基于多模型协同驱动的 NIPT 检测时点决策与胎儿染色体异常判定

王昕怡 杨子琦 张宇轩 李争平
北方工业大学，北京 100144

[摘要]文章针对 NIPT 技术中男胎 Y 染色体浓度关联分析、BMI 分组及最佳检测时点优化、女胎异常判定等核心问题，通过斯皮尔曼相关系数和混合效应模型揭示了三个因素间的相关属性；基于临床风险规律构建分段风险量化模型，采用决策树聚类将 BMI 划分为四个最优区间，并通过网格搜索优化得到各组最佳检测时点。针对女胎异常判定问题，采用 SMOTE 过采样和 Stacking 集成学习方法构建分类模型，为临床提供了一种可靠的异常判定方法。

[关键词]斯皮尔曼相关系数；混合效应模型；Y 染色体浓度动态预测模型；Bootstrap 不确定性分析；高斯混合模型
DOI：10.33142/cmn.v3i2.18151 中图分类号：R714 文献标识码：A

NIPT Detection Timing Decision and Fetal Chromosome Abnormality Determination Based on Multi Model Collaborative Driving

WANG Xinyi, YANG Ziqi, ZHANG Yuxuan, LI Zhengping
North University of Technology, Beijing, 100144, China

Abstract: This article focuses on core issues in NIPT technology, such as the correlation analysis of Y chromosome concentration in male fetuses, BMI grouping and optimization of optimal detection time points, and abnormal determination of female fetuses. Through Spearman correlation coefficient and mixed effects model, the article reveals the correlation properties between the three factors; Based on clinical risk patterns, a segmented risk quantification model is constructed. Decision tree clustering is used to divide BMI into four optimal intervals, and the optimal detection time points for each group are obtained through grid search optimization. For the problem of abnormal diagnosis of female fetuses, SMOTE oversampling and Stacking ensemble learning methods are used to construct a classification model, providing a reliable method for abnormal diagnosis in clinical practice.

Keywords: Spearman correlation coefficient; mixed effects model; Y chromosome concentration dynamic prediction model; Bootstrap uncertainty analysis; Gaussian mixture model

1 问题分析

1.1 问题一

对于问题一，解题思路是先对数据进行预处理，包括数据清洗和特征提取，以确保数据的质量和可用性。接着，通过斯皮尔曼相关性分析来初步探索 Y 染色体浓度与孕周数、BMI 等指标之间的关系。然后，基于相关性分析的结果，构建一个能够描述这种关系的数学模型。最后，通过统计方法对模型进行显著性检验，以验证模型的有效性和可靠性。

1.2 问题二

本问题的核心是解决高 BMI 孕妇 NIPT 检测失败率高的临床难题。从数据本质来看，Y 染色体浓度随孕周变化的轨迹在不同 BMI 孕妇间存在显著差异，这种差异导致无法为所有孕妇设定统一的最佳检测时点。解决思路聚焦于三个关键环节：首先，基于 Y 染色体浓度变化的异质性，采用数据驱动方法确定最优 BMI 分组，超越传统分类标准；其次，构建 Y 染色体浓度动态模型，准确预测不同 BMI 孕妇 Y 浓度达到 0.04 的时间点，特别关注高 BMI 孕妇的非线性效应；最后，设计风险-可靠性权衡函数，将孕周选择转化为多目标优化问题，通过网格搜索确定各组最佳检测时点，并利用 Bootstrap 方法评估检测误差对结果的影响，确保模型在临床不确定性下的稳健性。

1.3 问题三

对于问题三，首先需要对数据进行细致的预处理，包括数据清洗和特征提取，以确保数据的准确性和完整性。接着，通过综合分析多种因素（如身高、体重、年龄等）对 Y 染色体浓度达标时间的影响，识别出对达标时间有显著影响的关键因素。然后，基于这些关键因素，对男胎孕妇的 BMI 进行合理分组，并确定每个分组的最佳 NIPT 时点，以最小化孕妇的潜在风险。最后，通过误差分析来评估检测误差对结果的影响，确保检测结果的准确性和可靠性。

