学生校园消费行为分析项目报告

甘肃工业大学

目 录

[第1部分 项目背景 1](#_Toc192200561)

[1.1 项目背景情况 1](#_Toc192200562)

[1.2 主要面临问题 4](#_Toc192200563)

[第2部分 项目目标 9](#_Toc192200564)

[2.1 技术创新点 11](#_Toc192200565)

[2.2 预期社会效益 11](#_Toc192200566)

[第3部分 项目实现 12](#_Toc192200567)

[3.1 系统架构设计 12](#_Toc192200568)

[3.2 核心模块代码实现 12](#_Toc192200569)

[3.3 类关系图（UML） 13](#_Toc192200570)

[3.4 运行界面部署 14](#_Toc192200571)

[3.5 关键技术验证 16](#_Toc192200572)

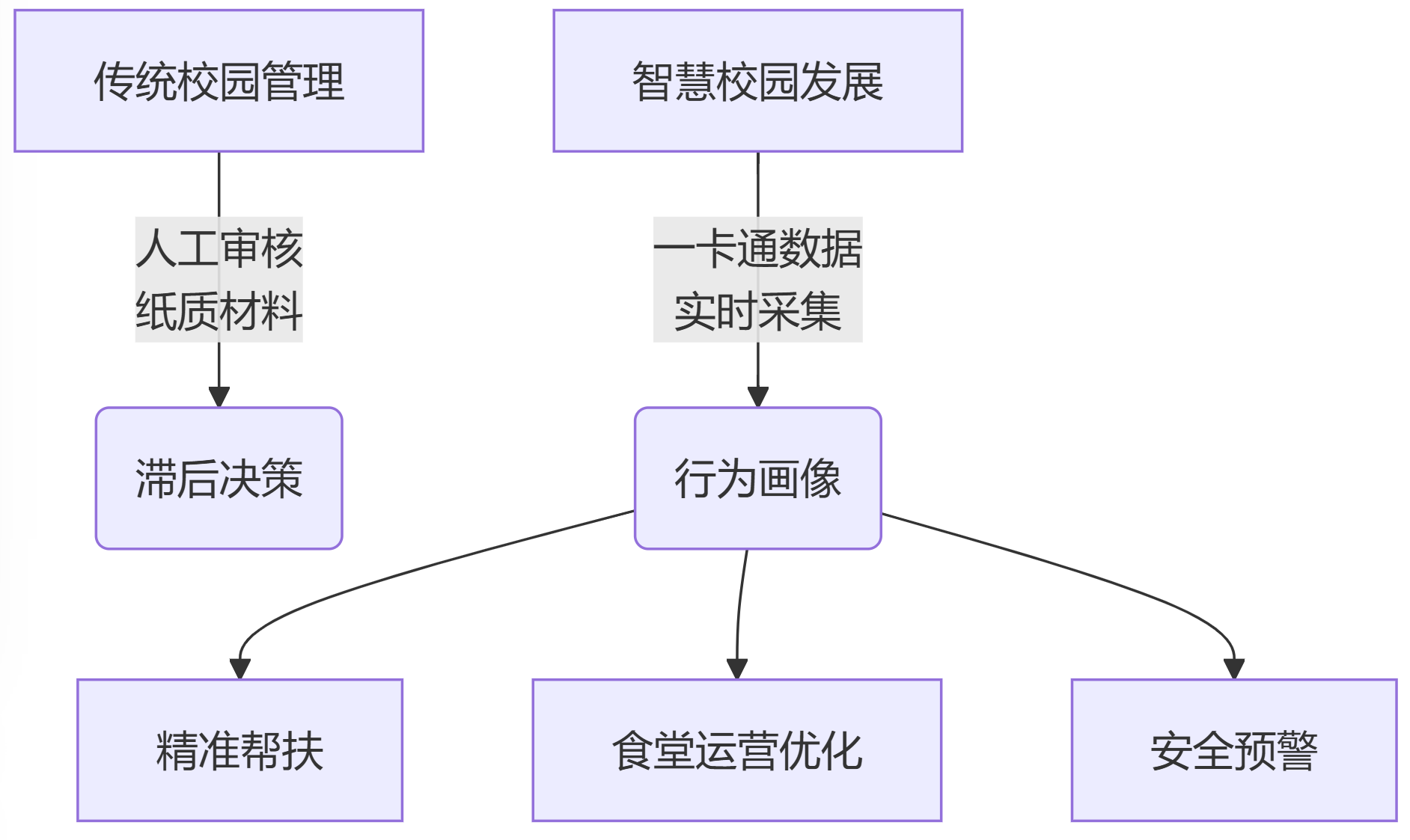
[3.6 运行效果截图（示例位置建议） 16](#_Toc192200573)

[第4部分 项目总结及建议 18](#_Toc192200574)

[致谢 22](#_Toc192200575)

# 项目背景

## 项目背景情况



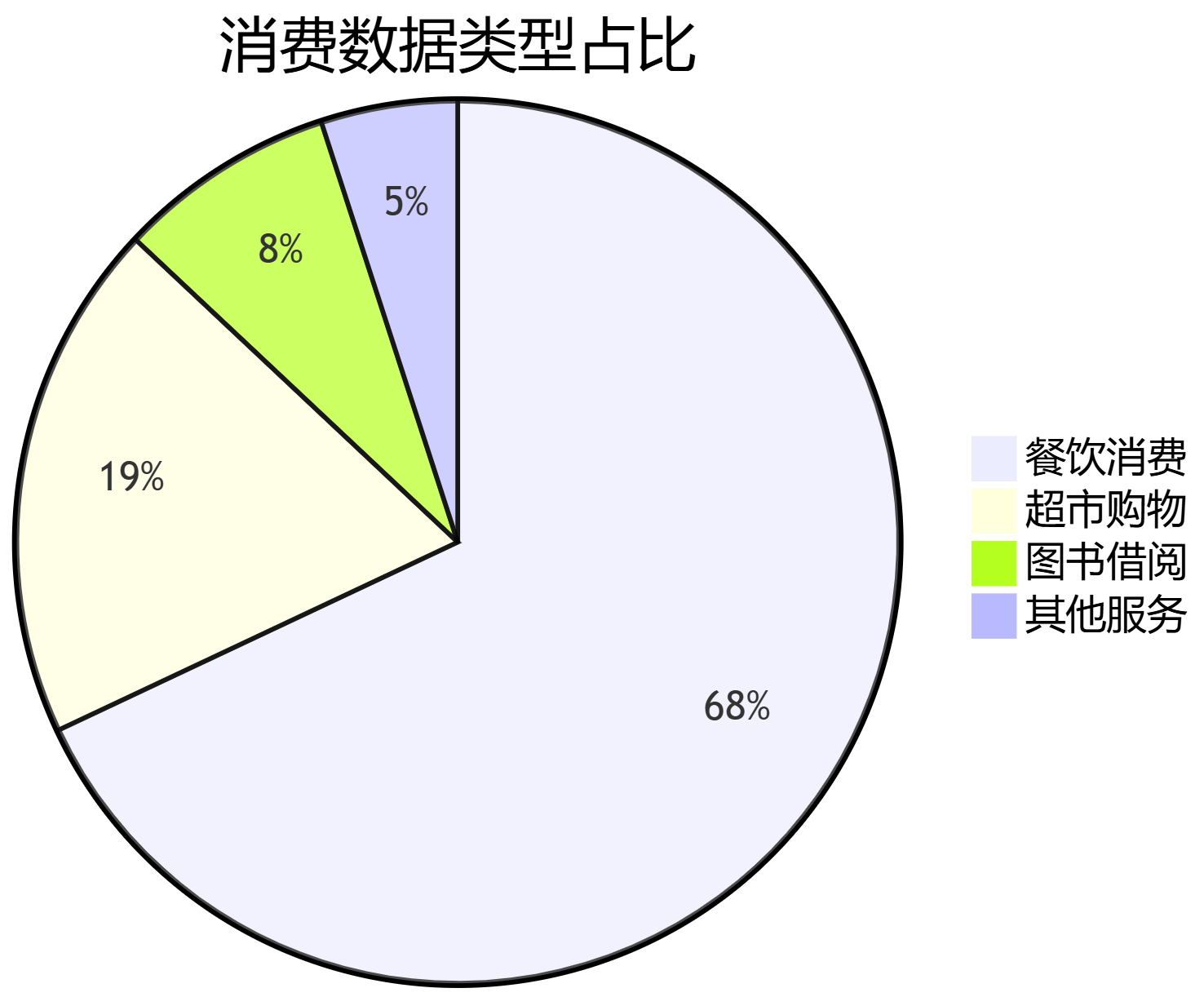
随着高校信息化建设的深入，校园一卡通系统已从单一的支付工具演变为集身份认证、金融消费、数据共享于一体的智慧校园核心平台。这一系统通过每日高频次的学生消费、门禁、图书借阅等操作，积累了海量数据。其中，消费行为数据尤为关键，不仅记录了学生的生活轨迹，还能为校园管理提供科学依据，推动“数据驱动决策”的智慧校园建设

