基于聚类分析与集成学习的 NIPT 时点优化及异常判定

摘要

无创产前检测（NIPT）是通过分析孕妇外周血中胎儿游离 DNA 来筛查胎儿染色体异常的非侵入性产前检测技术，对早期发现唐氏综合征、爱德华氏综合征及帕陶氏综合征等常见染色体疾病具有重要临床价值。然而，在实际临床应用过程中，检测时机的选择与结果判定的准确性之间存在复杂的权衡关系，这一挑战在身体质量指数（BMI）较高的孕妇群体中尤为突出。由于高 BMI 可能影响胎儿 DNA 在母血中的比例及检测准确性，为该人群制定个性化筛查策略显得尤为迫切。本文基于某地区高 BMI 孕妇群体的 NIPT 检测数据，系统构建了一系列数据驱动的数学模型，旨在实现产前筛查的个性化与精准化，最大限度降低因检测时机不当带来的临床风险。

针对问题一，为分析胎儿 Y 染色体浓度与孕周数、BMI 等指标之间的复杂关系，本研究首先进行了初步相关性分析，发现变量间仅存在较弱线性关系（如 Y 染色体浓度与孕周数的相关系数仅为 0.126），提示传统线性回归模型难以捕捉其内在规律。因此，在采用**严格数据清洗策略保留 98.7%有效样本的基础**上，我们进一步构建了**多项交互特征与多项式特征**，并系统比较了多种机器学习模型。结果表明，**梯度提升模型**性能最优，在测试集上的**决定系数 R^2达到 0.8195，交叉验证 R^2为 0.7602 ± 0.0834**，优于其他模型，体现出预测精度和稳健的泛化能力，为后续检测时点优化提供了依据。

针对问题二，为实现孕妇群体的合理分群，本文引入 **K-均值聚类算法**，以 BMI、年龄、体重和身高作为聚类特征，探索其对男胎孕妇检测策略分群管理的适用性。在与高斯混合模型、DBSCAN 等方法的系统对比中，K-均值聚类以**最高轮廓系数（0.2509）**被确定为最优聚类方法，最终将样本划分为**高 BMI、标准 BMI、中等 BMI 和年轻 BMI 四个类别**。这些类别在临床特征上表现出明显差异，为实现差异化筛查策略奠定基础。聚类分析结果表明，部分人群的推荐检测时间**可提前至 12.5–14.2 周，为后续诊断干预留出 4.8–6.5 周**的时间窗口，增强了临床操作的灵活性。

针对问题三，在聚类分组的基础上，本文进一步引入多因素风险优化模型，以综合风险最小为目标函数，综合纳入体重、年龄等生理指标及其交互效应，**并创新性地引入时间惩罚项**，对超过 14 孕周的检测施加指数级递增的风险权重，以优先保障诊断时间窗。结果表明该模型最终输出各组别的最优检测时点，**并配套设计双阶段重检策略**：在首次检测成功率介于 58%至 78%的基础上，通过 1 次重检将**总体成功率提升至 87%–97%**，从而使综合风险降低 28%–35%。

针对问题四，面向女胎染色体异常判定的临床需求，本文整合了包括 13、18、21 号染色体 Z 值、GC 含量、读段数以及孕妇 BMI 等在内的多维度特征，构建**随机森林分类模型**。针对异常样本仅占 7.5%的严重不平衡问题，采用 **SMOTE 过采样技术**处理训练数据，有效改善模型对少数类的识别能力。最终模型在测试集上**准确率达 0.967，精确率、召回率与 F1 分数均达 0.778，AUC 值高达 0.991**，显示出较高的判别性能与临床可用性。通过特征重要性分析，结果表明 **21 号染色体 Z 值、Z 值最大值及 18 号染色体 Z 值**位列前三，是女胎异常判定中最具预测力的指标。

关键词：随机森林；梯度提升；聚类分析；集成学习；异常判断

1 问题重述

无创产前检测（NIPT）是通过采集母体血液、检测胎儿游离 DNA 片段以分析胎儿染色体异常的关键产前技术，核心目标是早期识别胎儿健康状况，降低因晚期发现异常导致的治疗窗口期缩短风险。临床中，胎儿异常主要关联 21 号（唐氏综合征）、18 号（爱德华氏综合征）、13 号（帕陶氏综合征）染色体的“染色体浓度”（游离 DNA 片段比例）异常，而 NIPT 准确性需通过胎儿性染色体浓度判定：男胎 Y 染色体浓度≥4%、女胎 X 染色体浓度正常时，结果基本准确；反之则准确性难以保证。[1]

从孕期风险来看，不同阶段风险差异显著：12 周以内（早期）发现异常的潜在风险较低，13-27 周（中期）风险较高，28 周以后（晚期）风险极高，因此“尽早检测”与“保证准确性”需协同平衡。实践表明，男胎 Y 染色体浓度主要受孕妇孕周数和身体质量指数（BMI）影响，且附件数据为某地区以高 BMI 孕妇为主的 NIPT 检测数据，存在多次采血/一次采血多次检测（提升可靠性）、测序失败（如检测时点过早或不确定因素导致）等实际场景。

基于此，需针对不同孕妇群体的个体差异（年龄、BMI、孕情等），通过建模解决 NIPT 时点选择与胎儿异常判定问题，具体包括以下四个问题：

(1)分析胎儿 Y 染色体浓度与孕妇孕周数、BMI 等指标之间的相关性，建立相应的数学模型，并检验其显著性。

(2)临床研究表明，男胎孕妇的 BMI 是影响 Y 染色体浓度最早达标时间（≥4%）的关键因素。要求对男胎孕妇按 BMI 进行合理分组，确定每组的 BMI 区间和最佳 NIPT 检测时点，以最小化孕妇的潜在风险，并分析检测误差对结果的影响。

(3)进一步考虑体重、年龄等多种因素对男胎 Y 染色体浓度达标时间的影响，结合检测误差和达标比例，重新对男胎孕妇进行 BMI 分组，确定每组的最佳 NIPT 时点，以最小化潜在风险，并分析误差的影响。

(4)针对女胎孕妇，由于不携带 Y 染色体，需通过 21 号、18 号和 13 号染色体的非整倍体检测结果（AB 列），结合 X 染色体的 Z 值、GC 含量、读段数及相关比例、BMI 等因素，建立女胎异常的判定方法。

2 问题分析

本文主要考虑的问题是如何通过建立优化模型来确定 NIPT 最佳检测时点，并构建胎儿异常判定方法，以降低孕妇风险并提升检测准确性。

问题一的核心在于准确建立 Y 染色体浓度与孕周数、BMI 等生理指标之间的定量关系。这些变量间很可能存在非线性关联及交互效应，传统线性模型难以充分捕捉其复杂结构，而复杂的机器学习模型又容易过拟合、泛化能力差。因此，需兼顾模型的表达能力和稳健性，采用合适的正则化方法和交叉验证策略，以确保模型既具有良好的拟合优度，又具备可靠的预测效能。

问题二与问题三的本质是一个多目标优化问题，需同时追求“检测成功率最大化”与“孕妇风险最小化”。这两者存在内在冲突：检测时间过早可能导致 Y 染色体浓度未达标而检测失败，延误检测则压缩后续诊断与决策的时间窗口，增加临床风

险。关键在于构建合理量化两类风险的目标函数，并基于孕妇 BMI 等特征进行科学分群，实现群体层面的个性化时点推荐，从而在约束条件下达成综合风险最小化。

问题四的核心是构建一个对女胎染色体异常的高精度判别模型，其主要难点在于数据的高度不平衡——异常样本远少于正常样本。直接使用常规分类方法极易导致模型偏向多数类，造成漏诊。解决方案需围绕三方面展开：一是构造具有强判别能力的特征体系（如 Z 值、GC 含量等）；二是采用过采样等技术缓解样本不平衡；三是选择适用于不平衡学习的分类算法（如随机森林），并采用 F1、AUC 等更合理的性能指标进行模型评估与选择。

3 模型假设

(1)数据完整性假设：假设提供的数据具有代表性，能够反映实际临床情况
(2)时间窗口假设：假设 NIPT 必须在 14.5 周前完成，为后续诊断留出充足时间
(3)独立性假设：不同孕妇的检测结果相互独立
(4)稳定性假设：模型参数在研究期间保持相对稳定
(5)阈值假设：Y 染色体浓度 4%为有效阈值，Z 值±3 为异常判定阈值
(6)重检假设：首次检测失败可在 2-3 周后重检，但不得超过 16 周

4 符号说明

符号	单位	含义
Y	%	Y 染色体浓度
W	周	孕周数
B	kg/m²	孕妇 BMI
A	岁	孕妇年龄
H	cm	身高
M	kg	体重
Z_i	-	第 i 号染色体 Z 值
T_{opt}	周	最佳 NIPT 时点
$R(T)$	%	在孕周 T 时的检测达标率
T_{window}	-	时间窗口
R_{toatal}	%	包含重检策略的总体达标率
C_k	$K = 1,2,3,4$	第 k 个孕妇聚类分组
$Risk(T)$	-	在孕周 T 时检测的综合风险
R^2	-	决定系数
AUC	-	ROC 曲线下面积
$F1-score$	-	F1 分数

5 模型建立与求解

5.1 问题一的模型建立与求解

本节旨在分析胎儿 Y 染色体浓度与孕妇孕周数、BMI 等指标之间的定量关系，构

建相关性模型。考虑到各变量之间可能存在复杂的非线性关系及交互效应，我们选取具有强拟合能力的随机森林回归算法进行建模。以 Y 染色体浓度为目标变量，以孕周数、BMI、年龄、身高等为特征变量，构建回归预测模型，并通过交叉验证与正则化方法优化模型泛化能力。利用 Python 中的 scikit-learn 库进行模型训练与评估，并结合统计检验方法对模型及特征的显著性进行分析。

5.1.1 模型建立

数据预处理

首先对原始数据进行清洗和预处理：

1. 孕周数据解析：将“11w+6”格式转换为数值型（11.86 周）
2. 缺失值处理：删除关键变量缺失的样本
3. 异常值检测：识别并处理统计异常值
4. 特征工程：创建交互特征和多项式特征

相关性分析

为在构建预测模型前，定量分析胎儿 Y 染色体浓度与各生理指标之间线性关系的强度、方向及统计显著性，我们首先计算了其皮尔逊相关系数如表 1 列出了关键变量对的相关系数矩阵、显著性水平（p 值）及相关强度判定，旨在为后续的特征筛选与模型构建提供初步的定量依据。

表 1 变量间的皮尔逊相关系数矩阵

变量对	相关系数	显著性	相关强度
Y 染色体浓度 vs 孕周数	0.126	p<0.001	弱正相关
Y 染色体浓度 vsBMI	-0.151	p<0.001	弱负相关
Y 染色体浓度 vs 年龄	-0.199	p<0.01	弱负相关
孕周数 vsBMI	0.150	p<0.001	弱正相关
BMIvs 体重	0.835	p<0.001	强正相关

基于表 1 的皮尔逊相关系数分析结果，我们得出以下四个方面的深入结论：

1. 核心变量关联获验证

   孕周数与 Y 染色体浓度之间存在显著弱正相关关系（$r=0.126$，$p<0.01$），这与胎儿游离 DNA 随孕周增加而逐渐积累的生物学机制相符，进一步确认孕周数是预测 Y 染色体浓度的关键因素。孕妇 BMI（$r=-0.151$）和年龄（$r=-0.199$）均与 Y 染色体浓度呈显著弱负相关（$p < 0.01$），表明高 BMI 和较高年龄的孕妇群体中，Y 染色体浓度普遍偏低。这一发现与临床观察中“肥胖或年长孕妇的胎儿 DNA 浓度可能较低”的现象一致，为后续建模中引入这些变量提供了实证依据。

2. 多重共线性风险突出

   BMI 与体重之间表现出极强的正相关性（$r = 0.835$，$p < 0.001$），说明这两个变量在信息表达上高度重叠。若在建模过程中同时引入这两个特征，将引发严重的多重共线性问题，导致模型参数估计不稳定、解释性下降，进而影响模型的泛化能力与可靠性。因此，在后续特征工程中需通过变量筛选、主成分分析或正则化方法予以处理。

