基于数据驱动的NIPT时点选择与胎儿异常判定研究

摘 要

近年随着三孩政策的落实，国家逐步鼓励生育应对严峻的人口老龄化问题。伴随着剩余数量的增加，对于孕妇的产前检测变得愈发重要，NIPT作为新兴无创技术自然受到更多人的青睐，也引发了对其影响因素的探究。

针对问题一，探究男胎Y染色体浓度预期影响因素之间的关系。本文首先运用**多元非线性回归模型**建立整体函数关系，选择显著性水平并检验其显著性。接着我们探究不同影响因素之间的相关性程度，进一步探究其对Y染色体浓度的影响。最后我们得出孕妇BMI、年龄和检验孕周对男胎Y染色体浓度影响最大。

针对问题二，探究BMI与最佳NIPT检测时间的联系以防减少治疗窗口期。本文首先运用肘部法和轮廓系数寻找最佳聚类数即确定所分组数，其次运用**K-均值算法**和分类树回归算法评估风险以确定最佳NLPT时点，最后运用蒙特卡洛模拟和高斯噪音进行误差分析，减少其对实验结果的影响。

针对问题三，探究除BMI之外的其他因素对最佳NIPT检测时间的联系以防减少治疗窗口期。本文首先运用分位数回归模型预测Y浓度并计算达标孕周，其次运用**超图聚类算法**融合身高、体重、年龄和BMI等特征进行合理分组，并结合风险评估函数确定每组的最佳NIPT时点，最后运用蒙特卡洛模拟和高斯噪声进行误差分析，减少其对实验结果的影响。

针对问题四，探究判断女胎异常的影响因素及其预测方法。基于包含孕周、孕妇BMI、GC含量等17个特征的数据集，我们采用四种采样方法与四种**随机森林**变体进行两两组合，共生成16组模型。通过交叉验证评估F1分数筛选出最佳模型组合。为进一步提升预测性能，我们通过优化分类阈值提高模型的分类精度，最终模型通过特征重要性分析揭示关键影响因素，为女胎异常检测提供可靠支持。

本文就各特征值对于男女胎儿的影响进行分析，为产前检测的最佳NIPT时点提供良好的参考价值。

关键词：NIPT，多元非线性回归模型，K-均值算法，超图聚类算法，随机森林

一、问题背景与重述

1.1 问题背景

在医疗条件趋于完善的当下，对于一名孕妇来说，产前检查是不可或缺的一个环节。它可以综合评估孕妇生理状态和胎儿的发育情况，监测潜在风险并为针对问题提供相关指导。初期，羊膜穿刺为主要诊断方法，但其属于侵入性检查，增加了感染的风险[1]。而 NIPT 作为过去十余年内产前诊断领域的巨大突破，其对于染色体异常的检测更为灵敏，实用价值也更高[2]。

1.2 问题重述

NIPT 这种产前检测技术通过检测孕妇的血液情况来分析胎儿的生理状况。NIPT 的结果可信度取决于胎儿的染色体浓度。不同孕妇因其生理状态的差异而需要按照特定分组确定染色体浓度达标时间。附件给出了某 BMI 较高地区 267 名男胎孕妇及 147 名女胎孕妇的产前检测数据。

问题一：

对于男胎 Y 染色体浓度与其影响因素如孕周数和 BMI 的关系建立数学模型，并检验影响因素的显著性，保证其有效性。

问题二：

实验结果显示，男胎的 Y 染色体浓度最早达 4%的时间受其母亲 BMI 的影响，需要划分合理的 BMI 区间、进行误差检验并寻找该区间对应的 Y 染色体浓度优先达标时间，利于更早了解胎儿情况并监测孕妇可能面临的风险。

问题三：

除 BMI 外，男胎 Y 染色体浓度最早达 4%的时间也会受其母亲年龄等因素影响，考虑影响因子、实验本身误差和满足条件的 Y 染色体数量，划分合适的 BMI 区间、进行误差检验并确定其对应的最早 Y 染色体浓度时间，为后续检查提供充足时间。

问题四：

通过 21 号、18 号和 13 号染色体数量情况判断女胎异常与否，结合常染色体和性染色体各参数及 BMI 等因子寻找判定女胎问题的具体方法。

二、模型假设

1. 假设训练样本之间独立。
2. 假设输入数据（如孕妇 BMI、检测孕周）测量准确，无系统性偏差。
3. 采样方法（如 SMOTE、ADASYN）假设合成样本能够代表真实数据分布。
4. 不考虑外界因素如妊娠方法和生产次数对孕妇生理状况的影响。

