**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随机森林模型开发与测试

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2.2.2.ipynb代码。选择合适的特征，开发一个燃油效率预测模型。利用测试工具对模型进行测试，并对测试结果进行分析，完成测试报告，并运用工具对错误原因进行纠正。

（1）正确加载数据集，显示前五行的数据。

（2）使用线性回归模型进行模型训练，要求设定自变量和因变量，并根据自变量特征进行模型训练，最终将训练好的模型以“2.2.2\_model.pkl”命名保存到考生文件夹，结果文件以“2.2.2\_results.txt”命名保存到考生文件夹。

（3）使用测试工具对模型进行测试，并记录测试结果，以“2.2.2\_report.txt”命名保存到考生文件夹。

（4）运用工具分析算法中错误案例产生的原因并进行纠正，重新进行模型训练，并以“2.2.2\_results\_rf.txt”命名保存到考生文件夹。

（5）对测试结果进行详细分析，并编写测试报告，包括模型性能评估、错误分析及改进建议，将答案以“2.2.2.docx”命名写到答题卷文件中，并保存到考生文件夹。

（6）将以上代码以及运行结果，以html格式保存并命名为2.2.2.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”。

数据集说明：

mpg : 燃油效率，每加仑油可以行驶的英里数。

cylinders: 发动机气缸的数量。

displacement: 这是指发动机所有气缸的总容积。

horsepower: 发动机的马力。

weight: 车辆的重量。

acceleration: 加速。

model year: 车辆的生产年份。

origin: 指的是车辆的制造地或品牌所属国家。

car name: 每辆车的具体名称或型号。

**3.**技能要求

(1) 能维护日常训练集与测试集。

(2) 能使用工具对算法进行训练。

(3) 能使用测试工具对人工智能产品的使用进行测试。

(4) 能对测试结果进行分析，编写测试报告。

(5) 能运用工具，分析算法中错误案例产生的原因并进行纠正。

**4.**质量指标

(1) 深入理解业务，训练符合业务需求的模型。

(2) 数据预处理步骤完整，方法选择合理。

(3) 代码实现正确，结果符合预期。

(4) 测试结果分析全面，报告详细。

**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随机森林模型开发与测试

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 | 正确加载数据集得1分，显示前五行的数据得1分，总计得2分； | 根据数据 |  |  |
| M2 | 2 | 正确处理缺失值得2分； | 根据数据 |  |  |
| M3 | 2 | 正确定义自变量得1分，因变量得1分，总计得2分； | 根据数据 |  |  |
| M4 | 1 | 训练集与测试集的准确划分得1分； | 根据数据 |  |  |
| M5 | 2 | 初始化线性回归模型得2分； | 根据数据 |  |  |
| M6 | 1 | 训练线性回归模型得1分； | 根据数据 |  |  |
| M7 | 1 | 保存线性回归模型得1分 | 根据数据 |  |  |
| M8 | 1 | 使用线性回归模型预测得1分； | 根据数据 |  |  |
| M9 | 1 | 正确输出线性回归模型预测结果得1分； | 根据数据 |  |  |
| M10 | 1 | 初始化随机森林回归模型得1分； | 根据数据 |  |  |
| M11 | 1 | 训练随机森林回归模型得1分； | 根据数据 |  |  |
| M12 | 1 | 正确使用随机森林模型进行预测结果得1分； | 根据数据 |  |  |
| M13 | 1 | 保存随机森林模型预测结果得1分； | 根据数据 |  |  |
| M14 | 3 | 正确填写2.2.2.docx测试报告：其中模型性能评估得1分、错误分析得1分，改进建议得1分，总计得3分； | 根据数据 |  |  |
| 合计配分 | 20 | 合计得分 | |  |  |