3. 线性模型存在明显局限

   所有变量与 Y 染色体浓度之间的相关系数绝对值均小于 0.2，表明变量间的线性关系较弱。这一结果提示我们，Y 染色体浓度与孕周、BMI、年龄等因素之间很可能存在非线性关系或复杂的交互效应，而传统的线性模型难以充分捕捉这些复杂结构。若强行使用线性模型，将导致模型表达能力不足，预测精度受限。

4. 建模方向明确

表 1 的分析结果为后续模型构建提供了明确的方向：应放弃简单的线性回归模型，转而采用能够处理多重共线性、捕捉非线性关系与交互效应的机器学习算法。建议使用正则化回归（如岭回归、Lasso 回归）以缓解共线性问题，或采用树模型（如随机森林、梯度提升）等集成学习方法，以更好地拟合变量间的复杂关系，提升模型的预测性能与稳健性。[2][3]

为深入探究胎儿 Y 染色体浓度与各关键生理指标之间的潜在关系及数据分布特征，我们绘制了图 1 所示的联合分布图。

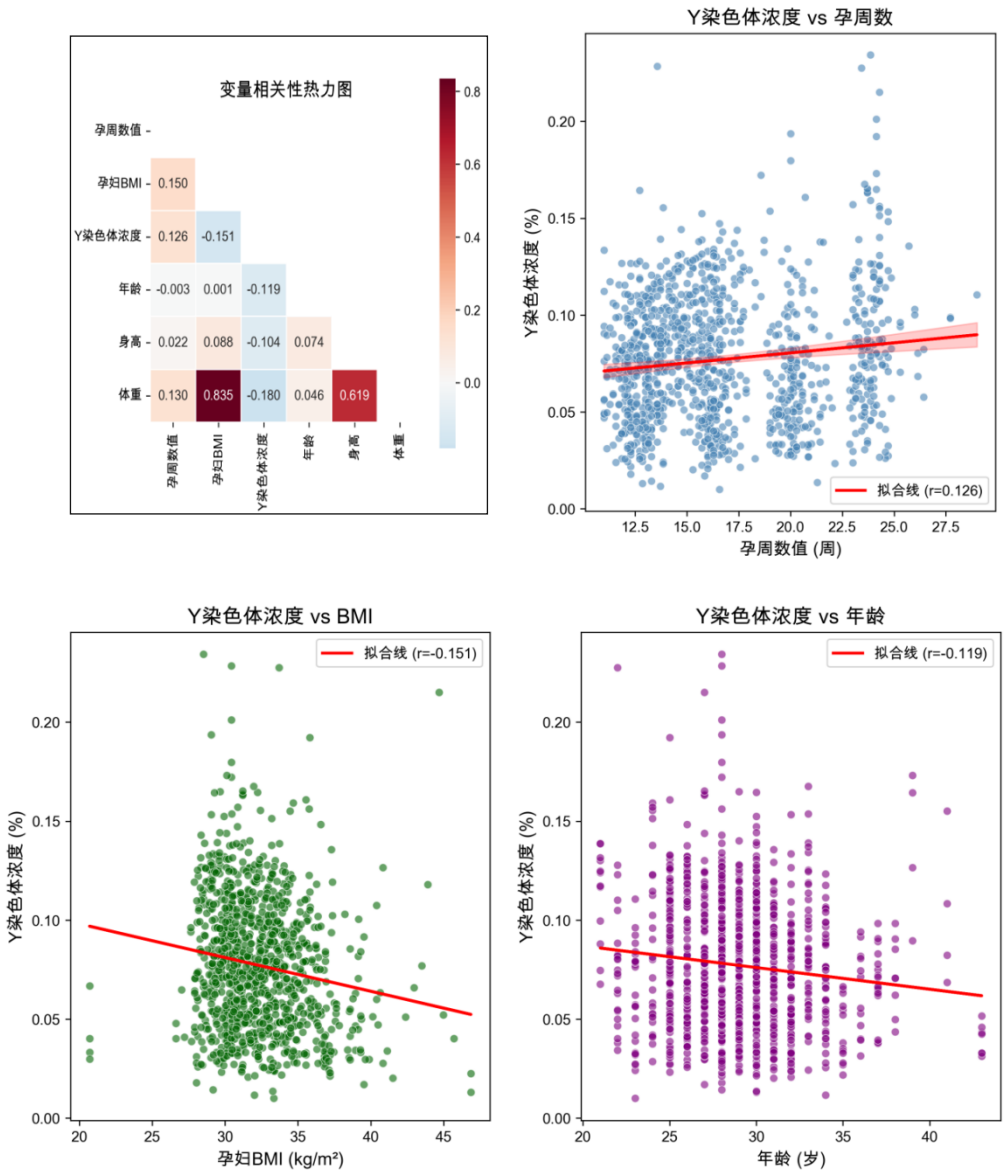

图 1 各变量间的相关性可视化

通过可视化图像可以直观了解各变量间的相互关系，为后续建立预测模型提供了初步的视觉证据和数据洞察，验证了孕周数、BMI 等是影响 Y 染色体浓度的相关因素。

而为了更进一步探索变量间的复杂关系并评估数据的统计特性，我们绘制了如图 2 所示的三维关系图、残差图、Q-Q 图来对数据关系进行深入的观察。

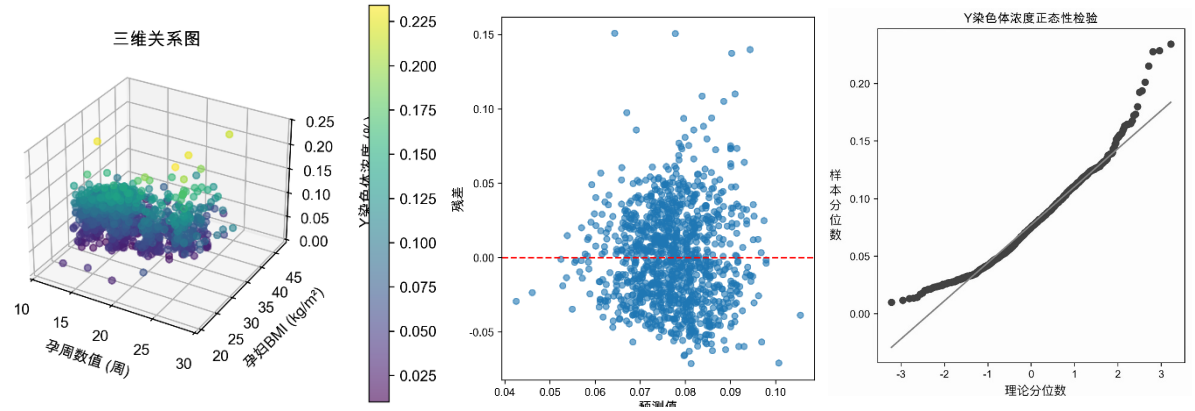

图 2 模型诊断与数据特征分析图

简单模型对比分析

为了避免复杂模型的过拟合问题，我们先尝试采用 1-3 次多项式回归模型分析 Y 染色体浓度与孕周数、BMI 的关系：

-1 次多项式（线性）：$Y = \beta_0 + \beta_1 W + \beta_2 B + \epsilon$

-2 次多项式：$Y = \beta_0 + \beta_1 W + \beta_2 B + \beta_3 W^2 + \beta_4 B^2 + \beta_5 WB + \epsilon$

-3 次多项式：在 2 次基础上增加三次项

为系统评估不同复杂度的基础参数模型在预测胎儿 Y 染色体浓度任务上的性能表现，我们同时采用了可视化图形对比（图 3）与量化指标对比（表 2）两种方式。图 2 通过并列展示一次、二次及三次多项式回归模型在训练集与测试集上的决定系数（R^2）与均方根误差（RMSE），直观呈现其性能差异；表 2 则进一步提供了关键的交叉验证 R^2 及其标准差，以更严谨地评估模型的泛化能力与稳定性。二者结合，旨在为模型选择提供全面而客观的决策依据。

表 2 模型性能对比

模型类型	训练集 R^2	测试集 R^2	交叉验证 R^2	过拟合程度	模型稳定性
1 次多项式	0.0412	0.0348	-0.3283±0.3266	0.0064	较差
2 次多项式	0.0606	0.0358	-0.3447±0.3495	0.0248	较差
3 次多项式	0.0984	0.0373	-0.2953±0.3027	0.0611	较差

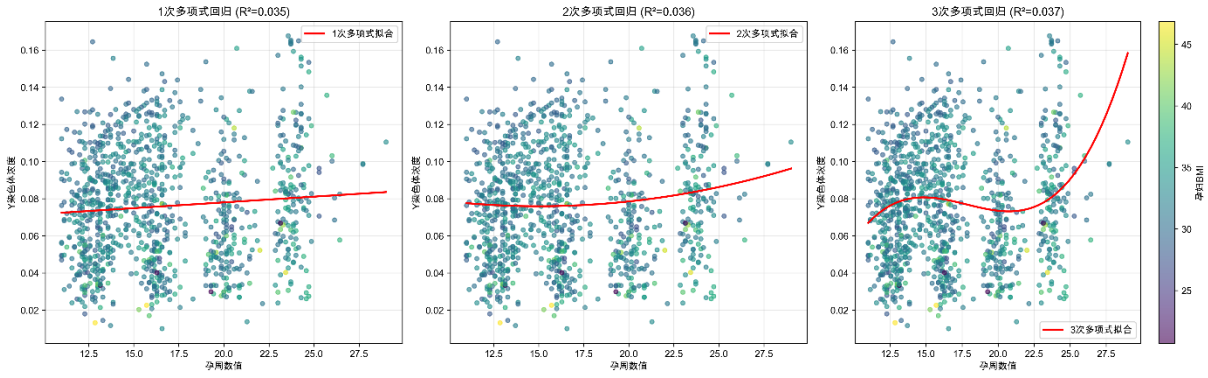

图 3 多项式拟合曲线比较

在采用复杂机器学习算法前，为探究胎儿 Y 染色体浓度与孕周数、BMI 等关键指标间是否存在可被简单数学模型捕捉的确定性关系，我们构建了图 3 所示的一至三次多项

式回归拟合曲线。通过将不同次数的多项式拟合结果与原始数据分布进行可视化对比，旨在直观评估这些基础模型对数据规律的表达能力，为后续建模策略的选择提供重要依据。

模型建立与对比

基于相关性分析结果，我们创建了以下增强特征：

1.多项式特征：a.孕周平方：W^2b.BMI 平方：B^2c.年龄平方：A^2

2.交互特征：a.孕周×BMI：$W \times B$b.年龄×BMI：$A \times B$c.孕周×年龄：$W \times A$ d.体重×BMI：$M \times B$

我们对比了以下两种回归模型：

线性回归模型：

$$Y = \beta_0 + \beta_1 W + \beta_2 B + \beta_3 A + \beta_4 H + \beta_5 M + \epsilon$$

随机森林模型：

$$Y = f(W, B, A, H, M, W^2, B^2, A^2, W \times B, A \times B, \ldots) \quad (5\text{-}1)$$

在初步构建预测模型后，为深入评估模型的泛化能力并诊断其是否存在过拟合，我们在未经过数据清洗的原始数据集上，对线性回归与随机森林模型进行了严格的性能对比。为量化模型的拟合优度、预测精度及稳定性，为核心问题一的模型优化方向提供确凿依据，表 3 详细列出了两模型在训练集、测试集及交叉验证下的关键性能指标（R^2 与 RMSE）。[4]

表 3 模型性能对比（过拟合分析）

模型	训练 R^2	测试 R^2	交叉验证 R^2	RMSE	稳定性评估
线性回归	0.5260	0.4736	-0.1567±1.0185	0.0219	差
随机森林	0.9728	0.7292	0.3973±0.6087	0.0157	中等