三、符号声明

符号	说明	单位
C_y	男胎 Y 染色体浓度	/
bm	孕妇 BMI 值	/
p	孕妇年龄	岁
we	检测孕周	周
k_1, k_2, k_3	特征系数	/
WCSS	簇内平方和	/
K	簇的数量	/
C_i	第 i 个簇	/
x	簇 C_i 所包含的数据点	/
μ_i	第 i 个簇的均值	/
i / j / k	簇 C_i / C_j / C_k 的样本	/
d	样本之间的距离	/
a	同一簇的样本间距离	/
b	数据点到最近邻簇的平均距离	/
s	轮廓值	/
R	孕妇潜在风险	/
α	延迟风险系数	/
dr	延迟风险	/
su	男胎 Y 染色体达标比例	/
β	时间风险系数	/
tr	时间风险	/
NOI	高斯噪声	/
μ	噪声均值	/
σ	噪声标准差	/
τ	分位数中位数	/
x_i	特征值自变量	/
y_i	特征值因变量	/
ϵ	避免分母为 0 的极小值	/
$\bar{d}$	超边中节点的平均距离	/
$dist\_weight$	距离权重	/
$feature\_weight$	特征权重	/
w	超边权重	/
D	度矩阵	/
W	邻接矩阵	/
U	嵌入矩阵	/
t	时间	/
Sm	噪声指数	/

四、模型的建立和求解

4.1 问题一模型的建立与求解

本题我们首先尝试运用多元线性回归模型和普通最小二乘法对数据进行分析和函数近似，但发现题目给的数据经此两种模型分析后拟合程度较差，因此我们选用拟合情况更好的多元非线性回归模型。

4.1.1 基于多元非线性回归模型的特征显著值分析

基础的特征值包括孕妇 BMI、孕周数，孕妇年龄和测序得出的 GC 含量。总体上有 14 种特征值组合：8（单一特征自身和平方变量）+6（不同特征两两组合变量）。经过决策优化，最后得出的特征及其显著性如下图 1：

特征	系数	标准差	t值	p值	显著性
孕妇BMI	-0.005362	0.001269	-4.226865	0.000026	非常显著
年龄	-0.003641	0.001221	-2.982963	0.002936	显著
检测孕周	0.003665	0.001339	2.737051	0.006329	显著
孕妇BMI^2	-0.001177	0.000585	-2.011685	0.044568	较显著
检测孕周^2	0.002426	0.001265	1.917149	0.055554	不显著
年龄 * GC含量	-0.001662	0.001291	-1.286918	0.198473	不显著
孕妇BMI * 检测孕周	0.001115	0.001118	0.997708	0.318705	不显著
孕妇BMI * 年龄	0.001260	0.001296	0.972064	0.331295	不显著
孕妇BMI * GC含量	-0.000971	0.001227	-0.791354	0.428958	不显著
年龄^2	0.000440	0.000687	0.639960	0.522371	不显著
检测孕周 * 年龄	-0.000736	0.001165	-0.632256	0.527390	不显著
GC含量^2	-0.000253	0.000507	-0.498431	0.618309	不显著
检测孕周 * GC含量	0.000113	0.001137	0.099120	0.921066	不显著
GC含量	0.000057	0.001169	0.048456	0.961365	不显著

图 1：特征及其显著性

由上图可得，Y 染色体浓度与孕妇 BMI，年龄，检测孕周和 GC 含量的关系：

$$C_y = k_1 * bm + k_2 * p + k_3 * we + \dots$$

已知 P 值大于 0.05 时，影响因子较为显著。因此孕妇 BMI、年龄、检验孕周这三个影响因子的显著性较强。

4.1.2 主要特征与其余特征的相关性分析

从上述 14 种特征值中筛选出 8 种相关性较强的因子，特征值距离圆心越远代表该特征对 Y 染色体浓度的影响越大，特征雷达图如下图 2：

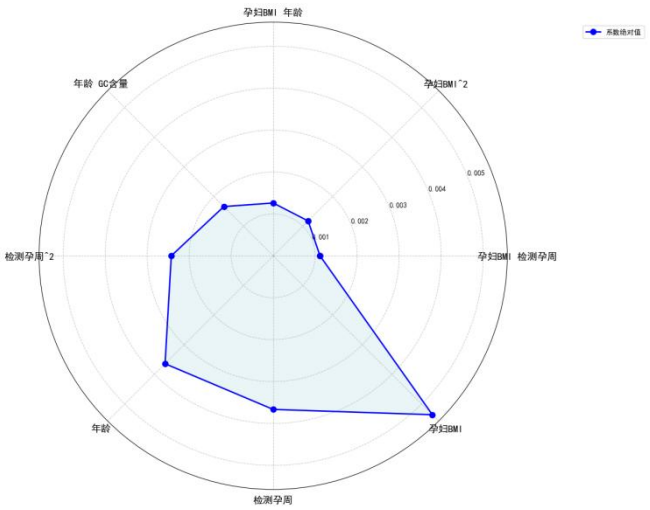

图 2：特征系数（绝对值）雷达图

通过上图可以看出，孕妇的 BMI 对其胎儿染色体浓度影响最大，检测孕周和年龄的影响其次，GC 含量的影响能力则远小于前三种。

其次建立不同影响因子之间的影响，与高度影响 Y 染色体的影响因子如孕妇 BMI，检测孕周等强相关的因子，其对 Y 染色体浓度的影响也会更大，特征彼此相关性如下图 3：

