摘要

生育率降低、人口老龄化加剧、人口红利逐渐消失等一系列问题正对我国的经济、社会、文化等各方面构成巨大挑战。为了优化人口结构、保持国家人力资源优势，我国政府不断调整生育政策，尽管先后颁布实施了“单独二孩”和“全面二孩”政策，我国居民的生育意愿并没有出现较大反弹，且生育率提升效果有限，因此 2021 年国家又出台了“三孩政策”。生育行为受许多因素影响，其中，生育意愿对生育行为起着先导作用。因此，研究生育意愿的影响因素对推动三孩生育政策与其它相关政策有效衔接、改善我国人口问题有着重要的现实意义。

本文采用 CGSS2018 数据，从问卷调查中选取超过 90 个问题，涵盖“社会人口属性”、“住房问题”、“社会保障”、“家庭”等 11 个模块，共 72 个自变量，结合统计与机器学习模型方法，研究影响居民三孩生育意愿的因素。在实证分析部分，本文首先对相关变量进行描述性统计分析，展现居民意愿孩子数量的基本情况，并从不同角度对居民意愿孩子数量的分布情况进行分析。然后，利用决策树模型和 XGBoost 模型对总体样本数据建模，寻找可能影响居民三孩生育意愿的重要变量，并按不同角度对样本数据划分建模，对比分析不同人群的三孩生育意愿的影响因素。最后，参考机器学习模型输出结果和已有相关文献，选取相关变量构建二项 Logit 模型，对居民三孩生育意愿的影响因素进行具体分析。

通过实证分析表明，房产数量对居民的三孩生育意愿产生显著正向影响。此外，现有孩子数量、养老保险、宗教信仰、受教育程度和年龄等因素均从不同方向对居民的三孩生育意愿产生影响。对比分析发现，女性的三孩生育意愿受年龄、社会保障情况和婚姻状况的影响较大，而男性的三孩生育意愿则受经济因素及父母的影响较大。现有孩子数量和经济投资因素对青年群体和中年群体的三孩生育意愿都有重要影响，青年群体的三孩生育意愿受知识

水平情况、工作状况等因素的影响较多，而中年群体的三孩生育意愿受户口情况、家庭拥有房产情况等因素的影响较多。无孩群体的三孩生育意愿受婚姻状况、生育自由观念等因素的影响较大；一孩群体的三孩生育意愿受出生或户口地、家庭汽车拥有情况、年龄等因素的影响较大；而二孩群体的三孩生育意愿则更多地受家庭房产情况、养老保险参与情况以及个人生活方式等因素影响。相较于无孩群体和一孩群体，二孩群体的三孩生育意愿更多地受到现实的家庭经济及稳定因素的影响。根据研究所得结论，本文提出相关政策建议，例如：政府可以将购房政策与生育相关政策结合考虑，针对不同人群精准制定相关措施等。

关键词：生育意愿；决策树；XGBoost；Logit 模型；人口红利

Abstract

A series of problems, such as the reduction of fertility rate, the aggravation of population aging and the gradual disappearance of demographic dividend, are posing great challenges to China's economy, society and culture. In order to optimize the population structure and maintain the advantages of national human resources, the Chinese government has continuously adjusted its birth policy. Although it has successively promulgated and implemented the policies of "selective two-child" and "universal two-child", the fertility willingness of Chinese residents has not appeared in big bounce and the fertility rate improvement effect is limited. Therefore, in 2021, the state introduced the "three-child" policy. Fertility behavior is affected by many factors, among which fertility willingness plays a leading role in fertility behavior. Therefore, it is of great practical significance to study the influencing factors of fertility willingness to promote the effective connection between the three-child birth policy and other related policies and improve the population problem in China.

In this paper, CGSS2018 data is used to select more than 90 questions from the questionnaire survey, covering 11 modules including "social population attributes", "housing problems", "social security" and "family", with a total of 72 independent variables. Combining statistical and machine learning model methods, this paper studies the factors affecting residents' fertility willingness of three children. In the part of empirical analysis, this paper firstly makes descriptive statistical analysis of related variables to show the basic situation of the number of residents' willing children and analyzes the distribution of the number of residents' willing children from different angles. Then the decision tree model and XGBoost model are used to model the total sample data to find out the important variables that may affect the fertility willingness of residents, and the sample data are divided and modeled from different angles to compare and analyze the influencing factors of the fertility willingness of different groups of people. Finally, referring

to the output of machine learning model and the relevant literature, we select relevant variables to construct a binary Logit model to analyze the influencing factors of residents' three-child fertility willingness in detail.

The empirical analysis shows that the number of real estate has a significant positive impact on residents' willingness to have three children. In addition, factors such as the number of existing children, pension insurance, religious belief, education level and age have an impact on residents' willingness to give birth to three children from different directions. The comparative analysis shows that women's three-child fertility willingness is greatly influenced by age, social security and marital status, while men's three-child fertility willingness is greatly influenced by economic factors and parents. The number of existing children and economic investment factors have an important impact on the fertility willingness of young people and middle-aged groups. The fertility willingness of young people is more influenced by knowledge level, work status and other factors, while the fertility willingness of middle-aged people is more influenced by household registration, family ownership and other factors. The fertility willingness of three children in childless groups is greatly influenced by factors such as marital status and concept of reproductive freedom; The fertility willingness of the three children in the one-child group is greatly influenced by factors such as family car ownership, age and birth or household registration; However, the fertility willingness of the two-child group is more affected by factors such as family real estate, pension insurance participation and personal lifestyle. Compared with the childless group and the one-child group, the fertility willingness of the three-child group is more affected by the realistic family economy and stability factors. According to the conclusion of the research, this paper puts forward relevant policy suggestions, such as: the government can combine the house purchase policy with the birth-related policy and consider accurately formulating relevant measures for different groups of people.

Key words: fertility willingness; Decision tree; XGBoost; Logit model; demographic dividend

目录

1.绪论

1.1 研究背景与研究意义

1.1.1 研究背景

人口问题一直都是国家和人民所关心的重大问题。从上个世纪 70 年代开始，我国提出按人口政策有计划的生育，并在 1982 年把计划生育确定为“基本国策”写入宪法。此后几十年间，少生、优生的思想深入人心，计划生育工作解决了我国人口增速过快问题，使得人口数量得到控制，缓解了人口与资源的紧张关系，为促进人口与经济协调发展作出重要贡献。但是，随着时间推移，计划生育政策导致了一系列新的问题产生。我国的总和生育率持续降低，到 2012 年仅为 1.26，远低于 2.1 的替代水平，“人口红利”情况逐步消退、人口老龄化加剧等一系列问题给我国社会经济的进一步深化发展造成了阻碍。在此背景下，2013 年我国开始实施“单独二孩”政策；2015 年 10 月 29 日，十八届五中全会通过决定：坚持计划生育的基本国策，完善人口发展战略，全面实施一对夫妇可生育两个孩子的政策。“二孩政策”在促进生育率提升方面具有一定成效，但较为有限。

