# **鸢尾花数据集数据分析与统计报告**

### **报告日期**

2025年2月27日

### **撰写人**

冯志彬

### **一、报告背景介绍**

鸢尾花数据集是机器学习和数据分析领域中经典的数据集之一，它包含了不同种类鸢尾花的多个特征测量值。通过对该数据集的分析，可以深入了解鸢尾花不同种类之间的特征差异，为植物分类、生态研究等提供有价值的信息。

### **二、目的陈述**

本次分析的主要目的是对鸢尾花数据集进行全面的探索性分析，了解数据的基本特征、分布情况以及不同种类鸢尾花之间的特征关系，为后续的分类模型构建和预测提供基础。

### **三、数据来源说明**

### 本报告使用的鸢尾花数据集为公开数据集，即iris.csv,包含了 150 条鸢尾花的观测记录。

### **四、数据概述**

### **数据结构**

鸢尾花数据集包含 5 个字段，具体如下：

* sepal\_length：萼片长度，数据类型为 float64。
* sepal\_width：萼片宽度，数据类型为 float64。
* petal\_length：花瓣长度，数据类型为 float64。
* petal\_width：花瓣宽度，数据类型为 float64。
* species：鸢尾花的种类，数据类型为 object。

### **数据质量**

通过对各字段空值数量的统计，发现 sepal\_length、sepal\_width、petal\_length、petal\_width 和 species 字段的空值数量均为 0，说明数据集不存在缺失值问题，数据质量较高。

### **数据预处理**

由于数据集中不存在缺失值和异常值，本次分析未进行额外的数据预处理操作。

### **五、数据分析方法**

### **描述性统计**

计算了各数值型字段（sepal\_length、sepal\_width、petal\_length、petal\_width）的基本统计量，包括总数（count）、均值（mean）、标准差（std）、最小值（min）、第一四分位数（25%）、中位数（50%）、第三四分位数（75%）和最大值（max），以了解数据的集中趋势和离散程度。

### **数据筛选**

* ****选择特定列****：选取 sepal\_length 和 sepal\_width 两列数据，查看其前五行，以聚焦特定特征的初步数据情况。
* ****条件筛选****：筛选出 sepal\_length 大于 5 的数据，分析满足该条件的数据特征。

### **简单统计操作**

* 计算了 sepal\_length 列的均值、sepal\_width 列的总和以及 sepal\_length 列的非空值数量，从不同角度对数据进行量化分析。

### **类别分析**

* ****唯一值查看****：查看 species 列中的唯一值，确定鸢尾花的种类。
* ****种类数量统计****：统计鸢尾花种类的个数。
* ****类别数量统计****：统计每个类别鸢尾花的数量，了解不同种类鸢尾花的分布情况。

## **六、数据分析结果**

### **数据浏览**

#### **前 5 行和前 10 行数据**

从数据的前 5 行和前 10 行可以看出，数据起始部分的鸢尾花种类均为 setosa，且各特征值相对较为集中，初步推测 setosa 种类的鸢尾花在这些特征上可能具有一定的稳定性。

#### **后 5 行数据**

数据的后 5 行鸢尾花种类均为 virginica，与起始部分的 setosa 种类相比，sepal\_length、petal\_length 和 petal\_width 等特征值明显较大，这可能暗示不同种类的鸢尾花在这些特征上存在显著差异。

### **基本信息**

数据集共有 150 条记录，涵盖了 5 个字段。各数值型字段均为 150 个非空值，species 字段也为 150 个非空值，数据集整体较为完整。

### **描述性统计**

| **特征** | **总数（count）** | **均值（mean）** | **标准差（std）** | **最小值（min）** | **第一四分位数（25%）** | **中位数（50%）** | **第三四分位数（75%）** | **最大值（max）** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| sepal\_length | 150.000000 | 5.843333 | 0.828066 | 4.300000 | 5.100000 | 5.800000 | 6.400000 | 7.900000 |
| sepal\_width | 150.000000 | 3.054000 | 0.433594 | 2.000000 | 2.800000 | 3.000000 | 3.300000 | 4.400000 |
| petal\_length | 150.000000 | 3.758667 | 1.764420 | 1.000000 | 1.600000 | 4.350000 | 5.100000 | 6.900000 |
| petal\_width | 150.000000 | 1.198667 | 0.763161 | 0.100000 | 0.300000 | 1.300000 | 1.800000 | 2.500000 |

* ****均值方面****：sepal\_length 的均值约为 5.84，sepal\_width 约为 3.05，petal\_length 约为 3.76，petal\_width 约为 1.20。这表明在整体数据中，萼片长度相对较长，而花瓣宽度相对较窄。
* ****标准差方面****：petal\_length 的标准差较大，为 1.76，说明该特征在数据集中的离散程度较高，不同鸢尾花之间的花瓣长度差异较大；而 sepal\_width 的标准差相对较小，为 0.43，表明萼片宽度相对较为稳定。

### **数据选择**

#### **特定列选择**

选择 sepal\_length 和 sepal\_width 两列数据的前五行，进一步观察这两个特征的初始数据情况，有助于后续对这两个特征的深入分析。

#### **条件选择**

筛选出 sepal\_length 大于 5 的数据，发现这些数据中包含了多种鸢尾花种类，这可能意味着 sepal\_length 大于 5 并不是某一种鸢尾花的专属特征。

### **简单数据统计操作**

* ****sepal\_length 均值****：sepal\_length 的均值为 5.843333333333334，与描述性统计中的均值一致，再次验证了计算结果的准确性。
* ****sepal\_width 总和****：sepal\_width 的总和为 458.1，可用于与其他相关指标进行对比分析。
* ****sepal\_length 非空值数量****：sepal\_length 列的非空值数量为 150，与数据基本信息中的记录数一致，说明该列数据完整。

### **类别分析**

#### **唯一值和种类数量**

species 列中的唯一值为 ['setosa', 'versicolor', 'virginica']，鸢尾花种类的个数为 3，这明确了数据集中包含三种不同种类的鸢尾花。

#### **类别数量统计**

| **种类** | **数量** |
| --- | --- |
| setosa | 50 |
| versicolor | 50 |
| virginica | 50 |

可以看出，三种鸢尾花的数量相同，均为 50 个，这表明数据集在种类分布上较为均衡，有利于后续对不同种类鸢尾花进行公平的比较和分析。

## **七、结论**

### **主要结论**

* 鸢尾花数据集质量较高，不存在缺失值，包含 150 条记录和 5 个字段，涵盖三种鸢尾花种类，且种类分布均衡。
* 不同种类的鸢尾花在萼片和花瓣的长度、宽度等特征上可能存在显著差异，尤其是 virginica 种类的鸢尾花在部分特征上数值较大。
* petal\_length 特征的离散程度较高，不同鸢尾花之间的花瓣长度差异较大；而 sepal\_width 相对较为稳定。