# 基于统计建模的糖尿病患者再入院风险与住院天数分析

## 统计分析与建模课程项目汇报

**小组成员**：

* 姜政言2353594
* 陈奕玮2350217
* 龙承佑2354176

## 摘要

本项目利用统计建模方法，对美国130家医院的糖尿病患者临床数据进行了深入分析，旨在解决医疗管理中两大核心痛点：预测患者的住院时长以优化床位资源配置，以及评估患者出院后短期内的再入院风险以改善医疗质量。通过对原始数据进行分层抽样、清洗及特征工程，我们构建了多元线性回归模型用于住院天数预测，以及基于类别权重优化的Logistic回归模型用于再入院风险分类。研究结果显示，既往住院史和多病共存是导致医疗资源消耗和再入院风险增加的关键驱动因素。最终，我们开发了一套基于R Shiny的交互式原型系统，能够直观展示风险评估结果，为临床辅助决策提供数据支持。

## 1. 项目背景与研究目标

### 1.1 项目背景

糖尿病作为一种全球性的慢性疾病，其复杂的并发症和长期的治疗周期给医疗卫生系统带来了沉重的经济与资源负担。在临床管理实践中，医疗管理者主要面临以下严峻挑战：

首先，医疗负担沉重。糖尿病患者往往伴随多种并发症，导致其再入院率显著高于普通患者群体，这不仅消耗了巨大的医疗资源，也增加了患者的经济压力。

其次，风险识别困难。目前的临床评估往往依赖医生的个人经验，难以准确、量化地识别出哪些患者在出院后极有可能在短期内再次入院。这种预测的缺失导致医院难以针对高危人群实施有效的干预措施。

最后，资源分配低效。由于缺乏对患者住院时长的准确预估，医院床位调度和护理资源分配往往处于被动应对状态，缺乏数据支持的精细化管理。

### 1.2 研究目标

基于上述背景，本项目旨在挖掘医疗大数据的价值，通过构建稳健的统计模型实现以下两大核心目标：

1. **构建分类预测模型（再入院风险预警）**：开发一个能够预测糖尿病患者在出院后30天内是否会再次入院的二分类模型。该模型的核心价值在于“预警”，帮助医生提前识别高风险患者，从而在出院前进行更具针对性的干预。
2. **构建回归预测模型（住院资源评估）**：开发一个能够估算患者本次住院持续天数（Length of Stay）的回归模型。该模型旨在辅助医院管理层进行更科学的床位周转管理和资源预估。

## 2. 数据来源与预处理方法

### 2.1 数据来源

本项目使用的数据集源自UCI机器学习知识库（UCI Machine Learning Repository）的“Diabetes 130-US hospitals for years 1999-2008”数据集。原始数据集包含101,766条记录和50个变量，涵盖了患者的人口学特征、入院信息、诊断编码、用药情况及检测结果等多维度信息。

### 2.2 数据抽样与清洗策略

由于项目对数据处理规模有严格限制（需控制在10MB以内），且原始数据量较大（约18MB），我们制定了严谨的抽样与清洗策略以确保数据的代表性和分析的可行性。

分层抽样方案：

为了防止简单随机抽样导致关键类别（如“30天内再入院”）的样本流失，我们采用了基于再入院状态（readmitted）的分层抽样法。我们从原始数据中按比例抽取了约16.5%的样本，最终获得16,777条记录（文件大小约为2.88MB）。验证结果显示，抽样后的数据完美保持了原始分布：30天内再入院率维持在11.16%，未再入院率维持在53.91%，确保了后续建模的统计有效性。

数据清洗流程：

在清洗阶段，我们首先剔除了重复的住院记录，保证每条数据的独立性。针对数据中的特殊符号“?”，我们将其统一转换为标准缺失值（NaN）进行处理。此外，考虑到研究目标是预测出院后的情况，我们剔除了出院去向为“死亡”或“临终关怀”的记录，因为这部分群体不再具有再入院的可能性。