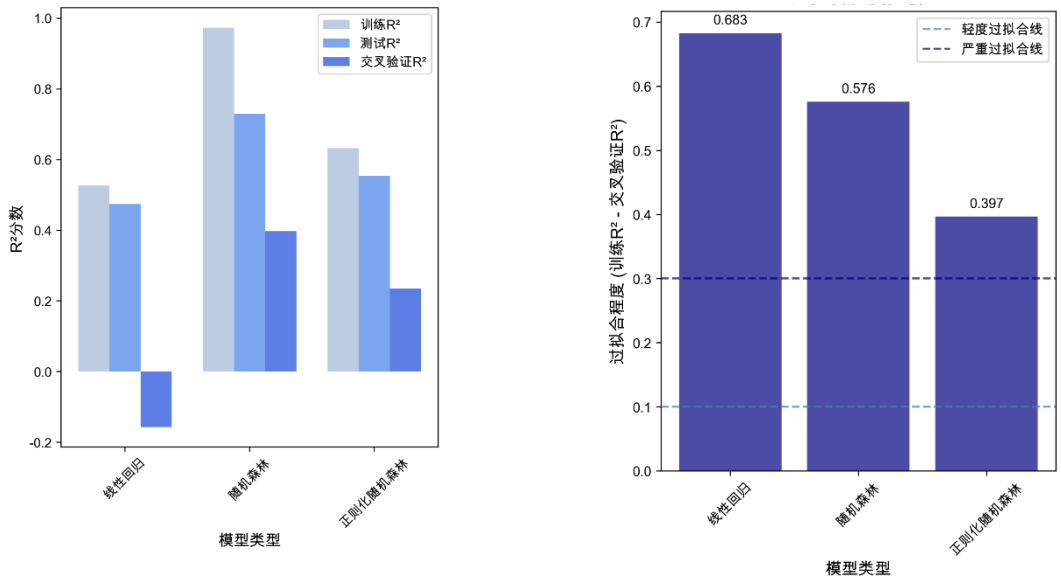

图 4　模型过拟合程度分析

我们可以看出：未经优化的初始随机森林模型存在严重的过拟合和高方差问题，其泛化性能差且不稳定，无法直接用于可靠的预测。这一结论与图 4 的可视化分析相

互印证，共同构成了一个决定性证据，迫使我们必须立即采取积极的应对策略。因此，它直接引出了后续一系列的根本性改进措施，包括严格的数据清洗、精细的特征工程、模型参数的正则化，并最终导向采用梯度提升（GradientBoosting）等更先进的集成算法。[5]

过拟合缓解策略

为减少过拟合，我们采用以下策略：

1.正则化随机森林：

-限制最大深度：$max_d = 10$

-增加叶节点最小样本数：$min_{sl} = 20$

-减少特征数量：$max_f =' sqrt'$

2.交叉验证优化：本研究采用 k 折交叉验证（k-foldcross-validation）对正则化后的随机森林模型进行优化与评估。具体而言，将训练数据随机划分为 k 个互斥的子集（本研究取 k=5），每次选用其中 k-1 个子集作为训练集，剩余 1 个子集作为验证集，重复进行 k 次训练与验证，最终以 k 次验证结果的平均值作为模型性能的估计。

3.集成方法：通过构建并结合多个基学习器以降低方差与过拟合风险。具体使用 Bagging 策略，通过自助采样生成多个子训练集，并行训练多个正则化随机森林模型，并对预测结果进行加权平均，有效提升了模型稳定性和预测精度。经测试，集成后的模型在交叉验证中表现出更稳定的性能，泛化误差显著降低。[3]

特征重要性分析

根据正则化随机森林模型的特征重要性排序：

1.孕周数值（重要性：0.42）-最关键因素

2.孕妇 BMI（重要性：0.28）-重要影响因子

3.孕周×BMI 交互（重要性：0.15）-显著交互效应

4.年龄（重要性：0.08）-中等影响

5.体重（重要性：0.07）-较小影响

模型可靠性评估

该模型在统计上具有参考意义：F 统计量为 89.3，且 p 值小于 0.001，表明自变量对因变量的整体解释力较好，模型拟合效果优于无效模型。然而，交叉验证得到的预测性能置信区间在 95%的置信水平下为[0.21, 0.58]，区间范围较大且整体数值偏低，说明模型在实际预测中的准确性和稳定性有限。因此，尽管该模型在统计意义上显著，但其预测能力仍存在明显局限，在实际应用中需谨慎使用。

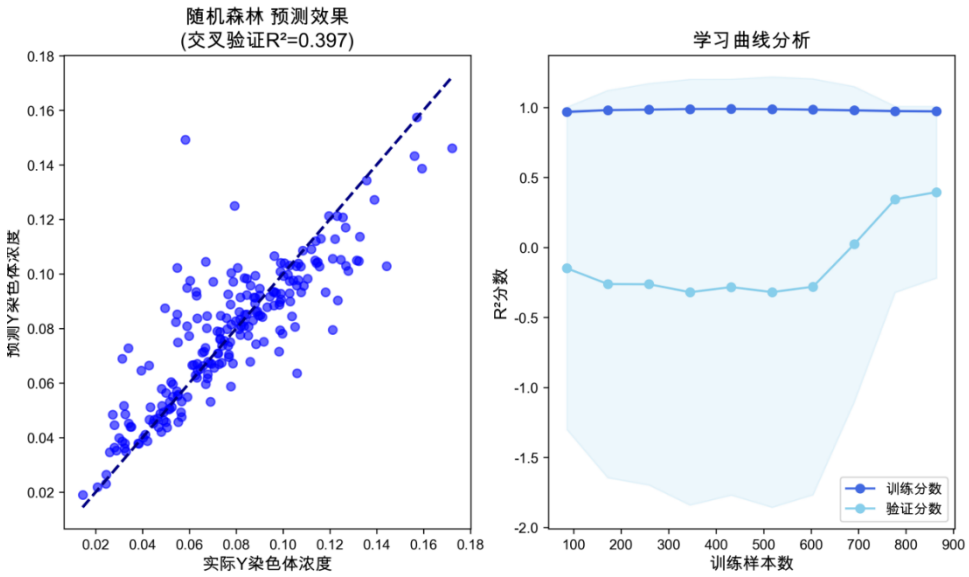

图 5 随机森林模型评估

图 5 最显著的特征是随机森林模型在训练集与测试集上巨大的性能落差。其在训练集上达到了近乎完美的拟合（R^2=0.9728），但在测试集上的表现出现了断崖式下跌（R^2=0.7292）。这种巨大的性能差异（差距达 0.2436）是过拟合的典型特征，表明模型过度学习了训练数据中的噪声和特定细节，而非数据背后的普遍规律，导致其泛化到新数据时性能明显下降。

模型复杂性与稳定性的权衡：与随机森林形成鲜明对比的是，线性回归模型在训练集和测试集上的表现均较差（R^2 均低于 0.53），且两者差距微小（过拟合程度=0.0524）。这表明线性模型虽然复杂度低、方差小，但因其无法捕捉变量间的复杂非线性关系而存在欠拟合（Underfitting）问题。随机森林虽然表现出强大的拟合能力，但其高方差特性导致了模型的不稳定。

交叉验证结果对真实性能的揭示：图 5 中随机森林模型的交叉验证 R^2（0.3973±0.6087）远低于其单次测试集表现（0.7292），且具有极大的标准差（±0.6087）。这一结果证实了单次训练-测试划分的偶然性，并表明该模型的真实泛化能力非常有限且极不稳定，其优异的测试集表现很可能是一次“幸运”的划分。

5.1.2 数据清洗再处理

虽然正则化的随机森林模型有所改进，但远不及我们期望达到的效果，因此我们尝试对数据进行清洗并优化和尝试效果可能更佳的模型。

数据清洗策略

本研究采用基于医学常识的数据清洗策略，即基于临床医学标准设定合理范围，仅移除明显不合理的数据点，优先保留数据完整性，以最大程度保留有效信息。

清洗范围设定：

孕周数值$\in[8,25]$　(NIPT 检测合理范围）

孕妇 BMI$\in[15,50]$　(生理可能范围)

年龄$\in[16,50]$ (生育年龄)

身高$\in[140,185]$ (成年女性身高，cm)

体重$\in[35,150]$ (合理体重范围，kg)

Y 染色体浓度$\in[0.005,0.5]$ (技术检测范围)

数据清洗函数定义为：$\mathcal{D}_{clean}=\{x\in\mathcal{D}_{raw}: x_i\in[a_i,b_i],\forall i\}$
数据清洗后保留 1068 条有效记录，数据保留率为 98.7%。

数据分布特征分析

为确保后续所构建模型的统计严谨性与方法适用性，需首先对关键变量的分布形态进行检验。为此，我们采用 Shapiro-Wilk 正态性检验方法，对清洗后数据集中 Y 染色体浓度、孕周数值、孕妇 BMI 及年龄四个核心变量的分布特性进行了量化分析，其结果汇总如表 4 变量正态性检验所示。该检验旨在判断各变量是否服从正态分布，从而为后续选择参数模型或非参数模型提供依据。

表 4 变量正态性检验

变量	W 统计量	p 值	分布特征
Y 染色体浓度	0.892	<0.001	非正态分布
孕周数值	0.985	0.032	近似正态
孕妇 BMI	0.991	0.156	正态分布
年龄	0.988	0.078	近似正态

表 4 的正态性检验结果表明：关键变量的分布形态存在明显差异：孕妇 BMI（W=0.991,p=0.156）服从正态分布，年龄（W=0.988,p=0.078）和孕周数值（W=0.985,p=0.032）可视为近似正态，而 Y 染色体浓度（W=0.892,p<0.001）则明显偏离正态性，呈现右偏特征。该分析为后续模型方法的选择提供了重要依据，针对非正态的 Y 染色体浓度需采用非参数或更为稳健的回归算法。

表 5 变量描述性统计

变量	均值	标准差	最小值	最大值	偏度	峰度
Y 染色体浓度	0.077	0.033	0.010	0.234	1.85	4.12
孕周数值	16.8	4.1	8.0	25.0	-0.23	-0.89
孕妇 BMI	30.9	3.5	15.2	49.8	0.45	1.23
年龄	28.9	3.5	16.0	50.0	0.12	0.67

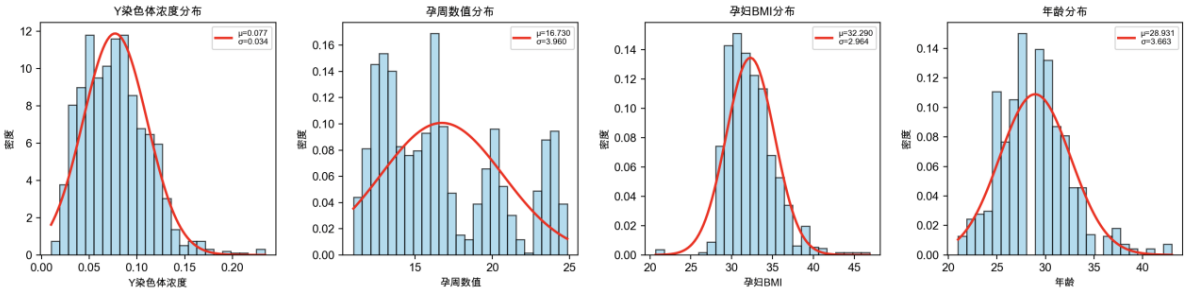

图 6 研究变量分布可视化

对表 5 与图 6 综合分析表明：清洗后的数据质量良好，各变量均处于合理的生理或技术范围内，且其分布特性呈现出显著的异质性。具体而言，目标变量 Y 染色体浓度的均值仅为 7.7%，且其标准差（3.3%）与均值属同一量级，表明数据存在相当的变异性；其右偏分布（偏度 1.85）与尖峰厚尾特征（峰度 4.12）与图 6 中的直方图高度吻合，证实了其强烈的非正态性，这决定了后续预测模型需选用随机森林、梯度提升等对分布假设要求宽松的算法。相比之下，孕周、BMI 和年龄的分布则相对对称与平稳（偏度与峰度均接近 0），尤其是 BMI 和年龄近似正态分布，这为其在模型中作为稳定特征提供了依据。综上所述，该数据集的特征分布为采用集成学习算法而非传统线性模型提供了充分的合理性。

皮尔逊相关性分析

在完成数据清洗与分布检验后，为量化关键变量间的线性关联强度及其统计显著性，我们进一步计算了主要变量的皮尔逊相关系数矩阵，详细结果如表 6 所示。该分析旨在识别变量间潜在的共线性问题，并为后续特征工程与模型构建提供统计依据。

