**人工智能训练师（三级）操作技能考核**

**试题单**

准考证号：

试题代码：

试题名称：医疗研究数据清洗和标注设计

考核时间：20min

**1.**场地设备要求

（1）人工智能训练师主机 1 台；   
（2）Python 编译环境；   
（3）职业人群体检数据集。

**2.**工作任务

随着医学技术的进步和医疗资源的丰富，医疗研究在改善患者治疗效果、提升医疗服务质量方面起到了重要作用。研究人员通过分析大量患者的治疗数据，能够评估不同治疗方案的效果，发现潜在的健康问题，并提出针对性的治疗建议。这不仅可以帮助患者获得更好的治疗效果，还能为医疗机构优化资源配置、提升服务水平提供重要依据。  
现提供一份医疗研究数据集，训练集样本数据一共5441条记录。请补全2.1.4.ipynb代码，完成下面的数据预处理任务：  
1、加载数据集，查看表的数据类型，表结构和显示每一列的空缺值数量；  
2、将“就诊日期”和“诊断日期”规范为“yyyy-mm-dd”格式，并将“病人ID”列名改为“患者ID”，显示修改后的表结构；  
3、增加“诊断延迟”（诊断日期-就诊日期）和“病程”（当前日期-诊断日期）两列，删除不合理的数据（如负数，年龄为几百岁等）；  
4、检查数据集中的重复值并删除所有重复值，并记录删除的行数；  
5、对数据段[年龄，体重，身高]进行归一化处理；  
6、统计不同疾病类型的治疗结果分布，并画出柱状图；  
7、分析年龄和疾病严重程度的关系，绘制出散点图；  
8、保存处理后的数据，并命名为：2.1.4\_cleaned\_data.csv，保存到考生文件夹；  
9、制定数据清洗和数据标注规范，将答案写到答题卷文件中，答题卷文件命名为“2.1.4.docx”，保存到考生文件夹；  
10、将以上代码以及运行结果，以html格式保存并命名为2.1.4.html，保存到考生文件夹，考生文件夹命名为“准考证号+身份证后6位”。

**3.**技能要求

（1）能结合人工智能技术要求和业务特征，设计数据清洗和标注流程；   
（2）能结合人工智能技术要求和业务特征，制定数据清洗和标注规范。

**4.**质量指标

（1）深入理解业务，训练符合业务需求的模型。

**人工智能训练师（三级）操作技能考核**

**试题评分表**

准考证号：

试题代码：

试题名称：医疗研究数据清洗和标注设计

考核时间：20min

测量分评分表

| 细则编号 | 配分 | 评分细则描述 | 规定或  标称值 | 结果或  实际值 | 得分 |
| --- | --- | --- | --- | --- | --- |
| M1 | 2 | 加载数据集并指定编码为gbk正确得1分，查看表结构基本信息正确得1分，总计得2分； | 根据数据 |  |  |
| M2 | 1 | 修改列名正确得1分； | 根据数据 |  |  |
| M3 | 2 | 增加列操作正确得1分，删除不合理数据得1分，总计得2分； | 根据数据 |  |  |
| M4 | 1 | 删除重复值正确得1分； | 根据数据 |  |  |
| M5 | 1 | 数据归一化处理得1分； | 根据数据 |  |  |
| M6 | 1 | 绘制柱状图正确得1分； | 根据数据 |  |  |
| M7 | 1 | 正确绘制散点图得1分； | 根据数据 |  |  |
| M8 | 1 | 保存处理后的数据正确得1分； | 根据数据 |  |  |
| M9 | 2 | 回答数据清洗规范：每回答正确1个规范点，得1分，最高得2分； | 根据数据 |  |  |
| M10 | 3 | 回答数据标注规范：每回答正确1个规范点，得1分，最高得3分； | 根据数据 |  |  |
| 合计配分 | 15 | 合计得分 | |  |  |

**参考答案**：

（1）2.1.4.ipynb：

import pandas as pd

# 加载数据集并指定编码为gbk

data = pd.read\_csv('medical\_data.csv', encoding='gbk')

# 查看数据类型和表结构

print(data.dtypes)

# 查看表结构基本信息

print(data.info())

# 显示每一列的空缺值数量

print(data.isnull().sum())

# 修改列名

data.rename(columns={'病人ID': '患者ID'}, inplace=True)

