第 1 次平时作业

1. 数据预处理

a) 数据规范化

某数据集记录了某单位成年人的身高和体重数据。身高范围为 1.4-1.9 m，体重范围为 40-90kg。请结合表 1，回答下列问题：

表 1. 体型数据表

ID	身高（m）	体重（kg）
1	1.70	50
2	1.60	50
3	1.70	60
4	1.65	45
……	……	……

1） 请以“判断用户体型相似性”为例，说明数据规范化的必要性。

2） 请采用 Min-Max 规范化，将表 1 用户的身高和体重数据规范化到[0,1].

3） 用户 2、3、4 中，谁和用户 1 的体型更相似？请给出结论和计算依据。

补充知识：

体型相似可以通过身高和体重的欧几里得距离判断。距离越小，体型越相似。用户i和j的体型距离$\mathrm{D}(i,j)$的计算公式如下：

$$\mathrm{D}(i,j)=\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2}$$

其中，x_i和x_j表示用户i和j的身高，y_i和y_j表示用户i和j的体重。

b) 数据离散化

分别利用等宽离散化和等深离散化将表 2 中的属性“年收入(万元)”转换为“低收入”、“中收入”和“高收入”三档。

表 2. 个人信息表

年龄(岁)	性别	年收入(万元)
25	男	10
27	女	25
30	男	30
45	女	60
28	男	40
32	男	20
52	男	50
35	女	30
55	男	100
48	女	120

2. 分类分析

利用 Titanic 数据集（详见 train.csv 和 test.csv），预测乘客及船员的生存情况。通过数据预处理，开展分类分析，主要进行数据集划分、模型参数选择和实验结果比较，重点实践决策树、逻辑回归和朴素贝叶斯三个算法。作业要求提交完整的分类分析报告，包括文字叙述、代码和运行结果。

作业说明：

1. 数据预处理不要求编程，给出求解结果和思路的文字表述。
2. 分类分析要求编程解决，请同时提交.ipynb 和.html 版本，.html 是.ipynb 运行结果的导出版本。
3. 截止日期：2024/04/07 23:59