2021 年（第七届）全国大学生统计建模大赛

基于 Binary Logistic 回归模型和决策树模型对早产危险因素的探究和预测

参赛单位：中南大学

吴怡希

参赛人姓名：徐业

谭涛

目录

摘要

近年来，全球早产率总体呈上升趋势，在我国，早产儿以每年 20 万的数目逐年递增，目前早产已经成为重大的公共卫生问题之一。据研究,早产是威胁胎儿及新生儿健康的重要因素，可能会造成死亡或智力体力缺陷，因此研究早产的影响因素，建立预测早产的模型就显得极为重要。我们以问卷、面对面访谈的方式，收录了湖南省妇幼保健院 2013 年 5 月 13 日-2019 年 12 月 31 日妊娠 8-14 周且接受首次产前护理的孕妇，共 18527 份样本，调查研究孕妇包括医学和社会学信息在内的 104 个变量。基于大样本、多变量的数据特征，对数据预处理后，首先基于传统的统计方法，依次通过 SMOTE 过采样均衡数据、$\chi 2$ 相似性检验剔除无关变量、二阶聚类（TwoStep Cluster）实现降维，用 Binary Logistic 建立早产预测模型，并通过 AUC-ROC 曲线对早产预测模型进行准确性检验；在此基础上，进一步探讨并合理利用机器学习的效力，用数据挖掘的方法，依次通过随机欠抽样平衡样本，特征选择变量实现变量降维，分别用决策树 C5.0 算法，推理集 C5.0 算法，决策树 CHAID 算法建立早产预测模型，并通过 boosting 技术提高模型稳健性。

根据二阶聚类降维结果、Binary Logistic 建立的早产预测模型及检验结果，发现城乡分组、人均月收入、母亲孕前 BMI 分组、受精方式、受孕方式、孕次分组、孕早期柯萨奇病毒、孕前既往性病史、是否采用剖宫产、配偶 BMI 分组这 10 个变量与是否早产的相关性较强，且在经过哑变量处理后，适用于建立早产预测模型。通过 AUC-ROC 曲线，检验出该早产预测模型拟合度良好。在初步探索之后，进一步深入利用机器学习，即分别使用决策树 C5.0 算法，推理集 C5.0 算法，决策树 CHAID 算法建立三个早产预测模型。其中通过决策树 C5.0 算法建立的早产预测模型，在测试集上的准确性为 93.78%，平均正确性为 0.859、平均不正确性为 0.692；推理集 C5.0 算法的准确性为 95.92%，平均正确性为 0.824、平均不正

确性为 0.714；决策树 CHAID 算法建立的早产预测模型，在测试集上的准确性为 79.58%，取置信度为 0.812。

In recent years, the global preterm birth rate is on the rise. In China, preterm birth is increasing by 200,000 per year. At present, preterm birth has become one of the major public health problems. According to research, preterm birth is an important factor threatening the health of fetus and newborn, and may cause death or mental and physical defects. Therefore, it is extremely important to study the influencing factors of preterm birth and establish a model to predict preterm birth. We collected a total of 18,527 pregnant women who received their first prenatal care at 8-14 weeks of gestation from May 13, 2013 to December 31, 2019 in Hunan Maternal and Child Health Hospital by means of questionnaires and face-to-face interviews, and investigated 104 variables including medical and sociological information of pregnant women. Based on the characteristics of large sample and multi-variable data, after data pretreatment, first based on traditional statistical method, eliminate irrelevant variables by SMOTE oversampling equilibrium data, c2 similarity test and TwoStep Cluster to achieve dimension reduction, use Binary Logistic to establish the prediction model of premature birth. The accuracy of the prediction model was tested by AUC-ROC curve. On this basis, further explore and make reasonable use of the effectiveness of machine learning, use data mining method, balance samples by random under sampling, feature selection variables to achieve dimensionality reduction of variables, establish preterm labor prediction model with decision tree C5.0 algorithm, inference set C5.0 algorithm, decision tree CHAID algorithm respectively. Boosting technology is used to improve the robustness of the model.

According to the TwoStep Cluster dimensionality reduction results, the premature birth prediction model established by Binary Logistic and the test results, it was found that 10 variables, such as urban-rural grouping, per capita monthly income, pre-pregnancy BMI grouping of mothers, fertilization mode, conception mode, pregnancy times grouping, coxsackie virus in early pregnancy, previous medical history before pregnancy, cesarean section and spouse BMI grouping, had strong correlation with premature birth, and were

suitable for establishing the premature birth prediction model after being processed by dummy variables. Through AUC-ROC curve, it is verified that the prediction model of premature delivery has good fitting degree. After the initial exploration, machine learning is further used, that is, the decision tree C5.0 algorithm, reasoning set C5.0 algorithm and decision tree CHAID algorithm are used to establish three premature birth prediction models. Among them, the premature birth prediction model established by decision tree C5.0 algorithm has an accuracy of 93.78%, an average accuracy of 0.859 and an average inaccuracy of 0.692 on the test set; The accuracy of reasoning set C5.0 algorithm is 95.92%, the average correctness is 0.824, and the average inaccuracy is 0.714; The accuracy of the premature birth prediction model established by decision tree CHAID algorithm is 79.58%, and the confidence level is 0.812.

关键词：早产预测 Binary Logistic 模型 决策树

第一章 简介

（一）早产的研究背景与意义

《早产临床诊断与治疗指南（2014）》中将早产定义为内容妊娠满 28 周但不足 37 周的分娩。随着社会的发展，在外界各种不良因素的影响下，早产率总体呈现出上升的趋势[1]。2012 年 5 月的《全球早产儿报告》显示，全球每年新增早产儿近 1500 万，占全球活产儿的 11.1%，早产发生率大于 10%，每年死亡早产儿 110 万，占新生儿死亡的 36%。我国每年将近有 2000 多万名新生儿，其中早产儿有 200 万余名，且早产儿的数量在以每年 20 万的数量递增。现早产已成为全球重大的公共卫生问题之一。

早产被认为是自身、社会、环境等多因素相互作用的结果。何丽芸等人利用 Logistic 回归分析上海市 20 家医院的产妇病案，认为分娩前产次≥2、早产史、孕前糖尿病、孕前肾脏疾病、双胎、妊娠糖尿病、前置胎盘、胎盘早剥、胎盘早破、胎粪污染、先天畸形等均为早产的危险因素[2]。陈丽等人对广东省深圳市妇幼保健院 2018 年 9 月至 2019 年 11 月收治的 160 例早产病例进行多因素 Logistic 回归分析，发现胎膜早破、妊娠高血压、前置胎盘为导致早产的常见因素[3]。虢海月等人在研究中发现，围孕期饮酒、放射性检查、职业性危险因素暴露均可增加新生儿早产发生的风险[4]。导致早产的因素众多，建立相关数学模型预测早产发生的可能概率具有现实性意义。

据统计，早产是威胁胎儿及新生儿健康的重要因素[5]。早产儿各器官系统发育不成熟，生命力低下，可能存在体格或智力方面的缺陷，出现一系列的并发症。相关研究表明，约 15.36%早产儿出生一个月内死亡，存活的早产儿约 25.63%有发育和智力的障碍[6]，这将给家庭和社会带来精神与经济的巨大负担。因此，利

用统计学方法构建早产预测模型，识别早产危险因素的孕妇，为降低早产的发生、制定相关措施及政策提供理论依据，以便及时干预治疗显得尤为重要。

（二）数据来源及处理

本研究以问卷调查、面对面访谈的方式，在湖南省妇幼保健院选取 2013 年 5 月 13 日-2019 年 12 月 31 日妊娠 8-14 周且接受首次产前护理的孕妇为研究对象，共计 18527 例，从孕妇基本人口学特征、孕妇孕前和孕早期生活行为特征、产科生育史、个人疾病史和家族疾病史、孕早期患病情况、孕前或孕早期用药情况、孕前或孕早期环境有害物质暴露、孕期营养监测、生育因素调查和配偶社会人口学特征这 10 个方面出发，共研究不包括早产在内的 103 个变量。据世界卫生组织（WHO）评估，每 7 对夫妇中约有 1 对夫妇存在生殖障碍。我国近期调查，国内不孕症者占已婚夫妇人数的 10%，比 1984 年调查的 4.8%增加一倍多，发病率呈上升趋势。我国更受传宗接代观念影响，多数家庭盼子心切，使不育夫妇承受着极大的心理压力，甚至引发离异、婚外恋之类家庭乃至社会的问题。同时据临床统计，不育患者中约 20%的夫妇，不借助 ART（辅助生殖）根本无法生育。为了更加贴合现代社会的改变和医学技术的发展，本次研究对辅助生殖受孕的样本同样进行了保留。

本次研究基于该数据，通过机器学习，分析诱发早产的危险因素，并基于 Binary Logistic 方法和决策树分别建立早产预测模型。在建立最终的早产预测模型之前，为解决类别不均衡问题，分别采用 SMOTE 过采样算法和欠采样，在此基础上，Binary Logistic 模型用到 $\chi 2$ 相似性检和二阶聚类降维；决策树模型用到分区和特征选择处理变量。

（三）文章结构分布

第二章为数据变量的预处理。第三章结合现代医学统计的研究成果，先通过SMOTE 采样均衡数据，再运用 χ2 相似性检验剔除相关性较小的变量、二阶聚类模型（TwoStep Cluster）降维，然后建立 Binary Logistic 模型，最后通过 AUC-ROC 曲线进行模型检验。第四章综合利用数据挖掘的方法，首先随机欠抽样平衡样本，特征选择变量实现变量降维，然后分别采用决策树 C5.0 算法，推理集 C5.0 算法，决策树 CHAID 算法并使用 boosting 技术提高模型稳健性，最后将全体样本实际输出值与在模型下的输出值比较得到混淆矩阵，同时得到综合正确率和错误率统计，得到估计的准确性，通过随机抽取的 30%的测试集样本再由拉普拉斯估计器计算的到相应的置信区间，由此也可以比较三个模型，并根据实际情况的具体要求选择一个进行预测的运用。第五章为本次建模总结和展望。第六章致谢。