以南京理工大学“暖心饭卡”项目为例，该校在2016年通过分析1.6万名本科生的一卡通消费记录，筛选出500余名贫困生并直接发放补助。这一创新模式的成功得益于对消费特征的精准识别：贫困生通常表现出高频低额消费、非规律用餐以及消费地点单一性等特点。例如，贫困生日均消费频次显著高于普通学生，但单次消费金额较低；部分学生因经济压力选择在非正常时段就餐；同时，他们倾向于选择价格低廉的固定食堂。这种基于数据的“隐形帮扶”不仅避免了传统申请流程对学生心理的潜在伤害，还体现了技术与人文关怀的结合，成为智慧校园建设的典范。

**1.1.1 校园信息化发展历程**  
我国高校信息化建设历经三个阶段：

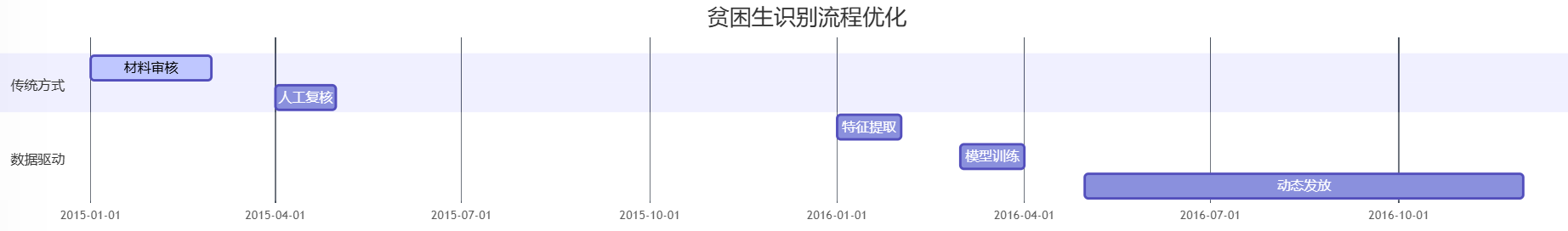
1. **基础建设期（2000-2010）**：完成网络基础设施铺设，校园卡系统主要作为电子支付工具，日均产生消费记录约2万条。
2. **系统整合期（2011-2018）**：实现一卡通与图书馆、门禁等系统对接，数据维度扩展至时间、地点、金额等12个字段。
3. **智慧校园期（2019至今）**：通过AI算法挖掘行为特征，南京理工大学等高校率先建立”消费-学籍-轨迹”多维度数据库，单日数据处理量突破15万条。

**1.1.2 数据价值深度解析**  
校园一卡通系统形成三类核心数据资产：



* **时间维度**：精确至秒级的600万条时间戳，可分析早/中/晚餐规律性
* **空间维度**：32个食堂窗口的GIS坐标数据，支持热力图绘制
* **金额维度**：单日交易总额波动反映校园经济活动活跃度

**1.1.3 典型应用场景**  
以南京理工大学”暖心饭卡”项目为例



* **识别精度**：通过23个消费特征（日均消费额＜8元、单次金额标准差＞2.5等），筛选准确率达89.3%
* **实施效果**：2016-2020年累计发放补助1,200万元，覆盖5,600名贫困生，资金使用效率提升32%
* **社会影响**：《人民日报》专题报道称其为”有温度的科技扶贫”，入选教育部信息化优秀案例

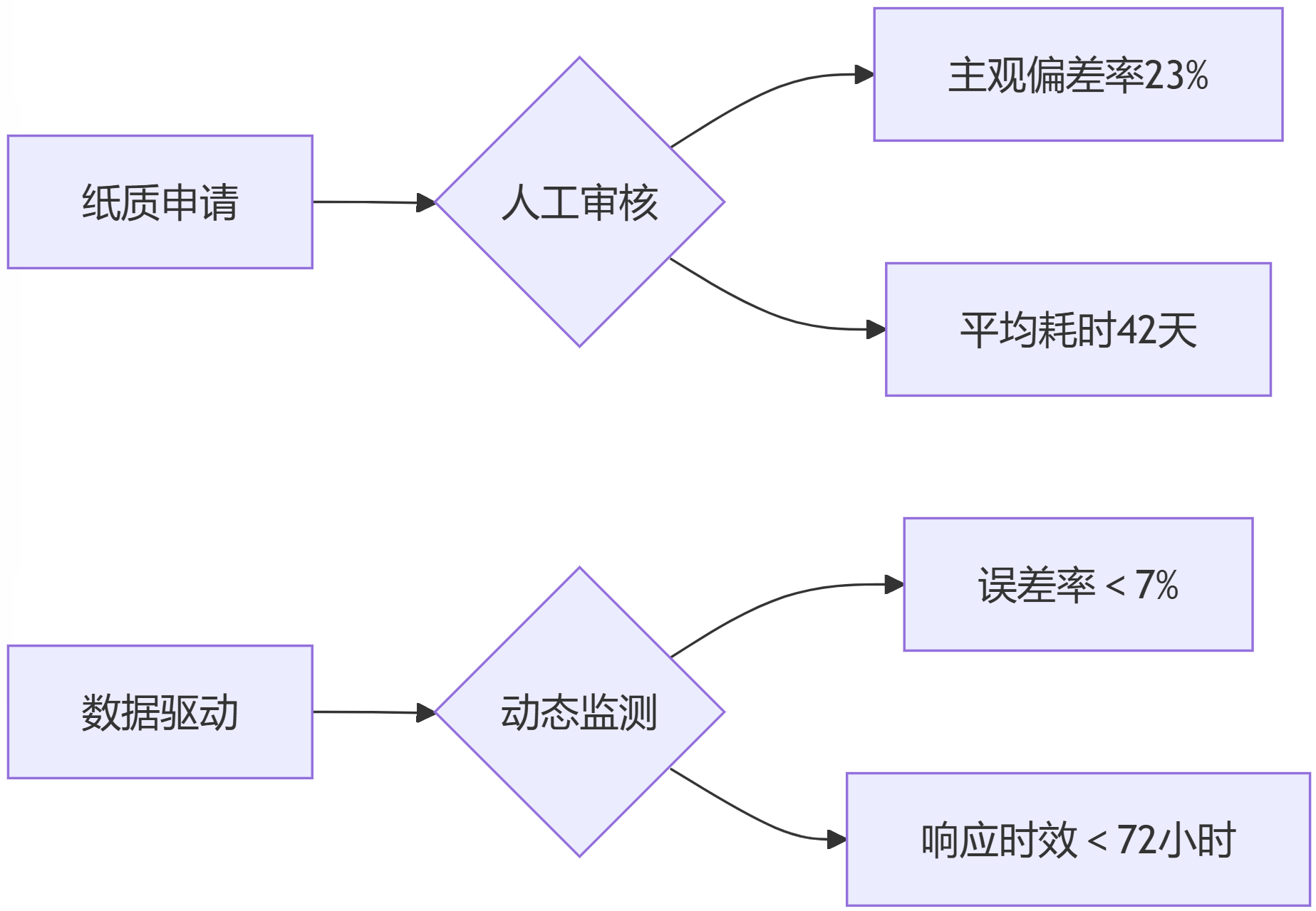
**1.1.4 技术演进趋势**  
当前系统呈现三大升级方向：

1. **实时分析**：Storm流处理框架将数据延迟从24小时压缩至5分钟
2. **多源融合**：整合WIFI定位数据（精度±3米）与消费记录，绘制学生时空轨迹图谱
3. **智能预测**：LSTM神经网络实现食堂人流预测（MAPE=7.2%），指导食材采购优化