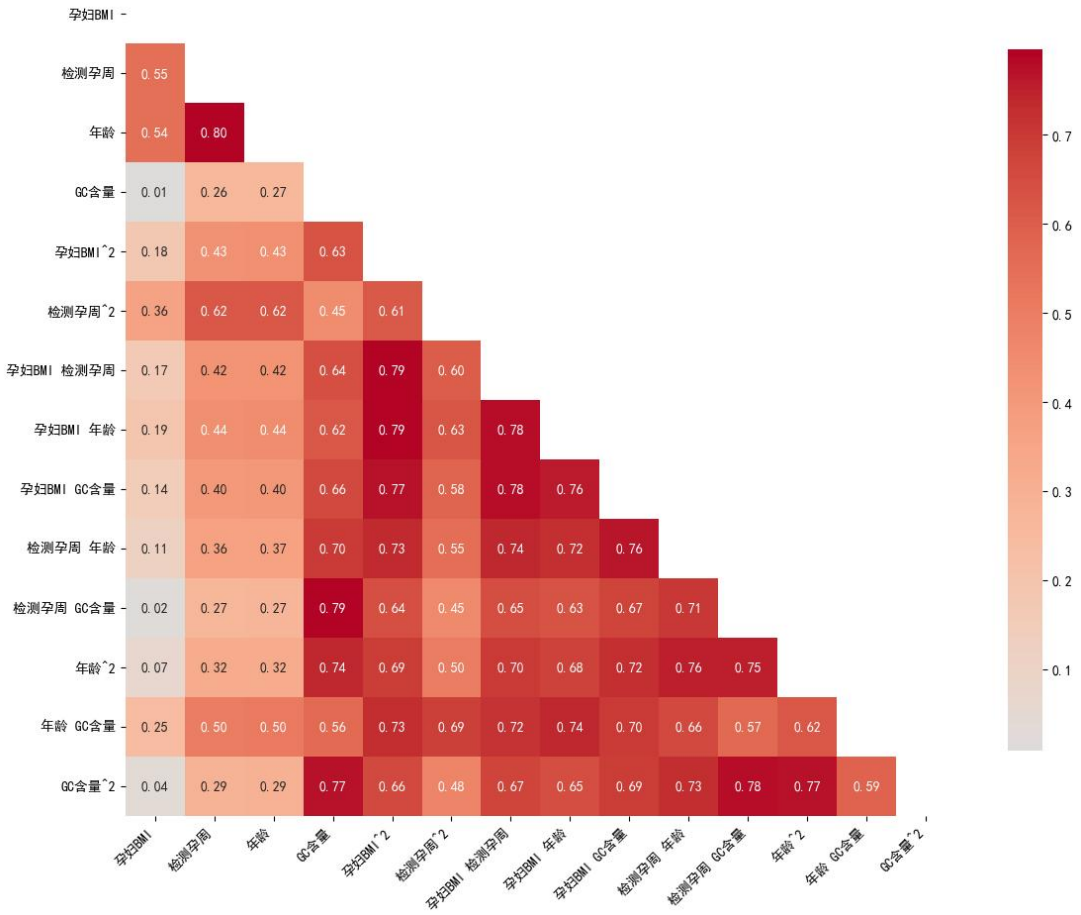

图 3：模拟特征相关性热力图

由上图可知，除孕妇 BMI、检测孕周和孕妇年龄外的其他影响因子与上述三个影响因子

的相关性均较小，因此其他因子对于 Y 染色体浓度的影响同样较小。

4. 2 问题二模型的建立与求解

在问题一的基础上，问题二进一步考虑影响 Y 染色体浓度的影响因子之间的关系，探究不同 BMI 对于 Y 染色体浓度达标时间的影响。

4. 2. 1 基于肘部法和轮廓系数的最佳聚类数求解

肘部法则，是机器学习中通过绘制不同 K 值（聚类数）簇内平方和的图像，寻找下降速率减缓的点（肘点），并将该点对应的 K 值视作最优聚类数量的常用方法。计算簇内平方和：

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2$$

轮廓系数即对不同的 K 值将聚类算法运用到数据集，用于评估聚类后簇与簇之间的离散程度。计算同一簇样本间平均距离：

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

数据点到最近邻簇的平均距离：

$$b(i) = \min_{k \neq C_i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

定义轮廓值：

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

分类讨论：

$$s(i) = \begin{cases} 1 - a(i)/b(i), if\ a(i) < b(i) \\ 0, \qquad\qquad if\ a(i) = b(i) \\ b(i)/a(i) - 1, if\ a(i) > b(i) \end{cases}$$