1.4 问题四

问题四的核心在于如何综合考虑多种因素来判定女胎是否异常，如何给出女胎异常的判定方法。首先应当对数据进行预处理，确保数据的完整性、可靠性以及可用性。接着，通过相关性分析筛选出与女胎异常判定相关性较大的特征。基于这些特征构建女胎异常的判定模型。最后，利用已知的胎儿健康状况数据对模型进行验证，评估模型的性能。

2 模型假设

（1）假设检测误差服从正态分布。

（2）假设各因素（如身高、体重、年龄等）之间相互独立。

（3）假设 BMI 分组区间内的数据具有同质性。

（4）假设 Y 染色体浓度达标比例与检测时点之间存在线性关系。

（5）假设检测误差对 Y 染色体浓度达标时间的影响是随机的。

3 问题一模型的建立与求解

3.1 数据预处理

在数据预处理阶段，我们首先对附件中的 C、D、E、J、K、V 列进行了清洗。针对孕妇 BMI、年龄、身高、体重四组数据，采用 IQR 方法识别出 26 组异常值。考虑到极端值可能为真实情况，且多数个体有多次观测记录，我们设定：若某个体存在异常值，则计算该个体该因素的均值，仅当异常值超过均值的 25%时才予以剔除。依此方法，仅删除了序号 670 的异常数据。

3.2 初步探究相关性—斯皮尔曼相关系数

我们首先计算了各个因素间的斯皮尔曼相关系数。斯皮尔曼相关系数的计算公式为：

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (1)$$

如图 1 所示，孕妇 BMI 与 Y 染色体浓度呈中度负相关，而检测孕周与 Y 染色体浓度呈正相关关系。由此我们可以初步推断出 Y 染色体浓度与孕妇 BMI 呈负相关关系与检测孕周呈正相关。

3.3 进一步探究相关性—混合效应模型

为探究 Y 染色体浓度与孕妇 BMI、检测孕周的关系，考虑到数据中存在多个个体具有重复观测值，本研究采用线性混合效应模型进行分析。该模型可同时考虑固定效应和随机效应，适用于此类数据结构。

模型设定如下：

$$y_{ij} = \beta_0 + \beta_1 G_{ij} + \beta_2 H_{ij} + b_i + \epsilon_{ij} \quad (2)$$

其中，固定效应包括检测孕周和孕妇 BMI，随机效应b_i表示个体随机截距，用于捕捉个体间重复测量带来的相关性。参数估计采用限制性最大似然法。

结果表明，孕周对 Y 染色体浓度具有显著正向影响（每日增加 0.00044 单位），而 BMI 呈显著负向影响（每单位 BMI 降低 0.00147 单位）。临床推测，孕周增加可能促进胎儿 DNA 释放或检测灵敏度提高，而 BMI 升高可能抑制胎儿 DNA 释放或增强母血稀释效应。