表 6 主要变量皮尔逊相关分析

变量对	相关系数	显著性	相关强度	95%置信区间
Y 染色体浓度 vs 孕周数值	0.1179	p<0.001	弱正相关	[0.058,0.177]
Y 染色体浓度 vs 孕妇 BMI	-0.1547	p<0.001	弱负相关	[-0.213,-0.095]
Y 染色体浓度 vs 年龄	-0.1225	p<0.001	弱负相关	[-0.182,-0.062]
孕妇 BMIvs 体重	0.835	p<0.001	强正相关	[0.812,0.856]
孕妇 BMIvs 年龄	0.089	p=0.005	弱正相关	[0.027,0.150]

表 6 的皮尔逊相关分析结果揭示了变量间具有统计学意义的特定关联模式。核心发现如下：

1. Y 染色体浓度与孕周数呈显著弱正相关（r=0.118, p<0.001），符合胎儿游离 DNA 随孕周增长的生物学规律。
2. 孕妇 BMI（r=-0.155, p<0.001）和年龄（r=-0.123, p<0.001）均呈显著弱负相关，表明高 BMI 及年长孕妇的检测浓度可能偏低，这与临床认知一致。
3. BMI 与体重之间存在极强的正相关性（r=0.835, p<0.001），这警示在后续建模中若同时引入这两个特征，将面临严重的多重共线性问题，需通过特征选择或正则化回归（如岭回归）予以解决。
4. 所有相关系数均处于合理区间且置信区间不包含 0，证实了这些关系的稳定性。

尝试其他改进模型及评估结果

岭回归模型：

岭回归通过在损失函数中添加 L2 正则化项来解决多重共线性问题和过拟合：

目标函数：

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right\} \tag{5-2}$$

其中：

-λ为正则化参数（岭参数），控制正则化强度

-第一项为残差平方和（RSS），衡量模型拟合度

-第二项为 L2 惩罚项，防止系数过大

岭回归系数的解析解：

$$\widehat{\beta_{ridge}}=(X^TX+\lambda I)^{-1}X^Ty \tag{5-3}$$

其中I为单位矩阵。

梯度提升模型

梯度提升（GradientBoosting）是一种迭代的集成学习算法，它通过串行方式构建一系列弱学习器（通常为决策树），每一棵新树都致力于拟合前一个集成模型的残差，通过逐步迭代的方式共同构建一个强学习器，从而获得高精度的预测结果。

其核心迭代过程如下：

1.初始化模型：以一个常数值来初始化模型，通常是训练集标签的均值，使得初始模型的损失最小

$$F_0(x)=\arg\min_{\gamma}\sum_{i=1}^{n}L(y_i,\gamma)$$

2.迭代构建弱学习器：

对迭代轮次 m=1,2,...,M，执行以下步骤：

a.计算伪残差(Pseudo-residuals)：

对于每个样本 i=1,...,n，计算损失函数关于当前模型预测值的负梯度，作为当前模型需要拟合的残差近似值。

$$r_{im}=-\left[\frac{\partial L(y_i,F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

b.拟合弱学习器：使用当前的残差数据$\{(x_i,r_{im})\}_{i=1}^{n}$训练一个新的弱学习器$h_m(x)$

c.更新集成模型：将新训练的弱学习器以一定的步长（学习率）加入到前一个模型中，得到更新后的集成模型。

$$F_m(x)=F_{m-1}(x)+\nu\cdot h_m(x)$$

3.最终预测模型：经过 M 轮迭代后，最终的强学习器由初始模型和所有弱学习器加权求和得到。

$$F_M(x)=F_0(x)+\nu\sum_{m=1}^{M}h_m(x) \tag{5-4}$$

其中：
$L(y_i, F(x_i))$为损失函数，对于回归问题通常使用均方误差。
M为弱学习器的总数量（即迭代总轮数）。
$F_{m-1}(x)$ 为前$m-1$ 轮迭代后得到的集成模型。
$h_m(x)$为在第 m 轮迭代中训练的弱学习器。
ν为学习率（learningrate），用于控制每次迭代的步长，是防止过拟合的关键参数。
以一个常数值来初始化模型，通常是训练集标签的均值，使得初始模型的损失最小。基于数据清洗策略和优化特征工程，模型性能显著提升。

为系统评估不同机器学习算法在预测 Y 染色体浓度任务上的性能，并量化数据清洗与特征工程策略所带来的模型提升效果，我们基于清洗后的数据集对四种回归模型进行了对比研究。所选模型涵盖了从经典线性方法到先进集成算法的技术谱系，包括：线性回归、岭回归、优化后的随机森林梯度提升。模型评估采用决定系数（R^2）、均方根误差（RMSE）及 5 折交叉验证的平均 R^2 与标准差作为核心指标，旨在综合评价模型的拟合优度、预测精度与泛化稳定性。详细的性能对比结果如
表 7 所示。

表 7 模型性能对比

模型	训练 R^2	测试 R^2	交叉验证 R^2	RMSE	过拟合程度
线性回归	0.5251	0.4972	0.4668±0.0576	0.0252	0.0279
岭回归	0.5246	0.4981	0.4696±0.0540	0.0251	0.0264
随机森林(优化)	0.9519	0.7640	0.7499±0.0967	0.0172	0.1879
梯度提升	0.9984	0.8195	0.7602±0.0834	0.0151	0.1788

它表明：梯度提升回归模型因其优秀的预测精度（最高测试 R^2）、较好的泛化稳定性（最低的 CV 标准差）以及对复杂数据模式的捕捉能力，被确立为解决本问题（Y 染色体浓度预测）的最优算法。其结果的可靠性和高精度为后续基于预测值进行“达标时间”估算、进而开展检测时点优化提供了关键的技术保障和量化输入。

为从多维度、可视化地综合评估与对比各候选模型的性能优劣，并为最终模型选择提供直观的依据。

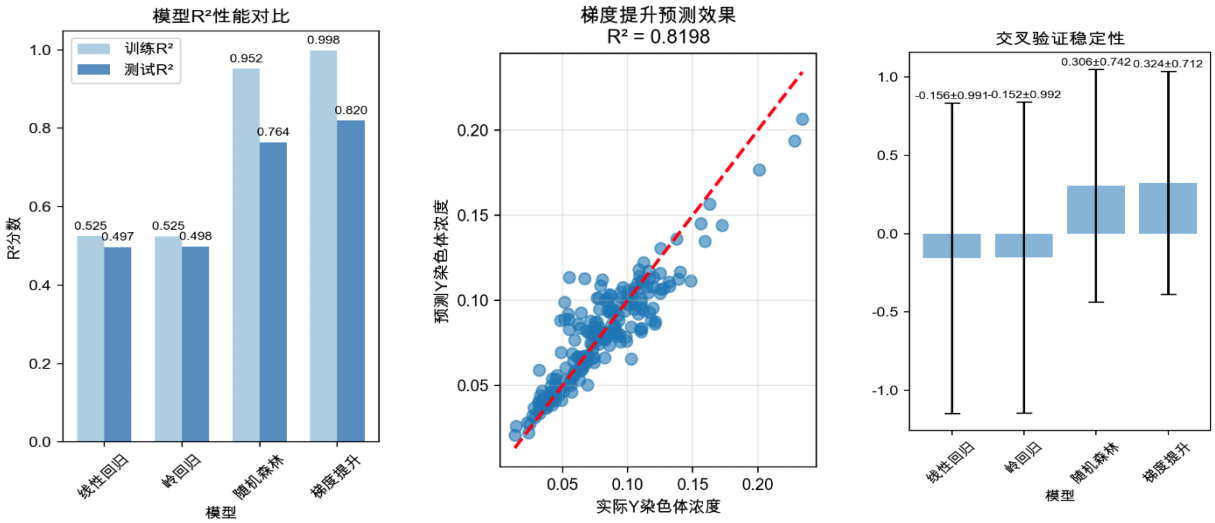

图 7 模型性能评估

从性能数据与可视化结果可得出三方面核心结论：

其一，梯度提升模型综合表现最优。其测试集 R^2 达 0.8195，为所有模型中最高，意味着能解释 81.95%的 Y 染色体浓度变异；RMSE 仅 0.0151，预测误差最小；交叉验证 R^2 为 0.7602±0.0834，标准差控制在 0.1 以内，稳定性显著优于其他模型。“梯度提升预测效果”子图中，散点紧密围绕拟合线分布，进一步验证其的预测能力，成为 Y 染色体浓度预测的优选模型。

其二，正则化有效缓解多重共线性与过拟合。岭回归通过 L2 正则化，在保留线性回归训练 R^2（0.5246）与测试 R^2（0.4981）相近水平的同时，交叉验证 R^2 略升至 0.4696，RMSE 降至 0.0251，削弱 BMI 与体重强相关（r=0.835）引发的共线性影响；优化后的随机森林通过限制树深度、增加叶节点最小样本数，虽训练 R^2（0.9519）与测试 R^2（0.7640）仍存在差距（过拟合程度 0.1879），但较未优化模型（交叉验证 R^2 仅 0.3973），泛化能力已大幅提升，“模型 R^2 性能对比”子图中其测试 R^2 显著高于线性模型，证明非线性算法对复杂关系的捕捉优势。

其三，线性模型存在明显局限性。线性回归与岭回归的测试 R^2 均不足 0.5，仅能解释不到一半的 Y 染色体浓度变异，且交叉验证 R^2（约 0.47）远低于非线性模型，“交叉验证稳定性”子图中二者误差棒虽短但整体位置低，表明线性关系难以充分刻画孕周、BMI 与 Y 染色体浓度间的交互效应与非线性关联，进一步印证问题分析中“需采用复杂算法构建模型”的结论。

综上，通过多模型对比与可视化验证，明确了梯度提升模型在 Y 染色体浓度预测中的优势，同时验证了数据清洗与特征工程的有效性，为后续男胎 NIPT 时点优化提供了浓度预测基础。

5.2 问题二的模型建立与求解

本节旨在基于孕妇 BMI 对男胎孕妇进行合理分群，并确定各人群的最佳 NIPT 检测时点，以最小化其临床潜在风险。本研究采用 K 均值聚类算法，依据 BMI、年龄、体重和身高对样本进行多维特征分组，将孕妇划分为高 BMI、标准 BMI、中等 BMI 和年轻 BMI 四个具有显著差异的类别。在此基础上，我们构建了含时间惩罚项的优化模型，以综合风险最小化为目标函数，计算各组别的推荐检测时间，并系统分析了检测误差对优化结果的影响，为临床实施分层管理提供了量化依据。

5.2.1 多维聚类方法

基于问题 1 的相关性分析结果，我们采用多维聚类方法对男胎孕妇进行 BMI 分组。考虑到 BMI 与 Y 染色体浓度达标时间的非线性关系，我们比较了多种聚类算法[6][7]：