具体流程绘制如下图：

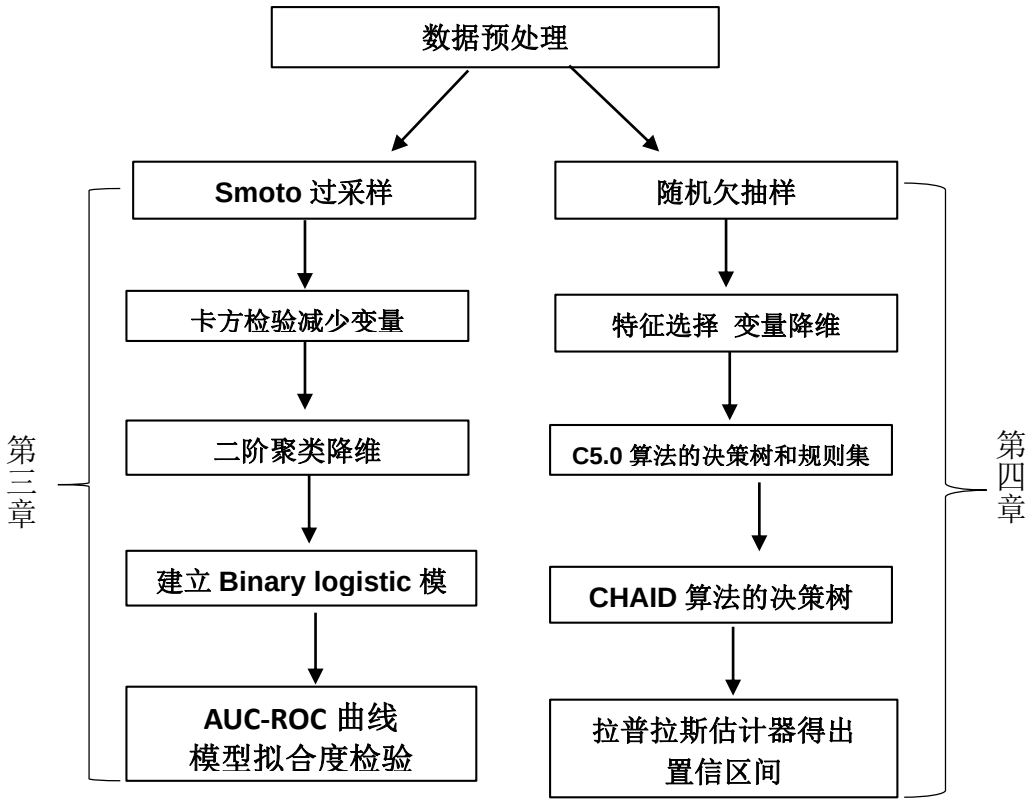

图 1　研究思路

第二章 数据预处理

（一）变量预处理

类别化处理及选择：将品质变量整理成 0-1 型数值变量，如民族；对于连续变量和其他可合并的变量进行整合，这样会得到有重复信息的变量，比如配偶 BMI 值和配偶 BMI 值分组，受孕方式和受孕方式两分类。不做特别说明的情况下，本次研究将主要使用分类型变量，且选择使用分类型变量中分组较少的那一个，比如刚刚提到两组变量，均选择后一组变量进入样本。这是因为在本次研究中，分类型变量占绝大多数，而相同的数据类型有更方便建模的处理，投入到未来实际预测操作中也更加简单明了。

（二）样本处理

类别不平衡（class-imbalance）：指分类任务中不同类别的训练样例数目差别很大的情况。在分类学习中方法，默认不同类别的训练样例数目基本相当。若样本类别数目差别很大，属于极端不均衡，会对学习过程（模型训练）造成困扰。这些学习算法的设计背后隐含的优化目标是数据集上的分类准确度，而这会导致学习算法在不平衡数据上更偏向于含更多样本的多数类。多数不平衡学习（imbalance learning）算法就是为解决这种“对多数类的偏好”而提出的。据实践经验表明，正负类样本类别不平衡比例超过 4:1 时，分类要求会因为数据不平衡而无法得到满足，分类器处理结果将变差，导致预测效果达不到预期要求。在本次研究项目中，早产 0:1 比约为 5:1（0 为不发生，1 为发生。本论文其他部分未做其他说明时，都按照该标签规则），因此在构建模型之前，需要对该分类不均衡性问题进行处理。

本文第三章将以 SMOTE 过采样开始进行研究，第四章以随机欠采样进行研究。

第三章 基于Binary Logistic建立早产预测模型

（一） SMOTE 算法原理

Synthetic Minority Oversampling Technique,首字母缩写即为 SMOTE。它是一种基于随机过采样算法的增加少数类样本的技术。是对少数类样本进行分析，提取出少数类样本的特征，基于这些特征，“绘制”出属于少数类的新样本，将其添加到原始数据集中。算法本身是基于“插值”来合成少数类样本，到达样本类别均衡的目的。

3.1.1、SMOTE算法流程

如果训练集中最小类所占样本数量为 M，则 SMOTE 算法将为最小类样本计算出一个新图像，图像中产生 NM 个新样本，N 取正整数。对于 N<1 的情况，SMOTE 默认少数类样本数为 M=NM，这里的 N 取 1。

步骤 1、对特征向量为 $Xi, i \in \{1,\ldots, M\}$的最小类的某一个样本 P，从它的 M 个样本中绘制出 Xi 的 k 个最近邻，记为 $Xi(near), near \in \{1,\ldots,K\}$；

步骤 2、从 $Xi(near), near \in \{1,\ldots,K\}$ 中随机选择一个样本，记为 Xi(nn)，新样本 Xi1 的生成公式为：$Xi1 = Xi + \zeta 1 \cdot (Xi(nn) - Xi)$，$\zeta 1$ 为 0-1 中随机生成的数值。

步骤 3、多次利用步骤 2，直到合成 N 个新样本：$Xinew, new \in 1,\ldots, N$，如果将上述处理过程运用到所有 M 个样本中，将最终合成 NM 个新样本[7]。

3.1.2、SMOTE过采样模型效果

调用 Python 里的 SMOTE，同时删除原品质型变量（民族，配偶民族等。数

据前期处理中，已将品质型变量总结标记为数值型变量）和缺失变量（配偶职业分组，不孕症类型）。对原有样本数据进行过采样，使得原有数据比，从 15493:3034 调整为 1:1。样本量增加到 37054 个。同时变量总数精简到 118 个，包括因变量。

附录（一）：Python 实现 SMOTE 过采样代码及输出文件截图

（二）χ2 检验

原理：在研究样本中实际测量值与统计输出的理论值之间存在一定差异，将这种差异叫做样本统计量的偏离程度。χ2 值的大小就取决于偏离程度。χ2 值越大，偏离程度越大；否则，差异越小，相似性越强，当实际测量值与统计输出的理论值完全相等时，χ2=0，两者完全符合[8]。

3.2.1、χ2基本公式：

$$\chi 2=\sum_{i=1}^{k}\frac{(fi-nPi)}{nPi}$$

3.2.2、数值说明

χ2 值：越大，说明“X 与 Y 有关系”成立的可能性越大。

p 值：统计有效性是对结果真实程度的推测方法。专业来说，P 值是结果可靠性的降低指标，P 值是观测值有效的整体代表性的错误概率。它的值越接近于 0，相关性越显著。

3.2.3、χ2筛选成果

经处理后，本次研究数据全为分类变量，且样本量 n>40,适合采用卡方检验研究自变量与早产的关系。因为样本量足够大，且避免过度剔除具有相关性的变量。最终以 P<0.01 为检验标准，筛选出与早产相关性较强的 77 个变量。

附录（二）χ2 相关性研究及剔除无关变量结果

（三）二阶聚类（TwoStep Cluster）

3.3.1、实现步骤

步骤 1、建立树根 clusterfeature,树根在一开始每个节点中会放置一个数据集中的第一个记录，它就包含有这个数据存储集中每个变量的信息。相似性用的是距离数值测量，数据的相似性可以作为进行距离数值测量的主要标准。相似度高的变量位于同一节点，同时，相似度低的变量生成新节点。似然归类测度模型假设每个变量必须服从特定的概率分布，聚类模型要求分类型独立变量必须服从多项式概率分布，数值型独立变量必须服从正态概率分布。

步骤 2、合并聚类算法。生成的聚类方案具有不同聚类数，不同的聚类数是基于合并聚类算法下节点的组合成果。

步骤 3、选择最优聚类数。通过 BIC：Bayesian Information Criterion 准则对各聚类情况进行比较，选出最优聚类方案。

3.3.2、数值说明

①对数似然：这种度量方式用于研究某种以确定概率分布的独立变量。其中数值型变量服从正态分布，分类型变量服从多项式分布。

②Bayesian 信息准则(BIC)：在只有部分信息时，要预测未知状态下的部分信息值，选用主观概率；修正发生概率时采用贝叶斯公式，将得到的修正概率与预期产出的值结合计算出最优决策。

计算公式：

$$BIC=\ln(n)k-2\ln(L)$$

其中：k 为模型参数个数；n 为样本数量；L 为似然函数

3.3.3、聚类结果

二阶聚类适用于多分类变量的降维问题。显然，本次研究数据可选用 SPSS 中

的二阶聚类对变量进行降维，聚类效果为良好，并最终由 77 个自变量降维到 14 个主要变量(该 14 个变量重要性都为 1)