### 2.3 特征工程与变量重组

为了提升模型的预测能力和解释性，我们对原始特征进行了深度的加工与重组：

1. ICD-9诊断编码的语义映射：

原始数据中包含成百上千个具体的ICD-9诊断代码，直接使用会导致特征维度过高且稀疏。我们根据医学领域知识，将这些代码归纳为8大具有临床意义的类别，包括：

* **Circulatory**（循环系统疾病，如高血压、心脏病）
* **Respiratory**（呼吸系统疾病）
* **Diabetes**（糖尿病及其并发症）
* **Digestive**（消化系统疾病）
* **Injury**（损伤与中毒）
* **Genitourinary**（泌尿生殖系统疾病）
* **Neoplasms**（肿瘤）
* Other（其他）  
  这种处理方式不仅降低了维度，还增强了模型对特定疾病类型的捕捉能力。

2. 类别变量的压缩与合并：

部分分类变量存在类别过多且分布极不均匀的问题。例如，“入院来源”（Admission Source）原始有25个类别，我们将其压缩为Referral（转诊）、Transfer（转院）、Emergency（急诊）和Other（其他）4类；“出院去向”（Discharge Disposition）从30类压缩为Home（回家）、Transfer（转院）、Other Facility（其他机构）等5类。

3. 缺失值与异常值处理：

对于缺失率较高的变量（如体重、血糖检测结果），直接填充可能会引入偏差，因此我们要么创建指示变量（如has\_weight）来标记缺失状态，要么将其归类为“Unknown”。对于数值型变量，我们进行了标准化处理（Z-score标准化），以消除不同量纲对回归系数的影响。

## 3. 探索性数据分析 (EDA)

通过对清洗后数据的深入可视化分析，我们获得了关于患者群体特征和变量关系的初步洞察。

### 3.1 关键指标的分布特征

从目标变量来看，样本中**30天内再入院率仅为11.16%**，而未再入院或30天后再入院的比例接近90%。这表明数据存在严重的类别不平衡问题（Imbalanced Data），提示我们在后续分类建模时必须采取特定的重采样或加权策略。

|  |
| --- |
|  |
| 图1：再入院情况分布图 |

从住院天数来看，数据呈现出典型的**右偏分布（Right-skewed）**。患者平均住院天数为4.40天，中位数为4天，绝大多数患者的住院时间在1周以内，但存在长尾效应，部分复杂病例住院时间长达14天。

|  |
| --- |
|  |
| 图2：住院天数分布直方图与箱线图 |

此外，对人口学特征的分析显示，样本涵盖了不同年龄、性别和种族的群体，其中高龄患者占比较大。

|  |
| --- |
|  |
| 图3：患者年龄、性别及种族分布 |

### 3.2 变量间的相关性分析

多变量分析揭示了几个显著的关联模式：

* **多病共存现象（Comorbidity）**：患者平均拥有7-9个诊断记录。相关性热力图显示，诊断数量（Number of Diagnoses）与住院天数呈现显著的正相关关系。这意味着患者患有的并发症越多，其病理状态越复杂，所需的治疗周期自然越长。

|  |
| --- |
|  |
| 图4：连续变量相关性热力图 |

* **不同群体的差异**：不同分组下的再入院率和住院天数存在显著差异。例如，不同年龄段和诊断数量的患者，其再入院风险表现出不同的趋势。

|  |
| --- |
|  |
| 图5：不同年龄、性别及诊断数量分组下的再入院率 |

|  |
| --- |
|  |
| 图6：不同分组下的平均住院天数 |

* **人口学与风险**：患者年龄主要集中在50-80岁区间。分层分析显示，70-90岁的高龄组患者，其30天再入院的风险略高于年轻组，且平均住院天数也更长。

## 4. 统计建模与评估结果

针对研究目标，我们分别构建了回归模型和分类模型，并采用逐步优化的策略，对比了仅包含基础人口学特征的“基准模型”与引入特征工程变量后的“改进模型”。

### 4.1 回归任务：住院天数预测

模型选择与构建：

我们采用了多元线性回归（Multiple Linear Regression）作为预测住院天数的基础算法。为了保证模型的稳健性，我们对所有连续变量进行了标准化处理，并通过计算方差膨胀因子（VIF）严格检查了多重共线性问题，结果显示所有变量的VIF均小于10，模型不存在严重的共线性干扰。