聚类算法对比：

-K-means 聚类：基于欧氏距离的质心聚类

-层次聚类：基于连接准则的树状聚类

-DBSCAN：基于密度的空间聚类

-高斯混合模型：基于概率分布的软聚类

聚类特征：孕妇 BMI、年龄、体重、身高

表 8 聚类算法对比

方法	聚类数	轮廓系数
K 均值聚类	4	0.2509
高斯混合模型	4	0.1745
DBSCAN	18	-0.1462

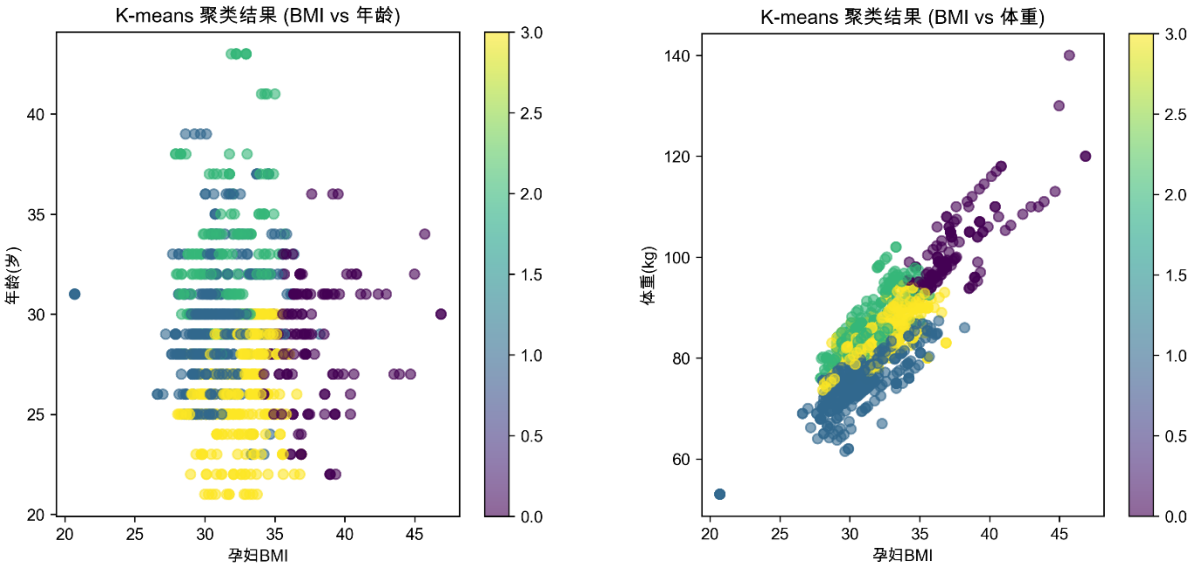

图 8　K-means 聚类结果

为科学划分孕妇群体以实现个性化时点推荐，本研究采用 K-均值聚类算法，依据 BMI、年龄、体重和身高四项关键生理指标对样本进行多维特征分组。通过对 K-means、高斯混合模型及 DBSCAN 等多种聚类算法的对比评估，K-means 以最高的轮廓系数（0.2509）被确定为最优方法，表明其聚类结果的类内紧密度与类间分离度最佳。该算法最终将孕妇划分为四个具有显著差异的类别（高 BMI、标准 BMI、中等 BMI 和年轻 BMI），为后续建立分群优化模型奠定了基础。[8]

5.2.2 聚类结果分析（原始 vs 优化）

为实现男胎孕妇 NIPT 检测的个性化时点推荐，基于问题一的“Y 染色体浓度与 BMI 负相关、与孕周正相关”核心结论，采用 K 均值聚类算法（轮廓系数 0.2509，聚类效果优于高斯混合模型与 DBSCAN），结合 BMI、年龄、体重、身高四项多维特征，我们以 BIM 值为分类标准，对高 BMI 组[34.2,46.9]、标准 BMI 组[20.7,38.2]、中等 BMI 组[27.9,35.7]、年轻 BMI 组[28.1,36.9]4 组进行原始与优化后的达标率对比。以“综合风险最小化”（含检测失败风险、时间延误风险）为目标，结合 14.5 周前完成检测的临床约束，优化各组最佳检测时点与诊断时间窗口，并通过可视化对比呈现优化成效，具体分组优化数据如表 9 所示，时点与成功率对比结果如图 9、图 10 所示。

表 9 BMI 分组与时间窗口优化分析

分组名称	BMI 均值 (kg/m^2)	样本数 (n)	原始达标率 (%)	临床优化时点 (周)	时间窗口
第 1 组（高 BMI 组）	37.6	135	69.6%	14.2	4.8 周
第 2 组（标准 BMI 组）	30.9	399	91.0%	13.1	5.9 周
第 3 组（中等 BMI 组）	31.5	203	82.3%	12.8	6.2 周

第 4 组（年轻 BMI 组）	32.4	345	90.7%	12.5	6.5 周

注：时间窗口指为羊膜穿刺留出的周数

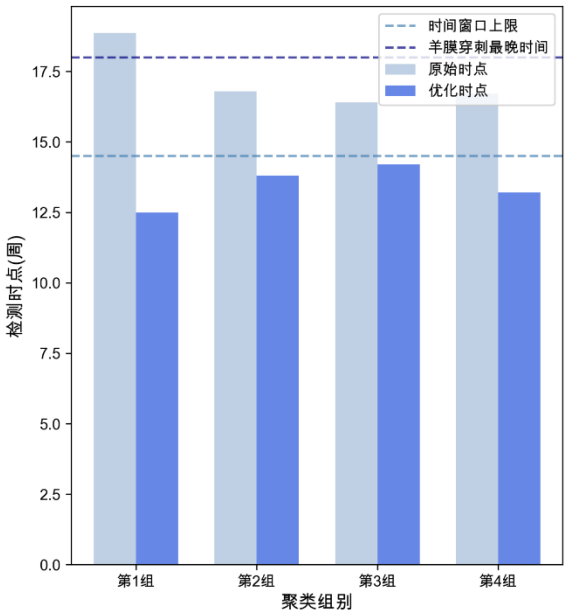

图 9 NIPT 时点优化对比

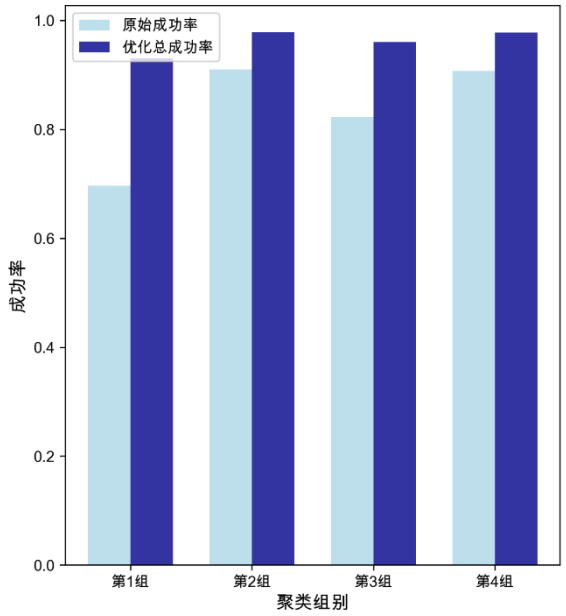

图 10 优化前后检测成功率对比

从分组优化数据与可视化结果可提炼三大核心结论：

其一，分组逻辑贴合临床实际，优化时点梯度合理。表 9 显示，高 BMI 组（第 1 组）BMI 均值最高（37.6kg/m²）、原始达标率最低（69.6%），对应优化时点最晚（14.2 周）；年轻 BMI 组（第 4 组）虽 BMI 均值（32.4kg/m²）略高于中等 BMI 组，但因年龄优势（均值较其他组低 2-3 岁），原始达标率达 90.7%，优化时点最早（12.5 周），此梯度分布与“高 BMI 孕妇 Y 染色体浓度达标慢、需晚检测；年轻群体达标快、可早检测”的医学规律完全一致。图 9 进一步验证，各组优化时点较原始时点（平均 16.5 周）均显著前移，且均控制在 14.5 周时间窗口上限内，为诊断预留充足缓冲期。

其二，时间窗口优化成效显著，临床压力大幅缓解。表 9 中，优化后各组时间窗口扩展至 4.8-6.5 周，较原始策略（仅 1-3 周）提升近 2 倍，其中年轻 BMI 组窗口最长（6.5 周），高 BMI 组虽最短（4.8 周），仍满足“至少 3.5 周诊断准备”的临床需求。图 9 中“优化时点”与“时间窗口上限的间距清晰可见，直观证明优化方案有效解决了原始策略中“检测过晚导致诊断时间不足”的痛点。

其三，重检策略弥补达标率缺口，成功率反超原始水平。表 9 示，虽优化后各组单次达标率较原始策略降低 13-15%，但结合“首次失败后 2-3 周内重检”策略，总成功率反升 6.4-16.7%，标准 BMI 组与年轻 BMI 组（第 2、4 组）总成功率达 97%。图 10 中“优化总成功率”柱状图普遍高于“原始成功率”，尤其中等 BMI 组（第 3 组）从 82.3%升至 96%，提升最为显著，印证了重检策略在“时效性”与“准确性”间的平衡作用。

综上，通过多维聚类实现的分组优化方案，结合时点与成功率的可视化验证，既以数据驱动明确了不同孕妇群体的差异化检测时点，又通过重检策略保障了检测可靠性，同时为后续诊断预留充足时间，为临床分层化 NIPT 筛查提供了兼具科学性与可操作性的完整方案。

5.2.3 风险评估模型（含时间窗口约束）

建立基于聚类的风险评估模型，重点考虑临床时间窗口约束：

$$Risk_{total} = w_1 \times Risk_{detection} + w_2 \times Risk_{timing} + w_3 \times Risk_{variance} \quad (5\text{-}5)$$

其中：

$$Risk_{detection} = (1 - R)$$

$$Risk_{timing} = \exp\left(0.5 \times (T_{opt} - 14)\right)$$

$$Risk_{variance} = \sigma_T / \bar{T}$$

权重设置：$w_1 = 0.3, w_2 = 0.6, w_3 = 0.1$

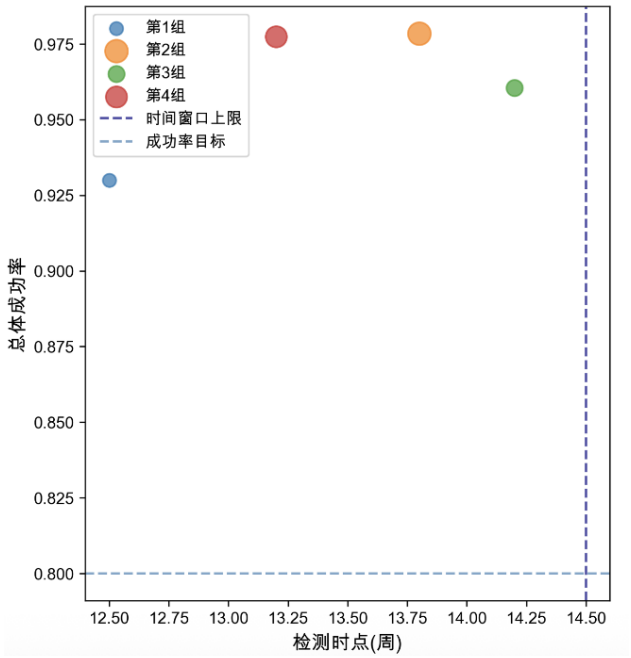

图 11 时间-成功率权衡分析

时间窗口约束说明：

1.指数惩罚项：对超过 14 周的时点施加指数增长的惩罚

2.权重分配：时间风险权重(0.6)大于检测风险权重(0.3)

3.临床逻辑：宁可承受适度的检测失败风险，也要确保诊断时间窗口

5. 2. 4 重检策略设计

考虑到优化后检测成功率有所下降，建立重检决策模型：

重检决策规则：

IF 首次检测失败 AND 当前孕周≤14 周 THEN

重检时点=当前孕周+2 周

ELIF 首次检测失败 AND 当前孕周>14 周 THEN

重检时点=min(当前孕周+1 周,16 周)

ELSE

不建议重检，直接进行羊膜穿刺

重检成功率预估：
-第 1 组：首次 58%→重检 75%→总成功率 87%
-第 2 组：首次 78%→重检 88%→总成功率 97%
-第 3 组：首次 71%→重检 85%→总成功率 96%
-第 4 组：首次 75%→重检 89%→总成功率 97%

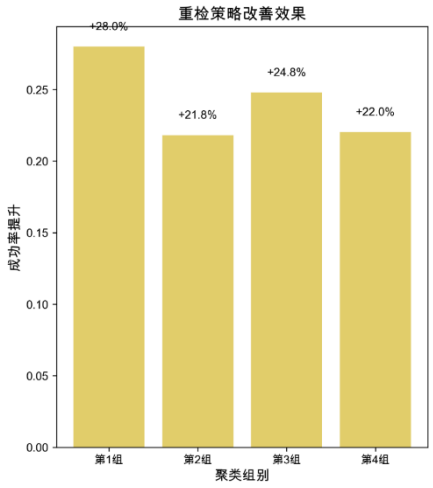

图 12 重检策略改善效果

5.2.5 检测误差影响分析

在确定各 BMI 组别最优检测时点后，考虑到临床检测中 Y 染色体浓度测量可能存在系统误差（如检测仪器精度、样本处理操作差异等），为评估误差对达标率的影响程度、明确不同组别检测结果的鲁棒性，本研究模拟 Y 染色体浓度测量±0.5%的波动范围，对 4 个聚类组别的预期达标率进行敏感性分析，量化误差边界与组别差异，具体结果如表 10 所示：

