**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：健康与营养咨询数据预处理与数据规范设计

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）健康咨询客户数据集。

**2.**工作任务

在健康与营养咨询领域，客户的健康数据是评估其饮食和生活方式建议的重要依据。通过对客户健康数据的分析，可以帮助健康咨询师更准确地评估客户的健康状况，并制定个性化的营养和健康管理计划。现提供一份健康咨询客户数据集。请补全2.1.5.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）缺失值处理：对于含有缺失值的行进行删除操作；  
（3）数据类型转换：将“Your age”列的数据类型转换为整数类型，并处理其中的异常值；  
（4）数据去重：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）数据归一化处理：对“如何形容你的当前健身水平？”（How do you describe your current level of fitness ?）列中的数据进行归一化处理；  
（6）绘制健身频率分布的饼图；  
（7）对数据进行标注划分；  
（8）保存处理后的数据，并命名为：2.1.5\_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“2.1.5.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2.1.5.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：健康与营养咨询数据预处理与数据规范设计

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 | 数据集正确加载的1分，查看表结构基本信息得1分，总共2分； | 根据数据 |  |  |
| M2 | 1 | 查看每一列的空缺值数量正确得1分； | 根据数据 |  |  |
| M3 | 1 | 删除含有缺失值的行正确得1分； | 根据数据 |  |  |
| M4 | 1 | 将“Your age”列的数据类型转换为整数类型并处理异常值得1分； | 根据数据 |  |  |
| M5 | 1 | 删除所有重复值正确得1分； | 根据数据 |  |  |
| M6 | 1 | 对数据进行归一化处理得1分； | 根据数据 |  |  |
| M7 | 1 | 绘制健身频率分布的饼图得1分； | 根据数据 |  |  |
| M8 | 1 | 对数据集进行划分正确得1分； | 根据数据 |  |  |
| M9 | 1 | 保存处理后的数据正确得1分； | 根据数据 |  |  |
| M10 | 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 |  |  |
| M11 | 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）2.1.5.ipynb：

import pandas as pd

# 加载数据集

data = pd.read\_csv( '健康咨询客户数据集.csv')

# 查看表结构基本信息

print(data.info())

# 删除含有缺失值的行

data\_cleaned = data.dropna()

# 转换 'Your age' 列的数据类型为整数类型，并处理异常值

data\_cleaned.loc[:, 'Your age'] = pd.to\_numeric(data\_cleaned['Your age'], errors='coerce')

data\_cleaned = data\_cleaned.dropna(subset=['Your age'])

data\_cleaned = data\_cleaned[data\_cleaned['Your age'] >= 0]

data\_cleaned.loc[:, 'Your age'] = data\_cleaned['Your age'].astype(int)

print(data\_cleaned['Your age'].dtype)

# 检查和删除重复值

duplicates\_removed = data\_cleaned.duplicated().sum()

data\_cleaned = data\_cleaned.drop\_duplicates()

print(f"Removed {duplicates\_removed} duplicate rows")

from sklearn.preprocessing import LabelEncoder

# 归一化 'How do you describe your current level of fitness ?' 列

label\_encoder = LabelEncoder()

data\_cleaned['How do you describe your current level of fitness ?'] = label\_encoder.fit\_transform(data\_cleaned['How do you describe your current level of fitness ?'])

print(data\_cleaned['How do you describe your current level of fitness ?'].unique())

from sklearn.preprocessing import LabelEncoder

import matplotlib.pyplot as plt

# 去掉列名中的空格

data.columns = data.columns.str.strip()

# 显示数据集的列名

print(data.columns)

# 删除包含缺失值的行

data\_cleaned = data.dropna(subset=['How often do you exercise?'])

# 统计不同健身频率的分布情况

exercise\_frequency\_counts = data\_cleaned['How often do you exercise?'].value\_counts()

# 绘制饼图

plt.figure(figsize=(10, 6))

exercise\_frequency\_counts.plot.pie(autopct='%1.1f%%', startangle=90, colors=plt.cm.Paired.colors)

plt.title('Distribution of Exercise Frequency')

plt.ylabel('')

plt.show()

import pandas as pd

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

import matplotlib.pyplot as plt

# 填充缺失值

data\_filled = data.apply(lambda x: x.fillna(x.mode()[0]))

# 划分数据（测试集占比20%）

train\_data, test\_data = train\_test\_split(data\_filled, test\_size=0.2, random\_state=42)

# 保存处理后的数据

cleaned\_file\_path = '2.1.5\_cleaned\_data.csv'

data\_filled.to\_csv(cleaned\_file\_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2点即可）**

1. 数据加载与初步检查，使用pd.read\_csv加载数据，info()显示列类型和非空值数量，isnull().sum()统计缺失值

2. 缺失值处理： 直接删除包含缺失值的行（dropna()），适用于缺失率较低的场景。。

3. 年龄数据处理： pd.to\_numeric强制转换数值类型，errors='coerce'将无效值转为NaN后删除，并过滤负值

4. 列名规范化：str.strip()去除列名首尾空格，确保字段引用准确性

5. 处理重复行：删除重复的行数

6. 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. 标签编码： 对分类变量（健身水平）使用LabelEncoder转换为数值标签，便于后续建模。

4. 数据划分：按8:2划分训练集和测试集，random\_state保证可复现性。

5. 绘制饼图展示健身频率分布，autopct显示百分比，colormap指定配色方案。

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