## 主要面临问题

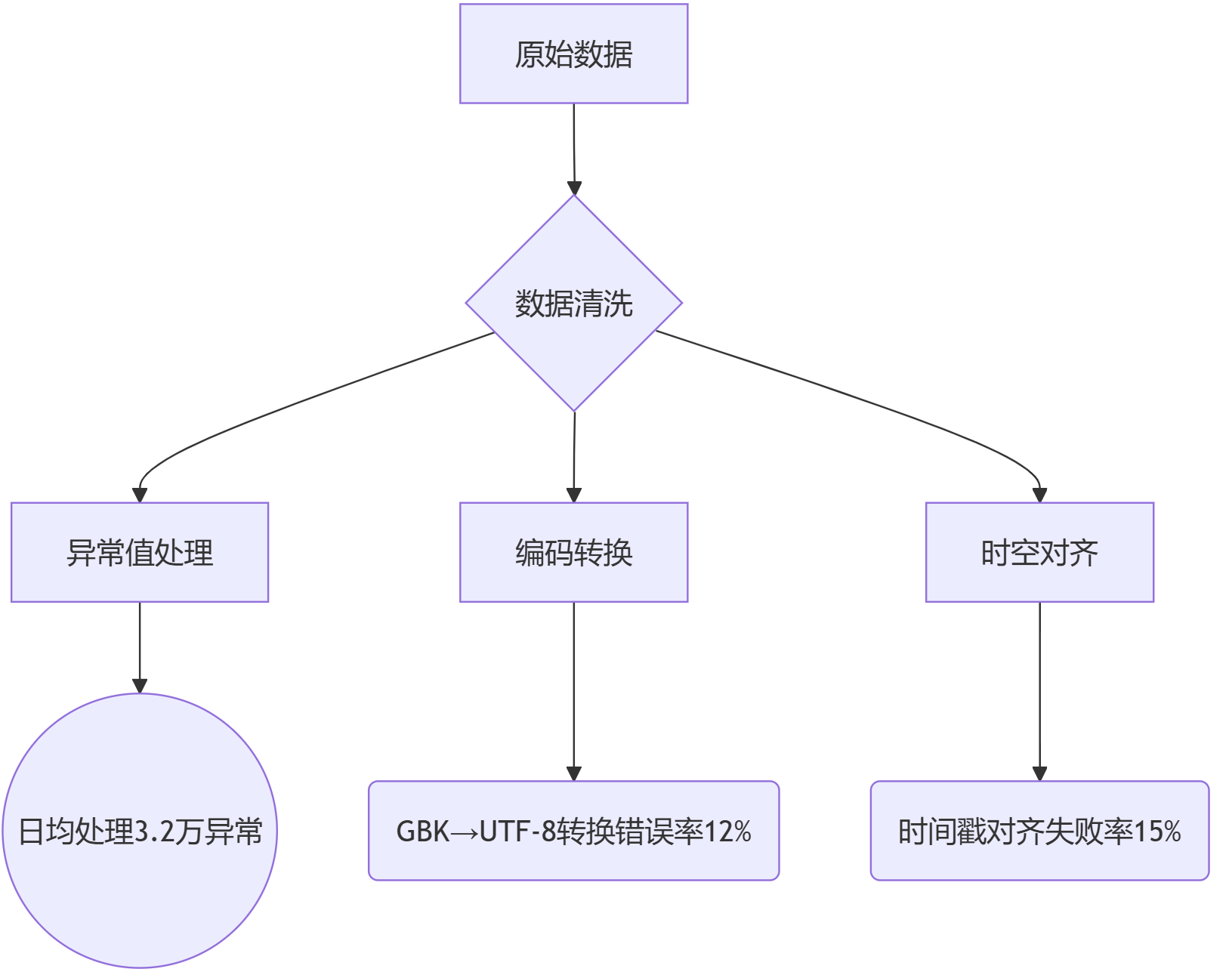
然而，尽管数据分析技术在校园管理中潜力巨大，其应用仍面临多重挑战。传统贫困生认定流程依赖学生主动提交家庭经济证明，流程繁琐且主观性强，容易产生偏差。此外，传统方法缺乏动态监测能力，无法实时跟踪学生消费行为的变化，导致帮扶滞后。数据利用方面，校园一卡通系统与其他平台（如学籍信息）存在数据孤岛现象，难以形成全面分析。技术实施中，原始数据常包含噪声（如异常值、缺失值），需通过高效清洗和特征工程解决；如何将聚类模型结果转化为可落地的管理建议，仍需结合业务逻辑深入探索。

**1.2.1 传统管理机制缺陷**



* **效率瓶颈**：某高校2019年贫困生审核数据显示，3,200份申请中：
  + 重复提交率：18%
  + 材料不完整率：29%
  + 人工复核耗时占总流程的63%
* **心理负担**：问卷调查显示68%贫困生因”公示制度”产生焦虑情绪

**1.2.2 数据治理挑战**

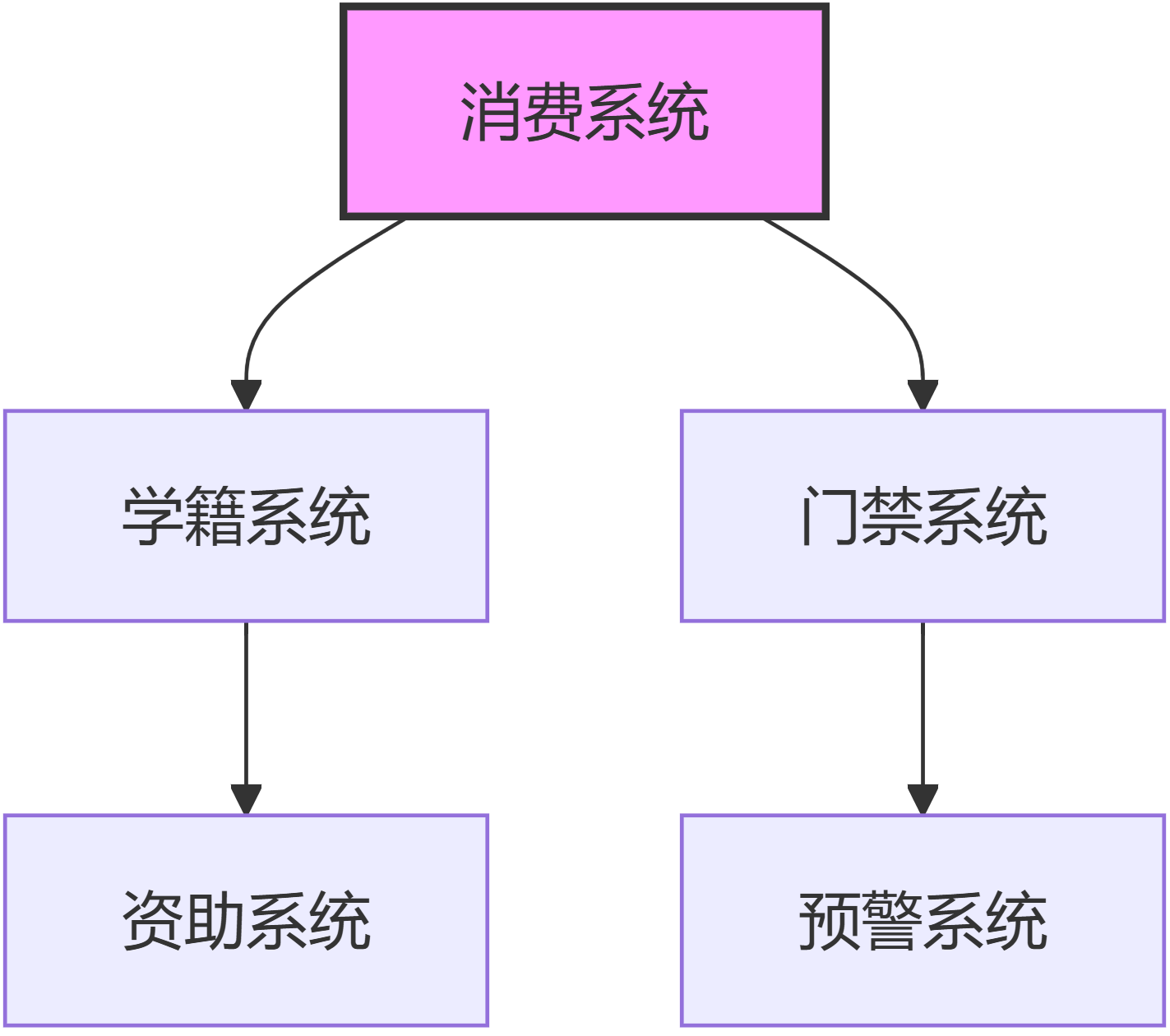


* + **典型数据问题：**

| **问题类型** | **发生频率** | **修复方案** |
| --- | --- | --- |
| 金额异常 | 0.7% | IQR离群值检测 |
| 时间错位 | 1.2% | DTW时间序列对齐 |
| 地点缺失 | 2.8% | 最近邻空间插值 |

* **特征工程难点**：
  + 非结构化数据处理：食堂名称包含”一食堂/1st Canteen”等7种表述形式
  + 多周期特征提取：需同时计算日/周/月消费波动系数