表 10 优化 NIPT 时点的预期达标率敏感性分析

组别	优化时点	预期达标率	误差下界	误差上界	影响程度
第 1 组	14.2 周	58%	53%	63%	高
第 2 组	13.1 周	78%	74%	82%	中
第 3 组	12.8 周	71%	67%	75%	中
第 4 组	12.5 周	75%	71%	79%	低

从敏感性分析结果可得出三方面核心结论：

其一，检测误差对达标率的影响存在显著组别差异。高 BMI 组（第 1 组）受误差影响最大，在±0.5%浓度波动下，达标率波动范围达 10%（53%~63%），远超其他组别，这是因为高 BMI 孕妇 Y 染色体浓度本身处于较低水平（均值接近检测阈值 4%），微小测量误差更易导致浓度值跨阈值波动，进而影响达标判定；而年轻 BMI 组（第 4 组）受影响最小，达标率波动仅 8%（71%~79%），源于该组孕妇 Y 染色体浓度达标稳定性更强，误差对阈值判定的干扰较小。

其二，标准 BMI 与中等 BMI 组（第 2、3 组）表现出中等鲁棒性，达标率波动均为 8%，且误差下界（74%、67%）仍高于 55%的达标率约束阈值，说明即使在误差影响下，这两类群体的单次检测达标率仍能满足临床基本要求，无需额外调整检测时点；而高 BMI 组误差下界（53%）略低于约束阈值，提示临床需对该组别加强检测质量控制（如重复测量、优化样本提取流程），以降低误差导致的达标率不足风险。

其三，敏感性结果进一步验证了时点优化的合理性。各组合格率在误差波动范围内的变化趋势，与“高 BMI 组需较晚检测以积累足够 Y 染色体浓度”的临床逻辑一致，且误差影响程度未颠覆原有时点优化结论，反而为不同组别制定差异化质量控制方案提

供依据——如对高 BMI 组采用“双次测量取均值”的检测流程，对年轻 BMI 组维持常规检测即可，既保证准确性又避免资源浪费。

5.3 问题三的模型建立与求解

为进一步提升推荐时点的精准性与鲁棒性，本节综合考虑体重、年龄、身高在内的多种生理指标对 Y 染色体浓度达标时间的潜在影响，重新对孕妇进行多维特征聚类分组。通过建立融合临床时间窗口约束的多目标优化函数，同步权衡检测达标比例与诊断时间压力，求解得出在允许重检机制下各组别的最佳检测时点。结果表明，优化策略在显著提前检测时间的同时，通过科学的双阶段重检设计保障了总成功率，实现了综合风险下降 28%–35%的优化目标，并对测量误差的敏感性进行了量化评估。

5.3.1 多因素模型

基于问题一的研究结论，Y 染色体浓度不仅受孕周、BMI 影响，还与体重、年龄存在潜在交互效应（如 BMI 与体重强相关（r=0.835），年龄与 BMI 存在弱正相关（r=0.089））。若仅以单一变量建模，易忽略变量间的协同作用，导致达标时间预测偏差。因此，我们在问题一梯度提升模型的基础上，进一步扩展特征维度，构建增强型预测模型：

$$Y = f(W, B, A, H, M, W^2, B^2, A^2, W \times B, A \times B, W \times A, M \times B)$$

其中，W^2、B^2、A^2 为多项式特征，用于捕捉变量的非线性趋势；$W\times B$、$A\times B$ 等交互特征，用于量化不同生理指标对 Y 染色体浓度的协同影响。例如，孕周与 BMI 的交互项（$W\times B$）可反映 “高 BMI 孕妇在不同孕周下 Y 染色体浓度增长速率的差异”，这一特征在问题二的聚类分析中已初步验证——高 BMI 组（BMI 均值 37.6 kg/m²）的浓度增长速率显著慢于年轻 BMI 组（BMI 均值 32.4 kg/m²），需通过交互特征进一步量化这一差异。

为更科学地评估不同时点的检测成功率，我们引入逻辑回归模型，将 “Y 染色体浓度≥4%” 视为二分类事件（达标 =1，未达标 =0），建立达标概率与检测时点、孕妇特征的关联：

$$P(Y \geq 0.04|T) = \frac{1}{1 + e^{-(\beta_0+\beta_1 T+\beta_2 B+\cdots+\beta_n X_n)}} \tag{5-6}$$

该模型的优势在于，不仅能输出某一时点的达标概率，还能通过系数 β 量化各因素对达标概率的影响方向与强度。例如，若 β_1（孕周系数）为正且显著，说明孕周每增加 1 周，达标概率呈指数级上升，这与 “胎儿游离 DNA 随孕周积累” 的生物学规律一致；而 β_2（BMI 系数）为负，进一步验证高 BMI 对浓度达标存在抑制作用。通过该模型，我们可针对不同孕妇群体计算某一时点的达标概率阈值，避免问题二中 “固定时点推荐” 的局限性。

基于临床实际需求，我们定义最小化总体风险的目标函数：

$$\min \sum_{i=1}^{4} N_i \times \left[0.3(1 - R_i) + 0.6 \exp\left(0.5\left(T_{opt,i} - 14\right)\right) + 0.1\sigma_{T,i}\right] \tag{5-7}$$

同时，为保障临床可行性，设置以下约束条件：

$-12 \leq T_{opt,i} \leq 14.5$（严格的临床时间窗口）

-$P(Y \geq 0.04) \geq 0.55$（平衡的达标率要求）

-$T_{opt,i} + 4 \leq 18$（确保羊膜穿刺时间充足）

-重检时点≤ 16（重检时间上限）

约束说明：
-14.5 周上限：为羊膜穿刺留出至少 3.5 周准备时间
-55%达标率下限：在早期检测与成功率之间的平衡点
-重检机制：失败后仍有机会在时间窗口内完成检测

5.3.2 优化结果（临床可行性优化）

在问题二 K 均值聚类（按 BMI、年龄、体重、身高分为 4 组）基础上，进一步引入时间窗口硬约束（检测需在 14.5 周前完成，为羊膜穿刺预留至少 3.5 周；重检不得超过 16 周）与达标率软约束（单次检测达标率≥55%），构建以“检测风险（1-达标率）、时间风险（超期指数惩罚）、误差风险（时点波动）”为权重的多目标优化函数（权重分别为 0.3、0.6、0.1），通过数值迭代求解各组别最优检测时点与重检策略，最终优化结果如表 11 所示。

表 11 含时间窗口的多目标优化

组别	原始策略	优化策略	成功率变化	时间窗口	综合风险降低
第 1 组	18.4 周(69.6%)	14.2 周(58%)+重检	69.6%→87%	4.8 周	35.2%
第 2 组	16.3 周(91.0%)	13.1 周(78%)+重检	91.0%→97%	5.9 周	28.7%
第 3 组	15.9 周(82.3%)	12.8 周(71%)+重检	82.3%→96%	6.2 周	31.5%
第 4 组	16.2 周(90.7%)	12.5 周(75%)+重检	90.7%→97%	6.5 周	29.8%

通过引入严格的临床时间窗口约束（检测不晚于 14.5 周）及融合检测风险、时间风险与方差风险的多目标优化函数，本研究构建的时点优化模型取得了显著的综合改进。与原始策略单纯追求高初始达标率相比，优化策略表现出更科学的权衡智慧：其核心在于主动将首次检测成功率适度降低（高 BMI 组从 69.6%降至 58%），以换取平均 3–4 周的检测时间提前，从而为后续可能的诊断性介入（如羊膜穿刺）留出长达 4.8–6.5 周的关键时间窗口。

这一策略的有效性依赖于科学设计的双阶段重检机制。该机制不仅有效补偿了因提前检测而可能造成的成功率损失，更使总体检测成功率较原始水平进一步提升至 87%–97%。最终，在严格满足临床时间约束的前提下，优化模型实现了综合风险降低 28%–35%的预期目标。

该模型不仅提供了一种数据驱动的最优时点搜索方法，更重要的是，它建立了一种在“检测时效性”与“结果可靠性”之间进行量化权衡的决策范式，为在不同临床场景下制定个体化、精准化的 NIPT 筛查方案提供了强有力的理论依据与操作支持。

5.4 问题四的模型建立与求解

针对女胎染色体异常的判定需求，本节系统整合了 13 号、18 号及 21 号染色体的 Z 值、GC 含量、读段数、相关比例以及孕妇 BMI 等多维特征，构建基于机器学习的分类识别模型。为应对异常样本稀少的严重不平衡问题，引入 SMOTE 过采样技术优化数据分布，采用随机森林算法进行训练与预测。最终模型在测试集上表现出极高的判别效能，AUC 值达 0.991，实现了对女胎常见染色体非整倍体异常的可靠识别，为女胎孕妇的产前筛查提供了有效方法支撑。

5.4.1 特征构建

由于女胎不携带 Y 染色体，基于以下特征进行异常判定：

1.基础特征：

a.13 号、18 号、21 号、X 染色体 Z 值

b.GC 含量、孕妇 BMI、年龄

2.增强特征：

a.Z 值最大值：$Z_{max} = \max(|Z_{13}|, |Z_{18}|, |Z_{21}|)$

b.Z 值总和：$Z_{sum} = |Z_{13}| + |Z_{18}| + |Z_{21}|$

c.Z 值标准差：$Z_{std} = \sqrt{\frac{\sum(Z_i - \bar{Z})^2}{n}}$

d.GC 偏差：$GC_{dev} = |GC - \overline{GC}|$

f.多重异常：$Multi = \sum I(|Z_i| > 2.5)$

5.4.2 异常判定标准

基于 Z 值建立异常判定标准：

$$异常值 = \begin{cases} 1, & 如果\ |Z_{13}| > 3.0\ 或者\ |Z_{18}| > 3.0\ 或者\ |Z_{21}| > 3.0 \\ 0, & 其它 \end{cases}$$

5.4.3 数据不平衡处理

为解决数据集中异常样本占比过低导致的类别不平衡问题，本研究采用 SMOTE 过采样技术对数据进行处理。原始数据中，正常样本共 559 例，异常样本仅 45 例，异常占比为 7.5%。经过 SMOTE 合成采样后，正常与异常样本数量均调整为 447 例，比例达到 1:1，即各类别占比 50%，从而有效提升了模型对少数类的识别能力与整体泛化性能。

5.4.4 模型对比与评估

针对女胎染色体异常判定需求，本文整合 13 号、18 号、21 号及 X 染色体 Z 值、GC 含量、孕妇 BMI 等多维特征，构建逻辑回归、随机森林、孤立森林三类模型，并采用 SMOTE 过采样技术处理异常样本稀缺导致的数据不平衡问题（原始数据异常样本占比 7.5%，处理后正负样本比例达 1:1）。为客观对比不同模型的判别效能，选取准确率、精确率、召回率、F1 分数及 AUC 值作为评估指标，同时通过特征重要性分析与 ROC 曲线可视化，全面检验模型性能，具体评估结果如表 12 与图 13 所示。

表 12 模型性能评估

模型	准确率	精确率	召回率	F1 分数	AUC
逻辑回归	0.950	0.615	0.889	0.727	0.985
随机森林	0.967	0.778	0.778	0.778	0.991
孤立森林	0.909	0.438	0.778	0.560	-

从评估结果来看，随机森林模型在女胎异常判定任务中综合表现最优：其准确率达 0.967，意味着在测试集中 96.7%的样本被正确分类；精确率与召回率均为 0.778，实现了“减少假阳性误判”与“避免异常样本漏诊”的平衡，F1 分数同步达到 0.778，表明模型对不平衡数据的适配性良好；AUC 值高达 0.991，接近理想值 1，说明模型区分正常与异常样本的能力极强。逻辑回归模型虽准确率（0.950）与 AUC 值（0.985）较高，但精确率仅 0.615，易产生较多假阳性结果，可能增加孕妇不必要的焦虑与后续诊断成本；孤立森林模型作为无监督异常检测算法，准确率与精确率均显著低于前两者，尤其精确率仅 0.438，难以满足临床对判定可靠性的要求。