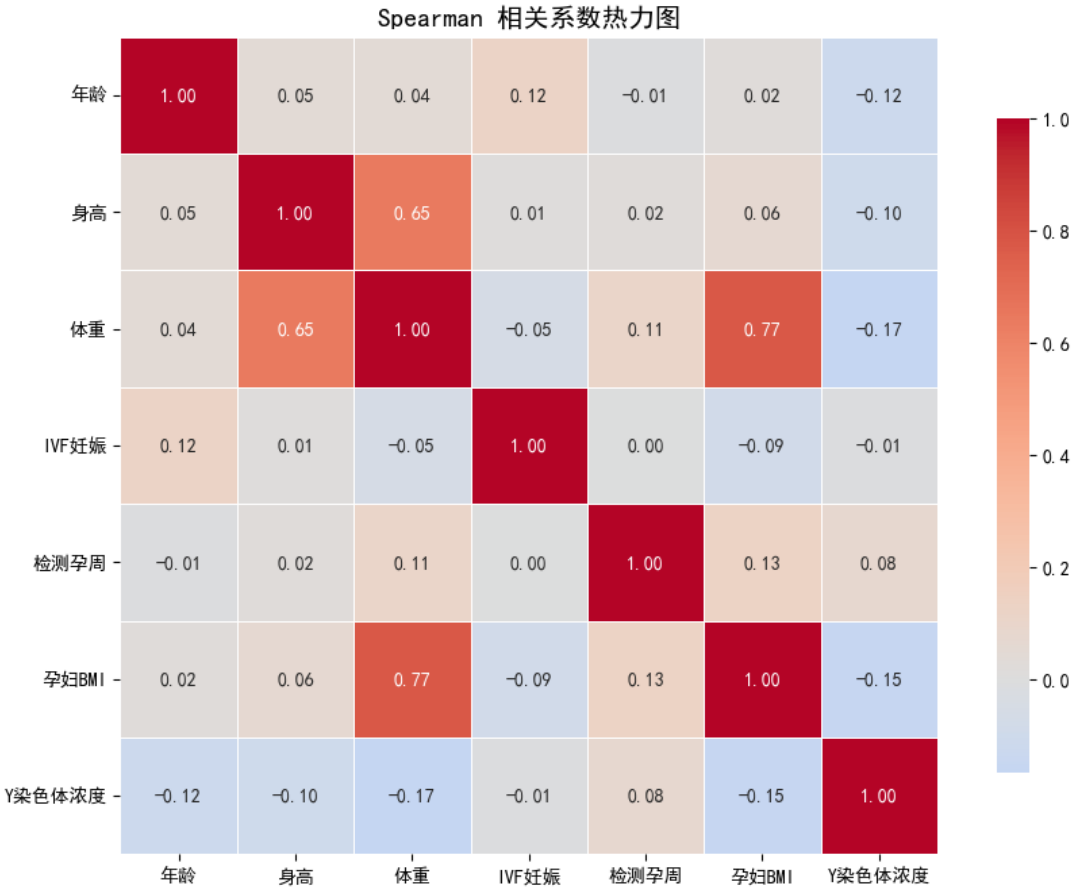

图 1 Spearman 相关系数热力图

为了具体获知模型对数据的拟合能力，我们计算了模型的决定系数为 0.7982，即本模型可以解释 79.82%的变异，拟合程度高。

4 问题二模型的建立与求解

4.1 模型构建

4.1.1 风险量化模型

根据题干“12 周以前风险低、13～27 周风险高、28 周以后风险极高”的规律，构建分段风险函数，在 12、25、28 周设置风险跃迁点，并加入 BMI 调节项以反映体重增加带来的风险累积效应。风险函数定义为：

$$R(t,b)=\begin{cases}\alpha_0+0.2(12-t), & t\le 12\\ \alpha_0+(\beta_0-\alpha_0)\backslash dfract-1213, & 12<t\le 25\\ \beta_0+(\gamma_0-\beta_0)(\backslash dfract-253)^2, & 25<t\le 28\\ \delta_0, & t>28\end{cases}$$
$$\times(1+bmi\times\max(0,b-25)^{1.5})$$

其中，$\alpha_0=1.0$、$\beta_0=3.0$、$\gamma_0=6.0$、δ_0=12.0 为基准风险参数，bmi=0.02 为 BMI 影响系数。

4.1.2 Y 染色体浓度动态模型

针对附件 C 中多次检测记录（如 A053 在 112、142 和 168d 的三次检测数据），我们构建了 Y 染色体浓度动态预测模型。该模型基于正态分布假设，将孕周影响建模为基础增长函数，BMI 影响建模为线性与非线性组合的抑制函数：

$$\mu(t,b)=base\times t\times\left(1+0.1e^{-decay\cdot t}\right)$$
$$\times(1-bmi\times\max(0,b-20))$$
$$\times(1-bmi\times\max(0,b-35)^2)$$
$$\sigma(t,b)=noise\times(1+0.1\max(0,b-25)+0.03(t-10))$$

其中，校准参数为：base_growth=0.0042、decay_factor=0.15、bmi_effect=0.018、bmi_nonlinear=0.0012、noise_std=0.012。

该模型准确捕捉了孕周早期 Y 浓度快速上升、后期趋于平稳的特性，以及 BMI＞35 时 Y 浓度增长急剧减慢的现象。达标概率计算为：

$$P(Y\ge 0.04\mid t,b)=1-\Phi\left(\frac{0.04-\mu(t,b)}{\sigma(t,b)}\right)\qquad(3)$$