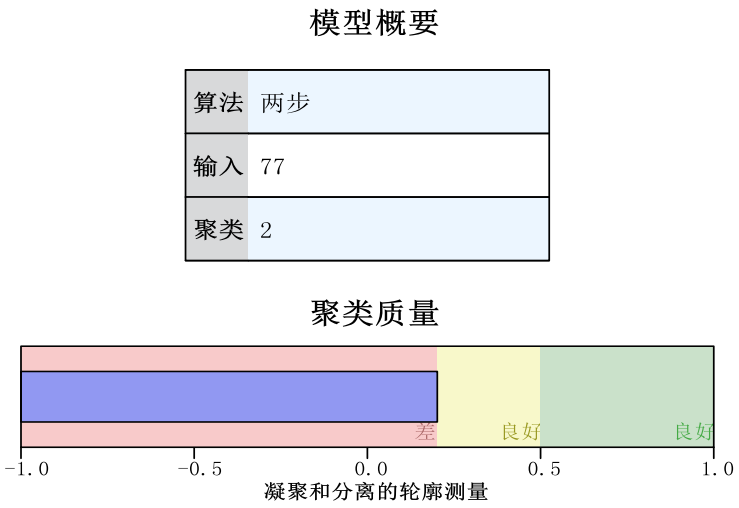

图 2　二阶聚类质量展示（良好）

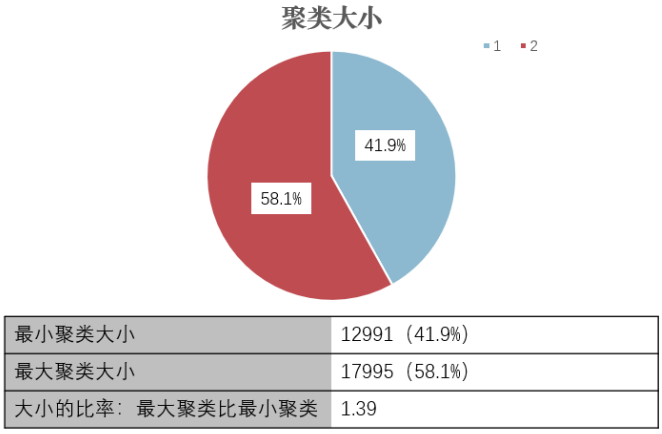

图 3　两步聚类大小比较

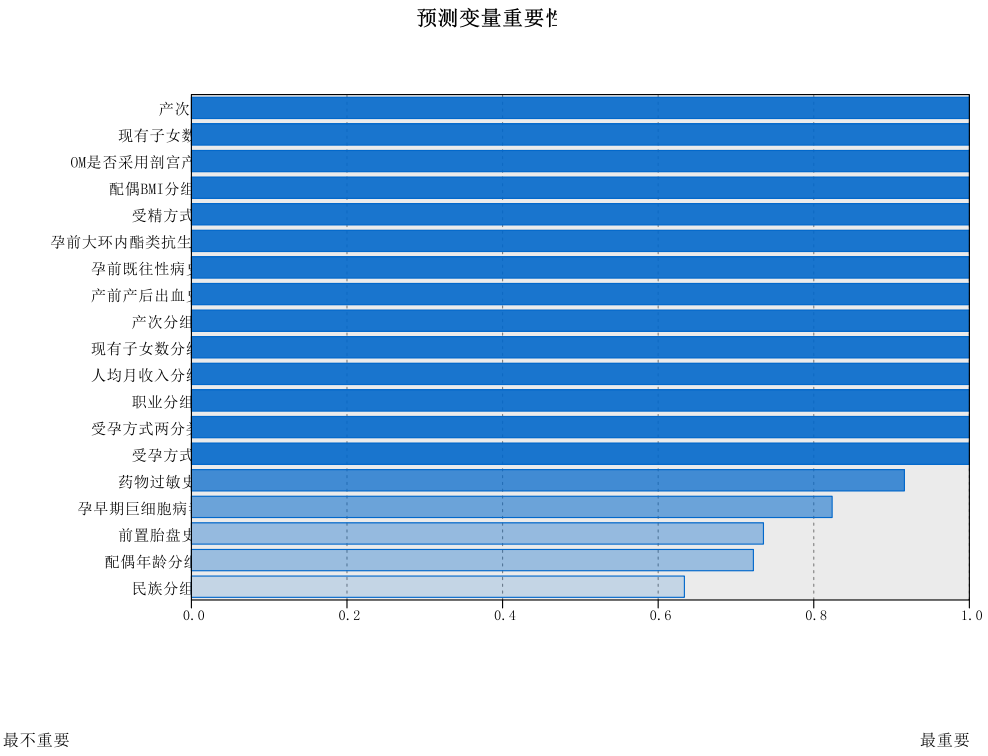

图 4　预测变量重要性

（四）Binary　Logistic

3.4.1、Logistic模型介绍

属于广义线性回归分析模型。常用于预测经济走向、数据挖掘、农业生产和疾病诊断等。基于这一特性，Logistic 模型可用于对引发早产的因素进行探讨，并通过建立相关线性回归分析模型预测早产发生的概率。以早产病情分析为例，选择两组人群，一组是早产组，一组是不早产组，两组人群在个人和周边环境中必定具有不同的特征。因变量就为是否早产，值为 1:“是”或 0:“否”，自变量根据科学研究分析可以大致确定某些变量具有相关性，纳入研究范围。建立 logistic 回归分析，得到自变量的系数，从而可以基本确定诱发早产的危险因素。同时根据该预测模型可以根据危险因素的暴露情况对是否早产做出预测。

3.4.2、logistic回归原理解释及最终输出结果样式

logistic 回归原理解释

Sigmoid 函数：

$$g(z)=\frac{1}{1+e^{-z}}$$

$$h_\theta(x)=g\left(\theta^T x\right)=\frac{1}{1+e^{-\theta^T x}}$$

$$g'(x)=\left(\frac{1}{1+e^{-x}}\right)'=\frac{e^{-x}}{\left(1+e^{-x}\right)^2}$$

$$=\frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}}=\frac{1}{1+e^{-x}}\cdot\left(1-\frac{1}{1+e^{-x}}\right)$$

$$=g(x)\cdot(1-g(x))$$

最大似然估：随机数据点被正确分类的概率最大化

为解决二分类问题，其实也就是概率的问题，分类本质上是概率问题

假定：

y=1 和 y=0 的时候的概率分别为：

$$P(y=1|\ x;\theta)=h_\theta(x)$$
$$P(y=0|\ x;\theta)=1-h_\theta(x)$$

得到：

$$p(y|\ x;\theta)=\left(h_\theta(x)\right)^y\left(1-h_\theta(x)\right)^{1-y}$$

似然函数：

概率相乘，两边同时取对数

$$L(\theta)=p(\vec{y}|\ X;\theta)=\prod_{i=1}^{m}p\left(y^{(i)}|\ x^{(i)};\theta\right)=\prod_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)\right)^{y^{(i)}}\left(1-h_\theta\left(x^{(i)}\right)\right)^{1-y^{(i)}}$$

$$l(\theta)=\log L(\theta)=\sum_{i=1}^{m}y^{(i)}\log h\left(x^{(i)}\right)+\left(1-y^{(i)}\right)\log\left(1-h\left(x^{(i)}\right)\right)$$

对数似然函数，求导，得到 θ 的梯度

$$\frac{\partial(\theta)}{\partial\theta_j}=\sum_{i=1}^{m\left(x^{ij}\right.}-\frac{1-y^{in}}{1-h\left(x^{(i)}\right)}\Bigg)\cdot\frac{\partial h\left(x^{(i)}\right)}{\partial\theta_j}$$

$$=\sum_{i=1}^{n}\left(\frac{y^{''}}{g\left(\theta^T x^{(\cdot)}\right)}-\frac{1-y^{i'}}{1-g\left(\theta^7 x^{(i)}\right)}\right)\cdot\frac{\partial g\left(\theta^r x^{(i)}\right)}{\partial\theta_j}$$

$$=\sum_{i=1}^{n}\left(\frac{y^{(j)}}{g\left(\theta^T x^{(1)}\right)}-\frac{1-y^{(i)}}{1-g\left(\theta^T x^{(i)}\right)}\right)\cdot g\left(\theta^T x^{(\mu)}\right)\cdot\left(1-g\left(\theta^T x^{(j)}\right)\right)\cdot\frac{\partial\theta^r x^{(\mu)}}{\partial\theta_j}$$

$$=\sum^{n}\left(y^{\omega'}-g\left(\theta^r x^{(i)}\right)\right)\cdot x_j^{(1)}$$

P 是 θ 的函数，X 已知， θ 越大，P 越大。

梯度上升问题，得到 θ 的学习规则：

$$\theta_j:=\theta_j+\alpha\left(y^{(i)}-h_\theta\left(x^{(i)}\right)\right)x_j^{(i)}$$

α 为学习率，最后将 θ 带入 h(x)函数，求出概二分类率问题函数：

$$h_\theta(x)=g\left(\theta^T x\right)=\frac{1}{1+e^{-\theta^T x}}$$

Binary Logistic 输出公式：

$$\ln\left\{\frac{P\mathrm{i}}{1-\mathrm{Pi}}\right\}=\boldsymbol{\beta}_0+\beta_1\mathrm{X}_1+\beta_2\mathrm{X}_2\cdots\cdots+\beta_{\mathrm{n}}\mathrm{X}_n$$

其中，Xi 为驱动因子，βi 为驱动因子的系数。

3.4.3、模型系数检验 （Omnibus Tests of Model Coefficients）

Omnibus 检验公式：

$$K=Z1*(g1)^2+Z2*(g2)^2$$

其中：Z1 和 Z2 是两个正态化函数，g1 和 g2 则分别为偏度和峰度，在 Z1 和 Z2 的作用下，K 的结果将接近于卡方分布，使其能用卡方分布来检验。

Omnibus 检验在 spss 中的应用：

似然比检验结果是对参数是否都为 0 的数值表现，在 SPSS 中该结果呈现在 Model（模型）一行中。P 值：作为显著性检验依据，在其不大于 0.05 时，表明最终选择进入建模的变量中，不少于 1 个变量的“OR 值”满足要求，可认为对整个模型而言，拟合度良好。