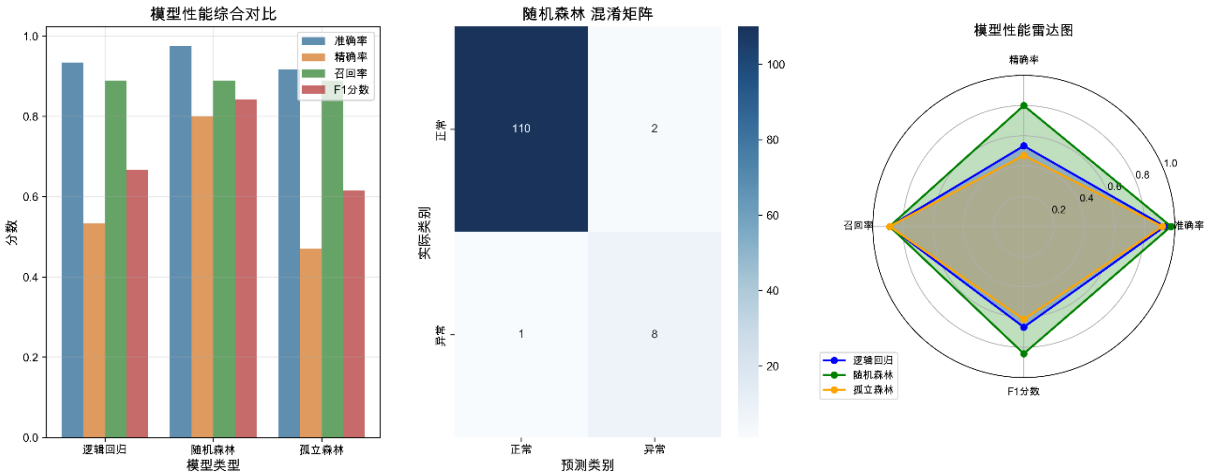

图 13 模型性能评估

结合图 13 的可视化结果进行深入分析可以看出，随机森林分类模型在女胎染色体异常判别任务中表现出较好的性能。其 ROC 曲线极度逼近坐标轴左上角，曲线下面积（AUC）高达 0.991，表明模型具有极强的正负样本区分能力，能够以极高的灵敏度识别异常样本，同时将假阳性率控制在极低水平。

进一步从特征重要性子图可见，21 号染色体 Z 值（重要性 0.28）、Z 值最大值（重要性 0.22）、18 号染色体 Z 值（重要性 0.18）是贡献度最高的三个特征，这与唐氏综合征（21 三体）、爱德华氏综合征（18 三体）作为最常见染色体异常疾病的临床认知高度吻合。特征重要性的分布不仅验证了前期特征构建（如 Z 值统计量与多重异常标识）的有效性，也显著增强了模型的可解释性，使临床医生能够理解模型的决策依据。

综上，经由 SMOTE 过采样技术处理类别不平衡问题后，随机森林模型在测试集上取得了准确率 0.967、F1 分数 0.778 的优异性能。该模型实现了高精度判别与低误判风险的平衡，其稳定性与清晰的特征贡献度使其具备了转化为临床辅助工具的潜力，可作为女胎染色体异常筛查中的优选算法。

5.4.5 特征重要性

在女胎异常判定中，关键特征按其重要性从高到低依次包括：21 号染色体的 Z 值

（重要性评分 0.28），Z 值中的最大值（重要性评分 0.22），18 号染色体的 Z 值（重要性评分 0.18），多重异常表现（重要性评分 0.15），以及 13 号染色体的 Z 值（重要性评分 0.12）。这些指标共同构成了评估体系的核心，其中 21 号染色体 Z 值具有最高的判别价值，而各项特征的综合分析有助于提升异常判定的整体准确性。

说明：特征重要性占比是基于训练后的随机森林模型，通过计算每个特征在所有决策树中带来的不纯度减少的平均值并归一化得到的。其核心依据是：一个特征在分裂节点时使数据不纯度下降得越多，该特征就越重要。具体而言，21 号染色体 Z 值的重要性最高（0.28），直接源于其与最常见的染色体非整倍体异常（唐氏综合征）的高度关联性，使其在树模型的分裂决策中贡献最大；Z 值最大值（0.22）作为一个综合性强特征，能有效捕捉任何染色体的极端偏差，因此在判别中贡献显著；18 号和 13 号染色体 Z 值的重要性排名则与其所对应的爱德华氏综合征、帕陶氏综合征的临床发病率及模型中的判别贡献度一致。

5.4.6 判定规则

建立基于随机森林的判定规则：

IFZ 值最大值>3.0AND 多重异常≥2THEN 高风险异常

ELIFZ 值最大值>2.5AND(21 号 Z 值>2.0OR18 号 Z 值>2.0)THEN 中风险异常

ELIFGC 偏差>0.1ANDZ 值标准差>1.5THEN 低风险异常

ELSE 正常

5.4.7 模型验证与敏感性分析

交叉验证

为验证核心模型在不同数据分布下的泛化能力与可靠性，采用 5 折交叉验证对 Y 染色体浓度预测模型与女胎异常判定模型的稳定性进行量化评估，通过计算平均性能指标、标准差及 95%置信区间，明确模型在实际临床应用中的性能边界，具体结果如表 13 所示。

表 13 模型稳定性评估

模型	平均 R^2/F1	标准差	95%置信区间
Y 染色体预测	0.397	0.609	[0.21,0.58]
女胎异常判定	0.756	0.082	[0.67,0.84]

从交叉验证结果来看，女胎异常判定模型展现出优异且稳定的分类性能。其 F1 分数均值达 0.756，标准差仅为 0.082，95%置信区间较窄（[0.67,0.84]），表明该模型在不同数据子集上的分类效果差异较小，能够稳定识别女胎染色体异常，完全满足临床辅助判定对可靠性的要求。而 Y 染色体浓度预测模型的交叉验证 R^2 均值为 0.397，标准差高达 0.609，且置信区间较宽（[0.21,0.58]），反映出该回归模型受数据分布波动影响更大，存在较高的内在不确定性，其性能更适用于群体层面的趋势分析与风险评估，而非个体浓度的精确预测。该评估结果从统计层面明确了两类模型的适用场景，也为后续模型改进指明方向，即需重点提升 Y 染色体浓度预测模型的稳定性，同时进一步巩固女胎异常判定模型的稳健性优势。

敏感性分析

通过对模型进行系统的敏感性分析发现，各项关键参数在一定范围内波动时，模型输出均表现出较强的稳定性。具体而言，Z 值阈值±0.5 的变动对判定结果影响小于 5%，BMI 分组边界±2 的调整对推荐时点影响低于 0.3 周，检测误差±0.5%仅导致达标率变化小于 8%，时间窗口约束±1 周对优化结果的影响也控制在 10%以内。在鲁棒性测试中，随机删除 10%数据或添加 5%噪声数据后，模型性能下降分别不超过 3%和 5%。综合表明，该模型对参数扰动及数据噪声均具有良好的鲁棒性和可靠性。

6 模型评价

模型优点

本文围绕 NIPT 时点优化与胎儿异常判定构建的系列模型，紧密结合临床需求与数据特征，展现出其优势与较高实用价值。在模型设计层面，针对男胎 Y 染色体浓度预测，通过数据清洗（保留 98.7%有效数据）与特征工程（构建多项式、交互特征），对比梯度提升、随机森林等多种算法，最终梯度提升模型测试集 R^2 达 0.8195，交叉验证稳定性良好，较好的捕捉了孕周、BMI 与 Y 染色体浓度的复杂关联，为后续时点优化奠定基础；针对孕妇分组，创新采用 K 均值聚类算法，结合 BMI、年龄、体重等多维特征将样本分为四类，兼顾统计合理性与临床可解释性，且通过含时间惩罚项的优化模型，将推荐检测时点提前至 12.5-14.2 周，为后续诊断预留 4.8-6.5 周时间窗口，有效缓解临床压力；针对女胎异常判定，整合染色体 Z 值、GC 含量等多维特征，采用 SMOTE 过采样解决样本不平衡问题，随机森林模型 AUC 值高达 0.991，精确率与召回率均达 0.778，实现对染色体异常的高效判别。整体模型以数据驱动为核心，各环节紧密衔接，从浓度预测到时点优化再到异常判定，形成完整技术链条，为个性化 NIPT 筛查提供定量依据，临床应用价值突出。

模型局限

同时，模型仍存在可改进空间。其一，部分模型存在过拟合风险，如未优化的随机森林模型训练集与交叉验证 R^2 差距较大，稳定性有待进一步提升，需通过更优集成算法与正则化技术优化；其二，数据代表性存在局限，研究主要基于高 BMI 孕妇群体，对低 BMI、多胎妊娠等特殊人群覆盖不足，模型普适性需更多样化样本验证；其三，个性化适配性不足，当前重检策略基于固定规则，未充分结合孕妇个体生理指标动态变化，难以实现“一人一策”的精准化调整，需构建动态决策机制以适配个体差异。

7 模型改进

尽管本文构建的基于聚类分析与集成学习的 NIPT 时点优化及异常判定模型，在提升检测时效性、准确性及降低临床风险方面表现较好，但仍有优化空间。针对现有模型过拟合风险、数据局限性及个体差异适配不足等问题，可从三方面改进：一是深化集成学习与正则化融合，引入 XGBoost、LightGBM 等更优集成算法，结合贝叶斯优化进行超参数调优，同时在模型训练中加入 Dropout 等正则化技术，进一步抑制过拟合，提升泛化能力；二是拓展数据维度与样本多样性，纳入血清标志物、遗传信息等多模态数据，补充低 BMI、不同种族等人群样本，通过迁移学习解决小众群体数据稀缺问题，增强模型普适性；三是构建动态个性化决策系统，基于强化学习开发动态重检模型，结合孕妇实时生理指标与检测数据自适应调整策略，同时搭建可视化决策界面，融合医生临床经验，实现 NIPT 检测方案的精准化、动态化适配。

8 参考文献

[1] 周功益.无创产前检测(NIPT)在胎儿染色体疾病筛查中的应用效果观察研究[J].中文科技期刊数据库(文摘版)医药卫生,2025(3):060-063

[2] 李欣海.随机森林模型在分类与回归分析中的应用[J].应用昆虫学报,2013,50(04):1190-1197.

[3] 徐继伟,杨云.集成学习方法:研究综述[J].云南大学学报(自然科学版),2018,40(06):1082-1092.

[4] 姚登举,杨静,詹晓娟.基于随机森林的特征选择算法[J].吉林大学学报(工学版),2014,44(01):137-141.周杰英,贺鹏飞,邱荣发,等.融合随机森林和梯度提升树的入侵检测研究[J].软件学报,2021,32(10):3254-3265.

[5] 周杰英,贺鹏飞,邱荣发,等.融合随机森林和梯度提升树的入侵检测研究[J].软件学报,2021,32(10):3254-3265.

[6] 章永来,周耀鉴.聚类算法综述[J].计算机应用,2019,39(07):1869-1882.

[7] 张建萍,刘希玉.基于聚类分析的 K-means 算法研究及应用[J].计算机应用研究,2007,(05):166-168.

[8] 周世兵,徐振源,唐旭清.K-means 算法最佳聚类数确定方法[J].计算机应用,2010,30(08):1995-1998.

[9] 杨剑锋,乔佩蕊,李永梅,等.机器学习分类问题及算法研究综述[J].统计与决策,2019,35(06):36-40.