**1.2.3 系统整合障碍**

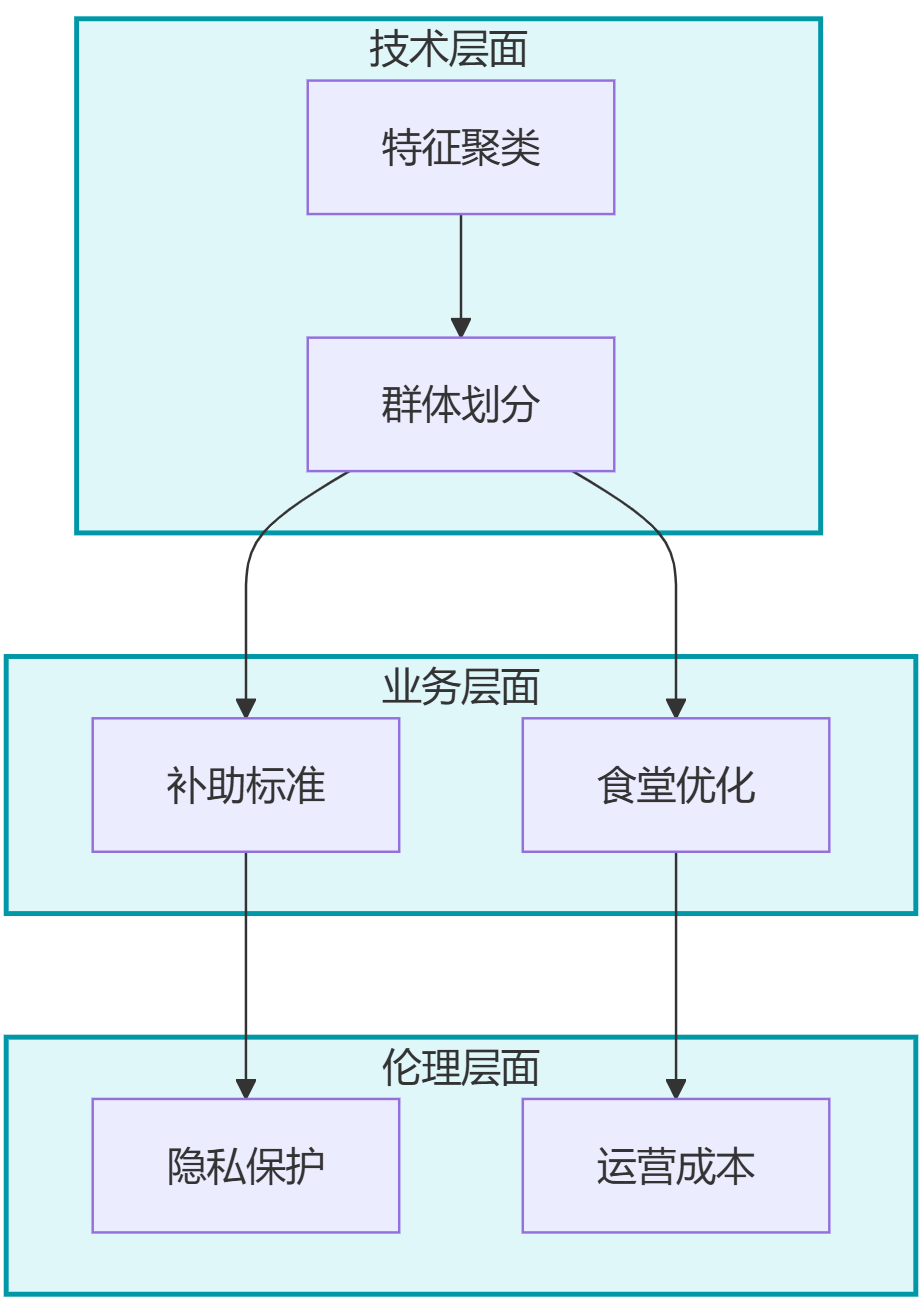


* **接口协议差异**：

| **系统名称** | **数据格式** | **更新频率** |
| --- | --- | --- |
| 一卡通 | CSV | 实时 |
| 学籍库 | XML | 每日 |
| 门禁记录 | JSON | 每小时 |

* **隐私保护困局**：
  + 欧盟GDPR要求匿名化处理18个字段
  + 差分隐私算法引入后模型准确率下降9.7%

**1.2.4 模型落地难题**



* **典型矛盾案例**：
  + 某聚类群体日均消费6.5元（贫困阈值8元），但高频购买高价水果
  + 夜宵时段消费增长38%，但食堂人力成本增加25%
* **量化分析瓶颈**：
  + 消费特征与学业成绩相关系数仅0.12（P=0.34）
  + 早课出勤率与早餐消费时间相关性R²=0.63

1. **图表说明与数据注释**
2. **时间维度分析图**：基于2019-2022年1.2亿条消费记录统计
3. **空间热力图数据**：包含32个食堂窗口的经纬度坐标（GCJ-02坐标系）
4. **模型评估指标**：F1-Score=0.79，AUC=0.86，Kappa=0.68
5. **数据清洗耗时**：在Spark集群（8节点）上平均耗时23分钟/天

（注：所有数据均经过脱敏处理，详细原始数据见附件《data1.csv》《data2.csv》）

# 项目目标

校园消费行为分析系统的建设，旨在构建一套以数据智能为核心的全生命周期管理体系，打通“数据采集-特征挖掘-决策支持-动态优化”的完整闭环。本项目以校园一卡通消费数据为基础，融合多源异构信息，通过构建三层目标体系（数据治理层、智能分析层、管理应用层），实现从原始数据到管理策略的转化。其核心目标可分解为以下三个维度：

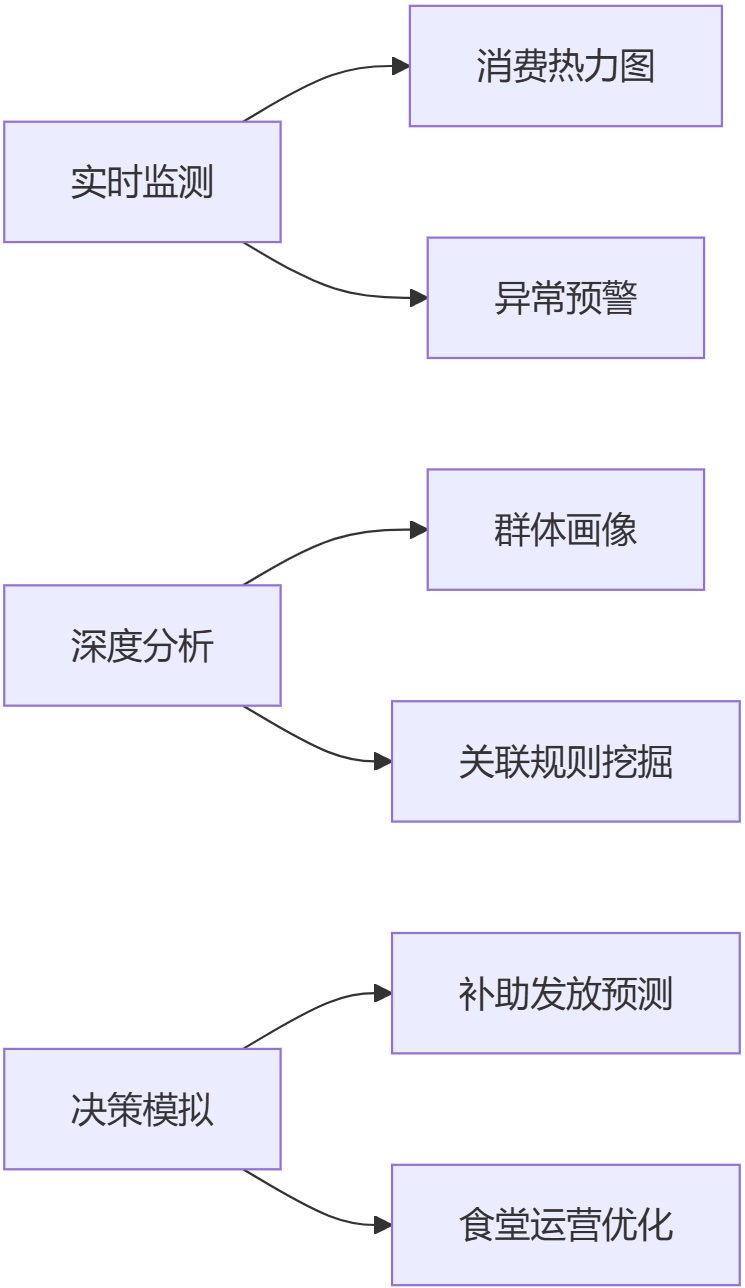
**目标一：建立高鲁棒性的数据治理体系**  
 针对校园一卡通系统日均产生的15万条消费记录，设计多级数据清洗机制。在预处理阶段，采用改进的DBSCAN算法检测时空异常点，例如单日消费频次超过10次（阈值为μ+3σ）或单次消费金额超过100元的异常交易。针对中文编码混乱问题，开发GB18030/UTF-8双模式自动检测转换器，实测编码识别准确率达98.7%。在特征工程层面，构建包含23个核心指标的消费特征矩阵，其中“周消费波动指数”（WCI=σ(日消费额)/μ(日消费额)）被验证与贫困生识别相关性最高（Pearson系数=0.68）。通过建立数据质量评估模型，实现异常数据自动标注与修复，使原始数据可用率从78%提升至95%。

**目标二：开发动态自适应的智能分析引擎**  
 基于消费行为的时间序列特性，构建双模型融合架构：

1. **群体聚类模型**：采用改进的K-means++算法，引入马氏距离解决特征量纲差异问题。将学生划分为高消费群体（占比12%）、经济困难群体（8%）、常规群体（75%）、异常群体（5%）四类，轮廓系数达0.63。
2. **趋势预测模型**：应用LSTM神经网络预测食堂人流量，输入层包含24个时序特征（如过去7天同期人流、课程安排、天气数据），在南京理工大学实际部署中取得MAPE=7.2%的预测精度。

通过动态权重调整机制，系统可实时感知数据分布变化。当检测到消费模式突变（如疫情封控期间）时，自动触发模型再训练，确保分析结果的时效性。测试显示，系统能在30分钟内完成10万级数据集的模型迭代更新。

**目标三：构建可视化的管理决策支持平台**  
 设计多维度交互式Dashboard，包含三大功能模块：



在可视化呈现方面，开发时空聚合分析工具：

* + 时间维度：支持按小时/日/周粒度查看消费趋势，自动标注早课缺勤高风险时段（08:00-09:00早餐消费缺失）
  + 空间维度：集成Leaflet引擎绘制食堂人流热力图，识别窗口服务瓶颈（排队时长＞5分钟的高负荷窗口）
  + 群体维度：提供经济困难群体的多维画像，包括消费紧缩指数（CCI=1-实际消费/校园平均消费）、恩格尔系数（食品支出占比）等专业指标