模型诊断：

对回归模型的残差进行了正态性检验和同方差性检验，结果显示模型假设基本成立（见下图）。

|  |
| --- |
|  |
| 图7：回归模型残差诊断图 |

模型表现评估：

通过对比基准模型与改进模型在测试集上的表现，我们发现特征工程显著提升了模型的解释力：

* **R² (决定系数)**：从基准模型的0.2645提升至改进模型的**0.2856**。虽然0.29的R²在工程领域可能不高，但在受人为因素和个体差异影响极大的医疗社会科学数据中，该模型能解释近30%的方差，已属于具有显著统计学意义的结果。
* **误差指标**：平均绝对误差（MAE）降低至**1.86天**。这意味着模型对大多数患者住院天数的预测误差控制在2天以内，具有实际参考价值。

### 4.2 分类任务：30天再入院风险预测

模型选择与策略：

针对二分类任务，我们采用了Logistic回归模型。鉴于数据中正负样本比例约为1:9的严重不平衡，常规模型极易倾向于预测“不再入院”以获得高准确率，从而漏掉高危患者。为此，我们引入了类别权重（Class Weights）策略，赋予正样本（再入院患者）更高的权重，强制模型关注这一少数群体。

核心评估导向：

在医疗预警场景下，漏报一名高危患者（False Negative）的代价远高于误报（False Positive）。因此，我们的评估核心是 “召回率优先”（Recall First）。

**模型表现评估**：

* **Recall (召回率)**：改进模型在测试集上的召回率达到了**67.65%**。这意味着模型成功识别出了约三分之二的实际再入院患者，作为初步筛查工具，其灵敏度令人满意。
* **AUC (曲线下面积)**：模型的AUC值为**0.6419**，表明模型在区分正负样本方面具备一定的预测能力，优于随机猜测。ROC曲线展示了改进模型在不同阈值下的表现。

|  |
| --- |
|  |
| 图8：分类模型ROC曲线 |

* **Accuracy (准确率)**：虽然为了提升召回率牺牲了部分特异性，但模型的整体准确率仍达到了53.50%，较基准模型有显著提升。

## 5. 关键发现与业务解读

### 5.1 住院天数的驱动因素（回归系数解读）

通过分析回归模型的标准化系数，我们量化了各因素对住院时长的影响程度：

1. **诊断数量 (Number of Diagnoses)**：这是影响最大的正向因子。系数表明，在其他条件不变的情况下，患者每增加一种并发症诊断，其住院天数就会显著增加。这提示医院应重点关注多病共存患者的床位预留。
2. **实验室检查次数 (Lab Procedures)**：检查越频繁，往往意味着病情越不稳定或越复杂，直接导致住院时间延长。
3. **胰岛素治疗 (Insulin Use)**：相比仅通过饮食控制或口服药治疗的患者，需要注射胰岛素的患者通常病情更重，模型显示其住院时间显著更长。

|  |
| --- |
|  |
| 图9：回归模型主要变量系数图 |

### 5.2 再入院的风险预警（OR值解读）

Logistic回归的优势比（Odds Ratio, OR）为我们揭示了再入院的风险图谱：

1. **既往住院次数 (Number Inpatient)**：这是最强的风险预测因子，其**OR值高达1.45**。这意味着，患者在过去一年中每多住院一次，其本次出院后再次入院的风险就增加45%。这为临床提供了一个简单而有力的筛查指标：关注那些“常客”。
2. **诊断数量**：**OR值为1.11**。每增加一个诊断编码，再入院风险增加11%，再次印证了复杂病情是导致反复入院的根源。
3. **出院去向的影响**：数据表明，被转院或转至专业护理机构（Skilled Nursing Facilities）的患者，其再入院风险显著高于直接回家的患者，这反映了这部分患者本身身体状况的脆弱性。