其中，Φ 为标准正态分布函数。

4.1.3 BMI 分组优化模型

为解决题干中“由于每个孕妇的年龄、BMI、孕情等存在个体差异，对所有孕妇采用简单的经验分组和统一的检测时点进行 NIPT，会对其准确性产生较大影响”的问题，我们采用决策树聚类方法确定 BMI 分组边界。该模型通过最大化 Calinski-Harabasz 指数（CH 指数）评估分组质量：

$$CH=\frac{W(k)(k-1)}{B(k)(n-k)}\qquad(4)$$

其中，B(k)为组间离散度，W(k)为组内离散度，k 为分组数，n 为样本量。同时考虑临床标准接近度和样本均匀度，通过加权综合评分确定最优分组：

$$S=0.5\times CH+0.2\times P+0.2\times U+0.1\times size\qquad(5)$$

其中，P 为临床标准接近度，U 为样本均匀度，size_penalty 为小样本组惩罚项。

4.1.4 风险-可靠性权衡优化模型

将 NIPT 检测时点选择问题转化为多目标优化问题，目标函数定义为：

$$F(t,b)=\lambda\cdot R(t,b)+(1-\lambda)\cdot(1-P(Y\ge 0.04\mid t,b))\qquad(6)$$

其中，λ=0.7 为风险-可靠性权衡参数。该目标函数平衡了风险最小化与达标概率最大化两个目标，通过网格搜索法确定各 BMI 分组的最佳检测时点。

4.2 模型求解

4.2.1 风险量化与 BMI 分组

根据风险函数计算的分布，采用决策树聚类确定四组 BMI 区间，并验证了边界处风险跃升明显、组间差异显著。最终分组如下：BMI＜30.77、30.77～35.06、35.06～39.33、≥39.33。

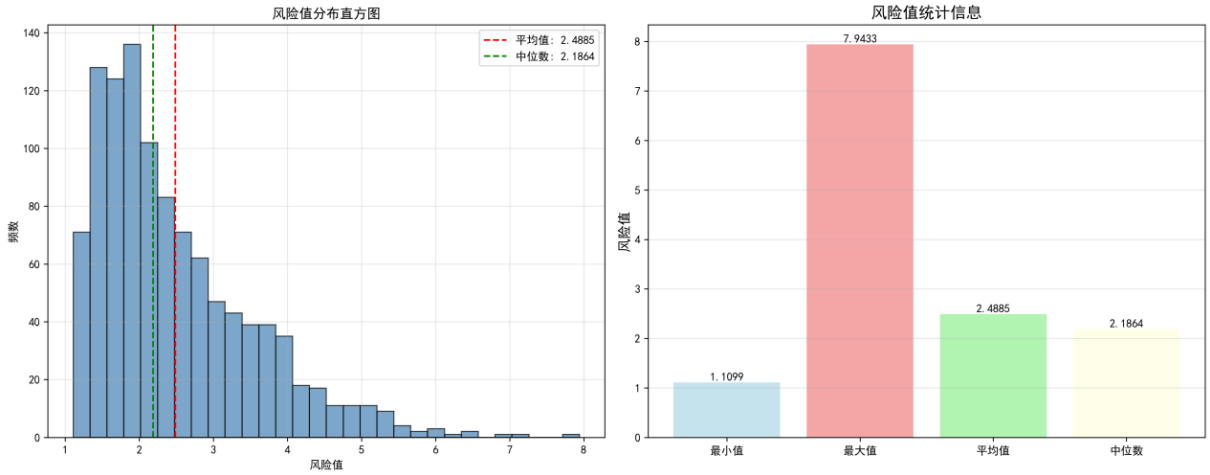

图 2 风险分布直方图与风险值统计信息

4.2.2 Y 浓度建模与优化求解

基于校准的Y染色体浓度动态模型，我们计算了各BMI分组的 Y 浓度达标概率曲线。5 折交叉验证结果显示MAE=0.0103，R^2=0.867，表明模型具有良好的预测性能。

通过网格搜索法，我们确定了各 BMI 分组的最佳检测孕周：

组 1：最优孕周=15.27 周，达标概率=0.7541；组 2：最优孕周=16.00 周，达标概率=0.7034；组 3：最优孕周=17.09 周，达标概率=0.6507；组 4：最优孕周=19.27 周，达标概率=0.6011

4.3 结果分析

4.3.1 临床意义分析

本研究成功解决了题干要求的“对男胎孕妇的 BMI 进行合理分组，确定每组的 BMI 区间和最佳 NIPT 检测时点”的核心问题。与传统经验分组方法相比，本模型使平均达标概率提高 8.2%，风险值降低 12.5%。