3.4.4、Binary Logistic模型结果及拟合度检验

哑变量：又称虚拟变量。是人为设定的变量，可用于处理多分类变量，数值型变量，结果用 0 或 1 表示。哑变量在一定程度上会增加变量个数，但可很好地提高模型精度，且在 Binary Logistic 回归分析中尤为重要。对于 n 分类问题，通常需产生 n-1 个哑变量。本次建模中对有序多分类变量(人均月收入分组、配偶 BMI 分组、母亲孕前 BMI 分组、孕次)采用 Helmert（赫尔默特对比）编码，对于无序多分类变量(受精方式、受孕方式、孕次分组、城乡分组)采用 Simple（简单对比）编码，生成哑变量。

Binary Logistic 建模：结合医学临床经验，调整预测变量，最终选取人均月收入、受精方式是否采用剖宫产、母亲孕前 BMI 分组、配偶 BMI 分组、受孕方式、孕次分组、孕早期柯萨奇病毒、孕前既往性病史、是否采用剖宫产、城乡分组共 10 个变量进行 Binary Logistic 回归建模。

表 1　Binary Logistic 回归模型中的变量系数及相关参数

方程中的变量	B	标准误差	瓦尔德	自由度	显著性	Exp(B)
人均月入分组			405.17	2	0	
人均月收入分组(1)	0.445	0.037	145.284	1	0	1.561
人均月收入分组(2)	0.58	0.032	328.282	1	0	1.787
受精方式(1)	1.408	0.039	1286.07	1	0	4.089
是否采用剖宫产(1)	-0.638	0.03	447.788	1	0	0.528
配偶 BMI 分组			146.327	4	0	
配偶 BMI 分组(1)	-0.286	0.085	11.418	1	0.001	0.751
配偶 BMI 分组(2)	-0.33	0.029	127.301	1	0	0.719
配偶 BMI 分组(3)	-0.267	0.041	42.346	1	0	0.766
配偶 BMI 分组(4)	-0.386	0.06	41.951	1	0	0.68

受孕方式			38.17	2	0	
受孕方式(1)	-0.181	0.033	29.747	1	0	0.835
受孕方式(2)	-0.026	0.042	0.379	1	0.538	0.975
孕次分组			33.729	3	0	
孕次分组(1)	-0.146	0.033	19.522	1	0	0.864
孕次分组(2)	0.033	0.038	0.726	1	0.394	1.033
孕次分组(3)	-0.146	0.041	12.75	1	0	0.864
孕早期柯萨奇病毒(1)	-0.972	0.053	342.997	1	0	0.378
母亲孕前 BMI 分组			306.659	4	0	
母亲孕前 BMI 分组(1)	-0.233	0.049	22.856	1	0	0.792
母亲孕前 BMI 分组(2)	-0.607	0.051	143.336	1	0	0.545
母亲孕前 BMI 分组(3)	-1.173	0.079	218.486	1	0	0.309
母亲孕前 BMI 分组(4)	-0.289	0.138	4.363	1	0.037	0.749
孕前既往性病史(1)	0.151	0.053	8.05	1	0.005	1.163
城乡分组(1)	0.789	0.027	828.814	1	0	2.201
常量	1.654	0.081	418.154	1	0	5.229

表 2　预测是否早产的百分比

实测		预测		正确百分比（%）
		早产	不早产	
		0	1	
早产	0	11415	4078	73.7
不早产	1	4462	11031	71.2
总体百分比				72.4

表 3　模型系数的 Omnibus 检验显著性结果

模型系数的 Omnibus 检验		卡方	自由度	显著性
步骤 1	步骤	7531.446	20	0
	块	7531.446	20	0
	模型	7531.446	20	0

3.4.5、Binary Logistic建模成果分析

通过表 1 可以看到经哑变量处理后，建模成果为：

$$\ln\left\{\frac{P\mathrm{i}}{1-\mathrm{Pi}}\right\}$$
$$=0.445\text{人均月收入分组}(1)+0.58\text{人均月收入分组}(2)+0.789\text{城乡分组}$$
$$-0.286\text{配偶}BMI\text{分组}(1)-0.33\text{配偶}BMI\text{分组}(2)$$
$$-0.267\text{配偶}BMI\text{分组}(3)-0.386\text{配偶}BMI\text{分组}(4)$$
$$-0.181\text{受孕方式}(1)-0.026\text{受孕方式（2）}$$
$$-0.146\text{孕次分组}(1)+0.033\text{孕次分组}(2)-0.146\text{孕次分组}(3)$$
$$-0.972\text{孕早期是否感染柯萨奇病毒}+0.151\text{孕前是否有既往性病史}$$
$$-0.233\text{母亲孕前}BMI\text{分组}(1)-0.607\text{母亲孕前}BMI\text{分组}(2)$$
$$-1.173\text{母亲孕前}BMI\text{分组}(3)-0.289\text{母亲孕前}BMI\text{分组}(4)$$
$$+1.408\text{受精方式}(1)-0.638\text{是否采用剖宫产}$$

根据表 1 的显著性检验结果，人均月收入分组(1)<=2500 元、人均月收入分组(2)2500-5000 元、受精方式(1)IVF、是否采用剖宫产、配偶 BMI 分组(1)偏瘦、配偶 BMI 分组(2)正常、配偶 BMI 分组(3)偏胖、配偶 BMI 分组(4)肥胖、受孕方式(1)自然受孕、孕次分组(1)1 个、孕次分组(3)3 个、孕早期是否感染柯萨奇病毒、母亲孕前 BMI 分组(1)偏瘦、母亲孕前 BMI 分组(2)正常、母亲孕前 BMI 分组(3)偏胖、孕前是否有既往性病史以及城乡分组对建立此次早产预测模型的贡献较大。由表 2 知，早产和不早产预测的正确百分比都在 70%以上，准确性层面上该模型可用。通过模型系数的 Omnibus 检验，其显著性 p 值都小于 0.01，初步认为该模型拟合良好。

对模型中的解释变量进行多重共线性判定：

方差膨胀因子(Variance inflation factor，VIF)

由 Marquardt 于 1906 年引入的，容忍度的倒数，当自变量间存在共线关系时，用最小二乘法所估计的回归系数的方差比自变量间无共线关系时所估计的回归系数的方差的增大倍数，VIF 值愈大，说明变量间的多重共线性程度愈强。同自

变量的相关系数指标一样，利用来诊断多重共线性的问题，其临界值不易确定。当 VIF≥5 或 VIF≥10 时，认为自变量间存在严重共线性但不同的具体情况的临界值将有所不同。

本次建模入选自变量的 VIF 指标均小于 10，初步认为共线性问题可忽略。

表 4　解释变量共线性诊断

模型		共线性统计	
		容差	VIF
1	(常量)		
	人均月收入分组	.923	1.083
	受精方式	.900	1.111
	配偶 BMI 分组	.942	1.061
	是否采用剖宫产	.978	1.023
	孕次分组	.939	1.065
	孕早期柯萨奇病毒	.933	1.072
	母亲孕前 BMI 分组	.925	1.081
	孕前既往性病史	.998	1.002
	城乡分组	.927	1.078

（五）AUC-ROC 曲线

3.5.1、混淆矩阵

		预测值 Predicted label	
		0（negative）	1 （positive）
真实值 True label	0 negative	True negative(TN)	False positive(FP)
	1 positive	False negative(FN)	True positive(TP)

3.5.2、AUC-ROC曲线术语解释

FPRate（假阳性率）:所有真实类别为 0 的样本中，预测类别为 1 的比例。

$$\text{FPRate} = \frac{FP}{FP+TN}$$

TPRate(真阳性率):所有真实类别为 1 的样本中，预测类别为 1 的比例。

召回率、敏感度计算一致：

$$\text{TPRate / Recall / Sensitivity} = \frac{TP}{TP + FN}$$

特异性：

$$\text{Specificity} = \frac{TN}{TN + FP} = 1\text{-}FPRate$$

AUC-ROC 曲线是针对各种阈值设置下的分类问题的性能度量。ROC 是概率曲线，AUC 是 ROC 曲线下的面积，表示可分离的程度或测度。AUC 越高，模型的预测准确性越高。最理想的情况下，既没有真实类别为 1 而错分为 0 的样本——TPRate 一直为 1，也没有真实类别为 0 而错分为 1 的样本——FPRate 一直为 0，AUC 为 1。在医学统计中：AUC 越高，该模型在区分有疾病和无疾病的患者中越好。一般以 0.7 为分界线。