两种方法得到对应 K 值的函数：

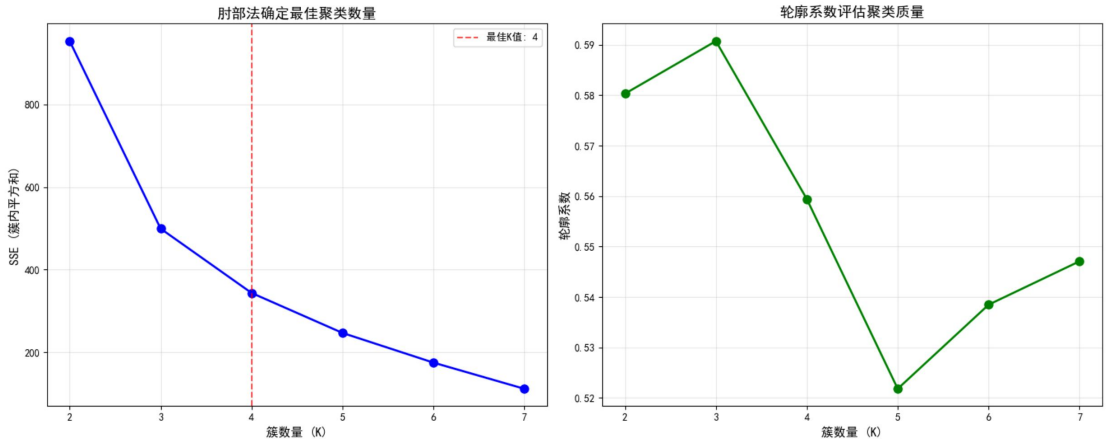

图 4：K 值函数

由上图可知肘部法的 K 值为 4，轮廓系数评估 K 值也为 4。根据 Calinski-Harabasz 指数，选择两者较大的 K 作为最终的结果。因此最终选择 4 为最佳 K 值，即将 BMI 分为 4 组。

4.2.2 基于 K-均值算法和分类树回归算法的最佳 NIPT 时点选择与风险检测

K-均值算法可将若干点划分到若干聚类中，利用高效的启发式算法将同一聚类的点快速收敛于一个局部最优解。因此使用 K-均值算法将孕妇按照最佳 K 值及其 BMI 分组：

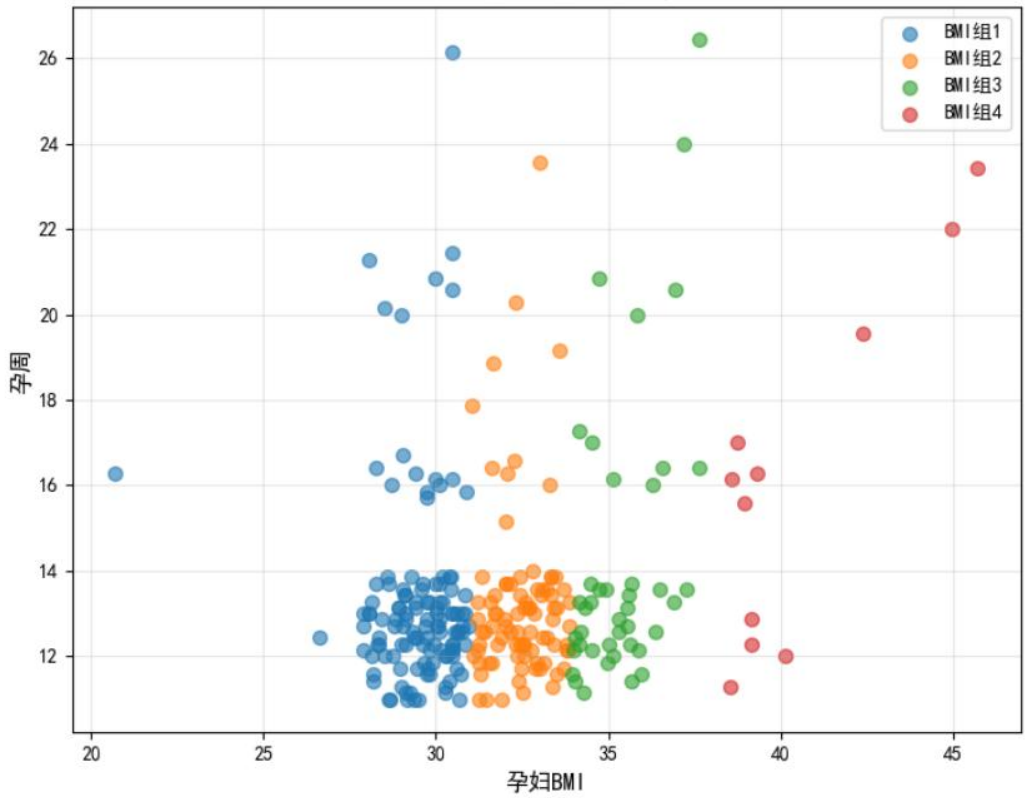

图 5：BMI 与达标孕周散点图

对于孕妇可能的潜在风险进行评估，可以分为延迟风险和时间风险。如果产前检测进行过早，可能因为 Y 染色体浓度不达标致使检测结果不准确，即时间风险；反之如果产前检测进行过晚，可能因为胎儿染色体数量异常情况发现不及时而影响治疗窗口期，即延迟风险。因此建立风险评估函数：

$$R = \alpha * dr * (1 - su) + \beta * tr * su$$

依据题干定义孕妇时间风险与最早检测时间的关系：

$$tr = \begin{cases} 0, & if\ 0 < t <= 12 \\ 5, & if\ 13 <= t <= 27 \\ 10, & if\ t > 27 \end{cases}$$

延迟风险受延迟时间影响，因此在原式上递归计算累计风险。

根据分类树回归算法，由于置信区间为 95%，即参数取值范围为前 95%，将四组孕周数的第 5 百分位数视作最早安全检测孕周，得到 BMI 分组及检测时点表格如下图 4：

