目录

[1.选题背景和项目概述 2](#_Toc159163177)

[1.1选题背景 2](#_Toc159163178)

[1.2项目概述 2](#_Toc159163179)

[2.大数据平台 2](#_Toc159163180)

[3.数据采集及探索 2](#_Toc159163181)

[3.1基础数据探索 2](#_Toc159163182)

[3.2数据质量分析 2](#_Toc159163183)

[4.数据预处理及存储 2](#_Toc159163184)

[4.1数据库设计 2](#_Toc159163185)

[4.2数据清洗 2](#_Toc159163186)

[4.3数据变换 2](#_Toc159163187)

[5. 模型构建 2](#_Toc159163188)

[5.1建立模型 2](#_Toc159163189)

[5.2模型评估与调优 3](#_Toc159163190)

[6. 数据分析 3](#_Toc159163191)

[7. 功能模块介绍 3](#_Toc159163192)

[8. 成果可视化 3](#_Toc159163193)

[9. 项目总结 3](#_Toc159163194)

[10.参考文献 3](#_Toc159163195)

# 1.选题背景和项目概述

## 1.1选题背景

随着社会的发展和经济的增长，人口结构的变化对于一个国家的社会经济发展至关重要。人口出生率是人口结构变化中的一个重要指标，直接关系到一个国家的人口数量、劳动力资源、社会保障、经济增长等方面。在中国，近年来，人口出生率逐渐下降的趋势引起了广泛关注。人口出生率下降可能会导致人口老龄化加剧、劳动力减少、社会保障压力增加等问题，对经济社会的可持续发展产生不利影响。

为了更好地了解和预测人口出生率的变化趋势，以便及时制定相应的政策来应对人口结构变化带来的挑战，本项目旨在建立一个预测婴儿出生率的系统。通过分析历史数据和相关因素，构建预测模型，以预测未来每年的人口出生率，为政府部门和决策者提供科学的依据，指导人口政策的制定和调整，促进人口结构的合理调整，推动经济社会的可持续发展。

随着信息技术的发展，大数据技术在人口学研究中的应用日益广泛，可以帮助处理海量的人口数据，发现其中隐藏的规律和趋势。在国内，已有一些学者和机构开展了关于人口出生率预测的研究工作，其中一些研究采用传统的时间序列分析方法，结合经济因素、社会政策等因素进行预测；另一些研究则采用机器学习方法，如神经网络、决策树等，建立预测模型。在国外，人口学研究也是一个重要的研究领域，已有许多研究针对人口出生率进行了预测和分析，其中一些研究借鉴了传统的人口学理论和统计学方法，另一些研究则采用了先进的数据挖掘和机器学习技术，取得了一些有价值的成果。本项目将应用目前国内外比较先进的大数据和机器学习技术进行研究。

## 1.2项目概述

本项目应用分布式数据存储技术存储数据以及分布式数据计算技术进行分析。Hadoop是一个开源的分布式存储和计算框架，用于处理大规模数据。它基于Google的MapReduce论文提出的思想，能够将数据分布式存储在多台计算机上，并通过MapReduce模型进行并行计算。Hadoop的核心模块包括Hadoop Distributed File System（HDFS）和Hadoop MapReduce。HDFS是Hadoop中用于存储大规模数据的分布式文件系统。它将大文件分割成多个块，并存储在多台计算机上，保证数据的高可靠性和高可用性。HDFS采用主从架构，包括一个NameNode（主节点）和多个DataNode（数据节点），用于管理文件系统的命名空间和存储数据块。Spark是一个快速、通用的分布式计算系统，提供了比Hadoop MapReduce更高层次的抽象，支持更多种类的计算任务。Spark的核心是基于内存的计算引擎，能够将数据加载到内存中进行高效的数据处理和分析。Spark提供了丰富的API，支持多种编程语言（如Scala、Java、Python和R），使得开发人员可以方便地进行大数据处理和分析。本项目中使用PySpark作为Python编程的工具，它提供了与Spark集成的Python API，可以方便地在Python中进行大数据处理和分析。同时，也会使用Jupyter Notebook等交互式开发环境来进行代码编写、调试和展示。

# 2.大数据平台

Hadoop是一个开源的分布式存储和计算框架，其部署过程如下：

1. 首先使用homebrew软件管理工具安装Hadoop；
2. 然后设置JAVA\_HOME和HADOOP\_HOME环境变量，指向Java和Hadoop的安装路径；
3. 修改Hadoop配置文件：包括core-site.xml、hdfs-site.xml、mapred-site.xml、yarn-site.xml等。这些配置文件包括了Hadoop集群的各种参数，如文件系统的URI、数据块大小、副本数量、资源管理器地址等；
4. 配置SSH免密码登录：确保所有节点之间可以通过SSH免密码登录，以便于集群节点之间的通信和文件传输。
5. 格式化HDFS文件系统：在NameNode节点上执行hdfs namenode -format命令，初始化HDFS文件系统。
6. 启动Hadoop服务：依次启动Hadoop的各个服务，包括NameNode、DataNode、ResourceManager、NodeManager等。

