**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：智慧交通中燃油效率模型的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)汽车燃油效率数据集（auto-mpg.csv）；

**2.**工作任务

在现代交通中，燃油效率（MPG）是衡量汽车性能和交通系统优化的重要指标之一。高效的燃油利用不仅能够降低车辆运营成本，还能减少碳排放，促进环保。开发一个用于预测汽车燃油效率的模型可以帮助智慧交通系统优化路线规划和车辆调度，从而提升整体交通效率和减少能源消耗。此外，这样的模型还可以帮助消费者做出更明智的购车决策，并帮助厂商优化汽车设计。  
现要求根据提供的汽车燃油效率数据集，补全2.1.1.ipynb代码。选择合适的特征，开发一个燃油效率预测模型。在开发预测模型之前，首先要对数据进行数据清洗和标注，请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及数据类型。  
(2)检查数据集中的缺失值并删除缺失值所在的行。  
(3)将“horsepower”列转换为数值类型，并处理转换中的异常值。  
(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，选择对燃油效率预测最有用的特征：选择以下特征：'cylinders'、'displacement'、'horsepower'、'weight'、'acceleration'、'model year'、'origin'  
(6)将“mpg”设为目标变量并标注；  
(7)对数据进行标注和划分；  
(8)保存处理后的数据，并命名为：2.1.1\_cleaned\_data.csv，保存到考生文件夹；  
(9)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“2.1.1.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2.1.1.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：智慧交通中燃油效率模型的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 | 检查缺失值得1分，删除缺失值得1分，总计得2分； | 根据数据 |  |  |
| M3 | 1 | 处理转换中的异常值得1分； | 根据数据 |  |  |
| M4 | 1 | 对数据进行标准化处理得1分； | 根据数据 |  |  |
| M5 | 1 | 选择特征得1分； | 根据数据 |  |  |
| M6 | 2 | 创建自变量得1分，目标变量得1分，总计得2分； | 根据数据 |  |  |
| M7 | 1 | 数据集划分得1分； | 根据数据 |  |  |
| M8 | 1 | 保存处理后的数据得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）2.1.1.ipynb：

import pandas as pd

# 加载数据集并显示数据集的前五行 1分

data = pd.read\_csv('auto-mpg.csv')

print("数据集的前五行:")

print(data.head())

# 显示每一列的数据类型

print(data.dtypes)

# 检查缺失值并删除缺失值所在的行 2分

print("\n检查缺失值:")

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

data = data.dropna()

# 将 'horsepower' 列转换为数值类型，并处理转换中的异常值 1分

data['horsepower'] = pd.to\_numeric(data['horsepower'], errors='coerce')

data = data.dropna(subset=['horsepower'])

# 显示每一列的数据类型

print(data.horsepower.dtypes)

# 检查清洗后的缺失值

print("\n检查清洗后的缺失值:")

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

from sklearn.preprocessing import StandardScaler

# 对数值型数据进行标准化处理 1分

numerical\_features = ['displacement', 'horsepower', 'weight', 'acceleration']

scaler = StandardScaler()

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

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

# 选择特征和目标变量 2分

selected\_features = ['cylinders', 'displacement', 'horsepower', 'weight', 'acceleration', 'model year', 'origin']

X = data[selected\_features]

y = data['mpg']

# 划分数据集为训练集和测试集 1分

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

# 将特征和目标变量合并到一个数据框中

cleaned\_data = X.copy()

cleaned\_data['mpg'] = y

# 保存清洗和处理后的数据

cleaned\_data.to\_csv('2.1.1\_cleaned\_data.csv', index=False)

# 打印消息指示文件已保存

print("\n清洗后的数据已保存到 2.1.1\_cleaned\_data.csv")

（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：统计每列的缺失值数量，并删除包含缺失值的行以确保数据完整性。

3. 转换与处理异常值: 将数值列（如“horsepower”）转换为数值类型，并处理无法转换的值（例如，将其变为缺失值）。

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“mpg”（燃油效率）。

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