生育率陷入超低水平、人口结构失衡、老龄化程度加深给我国经济社会可持续发展构成了极大的威胁与挑战，只有优化人口结构并提高人口素质才能有效应对这些挑战，因此，提高国民生育意愿和生育水平刻不容缓。2021 年 5 月，中共中央政治局召开会议，提出进一步优化生育政策，实施一对夫妻可以生育三个子女政策及配套支持措施。2021 年 6 月，中共中央国务院发布《关于优化生育政策促进人口长期均衡发展的决定》，提出了更详细的方案。“三孩政策”的出台即刻引起了社会热议，众多民众表达出低生育意愿，认为自己面临角色与生育困境，政策及其配套措施的有效性受到质疑。据国家统计

局发布数据显示，2022 年全国人口减少 85 万人，出现 61 年来首次负增长。四川自 2023 年 2 月 15 日起实施《四川省生育登记服务管理办法》，办法中的生育登记取消结婚限制与生育数量限制的消息引发热议。人口结构的“需要生育”、国家政策的“鼓励生育”与民众的“不想生育”构成了一组难以破解的矛盾。毋庸置疑，现实中影响生育意愿的因素极其复杂，单纯依靠放开生育政策未必就一定会提升育龄人群的生育意愿。目前，国内外对生育意愿的研究很多，但多集中于传统社会科学分析法，针对比较单一的方向，采用较少的变量。生育问题是一个综合性研究课题，应从多学科视角加大对居民生育意愿影响因素的全面研究。

1.1.2 研究意义

目前我国的人口问题日益突出，在经济发展的过程中，“人口红利”情况已经逐步消退，取而代之的是我国的老龄化越来越严重。与发达国家相比，我国社会进入老龄化具有时间短和速度快等特点，而且我国是发展中国家，国家的经济发展和社会保障还不够成熟，这使得我国面临“老龄化”的形势变得更加严峻。假若我国的人口发展现状没有得到较大改善的话，未来我国将是世界上人口老龄化最严重的国家之一。现实中影响居民生育意愿的因素极其复杂，目前单纯依靠放开三孩生育政策对于提高居民三孩生育意愿的效果并不明显。不同时代的居民对于生育的观念不同，新时代背景下，加大对居民生育意愿影响因素的全面研究，并根据相关结果及时出台与调整国家政策，对激发居民生育意愿具有重要现实意义。

影响居民生育意愿的相关研究已经很多，但大多数研究仅局限于某一学科以及某一领域，生育问题是一个综合性研究课题，应从多学科视角进行研究。同时，国内外对生育意愿的影响因素研究方法更多集中于传统社会科学分析法，实证分析过程采用的推理和模型更多的来自传统社会科学分析方法以及统计学模型。且以往的研究更多依赖学者背景，针对较单一的方向，采用较少的变量研究对生育意愿的影响。如今机器学习和深度学习在各个领域大放异彩，对不同行业数据都能做到很好的拟合与预测效果，因此，本文结合统计与机器学习模型方法对生育意愿影响因素进行研究。首先利用机器学

习模型，从更高的维度对影响居民生育意愿的因素进行更加全面的挖掘，挖掘出生育意愿的重要影响因素，发现更多可解释、可理解的相关关系，再利用统计模型研究重要影响因素对居民生育意愿的具体影响，以期根据研究结果提出相关政策建议。

1.2 研究设计

1.2.1 研究思路与方法

本文研究主要采用文献调查法、理论与实证分析相结合的方法和比较分析的方法。在理论储备方面，本文首先通过查阅大量国内外与生育意愿及其影响因素相关的文章，梳理相关的学术观点、研究方法及成果，整理出文献综述，总结前人的研究经验以及不足之处，思考可能的创新点。然后笔者结合自身已有知识体系，调研可挖掘影响居民三孩生育意愿因素的模型及所需相关数据，为本文研究的可行性与有效性奠定基础。在实证分析部分，本文在借鉴相关理论以及研究成果的基础上，选取适当的变量指标。首先，对CGSS2018 数据中的部分变量进行描述性统计分析，初步分析我国居民的意愿子女数量及其在不同人群中的分布情况。然后，运用决策树模型和 XGBoost 模型对样本数据进行建模，分析居民三孩生育意愿的重要影响因素。在对总体样本建模分析后，将样本人群按性别、年龄和现有孩子数量的不同进行划分，分别运用 XGBoost 模型建模，分析不同性别、不同年龄段和不同现有孩子数量的人群的三孩生育意愿的重要影响因素，并对比分析影响不同人群三孩生育意愿的重要因素的差别。基于机器学习模型的输出结果以及借鉴已有研究经验，选取可能对居民三孩生育意愿有影响的相关变量，利用二项 Logit 模型建模分析居民三孩生育意愿的影响因素及其影响方向。最后，根据本文研究结果，为提升居民生育意愿、改善人口结构提出针对性建议。

1.2.2 研究内容及框架

全文总共分为 5 个章节：

第一章为绪论。首先讲述了本文的研究背景和意义，然后介绍本文的研究思路及方法，并详细介绍主要内容及本文结构框架，最后给出本文可能的创新之处。

第二章为相关理论与文献综述。首先介绍了生育相关理论，包括：成本与效用理论、理性行动理论和计划行为理论、孩子的数量与质量转换理论。然后阐述了生育意愿的内涵与测度，接着分析国内外研究现状，并对已有文献进行述评。

第三章为模型理论概述。包含基于树模型的机器学习模型和二项选择模型的介绍。对于机器学习模型，首先介绍了决策树模型，包含决策树模型的学习策略、树的分裂方法和常见的决策树生成算法，其中重点介绍了所用到的 CART 算法。然后介绍了 XGBoost 模型，由于 XGBoost 模型是基于决策树模型的集成模型，属于提升方法，所以在该小节中介绍了提升方法，接着引入提升树和梯度提升算法，再阐述 XGBoost 模型的原理。同时该小节还介绍了 XGBoost 模型如何计算输出重要变量对模型建模的贡献得分，得到特征重要性。该方法可被用于寻找对分类决策影响最大的重要特征，即对三孩生育意愿影响的重要因素。对于二项选择模型，首先介绍二项选择模型的模型设定，从离散选择模型引入二项选择模型，并详细介绍本文采用的二项 Logit 模型，接着对二项 Logit 模型的参数估计和模型检验进行详细介绍。

第四章为影响居民三孩生育意愿因素的实证分析。首先对本文研究所用数据集的来源和数据处理情况进行介绍，同时详细讲解数据中变量的含义及类型，接着选取部分变量进行描述性统计分析。然后，利用经过预处理的样本数据进行建模分析，分别采用决策树模型和 XGBoost 模型对全部样本数据建模，分析三孩生育意愿的重要影响因素，接着划分出不同数据集，分别利用 XGBoos 模型建模，分析不同人群的三孩生育意愿的重要影响因素，并进行对比分析。最后，在借鉴机器学习模型输出结果的基础上，选取相关变量，建立二项 Logit 模型分析居民三孩生育意愿的影响因素及其影响方向。

第五章为研究结论、建议以及不足与展望。首先总结前一章实证分析中的各种结论，然后根据研究得出的结论结合实际情况给出相应建议，最后指出本文研究中的不足之处以及相关展望。