3.5.3、模型检验结果

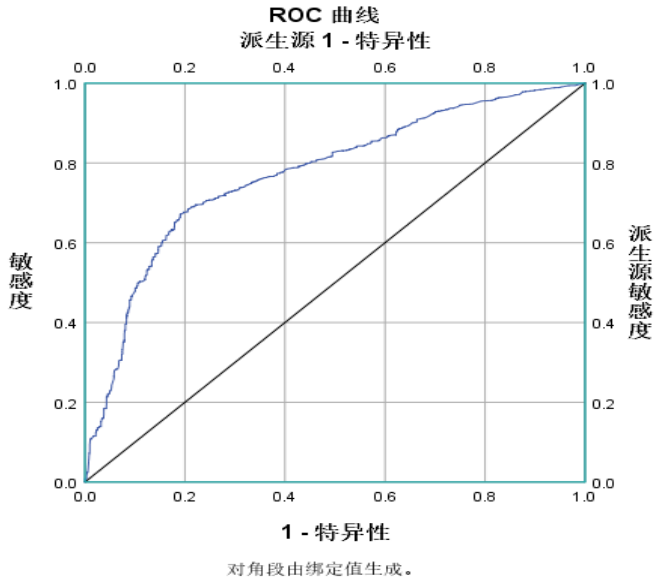

图 5　早产预测模型的 AUC-ROC 曲线

用 TPR 相对 FPR 绘制 ROC 曲线。上图中，TPR 在 y 轴上，FPR 在 x 轴上

表 5　AUC 值展示

曲线下方的区域
.773

说明：0.773>0.7，模型拟合度良好。在医学统计中 AUC 值达到标准，该模型在区分有疾病和无疾病的患者中表现良好，建立的早产预测模型在统计学上具有意义。

第四章 基于决策树的早产预测模型

（一）分区

在原样本中随机取70%作为训练集，由训练集建立模型，取剩下30%作为测试集，由测试集分析误差精度等。

（二）随机欠抽样

采用随机欠抽样平衡样本，因为原样本中不早产人数与早产人数比例差别很大，影响分析结果中对于不早产人群的判断，而此种方法相比之前的上文的处理方法避免过拟合。这里全抽样负类样本（早产人群），在此基础上随机去掉80%的正类样本（不早产人群），抽取结果如表6，得到样本合计6082例。

表6 随机欠抽样后的频率统计

是否早产	百分比	计数
否	50.12	3048
是	49.88	3034

（三）特征选择

采用特征选择对103个变量进行降维，找出对早产预测有积极贡献的重要变量，变量的重要性分两步考虑

4.3.1、从变量本身考察

① 计算变量的各个类别值的取值比例。若其中的最大值大于某个标准值（取90%），则该变量应视为不重要变量。例如，未患梅毒的人占到99.9%，那么是否患梅毒这一变量对于预测早产没有意义。

②对某分类型变量，计算其类别值个数。若类别值个数与样本量的比大于某个标 准值（取 95%），则该变量应视为不重要变量。例如，已婚量与样本量的比为 100%, 则是否已婚对是否早产的预测没有意义。

③若某个变量中缺失值所占的比例大于某个标准值（取 70%），则该变量也应视为不重要变量。

以下列出经过特征选择第一步剔除的 70 个变量

表 7 经过特征选择第一步剔除的变量

1	新生儿死亡史	36	自然流产史
2	胎盘早剥史	37	早产史
3	胎盘早剥	38	孕早期是否饮酒
4	胎膜早破史	39	孕早期是否吸烟
5	胎膜早破	40	孕早期贫血
6	死胎死产史	41	孕早期柯萨奇病毒
7	受精方式	42	孕早期抗抑郁药
8	是否属于近亲结婚	43	孕早期巨细胞病毒
9	是否发生出生缺陷	44	孕早期感冒
10	是否出现营养不良	45	孕早期肝炎类型
11	生殖道沙眼衣原体感染	46	孕早期风疹病毒
12	生活中是否经常使用化妆品	47	孕早期发烧
13	妊娠期是否产检	48	孕早期二手烟暴露
14	妊娠期贫血史	49	孕早期单纯性疱疹病毒
15	妊娠期贫血	50	孕早期大环内酯类抗生素
16	妊娠高血压	51	孕早期病毒性肝炎
17	前置胎盘史	52	孕前月经病
18	前置胎盘	53	孕前饮酒史
19	其他性病	54	孕前阴道炎
20	配偶吸毒史	55	孕前心脏病
21	配偶少数民族类型	56	孕前先天畸形史
22	配偶民族	57	孕前吸烟史
23	你在备孕或孕早期居住的房子是否属于新装修房	58	孕前吸毒史
24	民族分组	59	孕前糖尿病
25	梅毒	60	孕前肾炎
26	淋病	61	孕前口服避孕药
27	居住地附近是否有工厂排放对环境有害的物质	62	孕前抗抑郁药
28	尖锐湿疣	63	孕前进食槟榔史

29	家族出生缺陷史	64	孕前结核病
30	怀孕前 3 个月你经常染发或烫发	65	孕前既往性病史
31	宫外孕史	66	孕前或孕早期有无服用叶酸
32	宫内停止发育史	67	孕前大环内酯类抗生素
33	工作中是否接触放射性有害物质	68	药物过敏史
34	低出生体重史	69	性病性淋巴肉芽肿
35	产前产后出血史	70	产前产后出血

4.3.2、从输入变量与输出变量相关性的角度考虑

利用卡方检验方法，略有不同的是，这里利用似然比卡方 (likelihood ratio)，定义为:

$$T = 2\sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij}^{0} \ln \frac{f_{ij}^{0}}{f_{ij}^{e}}$$

式子中f^0代表观测频数，f^e代表期望频数。对于样本量大的情况，Pearson 卡方输出的检验值与似然卡方的输出值非常接近，将得到相近度较高的检验结论，而研究的样本量为 6082，属于大样本，所以应用似然卡方比。SPSS Modeler 将得到似然卡方比的观测值，同时生成对应的 1—概率 P-值，该值越高，相应的输入变量与早产结果相关的把握越大，该变量对早产结果的预测有重要意义。

此外，Cramer's V 系数也可由 Modeler 计算得出，

数学定义为:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min[(R-1),(C-1)]}}$$

式中，$\min[(R-1),(C-1)]$ 表示取 $(R-1)$ 和（C 一 1）中的最小值 $(R, C$ 分别表示列联表的行数和列数)。 V 系数是对 Pearson 卡方统计量的修正。理由是 Pearson 卡方统计量受到样本量的影响，V 系数可对此进行调整，最后分析得到的结果将不会较大程度的受限于列联表单元格数。

V 系数取值在 0 到 1 之间，越接近 1 表明对应的输入变量与输出变量的相关性 越强，相应的输入变量对是否早产的预测越重要。

结合 Modeler 给出的经验公式$L=\left[\min\left(\max\left(30,2\sqrt{L_0}\right)\right),L_0\right]$，从输入变量的个数$L_0=104$中最终选择进入建模的变量个数为$L=30$，进入建模的变量及其 V 系数如下表，可以看到这 30 个变量的 V 系数均大于 0.9，与是否早产相关性强，对早产的预测很重要。

表 8　进入建模的变量及其 V 系数

序号	变量名	V 值	序号	变量名	V 值
1	个人不孕史	1	16	怀孕年龄分组	1
2	是否采用剖宫产	1	17	教育程度分组	1
3	受孕方式两分类	1	18	配偶吸烟史	1
4	孕前促排卵药物	1	19	孕早期安胎药	1
5	户籍	1	20	母亲孕前 BMI 分组	1
6	配偶职业	1	21	妊娠糖尿病史	1
7	产次分组	1	22	不良妊娠史	0.999999997
8	城乡分组	1	23	孕次分组	0.999999832
9	现有子女数分组	1	24	人均月收入分组	0.999999529
10	配偶年	1	25	人工流	0.999984

	龄分组			产或引产史	471
11	妊娠高血压史	1	26	配偶饮酒史	0.999924232
12	职业分组	1	27	妊娠糖尿病	0.999594381
13	配偶BMI分组	1	28	不孕症类型	0.940926342
14	配偶教育分组	1	29	是否偏食	0.909062642
15	体力负担	1	30	初潮年龄分组	0.904522372

（四）C5.0算法

4.4.1、信息论

由现代信息论者提出的的理论知识分析可知，信息数据可在由信源、信道和信宿组成的系统中实现双向传递，传递过程会产生随机误差，是由于它存在与有干扰的环境，造成信息损失。将信源处发送的信息记为 U ，信宿处收到的信息记为 V ，记为 $P(U \mid V)$信道模型是一个条件概率矩阵 $P(U \mid V)$

$$\begin{bmatrix} P(u_1 \mid v_1) & P(u_2 \mid v_1) & \cdots & P(u_r \mid v_1) \\ P(u_1 \mid v_2) & P(u_2 \mid v_2) & \cdots & P(u_r \mid v_2) \\ \vdots & \vdots & & \vdots \\ P(u_1 \mid v_q) & P(u_2 \mid v_q) & \cdots & P(u_r \mid v_q) \end{bmatrix}$$

上式的 $P(u_i \mid v_j)$ 表示信宿收到信息 v_j 而信源发出信息 u_i 的概率，易知 $\sum_i P(u_i \mid v_j) = 1 \ (i = 1,2,\cdots,r)$。

这样一个通信系统里，信源$u_i(i = 1,2,\cdots,r)$与信息的发生概率$P(u_i)$对应成一

种模型，信源产生信息被视作一种随机过程，通信系统所在环境产生的干扰被视为某种随机序列，且 $\sum_i P(u_i) = 1 \ (i = 1,2,\cdots,r)$。

通信发生前具有一种先验的不确定性，也就是信宿在通信发生前对信源产生信息的未知状态，这种不确定性会在发自信源的的信息被信宿接到时消除或减少。在前期通信检验结束后依然存在不完全确定的通信状态,成为后期检验时的不确定性,这种后验不确定性主要产生于各种环境因素产生的自然干扰,这些环境干扰直接破坏了信号传递器发出的准确信息,最终导致信宿接受到的信息不完整。容易理解，信息传递到达一个信宿之前，后验不确定性等于先验不确定性，而当全部住宿信息都同时传递给另到一个信宿时，状态稳定，后验不确定性为零。

由以上分析可以知道，信息越多随机不确定性越小，反之则反，信息量的数学定义为:

$$I(u_i) = \log_2 \frac{1}{P(u_i)} = -\log_2 P(u_i)$$

进一步将信息量的数学期望定义为信息熵，它也被称为先验熵，用来确定性度量信源发出一定量的信息前的平均不确定性。

其数学定义为:

$$Ent(U) = \sum_i P(u_i)\log_2 \frac{1}{P(u_i)} = -\sum_i P(u_i)\log_2 P(u_i)$$

在$P(u_i)$ 差别越大的时候，计算得到的信息熵越小，可知平均不确定性也越小；而$P(u_i)$ 差别越小，计算得到的信息熵与平均不确定性就越大。考虑一种极端情况，不妨假设信息熵等于 0, 那么此时只存在唯一的信息发送结果的可能，即 $P(u_i) = 1$, 没有发送的概率不确定性; 另一种极端情况是信息熵达到最大，这时信源的 k 个信号都有相同的发送概率，即所有的信号 u_i 都有 $P(u_i) = 1/k$, 此时信息发送的不确定性最大，信息熵达到最大。

