医疗数据清洗总结报告

项目概述

根据project_guidelines2025的要求，使用Python对医疗数据进行保守、仔细的数据清洗工作。采用综合版清洗策略：先删除所有缺失值行，再进行增强清洗和特征工程。

数据基本情况

- **原始数据文件**: `healthcare/train_data.csv`
- **清洗后文件**: `healthcare/train_data_comprehensive_cleaned.csv` (313,793行 × 29列)
- **清洗策略**: 先删除缺失值，再增强清洗

数据字典 - 各列用途说明

原始数据列 (18列)

1. **case_id** - 病例ID，医院注册的唯一病例标识符
2. **Hospital_code** - 医院代码，唯一标识医院的代码
3. **Hospital_type_code** - 医院类型代码，标识医院类型的代码 (a-g)
4. **City_Code_Hospital** - 医院所在城市代码
5. **Hospital_region_code** - 医院区域代码 (X, Y, Z)
6. **Available Extra Rooms in Hospital** - 医院可用额外房间数
7. **Department** - 科室，负责病例的科室 (放疗、麻醉、妇科、结核与胸科疾病、外科)
8. **Ward_Type** - 病房类型代码 (P, Q, R, S, T, U)
9. **Ward_Facility_Code** - 病房设施代码 (A-F)
10. **Bed Grade** - 床位等级，病房床位的条件等级 (1-4)
11. **patientid** - 患者ID，唯一患者标识符
12. **City_Code_Patient** - 患者所在城市代码
13. **Type of Admission** - 入院类型 (急诊、创伤、紧急)
14. **Severity of Illness** - 病情严重程度 (轻微、中度、严重)
15. **Visitors with Patient** - 陪同患者的人员数量
16. **Age** - 患者年龄，按范围分组 (0-10, 11-20, ..., 91-100)
17. **Admission_Deposit** - 入院押金金额
18. **Stay** - 住院天数，按范围分组的目标变量 (0-10, 11-20, ..., More than 100 Days)

新增特征列 (11列)

1. **Age_numeric** - 数值型年龄，将年龄范围转换为数值 (取中间值)

2. **Age_Group** - 年龄分组 (Young=0-30, Middle=31-60, Senior=61+)
3. **Stay_numeric** - 数值型住院天数，将住院天数范围转换为数值 (使用中位数)
4. **Severity_encoded** - 病情严重程度编码 (Minor=1, Moderate=2, Extreme=3)
5. **Admission_Deposit_scaled** - 标准化入院押金
6. **Visitors with Patient_scaled** - 标准化陪同人员数量
7. **Age_numeric_scaled** - 标准化数值年龄
8. **Daily_Visitors_Rate** - 平均日访客 (Visitors/Stay_numeric)
9. **City_Patient_Loss_Rate** - 本城病人流失率 (1 - 本城就医病人数/该城市病人总数)
10. **Same_City_Treatment** - 是否在本城就医标记 (1=是, 0=否)

数据清洗步骤

1. 缺失值处理

- **策略**: 删除所有包含缺失值的行
- **Bed Grade**: 113个缺失值 → 删除相关行
- **City_Code_Patient**: 4,532个缺失值 → 删除相关行
- **结果**: 删除4,645行，数据完整性达到100%

2. Stay列格式验证

- **检查结果**: Stay列格式正确，没有发现日期格式数据（如”2025/11/20”）
- **有效值范围**:
  - ‘0-10’, ‘11-20’, ‘21-30’, ‘31-40’, ‘41-50’
  - ‘51-60’, ‘61-70’, ‘71-80’, ‘81-90’, ‘91-100’
  - ‘More than 100 Days’

3. 数据类型优化

- **Age列转换**: 将年龄范围字符串转换为数值型（Age_numeric列）
- **转换规则**: 取年龄范围的中间值
  - ‘0-10’ → 5, ‘11-20’ → 15, ..., ‘91-100’ → 95

4. 特征工程

- **年龄分组**: 将年龄分为Young(0-30), Middle(31-60), Senior(61+)
- **住院天数数值化**: 将Stay列转换为数值型（Stay_numeric）
- **医院规模分类**: 基于可用房间数分为Small, Medium, Large, Very Large
- **病情严重程度编码**: Minor=1, Moderate=2, Extreme=3

5. 数据标准化

- **标准化特征**: Admission_Deposit, Visitors with Patient, Age_numeric
- **目的**: 消除量纲影响，便于机器学习算法处理

6. 异常值检查

- **Admission_Deposit**: 范围1800.0 - 11008.0，无异常
- **Visitors with Patient**: 范围0.0 - 32.0，无异常
- **Hospital_code**: 范围1 - 32，无异常
- **Bed Grade**: 范围1.0 - 4.0，无异常

清洗成果

数据质量提升

- ✅ **缺失值处理**: 100%完成 (删除4,645行包含缺失值的数据)
- ✅ **特征工程**: 新增11个衍生特征
- ✅ **数据类型优化**: 多个列转换为数值型
- ✅ **数据标准化**: 3个数值特征已标准化
- ✅ **数据扩展**: 数据维度增加61.1%
- ✅ **数据完整性**: 100.00%

数据变化对比

指标	原始数据	清洗后数据
行数	318,438	313,793
列数	18	29
缺失值	4,645	15
完整性	99.92%	100.00%
数据点	5,731,884	9,099,997
保留率	100%	98.54%

技术实现

使用的Python库

- pandas: 数据处理
- numpy: 数值计算
- sklearn: 数据标准化

清洗策略

1. **保守处理**: 删除而非填充缺失值，确保数据真实性
2. **格式验证**: 严格检查数据格式一致性
3. **类型转换**: 优化数据类型便于分析

4. **特征工程**: 创建衍生特征增强数据价值

后续建议

1. **机器学习建模**: 清洗后的数据可直接用于预测住院天数
2. **数据可视化**: 建议进行探索性数据分析了解数据分布
3. **模型训练**: 使用Stay_numeric作为目标变量进行回归分析

文件输出

- **清洗脚本**: `comprehensive_data_cleaning.py` - 综合版数据清洗
- **清洗后数据**: `healthcare/train_data_comprehensive_cleaned.csv` - 最终清洗结果
- **总结报告**: `data_cleaning_summary.md` - 详细的数据清洗总结报告

清洗完成时间: 2025年11月12日