1.3 本文可能的创新

本文可能的创新之处主要有以下几点：

（1）目前对生育意愿的影响因素研究方法更多集中于传统社会科学分析法，实证分析过程大多采用描述性统计分析和 Logistic 回归方法。且以往的研究更多依赖学者背景，针对较单一的方向，采用较少的变量研究对生育意愿的影响。本文结合统计与机器学习模型方法对生育意愿影响因素进行研究，首先采用决策树模型和 XGBoost 模型从更高的维度对影响居民生育意愿的因素进行挖掘，挖掘出生育意愿的重要影响因素，再利用二项 Logit 模型研究重要影响因素的影响方向。

（2）现有研究多为观察某些因素对某类人群是否有影响，很少有各类人群的生育意愿影响因素的对比分析，本文从男性与女性、青年与中年以及现有孩子数量不同角度对比分析影响不同人群三孩生育意愿的因素，并提出针对性建议。

（3）生育意愿跟生育成本和生育支持密切相关，目前我国的住房成本和教育成本被视为生育成本的主要部分，关于住房对生育意愿的研究并不是很多，已有研究中有从房价视角对生育意愿进行分析的，但很少有研究涉及房产数量等具体住房情况对三孩生育意愿的影响分析。本文基于机器学习和统计模型方法共同发现，房产数量对居民的三孩生育意愿产生正向影响，根据结论提出相关政策建议。

2.相关理论与文献综述

2.1 生育相关理论

2.1.1 成本与效用理论

著名的经济学家与人口学家莱宾斯坦（Harvey Leibenstein）最早提出利用成本与效用来分析家庭生育决策，他在发表的著作《经济落后与经济增长》中提出，人们的意愿生育子女数量取决于孩子的成本与孩子的效用的对比关系。其中，孩子的成本分为直接成本和间接成本，直接成本指父母从怀有孩子开始到孩子生活自立期间所需花费的各种抚育费用；间接成本指父母由于抚育新增小孩所受到的教育和收入等机会损失。孩子的效用是父母从孩子身上获得的满足与收用，主要包括：消费效用、劳动-经济效用、经济风险效用、维持家庭地位的效用、对家庭的扩展作出贡献的效用和生活保障效用。莱宾斯坦从“经济人”假设出发，利用边际孩子效用分析父母是否作出新增孩子的决策，他认为孩子的边际效用同一般商品或劳务一样，遵循“边际效用递减规律”，且家庭的生育行为是符合经济理性和追求效用最大化的。父母通过比较第 n+1 个孩子（边际孩子）的效用与成本来做抉择，当边际成本大于边际效用时，决定不生育下一个小孩；当边际成本小于边际效用时，决定生育下一个小孩；当边际成本与边际效用相等时，是否生育下一个小孩则取决于随机因素。

此外，在莱宾斯坦的研究中还作出了收入变动下的孩子成本与效用分析。他认为，当家庭收入不断提高时，该家庭养育孩子的成本和获得的孩子所提供的效用会随之变化。具体表现为，当家庭收入水平上升，直接成本明显上升，间接成本也由于父母时间价值的增大而上升，但边际孩子的效用随着收入水平的提高反而下降，此时，家庭意愿生育数量会减少。

2.1.2 孩子的数量与质量转换理论

1960 年贝克尔（Becker）把古典经济学研究方法引入到家庭经济学中，系统探讨了家庭、生育等问题，他把孩子理解成耐用消费品，利用消费者行为理论来分析家庭生育决策。贝克尔认为，与一般消费品不同，孩子只能由家庭自己生产且生产具有不确定性，对于孩子的消费决策需要确定孩子的质量和生育孩子的数量，且对于孩子质量的选择会受到社会压力的影响。贝克尔为生育经济分析奠定了理论基础，此后，他和同事进一步提出了孩子的数量—质量模型，解释了孩子的数量与质量之间的替代关系。在贝克尔等人的研究中，增加孩子数量会造成质量成本的提高，而提高孩子质量会增加数量的影子价格，家庭会根据孩子的数量与质量的替代关系作出最有利的生育决策。随着收入水平的提高，家庭倾向于更注重孩子的质量而非数量，父母愿意花费更多经济成本在孩子的质量上，因此，在经济发达的高生活水平地区，父母更愿意提升孩子质量而减少对孩子数量的追求，即经济水平提高会导致生育意愿降低。

2.1.3 理性行动理论与计划行为理论

理性行动理论自 1970 年以来在心理学与人口学研究中被广泛运用，它强调意愿是有意识形成的，所有行为皆有意，并潜在假设人都是理性的。避孕技术的发展与传播极大地拓展了人们的生育决策范围，生育决策中纳入了更广泛的因素，更趋于理性。Ajzen 等人认为，意愿作为行为的先兆，是特定行为结果可能性信念的函数，意愿受这些信念的影响，而这些信念受个人因素和社会因素的影响。个人因素是对行为的态度，包含情感、行为和认知等概念上的特征。社会因素是个人感知自己有动机去遵守其他重要人对自己是否应该执行某行为的判断，它被称为主观规范。社会网络以社会学习、社会支持和社会压力等机制影响生育意愿。

Ajzen 在理性行动理论的基础上提出了计划行为理论，该理论中加入了个人感知行为控制因素，用以表达人们受经验和预期影响而对行为的看法，该因素来自对资源与障碍的信念。一个很好理解的例子便是收入，在社会压力

下，富人可能认为自己无法承担生育孩子的成本，而相对不富的人反而认为自己有能力养活孩子。生育意愿的研究中，感知行为控制包含个人对自己实施生育行为的能力、机会和资源的评价以及对这些能力、机会和资源等重要性的评价。人们认为自己拥有的资源和机会越多，生育意愿越强烈。

2.1.4 生育意愿的内涵与测度

生育意愿的说法源于舶来的 fertility desire，国内关于其定义和概念的讨论有很多。顾宝昌（1992）认为生育意愿直观体现生育观，他强调生育具有时间、数量和性别的三维性。席薇等（2000）认为意愿可以被视为行为最直接的决定因素，生育意愿是人们对生育行为的最佳设计，它对一个国家或地区的生育水平与人口发展起着决定性作用。谭克俭（2004）认为生育意愿是指个人在生育子女方面的愿望和要求，体现在对生育孩子的数量、时间、性别、素质等方面的期望，其中以数量为最重要的内容。郑真真（2011）认为生育意愿是个人对子女的偏好，包括期望生育的子女数量、性别、生育时间和间隔。并且郑真真（2014）指出，生育意愿可以通过测量生育意向与生育计划、理想子女数和期望生育子女数三个方面来代替。郑佳音等（2019）通过调查陕西省育龄妇女的理想子女数量来研究其二孩生育意愿及影响因素。邱幼云（2022）把对生育子女数量的打算作为因变量来研究城市女青年的生育意愿，发现在三孩政策放开后，城市已婚女青年的生育意愿并没有显著提升，仍然很低迷。

显然，学者对于生育意愿概念的界定都很一致，即认为它是一种愿望、态度或看法，多数学者认为生育意愿涵盖了理想子女数量、孩子性别和生育时间，尤其是前两个维度。生育意愿虽然不同于生育行为，但是对后者具有引领甚至决定作用，没有政策等强制约束因素影响时，人们的实际生育行为常常取决于生育意愿，是否生育多孩，更是与意愿子女数量密切相关。