# 3.数据采集及探索

## 3.1基础数据探索

本数据集包含样本总数45429条，标签数1条，特征总数53个。

A screenshot of a computer

Description automatically generated

数据中正负样本比例为1.057:1。

A red and blue stripes

Description automatically generated with medium confidence

数值型特征总数为30个，字符串型特征总数为23个。

## 3.2数据质量分析

空数据分析：

本数据集中所有数据都不为空，所有数据可用。

A screen shot of a computer screen

Description automatically generated

数值型特征分析：

通过查看数值型特征的均值、标准差、最大值、最小值、25分位数、50分位数以及75分为数，可以发现部分数值型特征分布差异明显，部分数值型特征分布差异不显著，因此有特征筛选的必要，后续将在数据清洗和数据变换部分详细介绍。

A black and white screen with numbers

Description automatically generated

A screenshot of a computer

Description automatically generated

字符串型（类别）特征分析：

将数据集分为正负样本分别查看类别型特征的分布，可以看出大部分特征都有2-3个类别，另外部分类别型特征标签差异较大，也为后续特征筛选提供了选择。

A screenshot of a black screen

Description automatically generated

# 4.数据预处理及存储

## 4.1数据库设计

原始文件为本地csv文件，本项目通过pyspark将原始文件读取后存储到hdfs，并基于数据量的考量分为2个partition。Hdfs中的文件以及数据结构保持不变，后续在数据清洗和变换过程中会生成新的中间文件。

## 4.2数据清洗

数据集中不含空数据，因此不需要对数据条目进行筛选。由于在数据质量分析中，发现特征与标签的相关性有肉眼可见差异，所以先进行相关性分析，来筛选有用特征。

首先画出相关性热度图如下，可以看出部分特征无效，可以将其筛选掉。

A screen shot of a graph

Description automatically generated

然后计算出具体的相关性数值，用于量化筛选特征，最终选择相关性大于0.05的特征进行模型训练。

A screenshot of a black and white screen

Description automatically generated

最终选择的特征总共39个，分别为：['FATHER\_COMBINED\_AGE', 'FATHER\_EDUCATION', 'MONTH\_PRECARE\_RECODE', 'CIG\_BEFORE', 'CIG\_1\_TRI', 'CIG\_2\_TRI', 'CIG\_3\_TRI', 'MOTHER\_HEIGHT\_IN', 'MOTHER\_BMI\_RECODE', 'MOTHER\_PRE\_WEIGHT', 'MOTHER\_DELIVERY\_WEIGHT', 'MOTHER\_WEIGHT\_GAIN', 'DIABETES\_PRE\_OHT', 'DIABETES\_GEST\_OHT', 'PREV\_BIRTH\_PRETERM\_OHT', 'NO\_INFECTIONS\_REPORTED', 'LABOR\_IND\_OHT', 'LABOR\_AUGM\_OHT', 'STEROIDS\_OHT', 'ANESTHESIA\_OHT', 'APGAR\_5\_RECODE', 'APGAR\_10', 'APGAR\_10\_RECODE', 'OBSTETRIC\_GESTATION\_WEEKS', 'INFANT\_WEIGHT\_GRAMS', 'INFANT\_ASSIST\_VENTI\_OHT', 'INFANT\_ASSIST\_VENTI\_6HRS\_OHT', 'INFANT\_NICU\_ADMISSION\_OHT', 'INFANT\_SURFACANT\_OHT', 'INFANT\_ANTIBIOTICS\_OHT', 'INFANT\_SEIZURES\_OHT', 'INFANT\_NO\_ABNORMALITIES', 'INFANT\_ANCEPHALY\_OHT', 'INFANT\_MENINGOMYELOCELE\_OHT', 'INFANT\_LIMB\_REDUCTION\_OHT', 'INFANT\_DOWN\_SYNDROME\_OHT', 'INFANT\_SUSPECTED\_CHROMOSOMAL\_DISORDER\_OHT', 'INFANT\_BREASTFED\_OHT']