# 查看修改后的表结构

print(data.head())

from datetime import datetime

# 增加诊断延迟和病程列

data['诊断延迟'] = (data['诊断日期'] - data['就诊日期']).dt.days

data['病程'] = (datetime(2024, 9, 1) - data['诊断日期']).dt.days

# 删除不合理数据

data = data[(data['诊断延迟'] >= 0) & (data['年龄'] > 0) & (data['年龄'] < 120)]

# 查看修改后的数据

print(data.describe())

# 删除重复值并记录删除的行数

initial\_rows = data.shape[0]

data.drop\_duplicates(inplace=True)

deleted\_rows = initial\_rows - data.shape[0]

print(f'\n删除的重复行数: {deleted\_rows}')

from sklearn.preprocessing import MinMaxScaler

# 对需要归一化的列进行处理

scaler = MinMaxScaler()

columns\_to\_normalize = ['年龄', '体重', '身高']

data[columns\_to\_normalize] = scaler.fit\_transform(data[columns\_to\_normalize])

# 查看归一化后的数据

print(data.head())

import matplotlib.pyplot as plt

import matplotlib.font\_manager as fm

# 统计治疗结果分布

treatment\_outcome\_distribution = data.groupby('疾病类型')['治疗结果'].value\_counts().unstack()

# 设置中文字体

font\_path = 'C:/Windows/Fonts/simhei.ttf' # 根据你的系统调整字体路径

my\_font = fm.FontProperties(fname=font\_path)

# 绘制柱状图

treatment\_outcome\_distribution.plot(kind='bar', stacked=True)

plt.title('不同疾病类型的治疗结果分布', fontproperties=my\_font)

plt.xlabel('疾病类型', fontproperties=my\_font)

plt.ylabel('治疗结果数量', fontproperties=my\_font)

plt.xticks(rotation=45, fontproperties=my\_font)

plt.legend(prop=my\_font)

plt.tight\_layout()

plt.show()

# 绘制散点图

plt.scatter(data['年龄'], data['疾病严重程度'])

plt.title('年龄和疾病严重程度的关系', fontproperties=my\_font)

plt.xlabel('年龄', fontproperties=my\_font)

plt.ylabel('疾病严重程度', fontproperties=my\_font)

plt.xticks(fontproperties=my\_font) # 设置x轴刻度标签的字体

plt.yticks(fontproperties=my\_font) # 设置y轴刻度标签的字体

plt.legend(prop=my\_font) # 设置图例字体

plt.show()

# 保存处理后得数据

output\_path = '2.1.4\_cleaned\_data.csv'

data.to\_csv(output\_path, index=False)

(2）制定数据清洗和特征工程规范  
**数据清洗规范（答对3点即可）**

1. 数据加载与初步检查，使用pd.read\_csv加载数据，指定encoding='gbk'解决中文编码问题。

2. 异常值处理： data.info()显示列类型和非空值数量，isnull().sum()统计缺失值。

3. 日期格式标准化： 将“就诊日期”和“诊断日期”转换为yyyy-mm-dd格式，确保后续时间差计算正确。重命名列名病人ID为患者ID，提升字段语义清晰度。

4.特征工程与异常值处理： 诊断延迟，计算诊断日期与就诊日期的天数差。病程：假设当前日期为2024年9月1日，计算病程天数（实际应用中应动态获取当前日期）。删除不合理数据：排除负诊断延迟、年龄<0或>120的异常记录。

5. 处理重复行：删除重复的行数

6. 对年龄、体重、身高进行Min-Max归一化，消除量纲差异。。

5. 保存清洗后的数据: 将经过清洗和处理后的数据保存为新的 CSV 文件，以便后续使用。  
**数据标注规范（答对3点即可）**

1. 数据来源：标注数据的来源，包括数据集的名称、获取日期和数据提供者。

2. 数据描述：提供详细的数据描述，包括每列数据的含义、单位和可能的取值范围。

3. 特征选择: 确定对目标变量预测最有用的特征。

4. 绘制柱状图：展示不同疾病类型的治疗结果分布，使用堆叠柱状图增强对比。绘制散点图：探索年龄（归一化后）与疾病严重程度的关系，检测潜在趋势

5. 保存处理后的数据：保存处理后的数据，并记录保存文件的路径和文件名。

6. 数据清洗和标注规范文档