本文重点从数量角度探讨居民的三孩生育意愿，研究居民是否有三孩的生育意愿及其影响因素，借鉴已有文献中对二孩或三孩生育意愿的研究，意愿子女数量是三孩生育意愿的重要量化体现，本文采用意愿子女数量作为三孩生育意愿的量化指标。

2.2 国内外生育意愿研究现状

2.2.1 国外生育意愿影响因素研究现状

进入二十一世纪后，西方学者对生育意愿及其影响因素进行了广泛研究，相关影响因素的研究涉及社会、经济、文化、国家政策、个人身体心理状况、家庭状况以及战争等。

Agadjanian V 和 Prata N（2002）利用 1996 年进行的一次具有全国代表性的调查的数据，研究了战争因素对安哥拉人的生育意愿及生育率的影响。他们发现了战时生育率下降和战后生育率恢复的证据，但这些趋势变化很大，取决于战争的类型和程度以及妇女的社会经济特征。同时，他们的研究还表明，无论居住在战争地区还是非战争地区，受教育程度较高和比较富裕的人在控制生育以应对战争方面更有执行力。

Adsera A（2006）利用 1985 年和 1999 年西班牙生育率调查的数据，探索经济状况与人们生育意愿及行为之间的联系。研究结果表明，西班牙劳动力市场紧缩和经济状况恶化是造成西班牙真实生育数量极大低于希望数量的重要决定因素。并且，研究发现 20 岁左右面临高失业率的女性往往会将生育率限制在理想水平以下。

Weeden J（2006）等学者对普通美国人和美国精英大学的毕业生进行调查研究。通过研究分析发现，无论针对哪种性别的个体，更高的教育程度与个体推迟生育都紧密相关，而且更高的教育程度与较低的生育率之间也有着一定程度的相关性。

Aassve A（2008）研究了欧洲的人口变化和政策因素对出生率的影响。他认为，政府政策的干预是欧洲人口出生率的持续下降的重要原因，同时，这也是各国实际人口行为差异的一个重要原因。新的和现代的人口行为形式，如同居、非婚生育和离婚，都与较高的出生率有关，但这些人口行为与政府的支持和福利提供的组织方式密切相关。

Billari F C（2009）等学者对保加利亚进行案例研究和抽样调查后发现，个人思想举止和对社会压力的态度与生育意愿显著相关，并且前者对生育意愿的影响更显著。同时，他们从社会心理学角度对生育行为进行研究后认为，

社会经济方面的、观念性的、心理方面的以及其它基于社会资本的因素是影响生育意愿的背景因素。

Lyngstad T H 和 Prskawetz A（2010）使用纵向人口范围的挪威行政登记数据，利用连续时间风险模型研究兄弟姐妹的生育决策、教育程度、收入和婚姻史等因素对个体生育决策的影响。研究结果表明，个体的生育决策不仅受自身特点和人生轨迹的影响，还受到与他人的社会交往的影响。兄弟姐妹的生育决策、教育程度、收入和婚姻史等因素对女性生育第一胎的决策影响较强，但对女性生育第二胎的影响相对较弱。

Behrman J A（2015）研究了学校教育对妇女期望生育率的影响。他认为，受教育程度的提高降低了妇女理想的家庭规模和期望生育率。

Erfani A（2017）利用 2012 年伊朗德黑兰生育意愿调查的数据，考察了影响生育意愿的直接因素。调查结果显示，德黑兰超过一半的年轻已婚成年人打算不生育孩子。他认为，态度和规范压力是影响生育第一个孩子意愿的主要因素，而生育第二个孩子的意愿主要受到态度和感知约束的影响。

Mönkediek B 和 Bras H（2018）探讨了家庭制度与生育意愿的联系。他们认为，家庭制度与生育意愿之间存在重要联系。家庭制度通过影响人们对孩子的态度和他们对现有生育标准的看法来确定人们的意图。这种影响一部分是通过影响家庭规模，另一部分是通过影响人们对生育要求的观念来实现的。

Kato T（2018）调查研究了日本育龄单身男女的性别角色态度和生育意愿之间的关系。他认为，与传统态度相比，对收入和家务持平等态度的日本男性和女性更倾向于有较低的生育意愿。然而，日本社会并没有摆脱传统的劳动分工，要想扭转低生育率形势，有必要通过制度支持来使人们的家庭与工作得以平衡。

Chen S M（2019）等学者通过逐步回归分析认为，育儿乐趣、健康风险、兄弟姐妹间的相互关心、家庭的兴旺、时间压力和机会成本对第二胎的生育意愿有显著的预测作用。与相对生育成本相比，生育福利对第二胎生育意愿的影响更大，生育效益比相对生育成本更应引起重视。

Jeon S（2021）等学者用定量分析方法研究了韩国新婚夫妇（婚后五年内）的生育意愿及其影响因素，他们发现非都市家庭和租房家庭的生育意愿较高，预期购房时间与生育意愿之间也存在显著关系。

Owoo N S 和 Lambon-Quayefio M P（2022）认为女性在工作和家庭之间的权衡很大程度上取决于一个国家的制度环境，他们利用加纳这一发展中国家的生活水平调查（GLSS）的数据进行研究，发现不确定的工作环境鼓励更高的生育率，这是由于拥有更多子女可以使未来更安全和更可预测。

2.2.2 国内生育意愿影响因素研究现状

目前国内研究生育意愿的文献数量较多，由于生育政策的多次调整和社会经济的不断发展，育龄人群的观念发生了变化，生育意愿也在变化。每次我国生育政策进行调整和修改后，都会有很多学者对新生育政策下的生育意愿进行研究。从研究方法上看，现有研究主要采取描述统计、Logistic 回归和负二项回归等方法。从研究内容上看，国内学者对生育意愿影响因素的研究多基于人口学特征和社会经济因素等。

从人口学因素来看，性别方面，周福林（2005）认为由于生理差别、付出时间和精力不同等原因，男性无论在意愿生育子女数量还是意愿生育性别比方面均高于女性。陈志华等（2014）认为性别、婚育状况和周边生育情况是对生育意愿影响较大的三个因素，女性比男性想少生育。谭江蓉（2018）认为全面二孩政策下，男性比女性流动人口的二孩生育意愿强烈，流动人口中女性打算生育二胎的可能性低于男性。年龄方面，石智雷等（2014）认为育龄妇女年龄越大，想要生二孩的比重越低。但牛亚冬等（2015）认为妻子年龄越大，生育二孩的可能越大。受教育程度方面，胡樱子等（2018）认为女性受教育程度的提升将导致女性生育意愿降低，女性受教育程度提高是中国生育率降低的影响因素之一。但是，任义科等（2016）认为受教育程度与收入水平息息相关，受教育程度越高的被访者收入水平相对较高，其生育二孩的意愿越强。婚姻家庭方面，任远等（2022）认为婚姻质量高的人口具有更高的平均期望生育数量，并且，在已经拥有了一个孩子的家庭中，婚姻质量对多孩生育意向具有积极影响。李晶（2019）认为幼儿照料问题是阻碍大部分家庭生育二孩的首要因素，隔代抚育在经济、劳务及精神支持等方面对家庭生育二孩意愿有着积极影响。田立法等(2017）认为第一胎为女孩的农村居民比第一胎为男孩的居民，有着更强的二胎生育意愿。工作和个人经历方

