data).fit()

# 从模型中提取参数
params = model_t.params
intercept, coef_ins, coef_ratio, coef_interaction = params['Intercept'], params['INS_g'], params['配比'], params['INS_g:配比']

# 定义目标函数，我们希望最大化焦油产率
def objective(x):
    # x 是配比
    return -(intercept + coef_ins * fixed_ins + coef_ratio * x + coef_interaction * fixed_ins * x)

# 找到最优配比，限定范围在 0 到 1 之间
r = minimize_scalar(objective, bounds=(0, 1), method='bounded')

# 输出优化结果
print("最优配比：", r.x)
print("预测的最大焦油产率：", -r.fun)

#%%
```