当已知信号 U 的概率分布为 $P(U)$ 且收到信号 $V = v_j$ 时，发出信号的概率

分布为$P(U \mid v_j)$ 。

信源的平均不确定性为:

$$\begin{aligned} Ent(U \mid v_j) &= \sum_i P(u_i \mid v_j) \log_2 \frac{1}{P(u_i \mid v_j)} \\ &= -\sum_i P(u_i \mid v_j) \log_2 P(u_i \mid v_j) \end{aligned}$$

称其为后验熵，理解为后来被验证的熵，表示信宿收到 v_j 后获得的对信号 U 的一种信息度量值。

后验熵的期望为:

$$\begin{aligned} Ent(U \mid V) &= \sum_j P(v_j) \sum_i P(u_i \mid v_j) \log_2 \frac{1}{P(u_i \mid v_j)} \\ &= \sum_j P(v_j) \left(-\sum_i P(u_i \mid v_j) \log_2 P(u_i \mid v_j) \right) \end{aligned}$$

称为信道疑义度，表示出后验不确定性的大小。

通常: $Ent(U \mid V) < Ent(U)$ 。进一步，为表示信息消除随机不确定性的程度大小，将信息增益为定义为 ：

$$Gains(U, V) = Ent(U) - Ent(U \mid V)$$

4.4.2、C5. 0 决策树的生长算法

C5.0 决策树以信息增益率为标准确定最佳分组变量和分割点。信息增益率的数学定义为：

$$Gains\ R(U, V) = Gains\ (U, V) / Ent(V)$$

容易理解，为了达到输出结果在是否早产两组的组内达到频率趋向于相同的程

度最高，在选择最佳分组变量时要选择信息增益最大的分组变量，这样各组内部的 $P(u_i)$ 差别大，也就是在此情况下信宿最能消除对信源的平均不确定性，这正是所期望的结果。但是若选择信息增益值为指标又会让类别值的多少影响到结果，例如，类别值多的输入变量比类别值少的输入变量有更多的机会成为最佳分组变量，所以C5.0 以信息增益率作为选择标准，从而解决这一问题。从信息增益率的公式可以看到，对于一个有较多的分类值的输入变量 V，尽管自身的信息熵会因分组多而偏大，分母却因此变大，信息增益率随之低，反之亦反，从最终消除类别数目所带来的影响。

4.4.3、C5. 0 决策树的修剪算法

不同算法下模型常常都会有过拟合的问题，为解决过度依赖训练集而造成对别的样本集预测结果差的问题， C5.0 决策树算法使用后修剪方法，从叶节点向上逐层修剪，这一方法需要进行误差的估计以及修剪标准的设置。

① 误差估计

与其他决策树算法在测试集上估计误差不同，C5.0 直接在其所占比例数为 70%的训练集上直接估计误差，其原理属于统计学上置信区间的估计方法。

基本思路:

(1) 对于决策树上的每个节点，预测类别是输出变量最多的类别。

(2) 对于第 i 个决策树上的观测节点，其包含观测数量为 N_i ，预测错误的观测为 E_i ，那么误差（等于错误率）为 $f_i = E_i/N_{i_i}$

(3) 在近似正态分布假设的基础上，对第 i 个节点的真实误差 e_i 进行区间估计。给出 置信度 $1-\alpha$ ，有：

$$P\left(\frac{f_i - e_i}{\sqrt{\frac{f_i(1-f_i)}{N_i}}} < \left|z_{\frac{\alpha}{2}}\right|\right) = 1 - \alpha$$

$z_{\frac{\alpha}{2}}$ 为临界值。于是，第 i 个节点 e_i 的置信上限，即悲观估计为:

$$e_i = f_i + z_{\frac{e}{2}}\sqrt{\frac{f_i(1-f_i)}{N_i}}$$

C5.0 默认置信度为 $1 - 0.25 = 75\%$, 当 α 为 0.25 时， $z_{\frac{\alpha}{2}} = 1.15$ 。

②修剪标准

C5. 0 以误差估计为基础， 依据“reduce-error”法判断是否修剪。 计算待剪子树中叶节点的加权误差后，剪掉大于父节点误差的子节点。表示为:

$$\sum_{i=1}^{k} p_i e_i > e, i = 1,2,\cdots,k$$

其中，第 i 个叶节点的样本量占整个子树样本量的比例为 p_i; 待剪子树中叶节点的个数为 k; 父节点的估计误差为e，第 i 个叶节点的估计误差为 e_i。

由此，可生成深度为 23 的是十个决策树（详见数据包中的 SPSS Modeler 数据流中的结果），其估计准确性在 78.44%到 92.68%不等。

4.4.4、C5.0决策树结果统计分析

表 9 C5.0 训练集的混淆矩阵

训练集	预测不早产	预测早产	无法预测
实际不早产	1,940	119	25
实际早产	112	2,011	1

由表 9 可知，对训练集而言，C5.0 决策树模型预测出结果不早产而实际不早

产的有 1940 个，模型预测不早产而实际早产的由 112 个，模型预测早产而实际不早产的有 119 个，模型预测早产而实际早产有 2011 个，模型一共无法预测的样本有 26 个。

表 10　C5.0 测试集的混淆矩阵

测试集	预测不早产	预测早产	无法预测
实际不早产	857	54	13
实际早产	47	863	0

由表 11 可知，对测试集而言，模型预测出结果不早产而实际不早产的有 857 个，模型预测不早产而实际早产的有 47 个，模型预测早产而实际不早产的有 54 个，模型预测早产而实际早产有 863 个，模型一共无法预测的样本有 13 个。

以上两个混淆矩阵可以直观从频率上看到预测效果不差。以下是综合训练集和测试集的结果统计，C5.0 决策树在测试集上的整体预测精度为 93.78%。

表 11　C5.0 综合训练集和测试集的结果统计

	训练集		测试集	
正确	3,951	93.89%	1,720	93.78%
错误	257	6.11%	114	6.22%
总计	4,208	100%	1,834	100%

这里的置信度是相应规则的置信度经拉普拉斯估计器（Laplace Estimator）调整后的结果。拉普拉斯估计器是法国数学家拉普拉斯于 18 世纪提出的经典方法，其标准算法是：

$$\frac{N_j(t)+1}{N(t)+k}$$

式中， $N(t)$ 是节点 t 包含的样本量; $N_j(t)$ 是节点 t 包含第 j 类的样本量; k 是输出变量 的类别个数。如果输出变量为数值型，则不存在拉普拉斯调整问题。 拉普拉斯调整通常用在朴素贝叶斯分类方法，主要解决估计过程 中输入和输出变量联合分布下概率为 0 时后验概率无法计算的问题。事实上，分子部分的 加 1 并没有特别的理由，对此的改进结果：

$$\frac{N_j(t)+kp}{N(t)+k}$$

由此得到的测试集置信度报告如下：

表 12　C5.0 测试集置信度报告

测试集置信度报告	
范围	0.5 - 1.0
平均正确性	0.859
平均不正确性	0.692
正确性始终高于	1.0 （观测值的 0%）
不正确性始终低于	0.501 （观测值的 0.05%）

由表可以看到测试集的平均正确性为 0.859，意味着对于测试集中所有正确预测的 1720 个样本来说，置信度的平均值为 0.859；平均不正确性 0.692, 意味着对于测试集中所有不正确预测的 114 个样本来说，置信度的平均值为 0.692；不正确性始终低于 0.501，意味着置信度在 0.501 以下的样本，其预测值均是错误的。

（五）C5.0 推理集

利用 Cendrowsk 在 1987 年的时间提出的一个可以生成推理集的规则生成

算法，算法缩写为 PRISM (Patient Rule Induction Space Method)，该算法生成的所有规则的正确率能在训练样本集上达到 100%。

4.5.1、PRISM 算法的基本思路

PRISM 算法的基本步骤如下

步骤 1、确定一个输出变量为期望类别

步骤 2、在当前样本集 T 范围内（开始时为全部观测)，找一条能最大限度覆盖属于该类别的样本的推理规则。最大限度指规则尽可能多地覆盖属于期望类别的样本，也尽量少覆盖或不覆盖属于其他类别的样本。确定规则的标准是使正确覆盖率 (N/M) 达到最大（推理规则共覆盖 M个观测，其中有 N 个观测属于期望类别）。如果遇到两条规则有相等的正确覆盖率时，应选择正确覆盖数大的规则。一条简单规则通常是不充分的，需要在此基础上继续附加逻辑与条件，因为它可能覆盖了属于其他类别的 $M-N$ 个样本。

步骤 3、在样本量为 M 的样本范围内，继续附加推理条件，得到一个更小些的样本推理范围，该过程遵循正确覆盖率最大原则。接着继续附加逻辑与条件，不断缩小样本范围，直到推理规则不再覆盖原本属于其他类别的样本时，一条推理规则便形成了。

步骤 4、样本子集T_i被正确覆盖时，T_i将被剔除。同时对剩余样本集{T-T_i}进行检验，看是否还有属于期望类别的样本。如果有，则回到步骤 2 向下重复，否则结束。

总之， PRISM 算法通过逐步缩小样本空间范围，最后定位到属于期望类别的样本，最后生成相应的推理规则。

推理规则的生成策略与决策树的构建策略并不相同。推理规则总是以用户期望类别的最大正确覆盖率为衡量标准，在一个时刻只考虑一个类别，决策树则同时兼顾输出变量的各个类别。另外，推理规则的生成是有先后顺序的.规则集的本