面，洪良华等（1984）认为服务性行业和工农业的已婚青壮年拥有更强的生育意愿，并且这三个行业的已婚青壮年的多育动机较强，而文教、机关的已婚青壮年的少育动机较强。周靖祥（2014）认为与职业相关的各项指标是影响女性生育意愿的重要因素，由于女性常常会在工作与家庭之间进行权衡，职业类型和单位类型均在不同程度上影响着女性的生育意愿。张勇等（2014）认为从个人迁移经历来看，丈夫和妻子均无迁移经历的城镇家庭更可能不愿意生育二孩。王殿玺等（2019）认为从代际职业流动的角度来看，子代相对于父代职业地位的上向变化影响着人们的生育意愿，即社会流动机制塑造着个体的生育意愿，可能的解释是子代为了获得比父代更好的上向流动机会和社会经济地位，而试图摆脱父代的大家庭模式，进而压缩和消解个体的生育意愿。

从社会经济因素来看，汤兆云等（2012）认为经济收入对生育意愿的影响有限，经济收入的提高对生育意愿变化的作用有限。范世明（2015）认为在经济欠发达地区，不同代际居民的生育意愿不存在较大的差异。而段洪波等（2014）认为经济状况与生育孩子数量之间存在着高度负相关关系，良好的经济状况和稳定的经济来源都对生育子女数量产生影响，且后者影响更大。但周晓蒙（2018）认为家庭经济地位提高将引起其意愿生育数量增多、不愿生育的概率减小，不过女性工资收入占比会反向影响家庭意愿生育数量。此外，住房和社会保障因素也被视为可能影响生育意愿的影响因素。杨克文（2019）认为房价上涨对生育意愿具有显著的负面影响，不过，高收入家庭因为房价上涨而受到的负面影响要小于低收入家庭。于勇等（2022）认为社会保障对农民三孩生育意愿具有显著影响，具体而言，医疗保险对农民三孩生育意愿产生“挤入”效应，而养老保险对农民三孩生育意愿产生“挤出”效应。王天宇等（2015）考察了新型农村合作医疗制度的建立对居民生育意愿的影响，认为参加新农合降低了居民的生育意愿。潘丹等（2010）针对农村家庭进行研究，认为家庭持久总收入对于农村家庭的生育意愿具有负向影响，但是该负向影响会由于收入结构变量的加入而降低。家庭中有医疗保险的人数比例和妇女的生育意愿成反向关系。

从其他影响因素来看，罗蓉（1996）认为家族文化、村落文化对农民的生育决策发挥着重要的作用，农民的生育决策在很大程度上受到风俗习惯、

社会规范和心理规范这些文化因素的影响。王化波（2005）认为虽然朝鲜族育龄妇女的自身因素和经济社会发展在一定程度上影响其生育意愿，但是朝鲜族独特的文化对其生育意愿的影响起着决定性作用。李波平等（2010）重点研究了 60 年代、70 年代、80 年代不同代际出生的育龄妇女的生育意愿，他们认为不同代际的育龄妇女的生育意愿存在显著差异。研究结果表明：传统思想和户口性质一直在影响着妇女的生育意愿，时代变化过程中，妇女生育意愿受父母意愿的影响逐渐减弱，受到教育和职业因素的影响逐渐显著。许多学者也从自己关注的角度研究了生育意愿的影响因素，梁土坤（2018）认为对流动人口来说，流入地城市生活水平的满意度、定居意愿等心理适应因素对其生育意愿具有显著反向作用，而居住质量、与本地人交往的频繁程度等社会适应对其生育意愿却具有显著的正向作用。彭铿（2020）认为幸福感越高的家庭生二孩的意愿更强烈，并且，农村女性认为社会越公平则会趋向于生二孩，而非农村女性认为社会越公平则会推迟生育二孩。李娟等（2022）认为生育意愿与公共服务满意度有关，公共服务满意度特别高或特别低的居民的生育意愿较高，而公共服务满意度处于中间水平的居民生育意愿较低。李红阳（2022）考察了非正规就业对已婚女性生育意愿的影响，认为非正规就业显著提高了已婚女性的生育意愿。黄君洁等（2022）认为社交网络上的亲子信息对居民的生育意愿产生影响，正向信息可以提高生育意愿，而负向信息会降低生育意愿。

2.2.3 对已有研究文献述评

通过梳理分析国内外已有文献，可以发现关于生育意愿的相关研究已有很多，但是受早期计划生育政策的影响，曾经国内对生育意愿的研究并没有很多，随着“二孩政策”的放开，我国关于生育意愿的研究逐渐增多，其中有许多是从“二孩”或“全面二孩”的政策背景对生育意愿进行研究。

从目前的研究内容看，多数学者基于自身背景选取较少变量进行研究，且研究方向较为单一，研究多为观察某些因素对某类人群是否有影响，很少有各类人群的生育意愿影响因素的对比分析。影响居民生育意愿的因素非常多，包括年龄、户口、性别、经济状况、健康状况等。生育意愿也跟生育成

本和生育支持密切相关，目前我国的住房成本和教育成本被视为生育成本的主要部分，关于住房对生育意愿的研究并不是很多，已有研究中有从房价视角对生育意愿进行分析的，但很少有研究涉及房产数量等具体住房情况对三孩生育意愿的影响分析。从研究方法上看，学者在研究生育意愿时多采用传统的统计模型和方法，例如：描述性统计分析和 Logistic 回归方法。如今机器学习算法模型在各个领域被广泛应用，它能很好地处理高纬度和线性不可分数据，非常适合运用于分类数据的判别。

因此，在新时代新的生育政策背景下，本文尝试结合统计与机器学习模型方法对居民的三孩生育意愿进行研究，分析居民三孩生育意愿的差异性受何种因素影响，以期为政府解决人口结构问题、提高生育意愿提出一点建议。

3.模型理论概述

3.1 基于树模型的机器学习模型

3.1.1 决策树

决策树是一种基础且常见的分类与回归方法。决策树模型具有易于理解、可解释性强且分类速度快等优点，在现实生活的不同领域中，决策树及以树模型为基础的集成算法模型被广泛应用。决策树模型呈树形结构，是一种有监督学习模型。在分类问题中，它主要采用 if-then 规则，基于特征对实例进行分类，可以把它看成从训练数据集中总结的一系列规则的集合。决策树模型的生成伴随着特征选择和树的生成与修剪，需要在损失函数意义下寻求局部和全局最优，我们选择的模型希望是能够很好地拟合训练数据，且对未知数据拥有良好预测效果的。