组别	BMI范围	BMI均值	样本数	最早安全检测孕周	推荐检测孕周	95%置信区间
1	(20.7,31.0]	29.5	121	11.1	12	12.9-13.8
2	(31.0,33.9]	32.5	86	11.3	12	12.8-13.7
3	(33.9,37.6]	35.4	42	11.4	12	13.3-15.4
4	(37.6,45.7]	40.5	11	11.6	12	13.5-19.0

图 6：BMI 分组及检测时点（优化前）

由上图可知，因时间风险函数的分段较为严格，其非连续性致使时间风险出现激增的情况，进而导致推荐检测孕周数固定。

因此，选取每一簇染色体浓度达标时的时间为集合，并取其中位数近似为该簇的最佳 NIPT 时点。同时运用分类树回归算法选取第 5 百分位数作为最早安全检测孕周，保证 95% 的孕妇胎儿 Y 染色体浓度达标，优化后 BMI 分组及检测十点表格如下图 5：

组别	BMI范围	BMI均值	样本数	最早安全检测孕周	推荐检测孕周	95%置信区间
1	(20.7,31.0]	29.5	121	11.1	12.7	12.6-13.0
2	(31.0,33.9]	32.5	86	11.3	12.8	12.4-13.1
3	(33.9,37.6]	35.4	42	11.4	13.2	12.6-13.6
4	(37.6,45.7]	40.5	11	11.6	16.1	12.3-19.6

图 7：BMI 分组及检测时点（优化后）

4.2.3 基于蒙特卡洛模拟和高斯噪声进行误差分析

蒙特卡洛模拟，以随机采样为基础，通过多次随机实验来近似实验结果。用于评估测量误差对于 NIPT 检验时间及其检验结果的敏感性。

高斯噪声，服从正态分布的随机噪声，定义其函数：

$$NOI = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

拟定模拟次数为 1000 次，设定高斯噪声相对标准差为 10%，BMI 分组及检测时点误差如图 6：

组别	原始推荐孕周	模板平均推荐孕周	推荐孕周平均推迟	推荐孕周标准差	检测失败率
1	12.7	12.7	0	0	0.00%
2	12.8	12.8	0	0.1	0.00%
3	13.2	13.2	0	0.1	0.00%
4	16.1	15.9	-0.3	0.4	0.00%

图 8：BMI 分组及检测时点（误差分析后）

误差影响后的数据仍在 95%置信区间范围内，因此误差对实验结果影响较小，在可控范围内。

4.3 问题三模型的建立与求解

问题三作为问题二的延伸，拓展探究除问题二提到的 BMI 外，其他因素如年龄等对 Y 染色体浓度达标时间的影响。进一步优化问题二得出的 BMI 分组和最佳 NIPT 时点。因此需要对问题一的模型进行扩展，将以上影响因子作为自变量纳入，重新构建 Y 染色体浓度与变量之间的关系模型。

4.3.1 基于分位数回归模型的函数优化

分位数回归模型，主要用于研究自变量条件与因变量条件分位数之间的关系，可进一步推断因变量的条件概率分布，分位数回归优化：

$$\min_{k_i} \sum_{i:y_i \geq X_i k_i} \tau |y_i - X_i k_i| + \sum_{i:y_i < X_i k_i} (1-\tau)|y_i - X_i k_i|$$

采用分位数回归模型，将身高和体重两个影响因子纳入问题一中的多元非线性回归模型。进一步分析其对 Y 染色体浓度的影响，并添加多项式特征（二次项）以捕捉所有的非线性关系。

4.3.2 基于二分法的初步筛选

为每个样本计算其达标孕周，按照 BMI 进行初步分组。用二分法寻找最小孕周，使得 Y 染色体浓度达标，初步分析 BMI 对达标孕周的梯度的影响，如下图 7：

孕妇BMI	样本量	中位数	均值	标准差
(20.677, 25.938]	4	20.43985182	20.19609717	1.837497522
(25.938, 31.172]	430	20.00903906	19.97670238	1.13770932
(31.172, 36.406]	557	20.0047661	19.96664967	1.143348794
(36.406, 41.641]	81	19.91121725	20.01271708	1.133419408
(41.641, 46.875]	9	20.4641999	20.61157941	0.954318129

图 9：孕妇 BMI 初期分组

4.3.3 基于超图聚类进行 BMI 分组

4.3.3.1. 特征准备：

首先选择特征并构造特征矩阵：['BMI', '年龄', '身高', '体重']，进而对各特征进行标准化。

4.3.3.2 构建超图：

首先为 BMI，身高，体重和年龄设置默认权重，其中 BMI 权重最大。其次对特征进行归一化。接着对于所有特征，提取其特征值，用 KNN 找邻域，生成其对应超边。

距离权重，代表特征本身的权重：

$$dist\_weight = \frac{1}{\bar{d} + \epsilon}$$

超边权重，反映超边所代表的节点集合的关联强度：

$$w(e) = dist\_weight \times feature\_weight$$

4.3.3.3 拉普拉斯矩阵，反映拉普拉斯算子的离散化：

$$\begin{cases} L = D - W \\ D_{ii} = \sum_{e \ni i} w(e) \\ W_{ij} = \sum_{e \ni i,j} \frac{w(e)}{|e|} (i \neq j) \end{cases}$$

