**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：低碳生活行为影响因素数据清洗和标注流程设计

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）大学生低碳生活行为的影响因素数据集。

**2.**工作任务

在应对气候变化的背景下，了解和促进低碳生活行为变得越来越重要。现要求根据提供的“大学生低碳生活行为的影响因素数据集”，选择合适的特征，开发一个预测大学生低碳生活行为的模型。提供的数据集样本数据包含300多个关于大学生低碳生活行为的自变量和因变量。自变量为主观规范、知觉行为控制、低碳行为态度，中介变量为行为意愿，因变量为低碳生活行为。同时，性别、生源地、月生活费作为控制变量进行研究。在开发预测模型之前，首先要对数据进行数据清洗和标注。补全2.1.2.ipynb代码。完成下面的数据预处理任务，并设计一套标注流程规范：  
（1）正确加载数据集，并显示前五行的数据；  
（2）检查数据集中的缺失值，使用删除包含缺失值的行的办法处理，记录缺失值处理后的数据行数；  
（3）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；  
（4）对数值型数据进行标准化处理，确保数据在同一量纲下进行分析；  
（5）根据业务需求和数据特性，选择对低碳生活行为预测最有用的特征，将“低碳行为积极性”设为目标变量并标注。选择以下特征：  
1.您的性别○男性 ○女性  
2.您的年级○大一 ○大二 ○大三 ○大四  
3.您的生源地○农村 ○城镇（乡镇） ○地县级城市 ○省会城市及直辖市  
4.您的月生活费○≦1,000元 ○1,001-2,000元 ○2,001-3,000元 ○≧3,001元  
5.您进行过绿色低碳的相关生活方式吗?  
6.您觉得“低碳”，与你的生活关系密切吗？  
7.低碳生活是否会成为未来的主流生活方式？  
8.您是否认为低碳生活会提高您的生活质量？  
（6）对数据进行划分（8:2）；  
（7）保存处理后的数据，并命名为：2.1.2\_cleaned\_data.csv，保存到考生文件夹；  
（8）制定数据清洗和标注规范，将答案写到答题卷文件中，答题卷文件命名为“2.1.2.docx”，保存到考生文件夹；  
（9）将以上代码以及运行结果，以html格式保存并命名为2.1.2.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）数据预处理步骤完整，方法选择合理；   
（2）代码实现正确，结果符合预期。

**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：低碳生活行为影响因素数据清洗和标注流程设计

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 | 1 | 数据集正确加载得1分； | 根据数据 |  |  |
| M2 | 2 | 删除缺失值所在行正确得1分，获取处理前和处理后的数据行数正确得1分，总计得2分； | 根据数据 |  |  |
| M3 | 1 | 删除重复的数据正确得1分； | 根据数据 |  |  |
| M4 | 1 | 对数据进行标准化处理正确得1分； | 根据数据 |  |  |
| M5 | 1 | 选择特征正确得1分； | 根据数据 |  |  |
| M6 | 1 | 创建目标变量正确得1分； | 根据数据 |  |  |
| M7 | 1 | 数据划分正确得1分； | 根据数据 |  |  |
| M8 | 2 | 合并处理后的数据正确得1分，保存处理后的数据正确得1分，总计得2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）2.1.2.ipynb：

import pandas as pd

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

#读取一个Excel文件，并将读取到的数据存储在变量data中

data = pd.read\_excel('大学生低碳生活行为的影响因素数据集.xlsx')

#打印出数据集的前5行

print(data.head())

#处理数据集中的缺失值

initial\_row\_count = data.shape[0] #处理前的数据行数

data = data.dropna() #删除缺失值所在行

final\_row\_count = data.shape[0] #处理后的数据行数

print(f'处理后数据行数: {final\_row\_count}, 删除的行数: {initial\_row\_count - final\_row\_count}')

#删除重复行

data =data.drop\_duplicates()

from sklearn.preprocessing import StandardScaler

numerical\_features = ['4.您的月生活费○≦1,000元 ○1,001-2,000元 ○2,001-3,000元 ○≧3,001元']

scaler = StandardScaler()

data[numerical\_features] = scaler.fit\_transform(data[numerical\_features])

selected\_features = [

'1.您的性别○男性 ○女性', '2.您的年级○大一 ○大二 ○大三 ○大四', '3.您的生源地○农村 ○城镇（乡镇） ○地县级城市 ○省会城市及直辖市', '4.您的月生活费○≦1,000元 ○1,001-2,000元 ○2,001-3,000元 ○≧3,001元',

'5.您进行过绿色低碳的相关生活方式吗?', '6.您觉得“低碳”，与你的生活关系密切吗？',

'7.低碳生活是否会成为未来的主流生活方式？', '8.您是否认为低碳生活会提高您的生活质量？'

]

X = data[selected\_features]

# 创建目标变量

y = data['低碳行为积极性']

from sklearn.model\_selection import train\_test\_split

# 数据划分

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# 保存处理后的数据

cleaned\_data = pd.concat([X, y], axis=1)

cleaned\_data.to\_csv('2.1.2\_cleaned\_data.csv', index=False)

（2）数据清洗和标注规范  
**数据清洗规范（答对2点即可）**

1. 数据加载：使用 pandas 库加载数据集，检查数据的基本结构和类型。

2. 检查缺失值：统计每行的缺失值数量，并删除包含缺失值的行以确保数据完整性。

3. 处理重复行：统计重复的行数，并删除重复的行数

4. 数据标准化: 对数值型数据进行标准化，以消除量纲影响，使用标准化方法。

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. 目标变量设定: 将数据集中用于预测的目标变量定义为“5.您进行过绿色低碳的相关生活方式吗?”。

5. 数据划分: 将数据分为训练集和测试集，通常采用 80/20 的比例，以便于模型的训练和评估。

6. 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。

7. 数据清洗和标注规范文档