特征选择是决策树学习中非常关键的一步，它在于选取对训练数据具有分类能力的特征。特征选择需要在样本的每一个特征上计算查看，选出一个最适合的作为树分裂的当前结点的划分特征，随着划分过程不断进行，要求决策树的分支结点尽量包含同一类别的样本，用“纯度”描述这一现象，即选择的特征要比其他特征更能实现高“纯度”。决策树模型的学习与建模过程可以采用不同的算法，这些不同的决策树算法在特征选择时采用不同的准则与计算方法，分类效果的优化衡量不同。常见的决策树算法主要有两类，一类是 ID3 算法和 C4.5 算法，它们的计算中引入了信息熵的概念，ID3 算法应用信息增益准则选择特征，C4.5 算法与 ID3 算法相似，但进行了一些改进，在生成过程中利用信息增益比选择特征；另一类是 CART 算法，在分类任务中，CART 算法采用基尼指数（Gini index）作为划分标准，克服了 ID3 算法在算法执行的过程子集非常纯时失效的缺点。与前两种算法只可以处理分类

任务不同，CART 算法能够处理分类和回归两种任务，并且它既可以处理离散属性，也可以处理连续属性。

CART 是由 Breiman 等人在 1984 年提出的应用广泛的决策树学习方法，全称为 classification and regression tree，即分类与回归树。本文探索的居民三孩生育意愿问题为二分类问题，所以所建的树为分类树，下面介绍 CART 树模型生成分类树的基本思路。

首先需要明确最优特征及其最优分割点的划分准则，那便是基尼指数。具体问题中，假设有M个类，样本点属于第m类的概率为p_m，则概率分布的基尼指数的计算公式为：

$$Gini(p) = \sum_{m=1}^{M} p_m(1-p_m) = 1 - \sum_{m=1}^{M} p_m^2 \qquad (3-1)$$

对于二分类问题，若记样本点属于第 1 个类的概率为p，则概率分布的基尼指数为：

$$Gini(p) = 2p(1-p) \qquad (3-2)$$

对于给定的样本集合D，其基尼指数为：

$$Gini(D) = 1 - \sum_{m=1}^{M} \left(\frac{|C_m|}{|D|}\right)^2 \qquad (3-3)$$

其中，C_m是D中属于第m类的样本子集，M是类的个数。

如果样本集合D根据特征A是否取值为a被分割成D_1和D_2两部分，即

$$D_1 = \{(x,y) \in D | A(x) = a\}, D_2 = D - D_1$$

则在特征A的条件下，集合D的基尼指数为：

$$Gini(D,A) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2) \qquad (3-4)$$

在分类树生成前，算法需要提前输入训练数据集并给定计算停止的条件，从树的根结点开始到生成各分支结点都需要利用式（3-4）计算每个特征在不同取值时的基尼指数，对比基尼指数值，选定最小的基尼指数值所对应的特征及分割点，随着特征结点确定和树的不断分支，每个结点下不同特征取值范围所包含的样本个数不断减少，当其数量少于预设值或基尼指数小于预设值时，算法计算停止，分类决策树生成。

3.1.2 XGBoost

在介绍 XGBoost 前首先介绍提升方法，提升（boosting）方法的基本思想为：面临复杂任务时，很难找到顶级专家来解决问题，把许多专家的解决方案进行综合，得到的效果往往好于一个专家的方案。可以理解为“三个臭皮匠顶个诸葛亮”。

对于复杂分类问题而言，寻求强分类器是非常难的，而寻求弱分类器则简单许多。提升方法以弱学习算法为基础，通过反复学习来得到一系列弱分类器，然后组合这些弱分类器，构成一个强分类器。它采用基本分类器组成的加法模型，学习算法为前向分布算法，以决策树为基函数的提升方法被叫做提升树（boosting tree）。提升树模型可以表示为多个决策树相加的加法模型：

$$f_M(x) = \sum_{m=1}^{M} T(x;\Theta_m) \tag{3-5}$$

其中，Θ_m为决策树的参数，$T(x;\Theta_m)$表示决策树，M为树的个数。

具体算法中，当确定了初始提升树$f_0(x) = 0$后，第m步的模型为：

$$f_m(x) = f_{m-1}(x) + T(x;\Theta_m) \tag{3-6}$$

其中，$f_{m-1}(x)$为当前模型，然后通过经验风险极小化来确定下一棵决策树的参数Θ_m：

$$\widehat{\Theta}_m = \underset{\Theta_m}{\operatorname{argmin}} \sum_{i=1}^{N} L\big(y_i, f_{m-1}(x_i) + T(x_i;\Theta_m)\big) \tag{3-7}$$

提升树被视为高功能学习算法，无论是否需要应对复杂的数据传输关系，树的线性组合对训练数据都会有良好的拟合效果。针对不同问题的提升树学习算法使用的损失函数不同，主要有平方误差损失函数、指数损失函数和一般损失函数。为了解决损失函数是一般损失函数时的优化困难问题，Freidman 提出了梯度提升（gradient boosting）算法。该算法利用损失函数的负梯度在当前模型的值

$$-\left[\frac{\partial L\big(y, f(x_i)\big)}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$$

作为残差的近似值，拟合一个回归树，核心思想是每轮通过拟合残差来降低损失函数。

XGBoost 是提升算法的一种具体实现，其算法是对传统的 GBDT（梯度提升树）的再次优化，使用牛顿法求解损失函数极值。XGBoost 在 GBDT 的基础上的最大优化是改写了目标函数，它显示的把树模型复杂度作为正则项加到优化目标中。XGBoost 训练时的目标函数由两部分构成，可以表示为：梯度提升算法损失+正则化项（模型复杂度），即：

$$Obj = \sum_{i=1}^{m} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (3-8)$$

其中，Obj代表目标函数，m为训练函数样本量，L是对单个样本的损失，为真实值y_i与预测值$\hat{y}_i$间的误差。正则化项代表模型的复杂程度，$\Omega(f) = \Upsilon T + \frac{1}{2}\lambda\|\omega\|^2$，其中$\Upsilon$和$\lambda$为人工设置的参数，$T$表示叶子结点数，$\omega$表示所有叶子结点值构成的向量。

本文重点运用 XGBoost 的特征重要度计算方法，在寻求特征重要性排序方面，XGBoost 是被广泛运用的模型。在决策树的分裂中，被选择作为树结点的特征被视为具有分类能力的特征，模型训练过程中会通过记录各特征的分裂次数和平均增益等信息来对特征的重要性进行量化。XGBoost 模型中包含着许多决策树，对特征重要性的量化用到类似思想，其评判中会用到特征重要性指标 Gain 得分：

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \qquad (3-9)$$

其中$\frac{G_L^2}{H_L+\lambda}$代表分裂后左子树的分数，$\frac{G_R^2}{H_R+\lambda}$代表分裂后右子树的分数，$\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$代表不分裂可以得到的分数，$\gamma$代表增加新叶子带来的模型复杂度变化。

随着模型的迭代生成，每个特征被选择的次数和特征结点分裂后的 Gain 得分都被记录，最终计算出平均值，便得到了不同特征对模型的贡献度，即特征重要性。

3.2 二项选择模型

3.2.1 二项选择模型的模型设定

在多数经济计量模型中，被解释变量常被假定为连续变量，然而现实生活中，我们往往会面临许多具有选择性的问题，我们要在有限的方案中进行抉择，这时的因变量只能是有限多个离散的数值。当被解释变量是离散的，而非连续的时，我们需要用到离散选择模型（Discrete Choice Model，DCM）。