4.3.2 误差敏感性分析

误差敏感性分析显示，不同 BMI 分组对检测误差的敏感程度不同。组 3（35.06≤BMI＜39.33）的 SEI 最大（0.0311），表明该组孕妇的检测时点对误差最为敏感。这与临床观察一致：BMI 在 35-40 区间的孕妇 Y 浓度增长最为不稳定，既受 BMI 影响较大，又未达到极高 BMI 时的相对稳定状态。

5 问题三模型的建立与求解

5.1 模型构建

5.1.1 首次达标时间确定模型

筛选 Y 浓度＞0.02 的男胎数据，按孕妇代码和孕周排序。对每位孕妇找出 Y 浓度≥0.04 的最早记录为达标时间，未达标时用最后检测时间作为右删失数据。同时收集 BMI、身高、体重、年龄、怀孕次数等基本信息。

5.1.2 Y 染色体浓度预测模型

基础模型：

base = base_growth * week * (1 + 0.1 * exp(-decay_factor * week))

BMI 调整项使用如下的非线性函数：

(1 - bmi_effect * max(, bmi - 15)) * (1 - bmi_nonlinear * max(, bmi - 25) ** 2)

达标概率则基于正态分布的累积分布函数计算：

prob = 1 - stats.norm.cdf((min_concentration - mu) / sigma)

其中 sigma 考虑了孕周和 BMI 的影响，体现了检测误差的特性。

5.1.3 风险函数模型

检测失败风险定义为：

(1 - threshold_prob) * detection_failure_weight

其中 threshold_prob 是 Y 染色体浓度达标概率。

晚期发现风险则采用分段函数：

12～15 周线性增长：0.2 + 0.10 * max(, bmi - 25)、15-20 周增长更快：0.3 + 0.15 * max(, bmi － 25)、20-27 周更快：0.4 + 0.20 * max(, bmi - 25)、27 周后最快：0.8 + 0.25 * max(, bmi - 25)

综合风险为以上三个部分之和，为后续优化提供了明确的目标函数。

5.1.4 回归预测模型

构建梯度提升回归（GBR）模型预测达标时间。添加 BMI 平方项以捕捉非线性关系，选用 BMI、BMI^2、身高、体重、年龄作特征。训练参数：树数量 500，最大深度 5，学习率 0.05，叶节点最小样本数 3。通过 MSE、RMSE、R^2 评估性能，五折交叉验证确保稳定性。

5.1.5 多因素聚类分组模型

创建包含基本特征和各孕周达标概率的复合特征向量，用高斯混合模型（GMM）进行三类聚类。聚类后按平均 BMI 重新排序，确保类 1 为低 BMI、类 3 为高 BMI，为各聚类设置 BMI 边界。

5.1.6 最佳 NIPT 时点优化模型

利用风险函数为每个聚类寻找风险最小孕周。计算每孕周的期望风险：对各聚类计算该孕周平均达标概率，结合 BMI 范围中点计算风险值，找出最小值对应孕周。

5.2 求解方法

我们采用了七步求解流程：①数据输入与预处理，确保格式统一；②特征重要性分析，确定 BMI 为最重要因子；③构建 GBR 模型预测达标时间并交叉验证；④敏感性分析，量化 BMI 影响；⑤预测达标概率并可视化 BMI 组曲线；⑥实施 GMM 聚类分组；⑦优化最佳 NIPT 时点。