|  |
| --- |
|  |
| 图10：分类模型主要变量OR值图 |

## 6. 原型系统展示

为了将复杂的统计模型转化为可用的临床工具，我们基于R语言的Shiny框架开发了一套交互式原型系统。该系统包含以下三个核心模块：

1. 数据概览仪表盘 (Dashboard)：  
   系统首页实时展示当前患者群体的关键KPI，包括总记录数、平均再入院率及平均住院天数。同时集成了EDA图表，医生可以直观查看患者的年龄分布、性别比例及合并症情况。
2. 模型评估实验室 (Evaluation Lab)：  
   该模块面向数据分析师或科研人员，可视化展示了模型的性能指标。用户可以查看ROC曲线、混淆矩阵热力图以及回归模型的残差分布图，从而评估模型的可靠性和适用范围。

|  |
| --- |
|  |
| 图11：分类模型混淆矩阵 |

|  |
| --- |
|  |
| 图12：分类模型预测概率分布图 |

1. 交互式风险预测器 (Interactive Predictor)：  
   这是系统的核心应用功能。临床医生只需在界面上方输入患者的特征信息（如年龄段、既往住院次数、本次实验室检查次数、诊断数量等），系统即可在下方实时计算并输出：
   * **预计住院天数**（例如：“预测值：6.5天”）。
   * 30天再入院风险概率（例如：“风险评估：高风险，概率 >20%”）。  
     该功能可直接辅助医生制定出院计划和随访策略。

## 7. 总结与未来展望

### 7.1 项目总结

本项目完整复现了医疗数据挖掘的全流程，从数据清洗、特征工程、探索性分析到统计建模与系统开发。我们成功通过分层抽样技术解决了大数据处理的算力限制，同时保证了数据的代表性。研究量化证实了 “既往住院史”**和**“多病共存”**是导致糖尿病患者医疗资源消耗和再入院风险增加的核心驱动力。我们构建的**“召回率优先”模型及配套的原型系统，为临床高危人群的早期筛查提供了一套行之有效的解决方案。

### 7.2 局限性与改进方向

尽管项目取得了一定成果，但仍存在局限性，这也是未来的改进方向：

* **数据维度的局限**：目前的模型仅基于结构化数据。未来若能引入非结构化的病历文本（如医生的入院记录、出院小结）以及具体的生命体征数值（如血压、详细的血糖波动曲线），将极大地提升模型的预测精度。
* **模型算法的局限**：线性模型虽然解释性强，但对非线性关系的拟合能力有限。未来可以尝试引入随机森林（Random Forest）、XGBoost或LightGBM等集成学习算法，以进一步提升R²和AUC表现。
* **应用深度的拓展**：目前的系统仅提供风险“预警”。未来可以结合规则引擎，增加个性化干预建议功能（例如，针对高风险患者自动推荐“出院后3天内电话随访”或“预约家庭医生上门服务”），从而实现从“预测”到“干预”的闭环管理。

# 8. AI工具使用情况

本小组确定选题时，从生成式AI处获取了选题方向相关建议，并最终确定项目选题。

本项目建模部分（特征工程、回归模型、分类模型及Shiny应用）使用Cursor AI编辑器进行辅助开发，协助完成代码编写、调试优化和技术文档撰写（建模技术说明、模型结果摘要），主要用于提升开发效率。核心建模思路、算法选择和参数调优均为独立完成。