离散选择模型的一般原理为随机效用理论，即当决策者i需要从J个方案中作选择时，其对其中的一种选择方案j的偏好可以用被选择对象的效用值U_{ij}来表示，该效用值由被选方案的属性X_j、决策者的特征X_k和不能直接观测的随机部分U_{ij}来描述。虽然U_{ij}是一个未知的函数，但在大多数应用中可以将其假设为线性函数：

$$U_{ij} = V_{ij} + \varepsilon_i = \beta_{ij}X_{ij} + \beta_{ik}X_{ik} + \varepsilon_{ij} \tag{3-10}$$

其中，V_{ij}表示可观测效用，也叫固定效用。此外，由经济学的基本假定，决策者总是追求效用最大化的，即当决策者i选择方案j时，有

$$U_{ij} \geq U_{id}(d \in J, d \neq j) \tag{3-11}$$

因此，根据随机效用理论和效用最大化原则可以将决策者i选择方案j的概率P_{ij}表示为：

$$\begin{aligned} P_{ij} &= Prob\left(V_{ij} + \varepsilon_{ij} \geq V_{id} + \varepsilon_{id}\right), (d \in J, d \neq j) \\ &= Prob\left(V_{ij} - V_{id} \geq \varepsilon_{id} - \varepsilon_{ij}\right) \end{aligned} \tag{3-12}$$

在实际应用时，通常会将随机部分ε_{ij}和ε_{id}假设为独立同分布。

当离散选择模型的备选方案集中只有两个选项时，模型被称为二项选择模型。假设个体只有两种选择，例如有三孩生育意愿（$y = 1$）或没有三孩生育意愿（$y = 0$）。是否有三孩生育意愿通常受个体的年龄、学历等许多变量影响，假设这些变量都包括在向量$\boldsymbol{x}$中，如果直接采用线性概率模型，拟合的结果与现实不符。为了使y的预测值总是介于[0,1]之间，在给定$\boldsymbol{x}$时，考虑y的两点分布概率：

$$\begin{cases} P(y = 1|\boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta}) \\ P(y = 0|\boldsymbol{x}) = 1 - F(\boldsymbol{x}, \boldsymbol{\beta}) \end{cases} \tag{3-13}$$

其中，函数$F(\boldsymbol{x},\boldsymbol{\beta})$称为连接函数，它连接起被解释变量$\boldsymbol{y}$和解释变量$\boldsymbol{x}$。

对于二项选择行为，通常是通过“潜变量”来概括该行为的净收益。当净收益大于 0 时，$y=1$；当净收益小于 0 时，$y=0$。假设净收益为：

$$y^{*}=\boldsymbol{x}'\boldsymbol{\beta}+\varepsilon \tag{3-14}$$

其中，净收益y^{*}是潜变量，它不可观测。个体的选择规则为：

$$y=\begin{cases}1，若y^{*}>0\\0，若y^{*}\leq 0\end{cases} \tag{3-15}$$

因此，可以得到：

$$P(y=1|\boldsymbol{x})=P(y^{*}>0|\boldsymbol{x})=P(\boldsymbol{x}'\boldsymbol{\beta}+\varepsilon>0|\boldsymbol{x})=P(\varepsilon>-\boldsymbol{x}'\boldsymbol{\beta}|\boldsymbol{x}) \tag{3-16}$$

假设ε服从逻辑分布，则

$$P(y=1|\boldsymbol{x})=P(\varepsilon>-\boldsymbol{x}'\boldsymbol{\beta}|\boldsymbol{x})=P(\varepsilon<\boldsymbol{x}'\boldsymbol{\beta})=\frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})} \tag{3-17}$$

此时，便得到了二项 Logit 模型。即当连接函数$F(\boldsymbol{x},\boldsymbol{\beta})$为逻辑分布的累积分布函数时，可以推出二项 Logit 模型的一般形式：

$$P(y=1|\boldsymbol{x})=F(\boldsymbol{x},\boldsymbol{\beta})=\frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})} \tag{3-18}$$

3.2.2 二项选择模型的参数估计与模型检验

显然，二项 Logit 模型是一个非线性模型，可以利用最大似然法（MLE）来对其进行估计。

二项 Logit 模型估计出的系数并非代表边际效应，为了揭示二项 Logit 模型系数的经济意义，我们需将其转换为线性回归模型，即由

$$P(y=1|\boldsymbol{x})=p=\frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})} \tag{3-19}$$

转换得到

$$\ln\frac{p}{1-p}=\boldsymbol{x}'\boldsymbol{\beta} \tag{3-20}$$

其中，$\frac{p}{1-p}$被称为机会比 odds ratio(OR)，机会比的半对数模型可记为

$$\ln(OR)=\boldsymbol{x}'\boldsymbol{\beta} \tag{3-21}$$

具体解释中，$\boldsymbol{\beta_j}$表示$\boldsymbol{x_j}$对机会比的半弹性。可以认为固定其它因素不变，$\boldsymbol{x_j}$每

增加一个单位，平均来说y的机会比增加原来的$\boldsymbol{\beta_j}$倍。

判断二项 Logit 模型最常用的拟合优度指标是 McFadden 提出的R^2_{McF}，也被称为伪R^2，该指标的计算公式为：

$$R^2_{McF} = 1 - \frac{\ln L}{\ln L_0} \qquad (3-22)$$

其中，$\ln L_0$为仅包含常数项的模型的对数似然函数之最大值，而lnL为包含所有解释变量以及常数项的模型的对数似然函数之最大值。

此外，也可以计算“正确预测的百分比”来判断拟合优度。该方法将预测值与实际值进行比较，算出正确预测的百分比。计算方式一般是用预测对的样本数量除以总样本数量。

在二项选择模型中，被解释变量并非服从正态分布，因此传统的 t 检验和 F 检验在针对二项选择模型的显著性检验时失效，我们只能依赖于三大检验，即：似然比检验、Wald 检验和拉格朗日乘数检验。

4.影响居民三孩生育意愿因素的实证分析

4.1 数据来源与处理以及描述

4.1.1 数据来源

本文数据采用中国综合社会调查（Chinese General Social Survey，CGSS）项目发表的 2018 年数据，即 CGSS2018。CGSS 由中国人民大学联合全国各地的学术机构共同执行，是中国第一个全国性、综合性、连续性的大型社会调查项目。项目组在中国大陆各省市自治区进行问卷调查并生成报告数据，其问卷问题具有层次性，能够体现社会变迁趋势，设计科学严谨，在学术界受到广泛的认可。CGSS 数据被广泛应用于科研教学和政府决策中，是研究中国社会的重要数据来源，对推动国内科学研究的开放与共享具有重要意义。

本文采用的 CGSS2018 是现已发布的最新版调查结果。该数据集包含 12788 个原始样本和 1029 列原始变量。

4.1.2 数据处理

结合数据情况以及已有相关研究并参考 2018 年 CGSS 居民问卷，本文选取超过 90 个可能跟生育意愿相关的问题进行研究，探究这些问题里蕴含的变量指标与居民生育意愿之间的联系。已有的生育意愿相关文献中，研究者由于自身研究目的不同，选取的研究对象不尽相同，从 18-22 岁到 16-60 岁都有。CGSS2018 问卷针对 18 岁以上的居民展开调查，本文根据自身研究目的和研究内容，选取 18-60 岁人群作为研究对象，再对变量作如下处理：