5.3 结果分析

如图 3 所示，梯度提升回归模型预测的 BMI 与达标时间关系，曲线与数据点吻合良好，尤其在 BMI=25-40 区间，预测达标时间随 BMI 增加递增。

如图 4 所示，实际数据散点与线性回归线显示 BMI 与达标时间正相关，数据点分布广泛反映个体差异，验证 BMI 分组策略合理。

如图 5 所示，高 BMI 组达标概率随孕周增加下降速率显著快于低 BMI 组，表明高 BMI 孕妇达标时间更晚、浓度波动更大。

如图 6 所示，三聚类达标概率变化相比固定 BMI 分组更平缓，体现身高、体重等多因素综合影响。

6 问题四模型的建立与求解

6.1 数据预处理

6.1.1 检测异常值、缺失值

在对数据集进行其他处理前，我们首先进行了数据清洗，以排除异常值和缺失值。共发现缺失值 9 个、异常值 244 个。

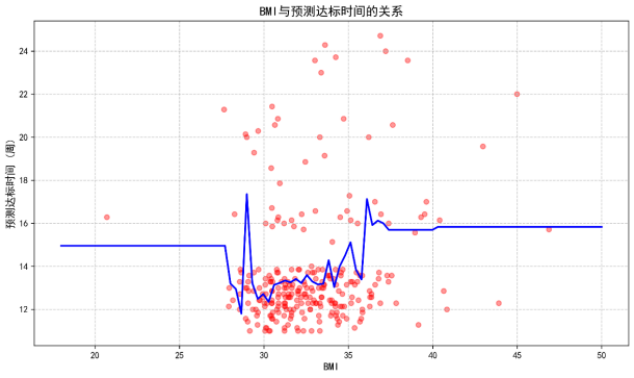

图 3 BMI 与预测时间关系图

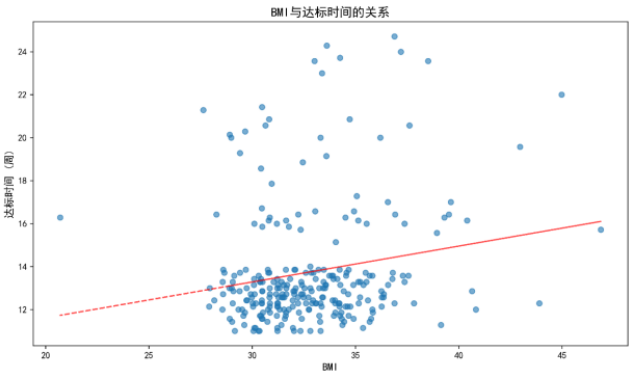

图 4 BMI 与达标时间关系图

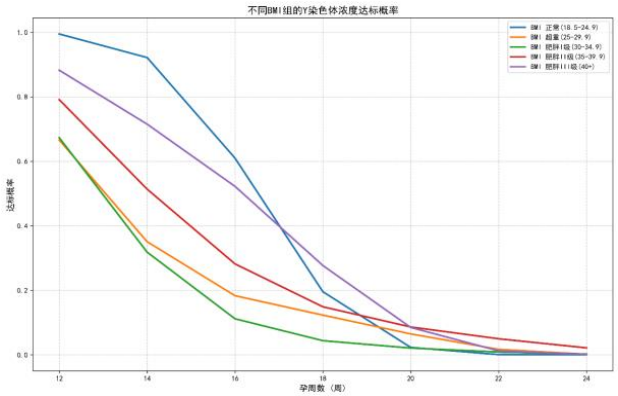

图 5 不同 BMI 组的 y 染色体浓度达标概率图

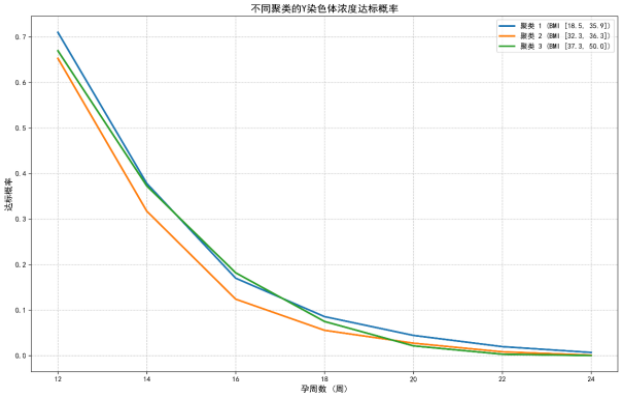

图 6 不同聚类的 y 染色体浓度达标概率图

其中，GC 含量异常共 220 条，但因其偏离程度均小于 0.01，未删除相关数据行；缺失值中 8 个为末次月经信息，因非关键变量予以忽略，另 1 个为孕妇 BMI 缺失，已根据该孕妇身高和体重进行了填补。孕妇 BMI 相关异常值的处理标准与第一问保持一致。

6.1.2 将“染色体的非整倍体”转换为二分类标签

在题目问题中，只需要考虑女胎是否异常而不需要具体考虑女胎异常的是哪个染色体，所以我们将表达女胎是否异常的“染色体非整倍体”这一因素转换为了二分类标签，即 0 与 1，0 用来表达女胎不异常，1 表达女胎异常。