**参考答案**：

（1）2.2.2.ipynb代码

import pandas as pd

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

from sklearn.linear\_model import LinearRegression

from sklearn.preprocessing import StandardScaler

from sklearn.pipeline import Pipeline

import pickle

from sklearn.ensemble import RandomForestRegressor

# 加载数据集

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

# 显示前五行数据

print(df.head())

# 处理缺失值

# 将 'horsepower' 列中的所有值转换为数值类型

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

# 删除包含缺失值的行

df = df.dropna()

# 选择相关特征进行建模

X = df[['cylinders', 'displacement', 'horsepower', 'weight', 'acceleration', 'model year', 'origin']]

y = df['mpg']

# 将数据集划分为训练集和测试集

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

# 创建包含标准化和线性回归的管道

pipeline = Pipeline([

('scaler', StandardScaler()),

('linreg', LinearRegression())

])

# 训练模型

pipeline.fit(X\_train, y\_train)

# 保存训练好的模型

with open('2.2.2\_model.pkl', 'wb') as model\_file:

pickle.dump(pipeline, model\_file)

# 预测并保存结果

y\_pred = pipeline.predict(X\_test)

results\_df = pd.DataFrame(y\_pred, columns=['预测结果'])

results\_df.to\_csv('2.2.2\_results.txt', index=False)

# 测试模型

with open('2.2.2\_report.txt', 'w') as results\_file:

results\_file.write(f'训练集得分: {pipeline.score(X\_train, y\_train)}\n')

results\_file.write(f'测试集得分: {pipeline.score(X\_test, y\_test)}\n')

# 训练一个随机森林回归模型作为替代模型

rf\_model = RandomForestRegressor(n\_estimators=100, random\_state=42)

# 训练随机森林回归模型

rf\_model.fit(X\_train, y\_train)

# 使用随机森林模型进行预测

y\_pred\_rf = rf\_model.predict(X\_test)

# 保存新的结果

results\_rf\_df = pd.DataFrame(y\_pred\_rf, columns=['预测结果'])

results\_rf\_df.to\_csv('2.2.2\_results\_rf.txt', index=False)

# 测试模型并保存得分

with open('2.2.2\_report\_rf.txt', 'w') as results\_rf\_file:

results\_rf\_file.write(f'训练集得分: {rf\_model.score(X\_train, y\_train)}\n')

results\_rf\_file.write(f'测试集得分: {rf\_model.score(X\_test, y\_test)}\n')

（5）对测试结果进行详细分析，并编写测试报告，包括模型性能评估、错误分析及改进建议，

测试报告分析

1. 模型性能评估

模型 训练集R² 测试集R² 过拟合风险

线性回归 0.82 0.81 低

随机森林 0.98 0.89 较高

线性回归：测试集R²为0.81，表明模型能解释81%的燃油效率方差。特征与目标变量间存在线性关系，但可能存在非线性未被捕捉。

随机森林：测试集R²提升至0.89，显著优于线性回归，说明燃油效率受复杂非线性因素影响。

2. 错误案例分析

主要错误来源：

线性模型局限性：排量（displacement）与MPG可能存在非线性关系（如二次关系），线性回归无法充分拟合。

特征交互缺失：气缸数（cylinders）与重量（weight）的交互效应未被线性模型捕捉。

异常值影响：数据中存在极端马力值，影响模型稳定性。

3. 改进建议

特征工程：添加非线性项（如排量的平方）或交互项（气缸数×重量）。对马力列进行对数变换，降低异常值影响。

模型优化：对线性回归使用正则化（Ridge/Lasso）防止过拟合。调整随机森林超参数（如max\_depth=5）以降低过拟合

数据增强：收集更多车辆技术参数（如变速箱类型、驱动形式）。加入环境因素（如温度、海拔）以提升泛化能力。

错误纠正与验证 纠正措施：采用随机森林模型后，测试集R²提升8%，证明其能更好捕捉非线性关系。

验证方法：通过SHAP值分析特征重要性，发现model year（车型年份）和weight（重量）对预测影响最大，符合实际物理规律。