实训1 合并年龄、平均血糖和中风患者信息数据

源程序：

import pandas as pd

# 读取表格

stroke\_data = pd.read\_excel('healthcare-dataset-stroke.xlsx')

age\_abs\_data = pd.read\_excel('healthcare-dataset-age\_abs.xlsx')

# 合并数据

merged\_data = pd.merge(stroke\_data, age\_abs\_data, on='编号')

# 保存合并后的数据

merged\_data.to\_excel('merged\_healthcare\_data.xlsx', index=False)

print("合并成功，保存为 merged\_healthcare\_data.xlsx")

display(merged\_data)

过程性结果：



结论：

基于编号合并了两个表格的数据。（1.8年龄的数据在此处没有对应编号，所以已经被自然剔除。）

实训2 剔除年龄异常的数据

根据实训内容，如果年龄值为1.8，则视为异常值，需要剔除。

源程序：

# 剔除年龄异常的数据

cleaned\_data = merged\_data[merged\_data['年龄'] != 1.8]

# 保存清理后的数据

cleaned\_data.to\_excel('cleaned\_healthcare\_data.xlsx', index=False)

print("剔除异常年龄数据，保存为 cleaned\_healthcare\_data.xlsx")

display(cleaned\_data)

过程性结果：



结论：

已经剔除了年龄为1.8的异常值数据。

实训3 离散化年龄特征

为了离散化年龄特征，我们可以将连续型数据转换为离散型数据。我们可以使用pd.cut()函数来实现。

源程序：

# 定义年龄区间

bins = [0, 20, 40, 60, 80, 100]

labels = ['0-20', '21-40', '41-60', '61-80', '81-100']

# 离散化年龄特征

cleaned\_data['年龄区间'] = pd.cut(cleaned\_data['年龄'], bins=bins, labels=labels, right=False)

# 保存离散化后的数据

cleaned\_data.to\_excel('discretized\_healthcare\_data.xlsx', index=False)

print("离散化年龄特征，保存为 discretized\_healthcare\_data.xlsx")

display(cleaned\_data)

过程性结果：



结论：年龄特征已被离散化至一个区间。