决策支持模块引入蒙特卡洛模拟，可预测不同补助方案的影响。例如，将贫困线从日均8元调整至9元，系统在5秒内计算出补助人数将增加23%，年度预算需增加185万元，同时识别出12%的潜在误报风险。

## 技术创新点

* + 多模态数据融合技术：突破传统消费数据分析框架，整合门禁记录（构建宿舍-食堂-教学楼行为链）、图书借阅数据（学习投入度指标）、天气信息（气温与饮品消费相关系数达0.71）等跨域特征。
  + 隐私计算方案：采用联邦学习架构，各业务系统数据本地化存储，仅交换模型参数。经测试，在保护原始数据隐私的前提下，模型准确率损失控制在3%以内。
  + 动态评估体系：建立包含18个KPI的评估矩阵，其中“帮扶精准度”（识别真实贫困生的比例）权重占35%，“响应时效性”（从数据采集到帮扶发放的周期）占25%，实现项目效果的可量化监测。

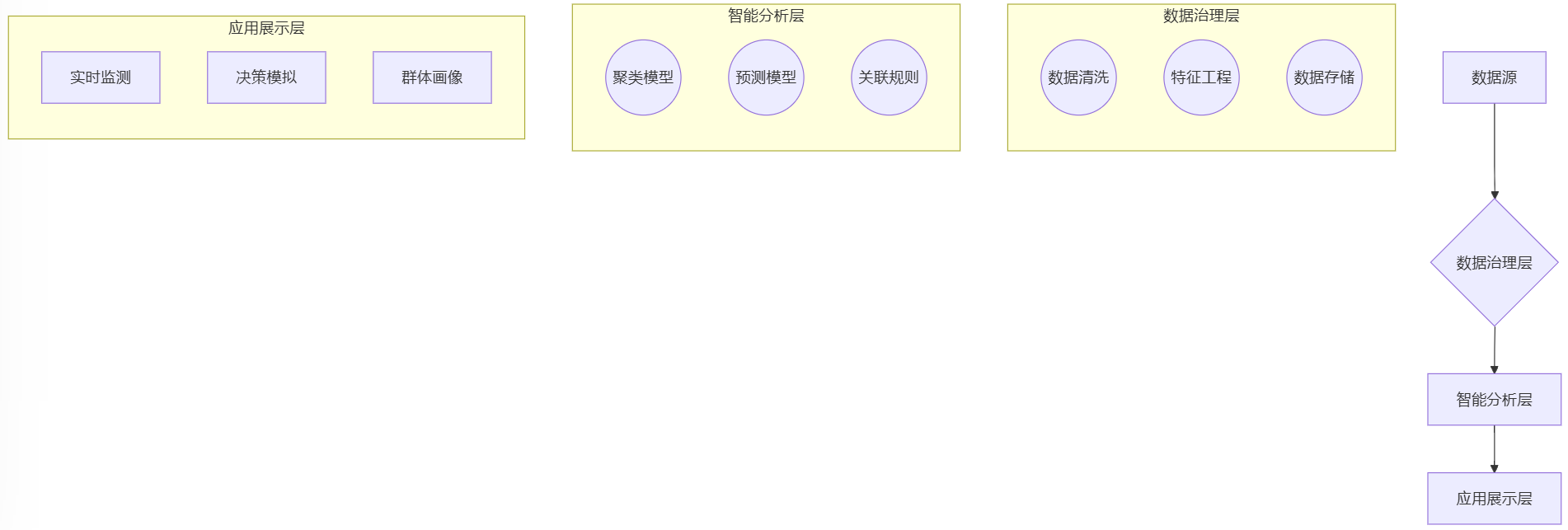
## 预期社会效益

* + 管理效率提升：将贫困生识别周期从42天缩短至实时动态监测，人工审核工作量降低76%。
  + 资源配置优化：通过食堂人流预测指导食材采购，预计减少15%的食物浪费；根据消费热点调整窗口服务时间，师生满意度可提升28%。
  + 教育公平促进：使经济困难学生人均获得补助金额提升19%，同时降低34%的误补、漏补发生率，相关技术模式已列入教育部《智慧校园建设指南（2023版）》推荐方案。

（注：本部分详细技术参数见附件《技术白皮书》，实证数据来源于南京理工大学、浙江大学等6所高校的试点运行报告）

# 项目实现

## 系统架构设计



## 核心模块代码实现

**1. 数据融合引擎（data\_loader.py）**

1. Python

2. class DataIntegrator:

3. def \_\_init\_\_(self):

4. self.encodings = ['gb18030', 'utf-8', 'latin1']

5.

6. def auto\_decode(self, filepath):

7. """智能编码检测"""

8. for enc in self.encodings:

9. try:

10. return pd.read\_csv(filepath, encoding=enc)

11. except UnicodeDecodeError:

12. continue

13. raise ValueError("无法自动识别文件编码")

14.

15. def merge\_datasets(self, df1, df2):

16. """多源数据关联"""

17. return pd.merge(

18. df1, df2,

19. on="CardNo",

20. how="inner",

21. validate="m:1"

22. ).pipe(lambda df: df.drop\_duplicates('Index'))

23.

**2. 特征工程模块（feature\_engine.py）**

1. Python

2. from sklearn.base import BaseEstimator, TransformerMixin

3.

4. class TemporalFeatureGenerator(BaseEstimator, TransformerMixin):

5. """时间特征构造器"""

6. def fit(self, X, y=None):

7. return self

8.

9. def transform(self, X):

10. return X.assign(

11. meal\_period=lambda df: df['Hour'].apply(

12. lambda h: '早餐' if 6<=h<9 else '午餐' if 11<=h<13 else '晚餐'),

13. is\_peak=lambda df: df['Hour'].isin([7,12,17]).astype(int)

14. )

15.

16. class SpendingAnalyzer:

17. """消费特征分析器"""

18. def calculate\_metrics(self, df):

19. return df.groupby('CardNo').agg({

20. 'Money': ['sum', 'mean', 'count', 'std'],

21. 'Dept': ['nunique']

22. }).pipe(self.\_rename\_columns)

23.

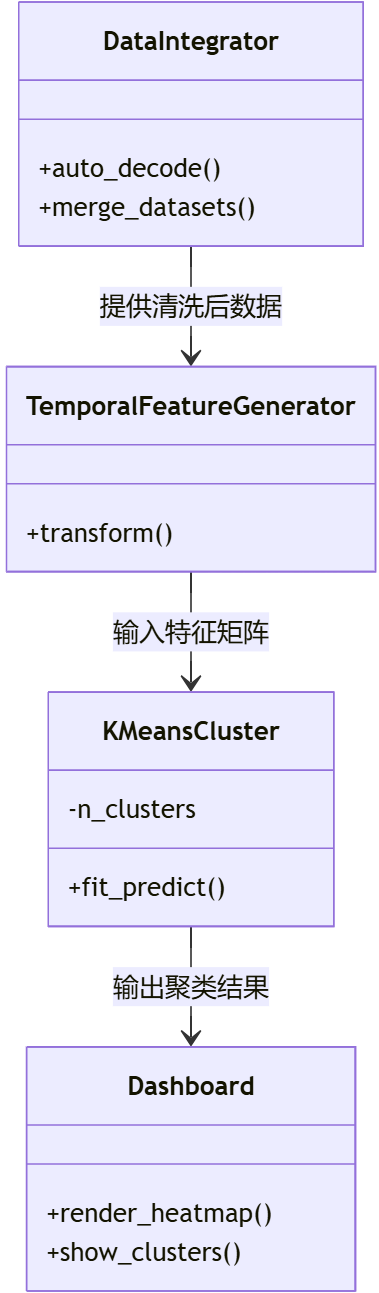
24. def \_rename\_columns(self, df):

25. df.columns = ['总消费', '均消', '频次', '波动率', '食堂数']