## 4.3数据变换

对于字符串类型的特征，模型无法识别字符串含义，因此需要进行编码，最常见的编码方式主要有以下两种。Label Encoding：将类别型特征映射为连续的整数编码，每个类别对应一个唯一的整数值。One-Hot Encoding：将类别型特征转换为二进制的稀疏向量表示，每个类别对应一个新的特征，仅有一个特征取值为1，其余为0。One-Hot Encoding消除了类别之间的大小关系，避免了模型的误导性，使得模型更加准确地学习和理解分类变量之间的关系，并且稀疏矩阵结构可以更好地表示分类特征，适用于大多数机器学习模型。因此此项目选用One-Hot Encoding。

从上述分析中可以看出，类别型特征的类别数普遍在2-3个，因此可以进行One-Hot Encoding编码并且不会对特征纬度产生巨大影响从而影响模型性能。

首先将类别型特征进行数字化编码，然后进行One-Hot Encoding。

A screen shot of a computer code

Description automatically generated

# 模型构建

## 5.1建立模型

本项目研究对象为婴儿存活率预测，因此可以定义为二分类任务。对于二分类任务，常用的机器学习模型有Logistic Regression（逻辑回归）、Decision Tree（决策树）、Random Forest（随机森林）和Gradient Boosting Trees（梯度提升树）。

Logistic Regression是线性模型，用于二分类问题，输出类别概率；Decision Tree基于树形结构的分类模型，易于理解和解释；Random Forest是集成学习方法，通过多棵决策树进行投票或平均来进行预测，准确性高；Gradient Boosting Trees也是集成学习方法，通过迭代训练多棵决策树来逐步减小损失函数，效果优秀。

本项目将通过pyshark的ml模块实现以上模型，并进行研究和分析。

## 5.2模型评估与调优

将数据集划分为训练集和测试集，比例为7:3，其中训练集有31978条数据，测试集有13451条数据。

使用相同的训练集分别训练4个模型，并在相同的测试集上进行测试，比较其预测结果的准确率。

其中Logistic Regression模型迭代10次，Decision Tree的最大深度为3层，Random Forest的最大深度为3层，Gradient Boosting Trees的最大深度为3层，其余均使用pyspark的ml模块的默认参数。

# 数据分析

实验结果如下：

A screen shot of a computer

Description automatically generated

实验结果分析：Decision Tree的准确率最低为0.954，其次是Logistic Regression模型为0.965。准确率最高的是Gradient Boosting Trees模型，准确率达到0.977，其次是Random Forest模型准确率为0.973。

从实验结果中可以看到，4个模型在本数据集上的表现都达到了比较高的准确率，从准确率来看，随着模型复杂度的提升，准确率也会随之提升，但是因为数据集复杂度并不高，所以准确率提升效率有限。

# 功能模块介绍

在init.ipynb文件中，主要通过pyspark实现本地文件与分布式文件系统hdfs的交互，并将原始数据分区存储到hdfs中。

在explore.ipynb文件中，主要通过pandas进行数据探索和特征分析，主要进行了数据质量分析、数据类型分析、数值型特征分析、类别型特征分析以及相关性分析，并进行了特征筛选。

在main.ipynb文件中，主要通过pyspark读取数据，并且基于特征分析的结果进行了特征筛选和，特征工程和数据集划分，并且通过ml模块实现了4中模型并进行训练和测试，最终比较测试结果。

# 成果可视化

在初始模型的基础上，比较迭代次数对模型准确率产生的影响，可以发现，随着迭代次数增加，Random Forest模型准确率随之提升，说明其有较好的抗过拟合能力和鲁棒性；Logistic Regression模型的提升并不明显；Decision Tree和Gradient Boosting Trees模型则产生了过拟合现象。

A graph with lines and points

Description automatically generated

# 项目总结

本项目使用了大数据技术对新生儿存活率进行分析，主要是在Hadoop分布式系统上，通过pyspark和pandas对数据进行分析，并使用机器学习技术进行建模和预测，达到了97.8%的准确率。本项目充分体现了大数据技术对数据分析和建模的有效作用，并给机器学习技术在其他领域的应用提供了参考。

# 10.参考文献

Chen, Yongjie. (2023). Model Analysis on the Birth Rate in China. Advances in Economics, Management and Political Sciences. 58. 220-227. 10.54254/2754-1169/58/20230915.

Haque, A., Hossain, T., & Nasser, M. (2015). Predicting the number of children ever born using logistic regression model. Biom Biostat Int J, 2(4), 00034.

Ahmadi, P., Alavimajd, H., Khodakarim, S., Tapak, L., Kariman, N., Amini, P., & Pazhuheian, F. (2017). Prediction of low birth weight using Random Forest: A comparison with Logistic Regression. Archives of Advances in Biosciences, 8(3), 36-43.