首先，对变量进行编码。数据中变量值多为文字，为了保证建立模型时数据具有可读性，先将文字型变量进行统一编码，取值用数字代替。例如变

量“a2”代表性别，其变量值有“男”、“女”，将其变量值编码为数字“0”、“1”；变量“a35”代表社会公平感，其变量值有“完全不公平”、“比较不公平”、“说不上公平但也不能说不公平”、“比较公平”、“完全公平”，将其变量值编码为数字“1”、“2”、“3”、“4”、“5”。

然后，对缺失值进行处理。拥有缺失情况的变量只有 4 个，分别为变量“a7a”、“a14a”、“a21”、“a29”，其缺失比例分别为 0.07%、4.03%、0.04%、2.33%。其中代表 BMI 值的变量“a14a”的缺失值可以由代表身高和体重的相关变量值计算补全。本文主要采用基于树模型的机器学习方法进行建模，通常在数据缺失量不大的情况下，对于顺序型数据和分类型数据常分别采用中位数和众数来对缺失数据进行代替，会得到比较好的建模预测效果。所以针对其余三个变量，考虑模型和数据特点，代表受教育程度的变量“a7a”的取值为顺序型，变量缺失值用中位数代替；代表目前户口登记地的变量“a21”和代表最主要信息来源的变量“a29”，它们的取值都为分类型，变量缺失值用众数代替。

最后，对部分变量进行合并处理。例如：代表报纸使用频繁程度的变量“a281”、代表杂志使用频繁程度的变量“a282”、代表广播使用频繁程度的变量“a283”、代表电视使用频繁程度的变量“a284”、代表互联网（包括手机上网）使用频繁程度的变量“a285”和代表手机定制消息使用频繁程度的变量“a286”，它们的取值在之前编码时都被赋值为 0 到 4，现将各变量取值相加后除以 6 并四舍五入取整获得取值为 0 到 4 的代表媒体使用情况的新变量“a28”。同时，为了减少过拟合，在数据处理中对部分变量取值范围进行归纳缩减，例如：将代表受教育程度的变量“a7a”的 14 个不同水平的取值归纳为代表“没有受过任何教育”、“初等教育”、“中等教育”、“成人高等教育”、“普通高等教育”的五个取值。

经过以上数据处理并且删除无效样本后筛选出一个有 8136 个样本、73 个变量的数据集，为了后续模型输出结果的易理解对各变量进行重命名。具体数据处理流程如图 4-1 所示：

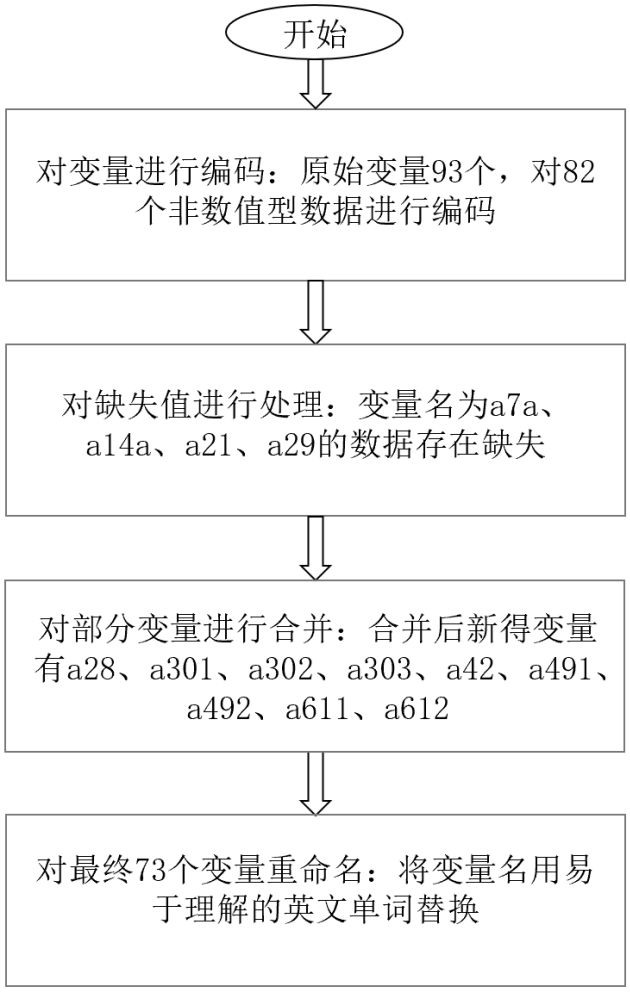

图 4-1 数据处理流程

4.1.3 数据描述

1、因变量

本文重点研究居民三孩生育意愿的影响因素，因此因变量为三孩生育意愿（fertility_desire）。该变量指标主要来自于 CGSS2018 问卷中的相关问题：“如果没有政策限制的话，您希望有几个孩子？”并将回答 0、1、2 视为未有三孩生育意愿，赋值为 0；回答 3 个及以上的视为具有三孩生育意愿，赋值为 1。在数据处理时剔除了回答“无所谓”、“不知道”和“拒绝回答”的异常样本。

2、自变量

本文自变量选自 CGSS2018 问卷核心模块中的 11 个模块，分别为受访者“社会人口属性”、“住房问题”、“健康”、“迁移”、“生活方式”、“阶层认同”、“政治参与行为与态度”、“个体认知能力”、“劳动力市场”、“社会保障”、“家庭”。表 4-1 展示了每个模块具体变量在建模过程中的变量名、变量解释和变量类型：

表 4-1 变量说明

模块	变量名	变量解释	变量类型
社会人口属性	gender	性别	分类型
	age	年龄	数值型
	nationality	民族	分类型
	religion	宗教信仰	分类型
	religi_activities	参加宗教活动的频繁程度	顺序型
	edu	受教育程度	顺序型
	student	是否在读	分类型
	income	个人全年总收入	数值型
	c_party	是否申请过加入中国共产党	分类型
	political	政治面貌	分类型
住房问题	living_space	目前住房套内面积	数值型
	house	个人拥有房产数量	数值型
健康	height	身高	数值型
	weight	体重	数值型
	bmi	BMI 值	数值型
	health	身体健康状况	顺序型
	exercise	体育锻炼每周次数	数值型
	myopia	是否近视	分类型
	h_effect	健康问题影响生活工作程度	顺序型
	depression	抑郁程度	顺序型
迁移	census_register	目前户口登记状况	分类型
	reg_place	目前户口登记地	分类型
	place_0	出生时母亲常居地	分类型
	reg_place_0	出生时户口登记地	分类型
	place_14	14 周岁时常居地	分类型
生活方式	media_use	媒体使用情况	顺序型
	message	最主要信息来源	分类型
	entertainment	娱乐休闲情况	顺序型
	sports_art	体育艺术休闲情况	顺序型
	together	聚会休闲情况	顺序型
	library	图书馆借阅	分类型
	phone	有无个人手机	分类型