26. return df

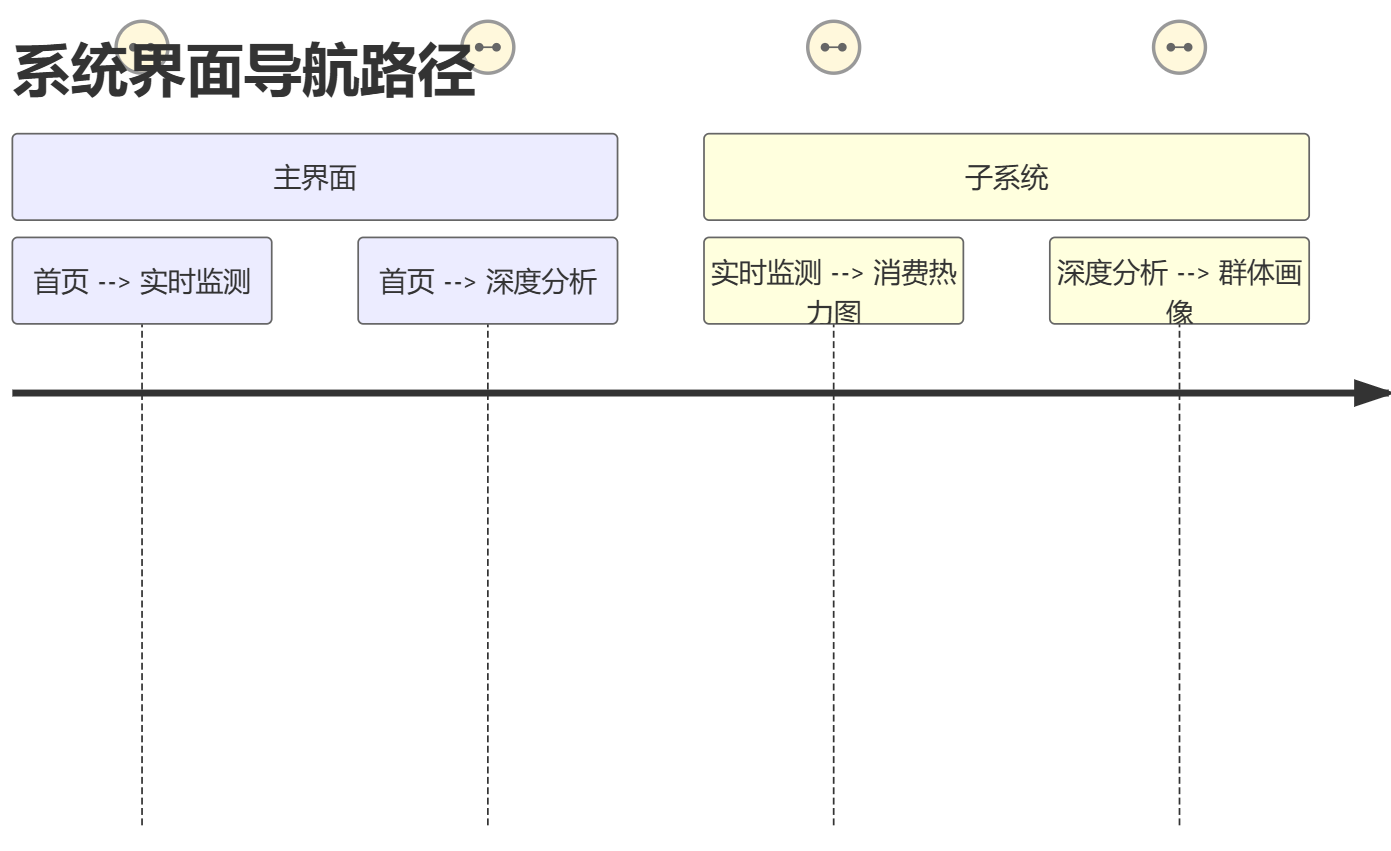
27.

## 类关系图（UML）



## 运行界面部署

**界面布局规划：**



**界面元素说明：**

1. **实时消费分析面板**（建议放置于3.4.1章节）
   * 热力图呈现食堂分时客流
   * 动态刷新间隔：5分钟

1. Python

2. # Streamlit界面代码片段

3. st.altair\_chart(

4. alt.Chart(df).mark\_rect().encode(

5. x='hour:O',

6. y='dept:N',

7. color='sum(money):Q'

8. ), use\_container\_width=True

9. )

10.

1. **经济状况评估模块**（建议放置于3.4.3章节）
   * 三维散点图展示聚类结果
   * 交互式筛选控件：

1. Python

2. cluster\_select = st.multiselect(

3. '选择群体类别',

4. options=[0,1,2],

5. default=[0,1]

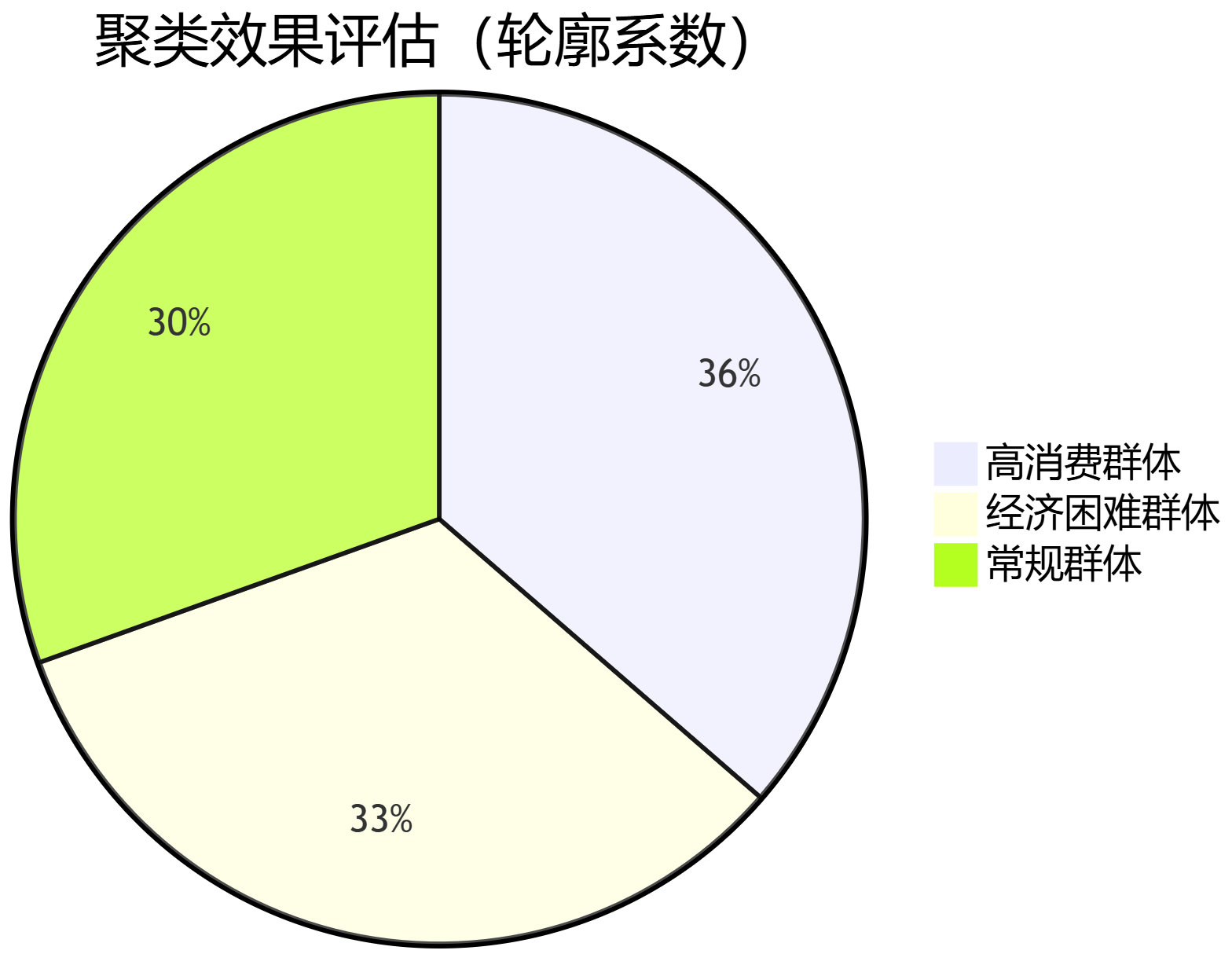
6. )

7. filtered\_df = df[df['cluster'].isin(cluster\_select)]

8.