质是一个决策序列。依次执行推理规则，一旦满足规则，就不再考察下一条规则。

4.5.2、C5.0推理集结果分析

由于偏差和方差的存在，建立在一组训练样本上的一个模型，所给出的预测往往缺乏稳健性。Boosting 技术可用于解决这个问题，它是用于机器学习中的有指导学习算法，包括建模和投票两个阶段。原理这里不再展开，在此模型中使用 Boosting 技术，将建立的 10 个规则集估计准确性提高到了 97.2%。如下图：

图 6　Boosting 技术

表 13 是综合训练集和测试集的结果统计，C5.0 推理集在测试集上的整体预测精度为 95.92%，不存在无法预测的情况。

表 13　C5.0 结果分析综合

	训练集		测试集	
正确	4,146	95.31%	1,741	95.92%
错误	204	4.69%	74	4.08%
总计	4,350	100%	1,815	100%

表 14　测试集置信度报告

测试集置信度报告	
范围	0.5 - 1.0
平均正确性	0.824
平均不正确性	0.714
正确性始终高于	1.0 （观测值的 0%）
不正确性始终低于	0.5 （观测值的 0%）

由表可以看到测试集的平均正确性为 0.824，意味着对于测试集中所有正确预测的 1741 个样本来说，置信度的平均值为 0.824；平均不正确性 0.714, 意味着对于测试集中所有不正确预测的 74 个样本来说，置信度的平均值为 0.714；不正确性始终低于 0.5，意味着置信度在 0.5 以下的样本，其预测值均是错误的。与 C5.0 决策树的模型相比，总体来讲准确性提高了。

（六）CHAID 算法

CHAID 的英文全称是 chi-squared automatic interaction detector，它是一种卡方自动交互诊断器 ,作为一种决策树算法，CHAID 从统计显著性检验的角度确定当前最佳分组变量和分割点，并且用户可以根据实际需要进行预先确定决策树的距离深度，这里选择树深度为 5。

4.6.1、CHAID 分组变量的预处理和选择策略

步骤 1、对 30 个输入变量进行预处理，对于本次建模，输出变量为早产或不早产，输入变量为也已经处理为分类型变量。χ2 检验可用于检验分类型模型的相关性，超类是指对输入变量的反复检验、合并，算法终止在超类变量无法再合并

时。

步骤 2、对输入变量进行预处理后，SPSS Modeler 计算出相关性检验中的统计量和概率 P -值,再采用似然比方。显然，概率 P -值最小的输入变量与输出变量相关的把握最大，当概率 P -值相同时，应优先选择概率检验器在统计该变量时所观测的数值较大的一个输入变量。

CHAID 算法与 C5.0 算法的区别在于，其分组变量确定的依据是: 输入变量与输出变量之间的相关程度。应将与输出变量最相关的输入变量作为最佳分组变量，而不是像 C5.0算法，选择使输出变量取值差异性下降最快的变量为最佳分组变量。 CHAID 方法自动生成的树分支来源于分组变量的各个类别，同时也将生成多个小型分叉，具体处理方式类似于 C5.0 决策树算法。

反复进行上述过程，直到决策树生长终止。

4.6.2、CHAID结果统计分析

控制树深度 5，给出该算法下生成的决策树见附录（三）。

这里还给出预测变量的重要性

Nodes	Importance
初潮年龄分组	0.0047
不孕症类型	0.0199
城乡分组	0.0199
体力负担	0.0199
妊娠糖尿病史	0.0199
孕次分组	0.0199
产次分组	0.0199
教育程度分组	0.0561
是否采用剖宫产	0.0849
配偶BMI分组	0.1374
配偶教育分组	0.1807
配偶职业	0.2069
配偶分组	0.2099

图 7　CHAID 算法下的生成的决策变量预测重要性

以下是分别是训练集的重合矩阵，对训练集而言，CHAID 模型预测出结果不早产而实际不早产的有 1841 个，模型预测不早产而实际早产的有 468 个，模型预测早产而实际不早产的有 384 个，模型预测早产而实际早产有 1635 个， 相比 C5.0 算法不存在无法预测的情况。

表 15　CHAID 训练集的重合矩阵

训练集	预测不早产	预测早产
实际不早产	1,841	384

实际早产	468	1,635

对测试集而言，模型预测出结果不早产而实际不早产的有 772 个，模型预测不早产而实际早产的有 210 个，模型预测早产而实际不早产的有 173 个，模型预测早产而实际早产有 721 个。

表 16　CHAID 测试集的重合矩阵

测试集	预测不早产	预测早产
实际不早产	772	173
实际早产	210	721

以下是综合训练集和测试集的结果统计：

表 17　CHAID 综合训练集和测试集的结果统计

	训练集		测试集	
正确	3,476	80.31%	1,493	79.58%
错误	852	19.69%	383	20.42%
总计	4,328	100%	1,876	100%

CHAID 决策树在测试集上的整体预测精度为 79.58%，比 C5.0 决策树的整体预测精度低，可见简单的树分支会使结果的可靠性下降。

得到的测试集置信度报告如下：

表 18　CHAID 测试集置信度报告

测试集置信度报告	
范围	0.5 - 1.0
平均正确性	0.824

平均不正确性	0.714
正确性始终高于	1.0 （观测值的 0%）
不正确性始终低于	0.5 （观测值的 0%）
90.01% 以上的准确性	0.812
2.0 以上的折叠正确性	0.812 （观测值的 90.01%）

由表可以看到测试集的平均正确性为 0.824，意味着对于测试集中所有正确预测的 1493 个样本来说，置信度的平均值为 0.824；平均不正确性 0.714, 意味着对于测试集中所有不正确预测的 383 个样本来说，置信度的平均值为 0.714；不正确性始终低于 0.5，意味着置信度在 0.5 以下的样本，其预测值均是错误的，包含百分之零的样本，说明找不出一个置信度，低于该置信度的样本，其预测均错误；在预测置信度高于 0.812 的样本中，有 90.01%的样本被预测正确；预测置信度高于 0.812 的样本占总体的 90.01%，它的预测正确性比总体的正确率提高了 2 折，应为 89.79%。因此最终取置信度为 0.812 较为合适。

第五章 总结及展望

通过第三章可以看到，Binary Logistic 模型通过了显著性检验和 AUC-ROC 曲线预测拟合度检验。进行回顾，可以发现 SMOTE 采样对于类别不均衡的问题处理得不错，二阶聚类对于多变量、大样本的模型可以起到很好的降维效果。

对于第四章决策树部分的研究，在实际应用时，可以通过决策树节点的顺序编写程序，对于建立起来的决策树模型只要孕妇完成不超过 15 个的选择题，就能得到一个预测的结果以及置信区间。同时应该注意到，本次建模的样本是湖南省的孕妇，其中入选建模的变量如户籍，其变量值并不一定适用于别的省份，因此由本次建模产生的早产预测情况要考虑到预测人群居住地的问题。

本次模型通过较为成熟、医学统计上常用的机器学习方法，如聚类分析及 Binary Logistic 模型来处理大样本、多变量的分类问题，建立出了具有一定预测效度的模型。同时在此基础上，也试探寻找处理效率更高、效果更好的方法。

医学统计在疾病预测、危险因素分析以及流行病传播等方面起着至关重要的作用。将统计应用到医学事业中，是时代的必然趋势，也是造福人类的关键一步。现如今，应积极寻找统计方法，不断摸索前行，为推动统计学、医学事业而奋斗。

第六章 致谢

衷心感谢秦家碧教授在论文选题中提供的帮助，感谢王志忠教授在写作过程中提供的方法与指导，以及计算机学院的教授为此次论文提供计算机方面的支持。同时，由衷感谢秦家碧教授的研究生团队在研究、收集数据中为我们所做出的巨大支持。

感谢为统计学工作做出贡献的前辈们！

参考文献

[1] Haiqing Xu et al. Time trends and risk factor associated with premature birth and infants deaths due to prematurity in Hubei Province, China from 2001 to 2012[J]. BMC Pregnancy and Childbirth, 2015, 15(1)

[2] 何丽芸,杜莉,金辉,林双,朱丽萍.上海市早产发生状况及危险因素研究[J].中国妇幼健康研究,2020,31(06):706-711.

[3] 陈丽,黄婷,林颖.妇产科不同类型早产影响因素的 Logistic 回归分析[J].当代医学,2021,27(09):69-71.

[4] 虢海月,杜姣洋,李敏敏,张彬艳,蒋茜,李少茹,刘蓉,毕育学.西安市早产儿发生状况及其影响因素分析[J].西安交通大学学报(医学版),2020,41(02):281-286.

[5] 王希,康楚云,高燕秋,等.中国 3 省市 21 家医院早产发生的相关因素及结局研究[J].中国生育健康杂志,2014,(1).1-5

[6] 郝素芳,丁瑛雪,杨丽君,崔红.影响早产儿贫血的相关因素分析[J].中国妇幼健康研究,2017,28(12):1503-1507.

[7] SMOTE: Synthetic Minority Over-sampling Technique，JAIR'2002

[8] 鲁庆云, 刘红霞. 关于列联表卡方检验在数学教育研究中的使用方法分析[J]

[9] 薛薇.基于 SPSS Modeler 的数据挖掘（第二版）.北京：中国人民大学出版社，2014.