4.3.3.4 特征嵌入与聚类优化

对 L 进行特征分解，获取特征值和特征向量；选取前 k 个特征向量（排除最小特征值对应的零向量），构建低维嵌入矩阵:

$$U = [u_2, u_3, \ldots, u_{k+1}]$$

用 K-Means 对嵌入矩阵聚类从而得到初始聚类标签；调整聚类结果，确保组间 BMI / 达标孕周差异足够大；计算聚类统计量并生成最终聚类标签，如下图 8：

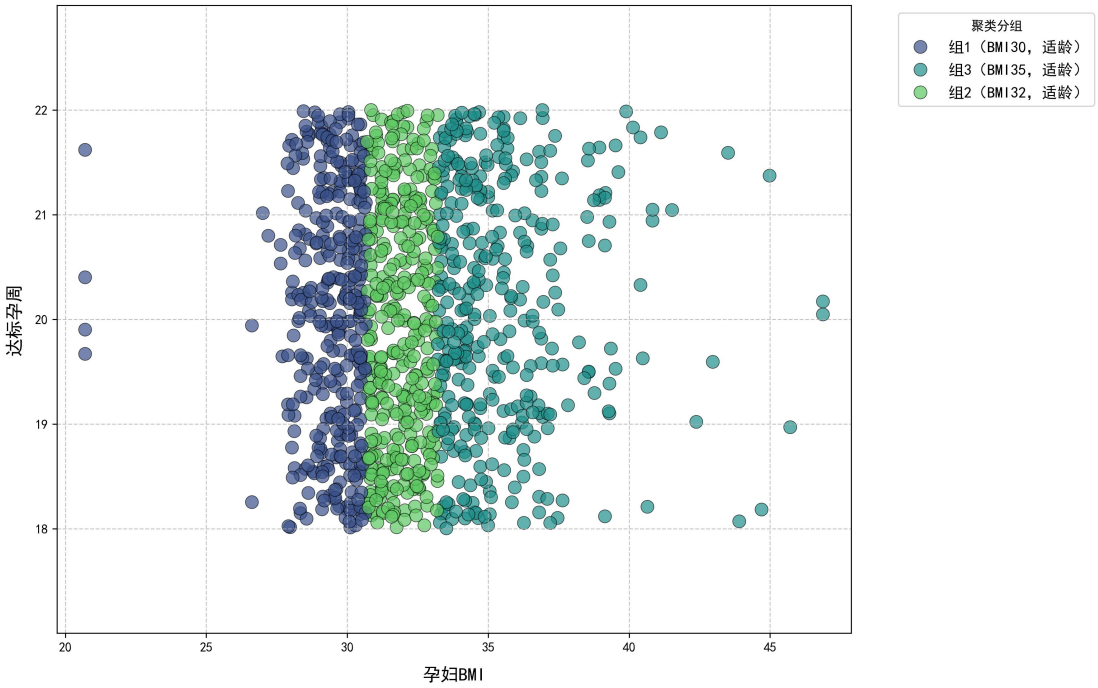

图 10：聚类后的 BMI 与达标孕周分布

4.3.4 风险函数

定义时间风险和延迟风险，量化过早检测的代价，但总体倾向于早期检测。

时间风险：

$$tr = \begin{cases} 1.0, & if\ t < 11 \\ 2.0, & if\ 11 \le t < 13 \\ 5.0, & if\ 13 \le t < 18 \\ 12.0 & if\ 18 \le t < 24 \\ 25.0 & if\ t \ge 24 \end{cases}$$

延迟风险：

$$\begin{cases} r_d(t,p,r) = \begin{cases} 30.0 & r \le 0\ 或 t > 40 \\ r_t(t) \cdot (1 - f_p) + r_d(t+2, p, r-1) \cdot f_p \cdot 2.0 & 其他 \end{cases} \\ f_p = \max(0.1, 1-p) \end{cases}$$

总风险：

$$\mathrm{R} = 0.6 \cdot r_t + 0.4 \cdot r_d$$

由公式可以看出，时间风险鼓励早期检测，延迟风险递归考虑失败后果，时间风险权重较大。

4.3.5 最佳检测时点优化

按聚类优化非侵入性产前检测时点，最小化总风险。采用多目标优化迭代候选孕周，结合风险函数选择最佳时点。

成功概率，即给特定孕周进行 NIPT 检测时，达到达标孕周的样本比例。

优化：

$$w^* = arg\min_{w} 总风险\ (w)$$

4.3.6 蒙特卡洛误差分析

通过 200 次噪声模拟，评估模型对数据变异的稳定性。

蒙特卡洛模拟：添加随机噪声，重新计算达标孕周和比例。

噪声：

$$Sm = bm \cdot (1 + N(0,0.08))$$

$$S_Y = C_Y + N(0, C_Y \cdot 0.7)$$
$$S_Y \in [0.0005,0.4]$$
$$Sm \in [16,55]$$

误差分析如下图 9：

BMI分组	原始达标比例	差后平均达标比	95%孕周波动范围	模拟结果中位数	模拟结果四分位距
组1	88.3%	70.6%	18.80-21.22周	20.00周	0.15周
组3	81.2%	65.1%	18.77-21.21周	20.00周	0.15周
组2	90.3%	70.2%	18.80-21.20周	20.01周	0.12周

