# 算法附加题：鸢尾花数据集分类

## 背景介绍

鸢尾花（Iris）数据集是机器学习领域最经典的数据集之一，由英国统计学家 Fisher 于 1936 年收集整理。 数据集包含 150 条样本，分为三种鸢尾花类别：

- Setosa（山鸢尾）
- Versicolor（变色鸢尾）
- Virginica（维吉尼亚鸢尾）

每个样本有 4 个特征：

1. 萼片长度（sepal length）
2. 萼片宽度（sepal width）
3. 花瓣长度（petal length）
4. 花瓣宽度（petal width）

---

## 任务要求

请使用 **Python** 及相关数据科学库（如 Pandas, Matplotlib, Numpy, Scikit-learn 等）完成以下任务。

### 第一部分：基础任务（共30分）

**1. 数据加载与探索**（5分）

- 从 `iris.csv` 文件加载数据。
- 查看数据的基本信息，包括：
    - 数据集的整体规模（样本数、特征数）。
    - 各特征的数据类型和是否存在缺失值。
    - 数据的基本统计摘要（如均值、方差、最值等）。

**2. 数据可视化与分析**（5分）

- **核心图表**：
    - 绘制**直方图**，观察不同类别下各特征的数据分布。
    - 绘制**散点图矩阵（配对图）**，探索特征之间的相互关系。
    - 绘制**箱线图**，对比不同类别在各特征上的分布差异。
- **分析**：根据你绘制的图表，简要分析哪些特征对区分类别最有效。

**3. 数据预处理**（5分）

- 检查并处理数据中可能存在的异常或缺失值。
- 将数据集划分为**训练集**和**测试集**（推荐比例 8:2 或 7:3），并确保划分过程可复现。

**4. 模型训练与评估**（15分）

- **模型选择**：至少使用 **两种** 不同的机器学习算法进行分类。
    - （推荐算法：逻辑回归、决策树、随机森林、支持向量机等）
- **性能评估**：
    - 计算并报告模型在测试集上的**准确率**。
    - 输出一份完整的**分类报告**，包含精确率、召回率等关键指标。
- **结果对比**：制作一个清晰的表格，用于比较不同模型的性能表现。

### 第二部分：进阶任务（共10分）

**5. 深入分析与优化**

- **相关性分析**：绘制**热力图**来展示特征之间的相关性。
- **特征处理**：对数据进行**标准化**或**归一化**处理，重新训练模型，并比较处理前后对模型性能的影响。
- **误差分析**：为表现最佳的模型生成**混淆矩阵**，分析模型主要在哪些类别上出现了分类错误。
- **稳定性评估**：使用**交叉验证**来更全面地评估模型的稳定性与泛化能力。

---

## 提交内容（共10分）

- Python 源代码(`.py` 或 `.ipynb`)。
- 结果报告，用word或markdown或者latex写都ok（用latex的话最终需导出pdf））内容包括
    1. 数据可视化、数据预处理
    2. 模型简要介绍
    3. 清晰的实验思路和步骤说明
    4. 对实验结果的分析和解释
    5. 遇到的问题和解决方案
    6. 总结和心得体会

上述python文件和报告打包成一个压缩包（zip或者rar都行）

---

## 学习资源参考

**1. Python 与数据科学基础**

- 菜鸟教程 Python 基础
- 廖雪峰 Python 教程

**2. 数据分析与可视化**

- matplotlib、numpy、pandas教程

**3. 机器学习入门**

- Scikit-learn 官方文档
- 吴恩达机器学习

**最终解释权归大云屋创想联盟组所有**