汇报用演示文档接受了Google Gemini对布局提出的美化建议，并采用了Nano Banana提供的装饰性图片。

本文档接受了Google Gemini给出的格式建议。

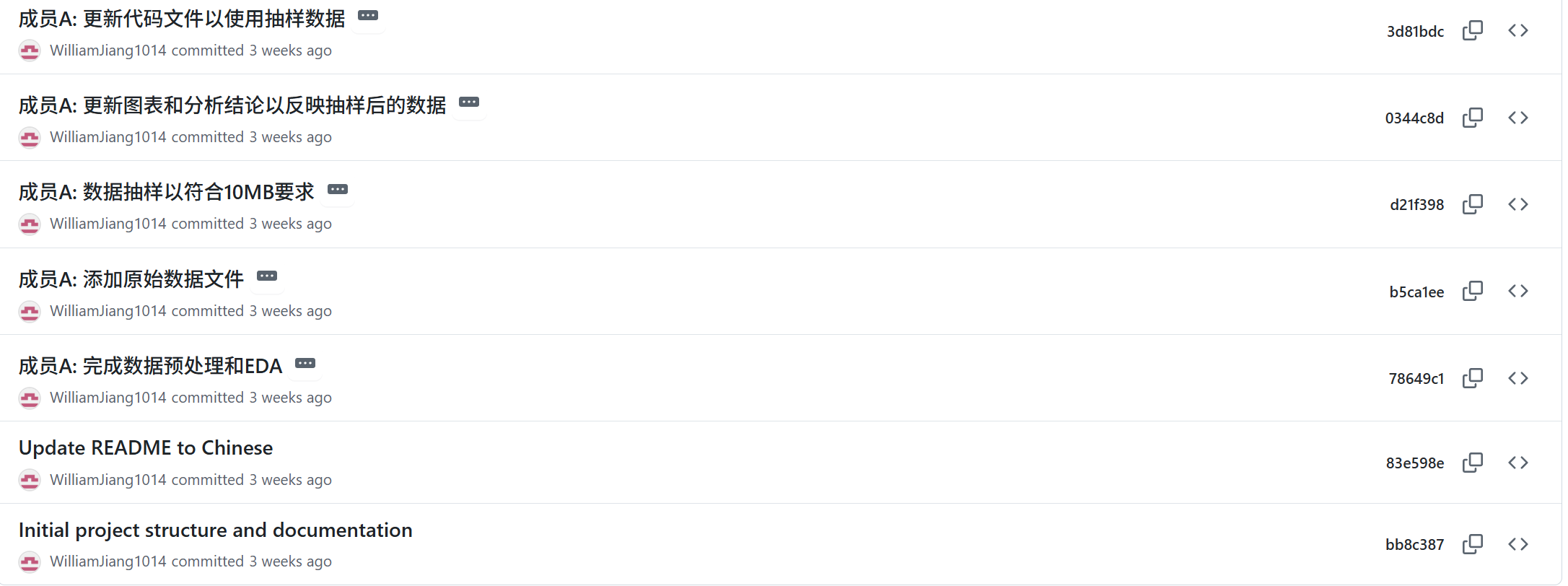
# 9. 项目链接

Github仓库链接：<https://github.com/WilliamJiang1014/statistical-modeling-course-project>

# 10. 团队分工

姜政言2353594

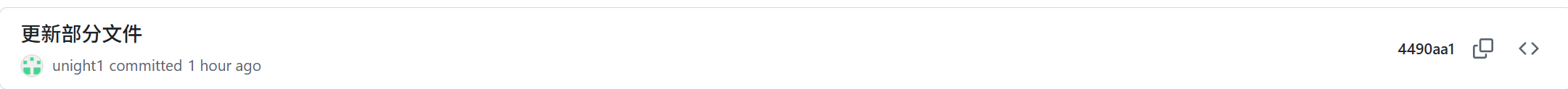
* 原始数据的收集与准备
* 数据预处理、数据清洗
* 进行探索性数据分析



陈奕玮2350217

* 建立住院天数预测模型
* 建立30天再入院风险预测模型
* 模型评估与优化
* 开发shiny应用





龙承佑2354176

* 汇总项目成果，形成演示文档及项目文档
* 制作项目展示视频
* 补充部分数据的可视化图像