图 11：误差分析

最终结论如下图 10：

BMI分组	BMI区间	BMI中位数	最佳检测时点(周)	最佳时点成功概率	时间风险	延迟风险	总风险
组1	20.7-30.7	29.7	21.5	84.7%	12	16.12	13.14
组2	30.7-33.2	31.8	21.5	90.0%	12	14.09	12.84
组3	33.2-46.9	34.8	21.5	88.7%	12	14.5	13.47

图 12：问题三结论

4.4 问题四模型的建立与求解

问题四侧重于探究女胎的染色体一场判定方法，由于女胎及其母亲自身都不携带 Y 染色体，异常情况与特定染色体数量、读段数等其他影响因子有一定的联系。

4.4.1 基于采样方法的数据完善

优先筛选 GC 含量在 40% ～ 60%的序列，其次通过 13 号，18 号，21 号染色体数量检测进行判定，染色体数量异常的记作 1，正常的则记作 0。最后将 17 个特征标准化后，对于正异常的数据，都保证其 8：2 的训练集与测试集比例。

由附件数据可知，605 条女胎数据中仅有 67 条为异常样本，发现其比例相对较小。为了解决这一数据不平衡问题，我们建立四种采样方法：SMOTE 欠采样，边界线 SMOTE，简单过采样和自适应合成采样。

4.4.2 基于随机森林模型的模型选择

查准率，指的是预测值和真实值均为 1 的样本在所有预测值为 1 的样本中所占的比例；

召回率，指的是预测值和真实值均为 1 的样本再真实值为 1 的样本中所占的比例：

	真实 1	真实 0
预测 1	True Positive(TP)真阳性	False Positive(FP)假阳性
预测 0	False Negative(FN)假阴性	True Negative(TN)真阴性

图 13：预测值真实值间衡量指标

真阳率与假阳率关系如下图 12，越偏向左上代表数据越准确：

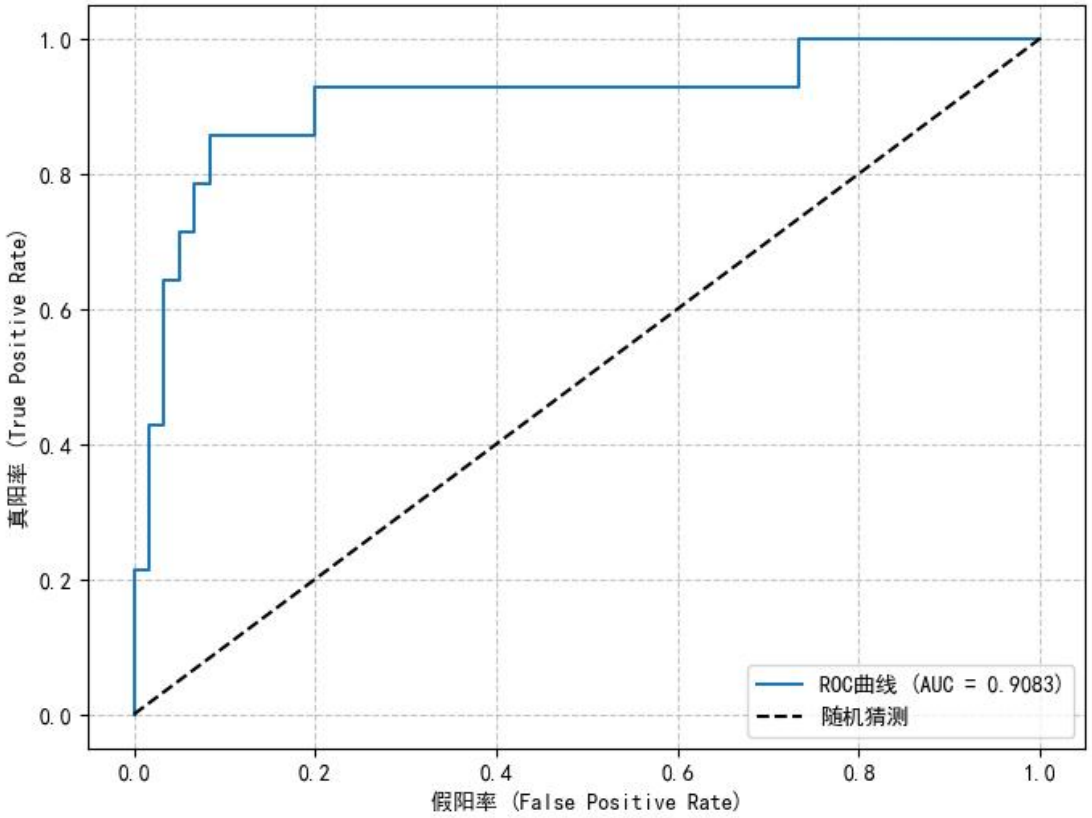

图 14：ROC 曲线

F1 分数，用于评价模型效果，定义为查准率和召回率的调和平均数：

$$F_1 = 2 * \frac{pr * re}{pr + re}$$

运用随机森林模型，分支产生四种不同的算法模型。将其与四种采样方法两两结合，进行交叉验证、训练和结果预测。得到每个模型和采样组合的 F1 值，如下图 8：

采样 / 模型	SMOTE欠采样	边界线SMOTE	简单过采样	自适应合成采样
加权随机森林	0.740741	0.740741	0.689655	0.592593
标准随机森林	0.692308	0.740741	0.714286	0.615385
平衡随机森林	0.645161	0.689655	0.642857	0.571429
简易集成	0.571429	0.600000	0.551724	0.606061