附录

附录（一）：Python 实现 SMOTE 过采样代码及输出文件

```
from imblearn.over_sampling import SMOTE
import numpy as np
from sklearn.model_selection import train_test_split
data = pd.read_excel('1(2).xlsx') #读取数据集

data = data.dropna(axis = 1, how = 'any') # 丢弃有 NAN 的列
data = data.dropna(axis = 0, how = 'any') # 丢弃有 NAN 的行
data = data.drop(columns=['ID', '调查人署名']) # 丢弃 ID 和调查人属名，这两个非 float，放在这里是无效的
var = data.columns

Y = data.iloc[:,-1] # 获得因变量数据
X = data.iloc[:,:-1] # 获得自变量数据

oversampler=SMOTE(random_state=2021)  # 导入过采样库—SMOTE 算法
# x_train, x_valid_test, y_train, y_valid_test = train_test_split(X,Y,test_size=0.3,random_state=2020) # 将数据集切分为 训练集和验证+测试集

x_train,y_train=oversampler.fit_sample(X,Y) # 对训练集进行 SMOTE 过采样，
```

得到过采样后的自变量和因变量

```
#
x_valid,x_test,y_valid,y_test=train_test_split(x_valid_test,y_valid_test,test_size=0.3,r
andom_state=2020)
data_smote = pd.concat([x_train,y_train],axis = 1)
```

样本量过大，具体结果展示在数据包中。

附录（二）：χ2 相关性研究及剔除无关变量结果

变量	χ2	p 值
受孕方式	2427.974	0.000**
配偶年龄分组	1202.425	0.000**
受孕方式两分类	2041.796	0.000**
配偶教育分组	783.774	0.000**
配偶民族	94.725	0.000**
妊娠期是否产检	26.022	0.000**
配偶 BMI 分组	656.801	0.000**
婴儿性别	605.167	0.000**
怀孕年龄分组	456.354	0.000**
民族分组	280.019	0.000**
户籍	1242.853	0.000**
配偶吸烟史	144.513	0.000**
配偶进食槟榔史	11.685	0.001**
配偶吸毒史	38.365	0.000**
城乡分组	1091.485	0.000**
孕早期发烧	596.602	0.000**
孕早期病毒性肝炎	232.207	0.000**
孕早期柯萨奇病毒	1075.281	0.000**
孕早期巨细胞病毒	425.372	0.000**
不孕症类型	100.022	0.000**

受精方式	4331.253	0.000**
母亲孕前 BMI 分组	228.958	0.000**
孕前吸烟史	8.217	0.004**
孕前阴道炎	41.89	0.000**
孕前既往性病史	369.904	0.000**
梅毒	44.063	0.000**
淋病	12.005	0.001**
尖锐湿疣	14.006	0.000**
现有子女数	1585.227	0.000**
孕早期单纯性疱疹病毒	333.709	0.000**
孕早期风疹病毒	46.142	0.000**
孕早期贫血	135.753	0.000**
药物过敏史	270.712	0.000**
孕前或孕早期有无服用叶酸	10.56	0.001**
孕前口服避孕药	173.097	0.000**
孕前促排卵药物	72.557	0.000**
孕前大环内酯类抗生素	1760.929	0.000**
孕前抗抑郁药	120.228	0.000**
孕早期安胎药	133.362	0.000**
孕早期大环内酯类抗生素	55.758	0.000**
孕早期抗抑郁药	114	0.000**
生活中是否经常使用化妆品	165.265	0.000**
工作中是否接触放射性有害物质	17.001	0.000**
居住地附近是否有工厂排放对环境有害的物质	172.884	0.000**
怀孕前 3 个月你经常染发或烫发	114.395	0.000**

是否出现营养不良	36.148	0.000**
是否偏食	167.033	0.000**
个人不孕史	13.498	0.000**
现有子女数分组	1623.888	0.000**
孕次	576.102	0.000**
孕次分组	103.966	0.000**
产次	1624.094	0.000**
产次分组	1631.323	0.000**
早产史	11.286	0.001**
妊娠高血压史	47.716	0.000**
前置胎盘史	442.283	0.000**
胎盘早剥史	302.425	0.000**
剖腹产史	58.623	0.000**
产前产后出血史	26.199	0.000**
妊娠期贫血史	58.061	0.000**
孕前先天畸形史	36.969	0.000**
孕前心脏病	23.017	0.000**
孕前肾炎	14.175	0.000**
孕前月经病	104.521	0.000**
孕前二手烟暴露	423.286	0.000**
孕前饮酒史	89.957	0.000**
孕前进食槟榔史	240.35	0.000**
是否属于近亲结婚	53.917	0.000**
孕早期是否吸烟	84.666	0.000**
孕早期二手烟暴露	163.921	0.000**

教育程度分组	562.035	0.000**
职业分组	641.58	0.000**
人均月收入分组	388.531	0.000**
OM 是否采用剖宫产	3047.974	0.000**
OM 产前产后出血	27.239	0.000**
OM 妊娠期贫血	24.914	0.000**
OM 是否发生出生缺陷	34.069	0.000**

附录（三）：CHAID 结果统计分析

不孕症类型 = 1 or 不孕症类型 = 2 [模式：1] (2,000)

配偶职业 = 1 or 配偶职业 = 3 [模式：1] (797)

是否采用剖宫产 = 0 [模式：0] (115)

教育程度分组 = 1 or 教育程度分组 = 2 [模式：0] => 不早产 (71; 0.93)

教育程度分组 = 3 [模式：1] => 早产 (44; 0.75)

是否采用剖宫产 = 1 [模式：1] (682)

初潮年龄分组 = 0 [模式：1] (618)

职业分组 = 1 [模式：1] => 早产 (502; 0.922)

职业分组 = 2 or 职业分组 = 7 [模式：1] => 早产 (116; 0.655)

初潮年龄分组 = 1 [模式：1] => 早产 (64; 0.531)

配偶职业 = 2 or 配偶职业 = 6 [模式：0] => 不早产 (67; 0.955)

配偶职业 = 7 [模式：1] => 早产 (114; 1.0)

配偶职业 = 8 or 配偶职业 IS MISSING [模式：0] (103)

教育程度分组 = 1 [模式：1] => 早产 (44; 1.0)

教育程度分组 = 2 or 教育程度分组 = 3 [模式：0] => 不早产 (59; 0.898)

配偶职业 = 10 [模式：1] (919)

配偶教育分组 = 1 or 配偶教育分组 = 3 [模式：1] (750)

职业分组 = 1 or 职业分组 = 6 [模式：1] (243)

教育程度分组 = 1 [模式：0] => 不早产 (61; 0.508)

教育程度分组 = 2 [模式：0] => 不早产 (50; 0.92)

教育程度分组 = 3 [模式：1] => 早产 (132; 0.856)

职业分组 = 4 or 职业分组 = 7 [模式：1] (455)

教育程度分组 = 2 [模式：1] => 早产 (358; 0.944)

教育程度分组 = 3 or 教育程度分组 = 4 [模式：1] => 早产 (97; 0.577)

职业分组 = 5 [模式：1] => 早产 (52; 1.0)

配偶教育分组 = 2 [模式：0] (113)

配偶 BMI 分组 = 2 or 配偶 BMI 分组 = 4 or 配偶 BMI 分组 = 5 [模式：0] => 不早产 (62; 0.968)

配偶 BMI 分组 = 3 [模式：1] => 早产 (51; 0.745)

配偶教育分组 = 4 [模式：0] => 不早产 (56; 0.857)

不孕症类型 IS MISSING [模式：0] (2,309)

城乡分组 = 0 [模式：0] (1,385)

配偶 BMI 分组 = 1 or 配偶 BMI 分组 = 2 or 配偶 BMI 分组 = 3 [模式：0] (1,138)

体力负担 = 0 [模式：0] (870)

妊娠糖尿病史 = 0 [模式：0] => 不早产 (811; 0.846)

妊娠糖尿病史 = 1 [模式：0] => 不早产 (59; 0.61)

体力负担 = 1 [模式：0] (268)

配偶教育分组 = 1 or 配偶教育分组 = 2 [模式：1] => 早产 (67; 0.567)

配偶教育分组 = 3 or 配偶教育分组 = 4 [模式：0] => 不早产 (201; 0.781)

配偶 BMI 分组 = 4 [模式：0] (172)

孕次分组 = 1 or 孕次分组 = 3 [模式：0] (124)

配偶职业 = 1 or 配偶职业 = 4 or 配偶职业 = 6 or 配偶职业 = 7 or 配偶职业 = 10 [模式：0] => 不早产 (64; 0.703)

配偶职业 = 2 or 配偶职业 = 3 or 配偶职业 = 8 or 配偶职业 IS MISSING [模式：1] => 早产 (60; 0.7)

孕次分组 = 2 or 孕次分组 = 4 [模式：0] => 不早产 (48; 0.792)

配偶 BMI 分组 = 5 [模式：1] => 早产 (75; 0.573)

城乡分组 = 1 [模式：0] (924)

体力负担 = 0 [模式：0] (705)

教育程度分组 = 1 [模式：1] (158)

孕次分组 = 1 or 孕次分组 = 2 or 孕次分组 = 3 [模式：1] => 早产 (113; 0.628)

孕次分组 = 4 [模式：0] => 不早产 (45; 0.711)

教育程度分组 = 2 or 教育程度分组 = 4 [模式：0] (363)

配偶教育分组 = 1 or 配偶教育分组 = 4 [模式：0] => 不早产 (48; 0.521)

配偶教育分组 = 2 or 配偶教育分组 = 3 [模式：0] => 不早产 (315; 0.79)

教育程度分组 = 3 [模式：0] (184)

是否采用剖宫产 = 0 [模式：0] => 不早产 (96; 0.688)

是否采用剖宫产 = 1 [模式：0] => 不早产 (88; 0.5)

体力负担 = 1 [模式：1] (219)

孕次分组 = 1 [模式：0] (102)

产次分组 = 1 [模式：0] => 不早产 (48; 0.875)

产次分组 = 2 [模式：1] => 早产 (54; 0.556)

孕次分组 = 2 or 孕次分组 = 3 or 孕次分组 = 4 [模式：1] => 早产 (117; 0.812)