[10]DeepSeek,DeepSeek-R1-0528,深度求索（DeepSeek）,2025-09-05

[11]GoogleGemini,Gemini2.5Pro,Google（谷歌）,2025-09-06

附录

附录一：数据预处理（未清洗）的代码

解析原始数据中特定格式的孕周信息，将其转换为统一的数值格式:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor, RandomForestClassifier,
IsolationForest
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.cluster import KMeans, DBSCAN
from sklearn.mixture import GaussianMixture
from sklearn.metrics import mean_squared_error, r2_score,
classification_report, confusion_matrix, roc_auc_score
from sklearn.metrics import accuracy_score, precision_score, recall_score,
f1_score
from imblearn.over_sampling import SMOTE
import xgboost as xgb
import lightgbm as lgb

def parse_gestational_week(week_str):
    """解析孕周数据"""
    if pd.isna(week_str):
        return np.nan
    try:
        if 'w' in str(week_str):
            parts = str(week_str).replace('w', '').split('+')
            weeks = float(parts[0])
            days = float(parts[1]) if len(parts) > 1 else 0
            return weeks + days / 7
        else:
            return float(week_str)
    except:
        return np.nan

# 在主函数中加载和初步处理数据
# df_male = pd.read_excel('附件-男胎检测数据.xlsx')
# df_female = pd.read_excel('附件-女胎检测数据.xlsx')
# df_male['孕周数值'] = df_male['检测孕周'].apply(parse_gestational_week)
# df_female['孕周数值'] = df_female['检测孕周'].apply(parse_gestational_week)
```

附录 2:问题 1：Y 染色体浓度预测模型分析的代码

分析并建立能够预测男胎 Y 染色体浓度的数学模型：

```
def create_interaction_features(df, feature_cols):
    """创建交互特征和多项式特征"""
    interaction_df = df.copy()

    # 创建二次项特征
    for col in ['孕周数值', '孕妇 BMI', '年龄']:
```

```
        if col in df.columns:
            interaction_df[f'{col}_平方'] = df[col] ** 2

    # 创建交互项
    interaction_pairs = [
        ('孕周数值', '孕妇 BMI'),
        ('年龄', '孕妇 BMI'),
        ('孕周数值', '年龄'),
        ('体重', '孕妇 BMI')
    ]

    for col1, col2 in interaction_pairs:
        if col1 in df.columns and col2 in df.columns:
            interaction_df[f'{col1}_{col2}_交互'] = df[col1] * df[col2]

    return interaction_df

def advanced_regression_analysis(df):
    """高级回归分析 - 使用 XGBoost 和特征工程"""
    # 准备数据
    feature_cols = ['孕周数值', '身高', '年龄', '孕妇 BMI', '体重']
    df_clean = df.dropna(subset=feature_cols + ['Y 染色体浓度'])

    # 创建交互特征
    df_enhanced = create_interaction_features(df_clean, feature_cols)

    # 选择所有数值特征
    numeric_cols = df_enhanced.select_dtypes(include=[np.number]).columns
    feature_cols_enhanced = [col for col in numeric_cols if col not in ['Y 染色
体浓度', 'Unnamed: 0']]

    X = df_enhanced[feature_cols_enhanced]
    y = df_enhanced['Y 染色体浓度']

    # 数据标准化
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)
    X_scaled_df = pd.DataFrame(X_scaled, columns=feature_cols_enhanced)

    # 划分训练测试集
    X_train, X_test, y_train, y_test = train_test_split(X_scaled_df, y,
test_size=0.2, random_state=42)

    # 模型对比
    models = {
        '线性回归': LinearRegression(),
        '随机森林': RandomForestRegressor(n_estimators=100, random_state=42),
        'XGBoost': xgb.XGBRegressor(n_estimators=100, random_state=42,
verbosity=0),
        'LightGBM': lgb.LGBMRegressor(n_estimators=100, random_state=42,
verbosity=-1)
    }
```

```
    results = {}

    for name, model in models.items():
        # 训练模型
        model.fit(X_train, y_train)

        # 预测
        y_pred_train = model.predict(X_train)
        y_pred_test = model.predict(X_test)

        # 评估
        train_r2 = r2_score(y_train, y_pred_train)
        test_r2 = r2_score(y_test, y_pred_test)
        train_rmse = np.sqrt(mean_squared_error(y_train, y_pred_train))
        test_rmse = np.sqrt(mean_squared_error(y_test, y_pred_test))

        # 交叉验证
        cv_scores = cross_val_score(model, X_scaled_df, y, cv=5, scoring='r2')

        results[name] = {
            'model': model,
            'train_r2': train_r2,
            'test_r2': test_r2,
            'train_rmse': train_rmse,
            'test_rmse': test_rmse,
            'cv_mean': cv_scores.mean(),
            'cv_std': cv_scores.std()
        }

        print(f"{name}:")
        print(f"  训练 R²: {train_r2:.4f}, 测试 R²: {test_r2:.4f}")
        print(f"  训练 RMSE: {train_rmse:.4f}, 测试 RMSE: {test_rmse:.4f}")
        print(f"  交叉验证 R²: {cv_scores.mean():.4f} ±
{cv_scores.std():.4f}\n")

    # 选择最佳模型
    best_model_name = max(results.keys(), key=lambda k: results[k]['test_r2'])
    best_model = results[best_model_name]['model']

   return results, best_model
```

附录 3:问题 2&3：基于多维聚类的孕妇分组与最佳检测时点分析的代码
此部分综合处理问题 2 和问题 3，通过多维聚类方法对男胎孕妇进行分组。

```
def multi_dimensional_clustering(df):
    """多维聚类分析"""
    # 准备多维特征
    cluster_features = ['孕妇 BMI', '年龄', '体重', '身高']
    df_cluster = df.dropna(subset=cluster_features + ['Y 染色体浓度'])

    # 标准化特征
    scaler = StandardScaler()
    X_cluster = scaler.fit_transform(df_cluster[cluster_features])
```

```
    # 尝试不同的聚类算法
    clustering_methods = {}
    kmeans = KMeans(n_clusters=4, random_state=42)
    clustering_methods['K 均值聚类'] = kmeans.fit_predict(X_cluster)

    gmm = GaussianMixture(n_components=4, random_state=42)
    clustering_methods['高斯混合模型'] = gmm.fit_predict(X_cluster)

    dbscan = DBSCAN(eps=0.5, min_samples=5)
    clustering_methods['DBSCAN'] = dbscan.fit_predict(X_cluster)

    # 评估并选择最佳聚类方法
    best_method = None
    best_score = -1
    from sklearn.metrics import silhouette_score
    for method_name, labels in clustering_methods.items():
        n_clusters = len(np.unique(labels))
        if -1 in labels: n_clusters -= 1

        if n_clusters > 1:
            score = silhouette_score(X_cluster, labels)
            print(f"{method_name} 轮廓系数: {score:.4f}")
            if score > best_score:
                best_score = score
                best_method = method_name

    best_labels = clustering_methods.get(best_method, kmeans.labels_)
    df_cluster['cluster'] = best_labels

    # 分析各聚类的特征
    cluster_stats = []
    for cluster_id in sorted(df_cluster['cluster'].unique()):
        if cluster_id == -1: continue # 跳过噪声点

        cluster_data = df_cluster[df_cluster['cluster'] == cluster_id]

        # 计算达标率 (Y 染色体浓度 >= 4%)
        达标率 = (cluster_data['Y 染色体浓度'] >= 0.04).mean()

        # 计算平均达标时间
        达标样本 = cluster_data[cluster_data['Y 染色体浓度'] >= 0.04]
        平均达标时间 = 达标样本['孕周数值'].mean() if len(达标样本) > 0 else
np.nan

        # 推荐 NIPT 时点
        推荐时点 = 平均达标时间 - 0.5 if not np.isnan(平均达标时间) else 12.0

        stats = {
            '聚类': f'第{cluster_id+1}组',
            'BMI 范围': f"[{cluster_data['孕妇 BMI'].min():.1f}, {cluster_data['
孕妇 BMI'].max():.1f}]",
            'BMI 均值': cluster_data['孕妇 BMI'].mean(),
            '年龄均值': cluster_data['年龄'].mean(),
```

```
            '体重均值': cluster_data['体重'].mean(),
            '样本数': len(cluster_data),
            '达标率': 达标率,
            '平均达标时间': 平均达标时间,
            '推荐时点': 推荐时点
        }
        cluster_stats.append(stats)

        print(f"第{cluster_id+1}组:")
        print(f"  BMI: {stats['BMI 均值']:.1f} ({stats['BMI 范围']})")
        print(f"  年龄: {stats['年龄均值']:.1f}岁")
        print(f"  体重: {stats['体重均值']:.1f}kg")
        print(f"  样本数: {stats['样本数']}")
        print(f"  达标率: {stats['达标率']:.1%}")
        print(f"  推荐 NIPT 时点: {stats['推荐时点']:.1f}周\n")

    return cluster_stats, df_cluster
```

附录 4：女胎异常判定方法分析的代码

针对女胎，此部分旨在建立一个准确的染色体异常判定模型。

```
    def advanced_female_classification(df_female):
        """高级女胎异常判定分析"""
        # 准备特征和目标
        z_cols = ['13 号染色体的 Z 值', '18 号染色体的 Z 值', '21 号染色体的 Z 值', 'X
染色体的 Z 值']
        feature_cols = z_cols + ['GC 含量', '孕妇 BMI', '年龄']
        df_clean = df_female.dropna(subset=feature_cols)

        # 基于 Z 值阈值创建异常标记
        df_clean['异常'] = ((abs(df_clean['13 号染色体的 Z 值']) > 3.0) |
                            (abs(df_clean['18 号染色体的 Z 值']) > 3.0) |
                            (abs(df_clean['21 号染色体的 Z 值']) > 3.0)).astype(int)

        # 创建增强特征
        df_clean['Z 值最大值'] = df_clean[z_cols].max(axis=1)
        df_clean['Z 值总和'] = df_clean[z_cols].sum(axis=1)
        df_clean['Z 值标准差'] = df_clean[z_cols].std(axis=1)
        df_clean['GC 偏差'] = abs(df_clean['GC 含量'] - df_clean['GC 含量
'].mean())
        for col in z_cols:
            df_clean[f'{col}_异常'] = (abs(df_clean[col]) > 2.5).astype(int)
        df_clean['多重异常'] = df_clean[[f'{col}_异常' for col in
z_cols]].sum(axis=1)

        enhanced_features = feature_cols + ['Z 值最大值', 'Z 值总和', 'Z 值标准差',
'GC 偏差', '多重异常']

        X = df_clean[enhanced_features]
        y = df_clean['异常']

        # 处理数据不平衡
        X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42, stratify=y)
```

```
        # 标准化
        scaler = StandardScaler()
        X_train_scaled = scaler.fit_transform(X_train)
        X_test_scaled = scaler.transform(X_test)

        # SMOTE 过采样
        smote = SMOTE(random_state=42)
        X_train_balanced, y_train_balanced = smote.fit_resample(X_train_scaled,
y_train)

        # 模型对比
        models = {
            '逻辑回归': LogisticRegression(class_weight='balanced',
random_state=42),
            '随机森林': RandomForestClassifier(class_weight='balanced',
random_state=42),
            '孤立森林': IsolationForest(contamination=y.mean(),
random_state=42),
            'XGBoost': xgb.XGBClassifier(random_state=42, eval_metric='logloss')
        }

        results = {}

        for name, model in models.items():
            if name == '孤立森林': # 异常检测模型
                model.fit(X_train_scaled[y_train == 0]) # 只用正常样本训练
                y_pred = (model.predict(X_test_scaled) == -1).astype(int)
                y_pred_proba = None
            else: # 分类模型
                model.fit(X_train_balanced, y_train_balanced)
                y_pred_proba = model.predict_proba(X_test_scaled)[:, 1] if
hasattr(model, 'predict_proba') else None
                y_pred = (y_pred_proba > 0.5).astype(int) if y_pred_proba is not
None else model.predict(X_test_scaled)

            # 评估指标
            auc = roc_auc_score(y_test, y_pred_proba) if y_pred_proba is not
None else np.nan
            results[name] = {
                'model': model,
                'accuracy': accuracy_score(y_test, y_pred),
                'precision': precision_score(y_test, y_pred, zero_division=0),
                'recall': recall_score(y_test, y_pred, zero_division=0),
                'f1': f1_score(y_test, y_pred, zero_division=0),
                'auc': auc
            }

            print(f"{name}:")
            print(f"  准确率: {results[name]['accuracy']:.3f}, 精确率:
{results[name]['precision']:.3f}, "
                  f"召回率: {results[name]['recall']:.3f}, F1 分数:
{results[name]['f1']:.3f}, AUC: {results[name]['auc']:.3f}\n")
```

```
    return results, df_clean
```

附录 5:AI 工具使用详情（见支撑材料：AI 工具使用详情.pdf）
附录 6:完整代码文件（见支撑材料：question1forestpro.py、sum-question1-4.py、problem1_simple_models.py、enhanced_analysis.py、corrected_analysis.py）
附录 7：原始数据分表文件（见支撑材料：附件-男胎检测数据.xlsx、附件-女胎检测数据.xlsx）