图 15：采样模型组合及 F1 分数

分析最佳 F1 分数的特征值，可以得到其重要性：

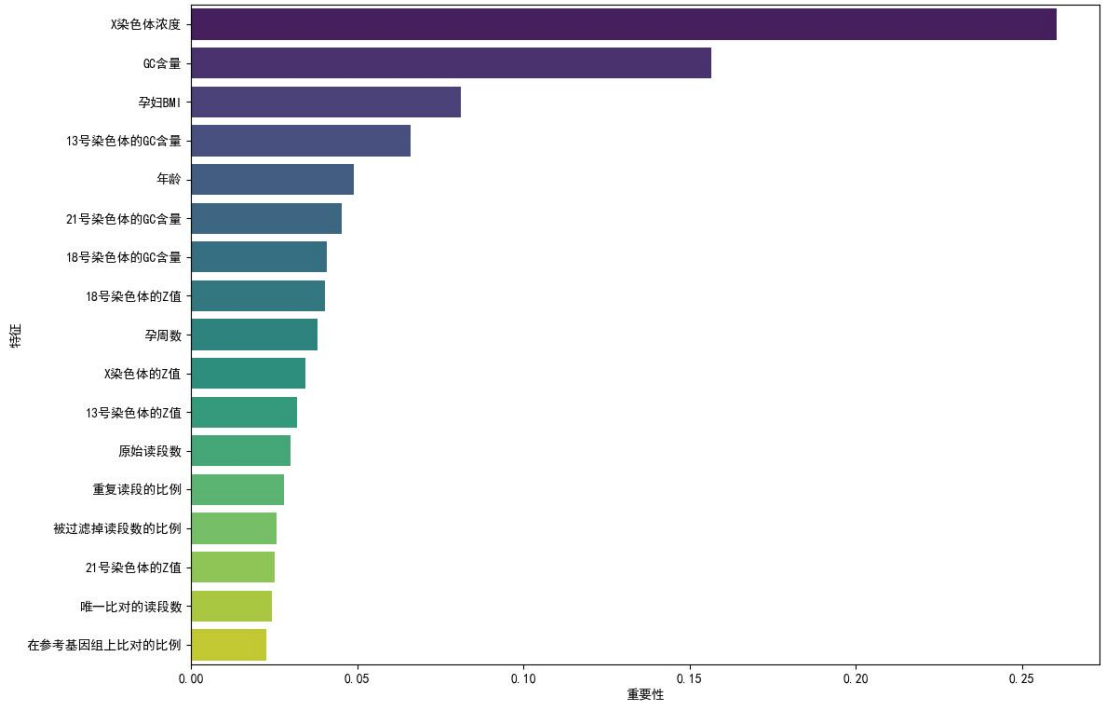

图 16：特征重要值分析

五、模型优缺点

优点：

1. 采用非线性模型，能够捕捉特征之间的交互和非线性关系，减小均方误差。
2. 超图聚类能捕捉多重关系，组间区分度高，灵活强调特定特征的贡献，适应不同数据集的特性。
3. 分位数回归对异常值和非正态分布数据更稳健，适合处理 Y 染色体浓度等可能存在偏态的生物学数据。
4. 通过多种采样方法有效处理了异常样本和正常样本的不平衡问题，改善了模型对少数类的预测能力。
5. 使用蒙特卡洛模拟（10%高斯噪声，1000 次模拟）评估了推荐孕周的稳健性，提供了检测失败率和平均延迟等指标，增强了模型的鲁棒性。

缺点：

1. 可能存在较多噪声，模型的决定系数（R^2）较小，表明模型可能无法有效解释 Y 染色体浓度的变异性。
2. 超图构建（通过 KNN 生成超边）和拉普拉斯矩阵计算涉及大量样本间的距离计算，尤其当样本量较大时，计算开销显著增加，影响效率。
3. 随机森林分类复杂性高，存在过拟合风险，性能会产生波动。

六、参考文献

[1] Qi, Q.-G. *et al.* (2021) 'Amniocentesis and Next Generation Sequencing (NGS)-Based Noninvasive Prenatal DNA Testing (NIPT) for Prenatal Diagnosis of Fetal Chromosomal Disorders', *International Journal of General Medicine*, 14, pp. 1811–1817. doi: 10.2147/IJGM.S297585.

[2] 潘澍青,潘小莉,葛丽莎等.无创产前检测在产前修复中的应用价值[J].浙江医学,2024,46(17):1881-1884.

七、附件目录

代码源文件：

问题一

问题二

问题二风险函数

问题三

问题四