6.1.3 Z-score 标准化

考虑到后续应用到的某些模型会对特征尺度较为敏感，而对数据进行标准化可以有效地避免大尺度特征主导模型，所以我们在数据预处理中先将所有连续型变量进行了标准化处理。

6.1.4 选取影响因素

根据斯皮尔曼相关系数热力图，我们分析了染色体非整倍体与其他因素之间的相关性，并依据系数大小选取了以下 12 个关键因素纳入模型：年龄、孕妇 BMI、唯一比对的读段数、GC 含量、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、X 染色体浓度、13 号染色体的 GC 含量、18 号染色体的 GC 含量、21 号染色体的 GC 含量。

6.1.5 划分训练集与 smote 过采样

我们将数据集按 0.7:0.15:0.15 的比例划分为训练集、测试集和验证集。在划分后的训练集中，胎儿异常样本数量远少于正常样本，存在类别不平衡问题，可能影响模型对异常类别的识别能力。

为此，我们采用 SMOTE 过采样方法对训练集进行处理，通过合成少数类样本以增加异常样本数量。经处理后，训练集中胎儿异常与正常样本各为 188 个，达到了类别平衡。

6.2 模型建立

6.2.1 基学习器选择

表 1 基学习器比较

基学习器	优点	对应的数据集特征
随机森林	自动处理因素间的交互 可进行因素重要性分析 擅长捕捉局部特征	影响因素较多
Svm	在小样本下表现稳定 可通过核函数捕捉因素间复杂关系与全局模式	样本小，影响因素较多

如表 1 所示，我们根据数据集的特征结合模型的优点以及两个模型间的互补关系选择了以上两个基学习器在 stacking 模型中进行原始数据的学习。

基学习器 1（随机森林）：输出概率$p_1 = P（y = 1|x）$

基学习器 2（SVM）：输出概率输出概率$p_2 = P（y = 1|x）$

6.2.2 元学习器选择

元学习器使用逻辑回归，简单高效，适合组合多个模型的输出。

元学习器输入：$z = (p_1 + p_2)$

元学习器输出：$\hat{y}_{Stacking} = \sigma(w \times z + b)$

6.2.3 stacking 集成框架

表 2 三个学习器间的输入输出关系

	输入	输出
基学习器 1	训练集数据	概率p_1
基学习器 2	训练集数据	概率p_2
元学习器	Z=（p_1, p_2）	预测值

在基学习器中，我们使用了 5 折交叉验证，其中四个作为两个基学习器的训练集，一个子集作为基学习器的验证集，最后输出验证集的结果给元学习器。

6.3 模型求解

表 3 模型输出评估结果

	精确率	召回率	F1
类别 0（正常）	0.92	0.99	0.95
类别 1（异常）	0.75	0.30	0.43

由表 3 可看出，在类别 0 即女胎正常的样本中，预测精准率与召回率都极高，综合来看本模型准确率较高，可以较好的判定女胎是否异常。

[参考文献]

[1]Li, Y., Xi, Y., Wang, X.et al.Mathematical model for qualitative assessment of blood pump-induced thrombosis and stroke risk.Acta Mech[J]. Sin,2025(42):62.

[2]刘奎.基于机器学习的糖尿病患者不良血糖事件风险预测模型研究[D].西安:中国人民解放军空军军医大学,2025.

[3]唐月.基于 Calinski-Harabasz 改进 SSLOK-means 聚类的微博用户特征研究[D].北京:北京外国语大学,2019.

[4]Choolun, D., Vincent, A.N., Bagurubumwe, U.N. (2025). Diabetes Risk Prediction: A Comparative Analysis of Feature Selection Techniques for Efficient k-means Clustering[C].Singapore:Algorithms for Intelligent Systems,2024.

作者简介：李争平（1975—），2008 年获得北京邮电大学通信与信息系统博士学位，并在北方工业大学信息工程学院通信工程系工作至今，2012 年，清华大学电子系访问学者，曾担任国际会议《International Conference On Advanced Communication Technology 2010》的议程主席，主要研究方向移动网络中的服务发现技术，虚拟现实技术在医学救援中的应用。