步骤（要求）：

1. 自行收集数据（爬虫、公共数据集），数据量不少于2000条，特征不少于8个，理解原始数据各个指标的具体含义，明确问题，要利用数据做什么；
2. 对原始数据进行数据检测（处理）、数据特征分布分析、对比分析、相关分析等操作，从原始数据中获取有用信息；
3. 基于研究问题，为初步处理后的数据选择合适的模型，进行建模分析（模型对比分析）；
4. 对建立的模型进行模型评估，若模型不能通过评估，继续优化模型（再次进行数据处理或调参等）；
5. 总结（总结全文具体做了什么操作，能得出什么结果）

* 注：最终课程论文以小组形式完成，一组1-4人，可单人成组，最多不超过4人。

1. 绪论
   1. 研究背景

阐述研究主题的重要性以及开展研究的必要性，比如它是社会热点？民

生热点？产生较大影响？同时提出研究问题。

* 1. 研究意义及目的

研究该主题能带来的好处？积极效果

* 1. 研究现状

文献分析，对于研究主题，现在研究到什么程度，做出了哪些成果，还

存在什么问题等。

* 1. 研究内容及框架结构

本文主要研究什么内容，预估达到什么目的，为了实现该目的，进行了

哪些操作。本文一共分为几个章节，每个章节具体做的事。

1. 数据采集及处理
   1. 数据介绍

本文基于UCI数据库中的葡萄酒质量数据集进行数据分析，旨在探索葡萄酒质量及其影响因素。该数据集包含4898个样本以及10个指标，各指标具体含义如表2-1：

表2-1 葡萄酒质量数据指标说明表

|  |  |  |
| --- | --- | --- |
| **指标名称** | **含义** | **类型** |
| fixed acidity | 固定酸度 | 连续数据 |
| volatile acidity | 挥发性酸度 | 连续数据 |
| citric acid | 柠檬酸 | 连续数据 |
| residual sugar | 残余糖分 | 连续数据 |
| chlorides | 氯化物 | 连续数据 |
| free sulfur dioxide | 游离二氧化硫 | 离散数据 |
| density | 密度 | 连续数据 |
| pH | pH值 | 连续数据 |
| alcohol | 酒精含量 | 连续数据 |
| quality | 质量 | 分类数据/离散数据 |

* 1. 数据处理

（1）描述性统计分析

图片包含 图形用户界面

AI 生成的内容可能不正确。

图2-1 描述性统计指标图

由图2-1各统计指标计算结果可知，9个指标均不存在缺失值数据，因此不需要对原始数据进行缺失值处理。虽然9个指标均值与中位数相差较小，但各指标极差数值较大，初步估计各指标大部分数据分布较为集中，但可能存在较小部分两端极值。

（2）重复值检测与处理

对原始数据进行重复值检测，结果显示4898个样本中存在943条重复记录，对重复记录进行删除处理后，利用剩下3955条记录进行后续的分析。

（3）缺失值检测与处理

（4）异常值检测与处理

表2-2 异常值结果分析表

|  |  |  |
| --- | --- | --- |
| **指标名称** | **含义** | **异常值数量** |
| fixed acidity | 固定酸度 | 106 |
| volatile acidity | 挥发性酸度 | 168 |
| citric acid | 柠檬酸 | 223 |
| residual sugar | 残余糖分 | 5 |
| chlorides | 氯化物 | 185 |
| free sulfur dioxide | 游离二氧化硫 | 39 |
| density | 密度 | 3 |
| pH | pH值 | 65 |
| alcohol | 酒精含量 | 0 |

表2-2异常值检测结果显示，fixed acidity、citric acid、chlorides等指标内部异常值数量相对较少，包含异常值的记录仅659条，相较于原始数据而言，占比较小，直接删除包含异常值的记录，不会导致原始数据丢失过多有用信息，对后续可视化分析或建模分析效果影响较小，因此，选择直接删除的方式处理异常数据，保证数据信息的完整。

（5）数据标准化

（6）数据变换

1. 数据特征分析
   1. 特征分布分析

（1）葡萄酒质量分布分析

图表, 条形图

AI 生成的内容可能不正确。

图3-1 葡萄酒质量分布柱状图

由图3-1分析可知，葡萄酒质量从3到9不等，其中质量为6的葡萄酒样本数量最多，达到了1555条，接近1600条，而质量为3和质量为9的葡萄酒样本数量最少，不足20条，说明大部分葡萄酒为中等质量葡萄酒，质量特别差和特别好的葡萄酒，相对较少，原始数据中存在样本不均衡问题。

（2）固定酸度分布分析

图表, 直方图

AI 生成的内容可能不正确。

图3-2 固定酸度分布图

由图3-2分析可知，样本数据的固定酸度大致呈现对称分布，在[6.4,7.2]区间的样本数量最多，接近1400条，约占总样本的1/3，而在[4.8,5.6]区间的样本数量最少，不足200条。

（3）密度分布分析

（4）PH值分布分析

* 1. 特征对比分析

1. 不同质量下酒精含量对比分析

图表, 直方图

AI 生成的内容可能不正确。

图3-3 不同质量的酒精含量分布图

图表, 直方图

AI 生成的内容可能不正确。

图3-4 不同质量的酒精含量分布图

图表, 直方图

AI 生成的内容可能不正确。

图3-5 不同质量的酒精含量分布图

由图3-3、3-4、3-5分析可知，质量为3的样本数据数量较少，分布较为

均匀；质量为4的样本数据酒精含量大多分布在[10.12,11.84]区间内；质量为5的样本数据酒精含量大多分布在[9.4,10.4]区间内；质量为6的样本数据酒精含量大多分布在[9.6,10.7]区间内；质量为7的样本数据酒精含量大多分布在[11.96,13.08]区间内；质量为8的样本数据酒精含量大多分布在[11.8,12.9]区间内；质量为9的样本数据数量较少，分布比较均匀。

（2）不同质量下PH值对比分析

（3）不同质量下密度对比分析

* 1. 相关性分析

图表, 树状图

AI 生成的内容可能不正确。

图3-6 相关系数热力图

由图3-6分析可知，固定酸度、挥发性酸度、柠檬酸、残余糖分等9个指

标与葡萄酒质量的相关系数绝对值均小于0.5，说明各指标与葡萄酒质量的相关程度不高。固定酸度、挥发性酸度、残余糖分、氯化物和密度这5个指标与葡萄酒质量呈负相关，体现在这些指标数值增加时，葡萄酒质量倾向于减小，但减小的趋势不明显；而柠檬酸、游离二氧化硫、PH值和酒精含量这4个指标与葡萄酒质量呈正相关，体现在这些指标数值增加时，葡萄酒质量更倾向于增加，但增加的趋势不明显。

* 1. 属性构造（降维处理/特征选择）

1. 模型构建与评估
   1. 机器学习模型

（1）模型介绍

（2）评价指标

* 1. 模型构建
  2. 模型评估

1. 总结

参考文献