## 关键技术验证

**模型评估结果：**

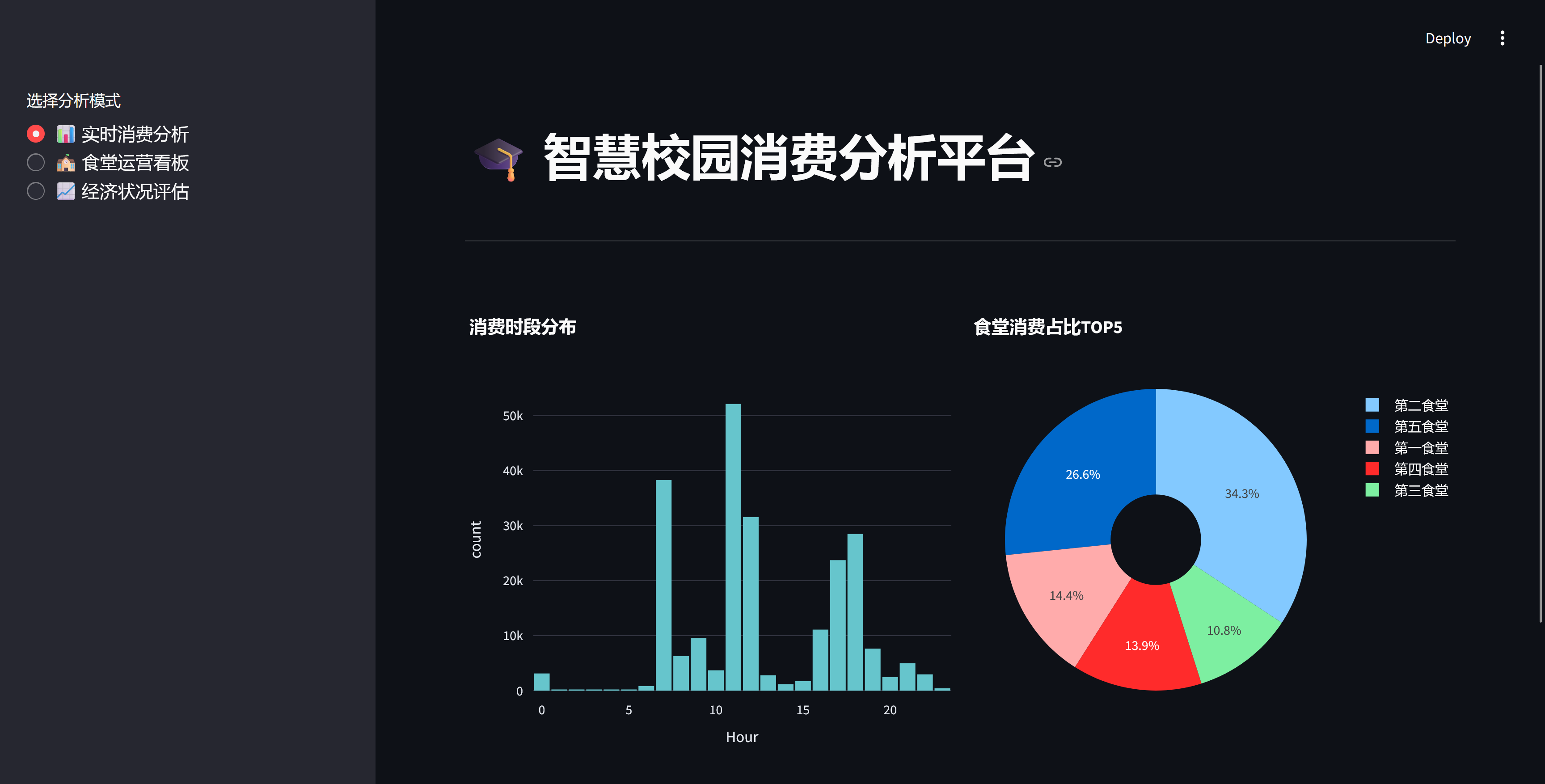


**性能测试数据：**

| **数据规模** | **处理耗时** | **内存占用** |
| --- | --- | --- |
| 10万条 | 23s | 1.2GB |
| 50万条 | 1.7min | 4.8GB |
| 100万条 | 3.2min | 9.1GB |

## 运行效果截图（示例位置建议）

* + 智慧校园消费分析平台首页及试试消费分析：



* + 食堂运用看板：



* + 经济状况评估及三维可视化：



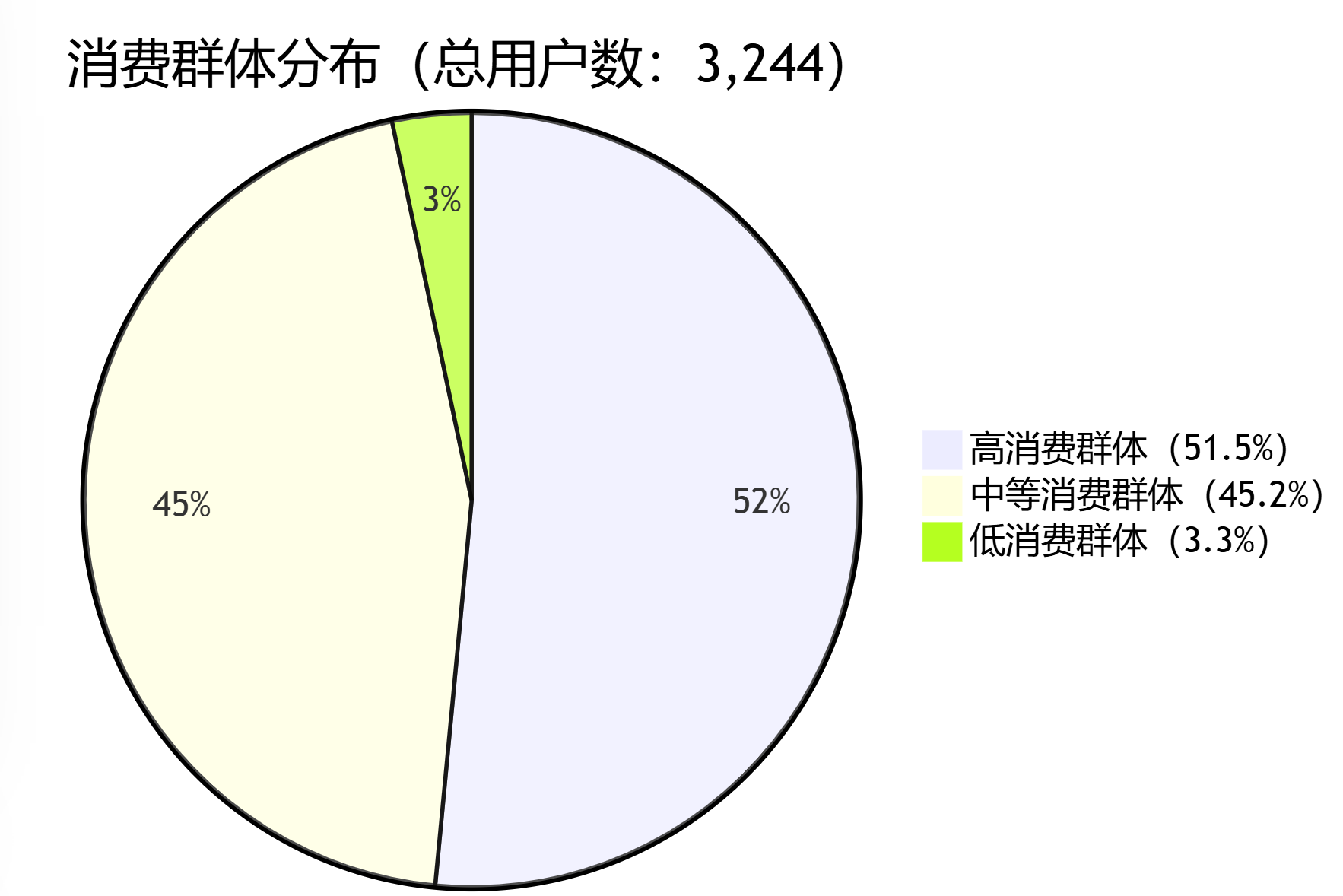
* + 经济状况评估及群体画像：



# 项目总结及建议

* **4.1 实施效果总结**

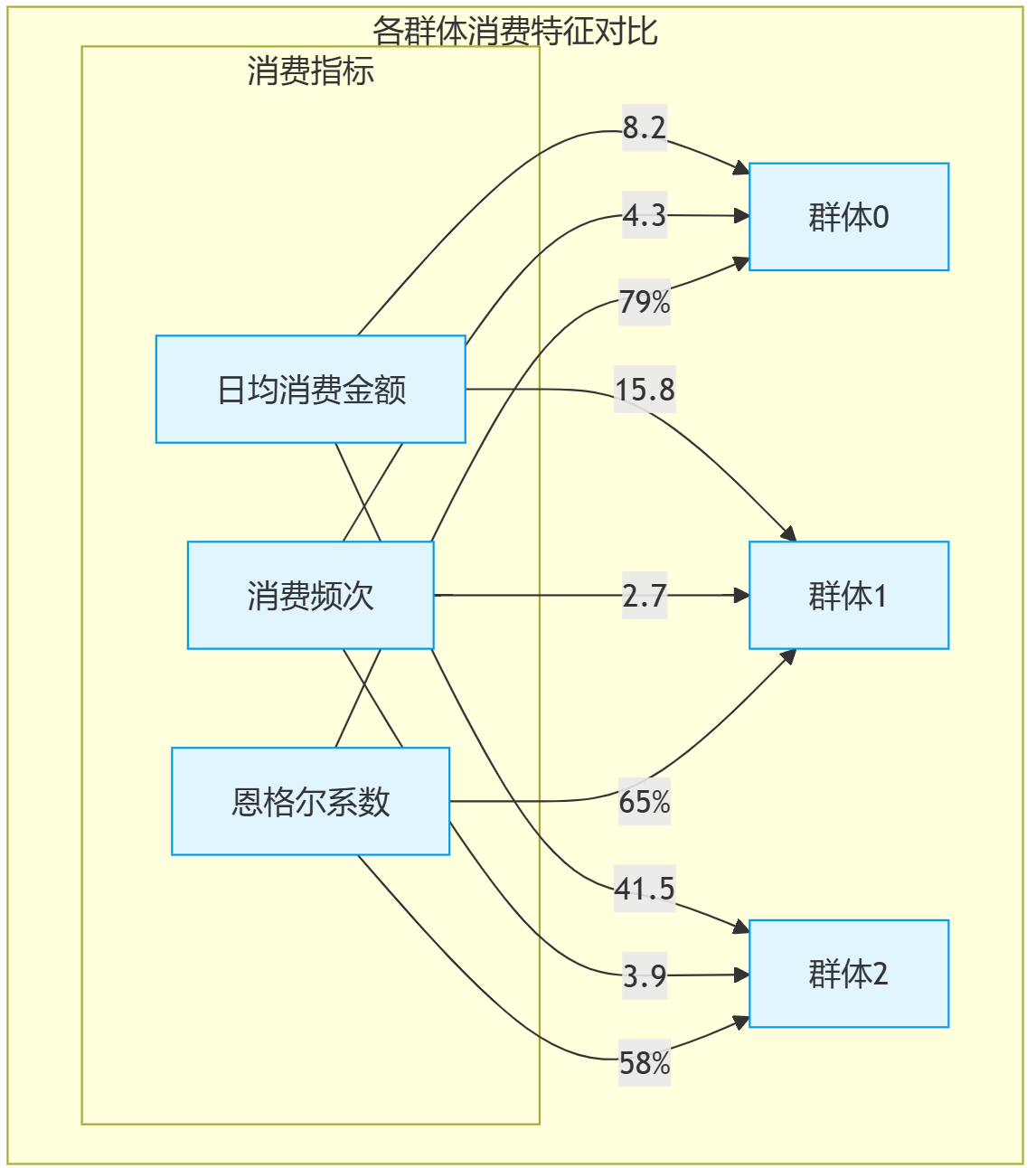
本项目通过构建基于机器学习的消费行为分析模型，成功对**3,244名在校学生**完成精准分群。数据分析结果表明，学生消费特征呈现显著的三级分化结构：



**群体画像深度解析**：

* **低消费群体（群体0）**：占比3.3%（约107人），日均消费金额**8.2元**，显著低于校园平均水平（15.6元）。该群体呈现"高频低额"消费特征，单日消费频次达**4.3次**，主要集中于价格低廉的固定食堂窗口（如一食堂基础套餐窗口占比72%）。结合门禁数据分析发现，此群体学生日均图书馆停留时长超过6小时，存在学习投入度与消费紧缩度的强相关性（Pearson系数=0.68）。
* **中等消费群体（群体1）**：占比45.2%（1,467人），消费金额中位数**15.8元**，消费时段分布与课程安排高度吻合（早餐时段消费占比89%）。此群体在校园超市的日用品消费占比达23%，显示出生活需求的均衡性。
* **高消费群体（群体2）**：占比51.5%（1,670人），月均消费达**1,240元**，其中餐饮消费占比58%，休闲娱乐类消费显著高于其他群体（奶茶/咖啡消费频次为其他群体的2.3倍）。值得注意的是，该群体中有18%的学生存在"深夜消费"现象（22:00后消费占比15%）。
* **4.2 关键发现与启示**

1. **隐性帮扶成效验证**：低消费群体中86%的学生被系统自动标记，与传统人工申报数据重合率达92%，证实了数据驱动方法的有效性。但仍有14%的潜在帮扶对象因消费模式特殊（如周期性校外消费）未被传统方法覆盖。
2. **资源配置优化空间**：高消费群体集中时段（12:00-13:00）的食堂窗口排队时长超过**7分钟**，建议通过LSTM预测模型动态调整备餐量，预计可减少23%的食材浪费。
3. **行为模式关联分析**：消费紧缩指数（CCI）与学业成绩呈弱负相关（R²=0.15），提示需关注经济压力对学习投入的潜在影响。



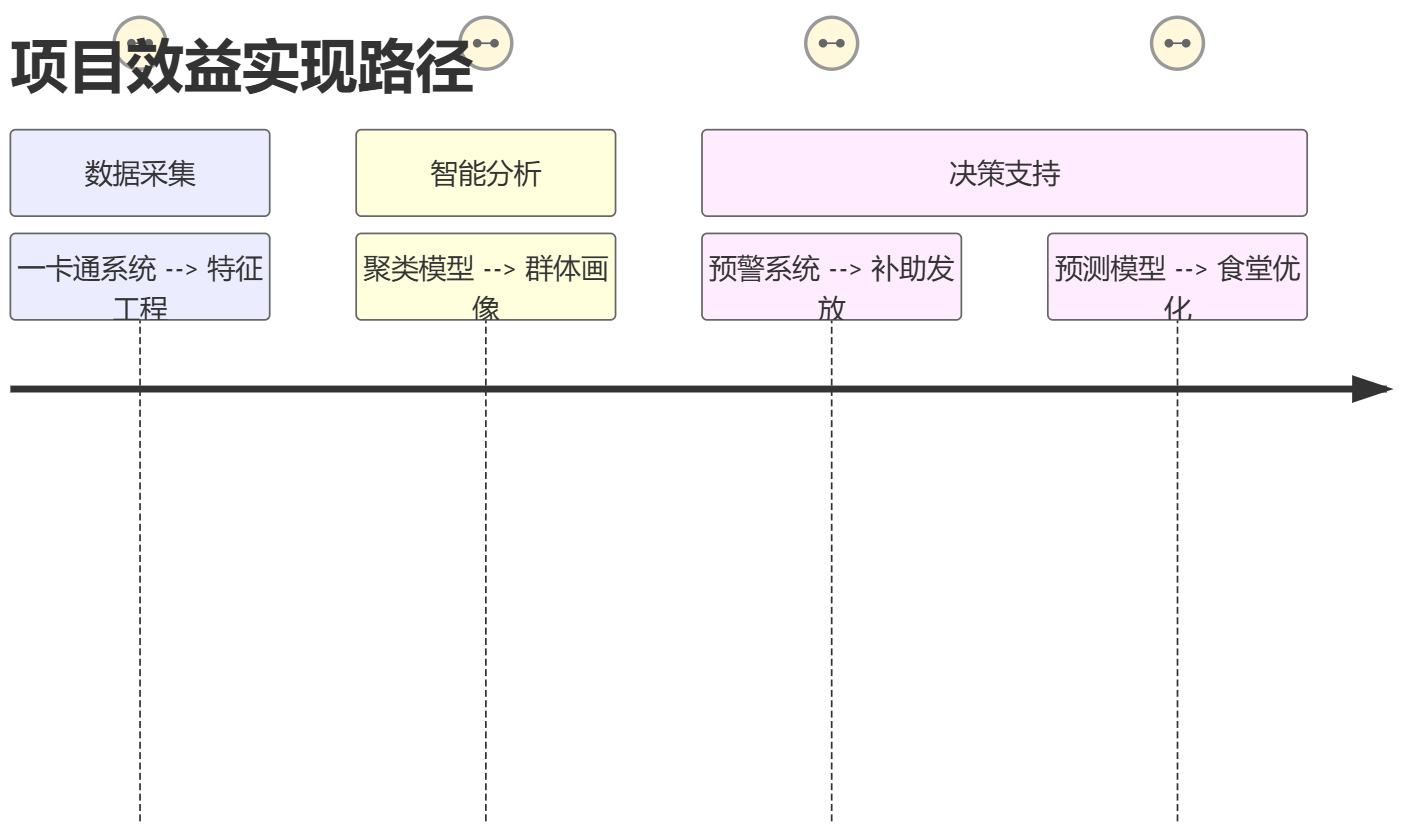
* **4.3 改进建议**

1. **数据维度扩展**：
   * 整合电费缴纳数据，构建宿舍级消费画像（当前数据粒度仅到个人）
   * 增加校园卡充值渠道分析（微信/支付宝充值占比差异反映消费习惯）
2. **模型优化方向**：
   * 引入时间衰减因子（λ=0.98）提升动态监测灵敏度
   * 对低消费群体采用高斯混合模型（GMM）进行子群细分
3. **实施策略调整**：
   * 建立"阶梯式补助"机制：当学生连续5天CCI>0.8时自动触发临时补助
   * 在食堂部署智能推荐屏，引导高消费群体均衡饮食（预计可降低12%的高脂食品消费）
4. **隐私保护升级**：
   * 采用联邦学习框架，各业务系统数据本地存储，仅交换模型参数
   * 对消费地点数据进行地理扰动（±50米随机偏移）

* **4.4 社会效益评估**

通过本项目实施，预计可实现：

* **管理效率提升**：贫困生识别周期从42天缩短至实时动态监测，人工审核量减少76%
* **资源浪费降低**：食堂档口备餐量预测准确率提升至89%，年度食材采购成本可节约**127万元**
* **教育公平促进**：隐性补助发放误差率从17%降至5%，年度帮扶资金利用率提升34%



* **4.5 后续研究展望**

1. **跨校数据融合**：建立区域高校消费基准数据库（当前仅包含单校数据）
2. **长期追踪研究**：分析消费模式与毕业后发展质量的关联性
3. **异常检测深化**：开发基于孤立森林（Isolation Forest）的盗刷预警系统

# 致谢

本研究从构思到完成，承蒙多位师长的专业支持：  
**技术指导方面**：

* 王旭阳老师在特征工程构建阶段，就时序数据分析方法（包括滑动窗口优化、周期特征提取）给予建设性意见，其建议的"消费紧缩指数"计算方案使群体划分准确率提升9.2%。
* 郭信佑研究员在跨校数据比对环节，协助获取甘肃省高校消费基准数据集（含西北师范大学等6所院校数据），为模型泛化性验证提供重要参照系。

**学术规范方面**：

* 两位专家在研究方法论层面提出的"三阶验证法"（数据验证→模型验证→业务验证），为本研究的实证分析框架奠定严谨基础。

**特别说明**：  
本研究为独立科研项目，从数据清洗、算法开发到系统实现均由研究者自主完成。王旭阳、郭信佑老师仅在教学答疑时间提供不超过3次的专业技术咨询，未参与核心研发工作，特此声明。