实验二：传统机器学习

实验目标

1. **掌握传统机器学习的基本流程和方法**
   - 熟悉数据预处理、特征工程、模型训练与评估展示等完整流程。
   - 理解传统机器学习算法的核心原理（如分类、回归、聚类等）。
2. **通过监督学习和无监督学习实验，理解不同算法的应用场景和效果评估**
   - **监督学习**：
     选择分类算法（如KNN、决策树、SVM等），对比分类性能。
     掌握模型评估指标（如准确率、召回率、F1分数、混淆矩阵等）。
   - **无监督学习**：
     选择聚类算法（如K-Means、层次聚类等），分析聚类效果。
     掌握聚类评估指标（如轮廓系数、肘部法则等）。
3. **通过可视化展示，深入理解数据和模型**
   - **数据可视化**：
     使用散点图、箱线图、平行坐标图等展示数据集特征分布和类别差异。
     通过可视化探索特征之间的相关性。
   - **模型可视化**：
     对监督学习模型，绘制决策边界（如决策树、SVM）或特征重要性（如随机森林）。
     对无监督学习模型，展示聚类结果（如K-Means的聚类中心、聚类分布）。
   - **效果可视化**：
     使用混淆矩阵热力图、ROC曲线等展示分类模型的性能。
     使用肘部法则图、轮廓系数图等展示聚类模型的效果。

实验环境

- **硬件**：Intel Core i5以上处理器，8GB以上内存，256GB SSD。
- **软件**：Python 3.8+，scikit-learn，Pandas，Matplotlib，Yellowbrick等。

实验题目

1. 鸢尾花分类（监督学习）
2. 客户细分（无监督学习）

实验内容与步骤

1. 监督学习（鸢尾花分类）

- **数据集**：
  - 内容：Iris数据集是机器学习领域经典的入门数据集，包含150个样本，分为三类鸢尾花（Setosa、Versicolor、Virginica），每个样本有4个特征（花萼长度、花萼宽度、花瓣长度、花瓣宽度）。
  - 获取：Python中Scikit-learn库，内置了Iris数据集，可直接加载使用。 或UCI Iris数据集页面，https://archive.ics.uci.edu/ml/datasets/iris下载。

- **数据预处理**：
  - 标准化：使用 `StandardScaler` 对数据进行标准化处理。
  - 缺失值处理：检查数据中的缺失值，并采用适当的方法进行填充或删除。
- **模型选择**：
  - 逻辑回归：使用 `LogisticRegression` 类实现。
  - 决策树：使用 `DecisionTreeClassifier` 类实现。
  - SVM：使用 `SVC` 类实现。
- **评估指标**：
  - 准确率：使用 `accuracy_score` 函数计算。
  - F1分数：使用 `f1_score` 函数计算。
  - 其它
- **可视化**：
  - 数据探索可视化。
  - 特征分布可视化。
  - 模型效果可视化。
  - 其它……

实验方法示例（逻辑回归）：

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

model = LogisticRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print(accuracy_score(y_test, y_pred))
```

2. 无监督学习（客户细分）

- **数据集**：

1.**Kaggle平台**：可直接访问Kaggle官网,https://www.kaggle.com/,搜索“Mall Customers”获取数据集。
2.**GitHub**：可GitHub仓库下载该数据集，
https://github.com/Karansingh1221/Mall_Customer_dataset。

- **数据处理**：
  - 数据加载与探索：读取数据集，检查数据类型、缺失值和异常值
  - 数据清洗：处理缺失值（如填充或删除），处理异常值（如通过箱线图识别并处理）等。

    - 特征编码与选择：将变量转换为数值型数据，选择与任务相关特征。
    - 数据标准化：对数值型特征进行标准化，以消除量纲影响。
- **降维**：
    - PCA：使用 `PCA` 类实现，降低数据维度以便可视化。。
    - t-SNE：使用 `TSNE` 类实现，进一步降低维度并保持数据结构。
    - 其它。
- **聚类分析**：
    - K-Means：使用 `KMeans` 类实现，设置合适的聚类数量。
    - 层次聚类：使用 `AgglomerativeClustering` 类实现。
    - 其它：如DBSCAN、谱聚类（Spectral Clustering）、高斯混合模型（GMM）等。
- **评估**：
    - 可视化：例如使用散点图展示不同聚类簇的分布情况。
    - 内部指标：使用轮廓系数、Calinski-Harabasz指数、Davies-Bouldin指数等评估效果。
    - 模型参数：分析不同模型参数对聚类结果的影响。
    - 其它……
- **可视化**：
    - 数据探索可视化。
    - 特征分布可视化。
    - 模型效果可视化。
    - 其它……

实验方法示例（K-Means聚类）：

```
from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=3)
kmeans.fit(X)
print(kmeans.labels_)
```

实验要求

- 完成题目要求，代码实现包括数据预处理、模型训练和评估展示等完整流程。
- 设计完成任务实现过程中各阶段可视化展示，并进行总结分析。
- 分析不同模型参数对结果的影响，如逻辑回归中的正则化强度等。
- 根据实验目的及要求，完成实验报告及代码，并按时提交。

实验评分标准

- **代码实现（40%）**：
    - 代码的正确性：能够正确运行并得出预期结果。
    - 代码的完整性：包含数据预处理、模型训练、评估和可视化的完整流程。
    - 注释的清晰度：代码中有足够的注释，便于理解。

- **结果分析（40%）**：
  - 对模型性能的评估：准确率、F1分数等指标的计算和解读。
  - 参数影响的分析：探讨不同参数对模型性能的影响。
  - 可视化结果的展示：混淆矩阵、聚类结果等可视化图表的清晰度和准确性。
  - 思考和结论：对整个实验过程，包括问题、方法、结论等进行思考总结。
- **报告撰写（20%）**：
  - 实验报告的结构：包含实验背景、目标、方法、结果和结论等部分。
  - 实验报告的逻辑性：内容连贯，逻辑清晰。
  - 实验报告的可读性：语言简洁明了，图表清晰美观。

实验输出示例

- **监督学习**：
  - 逻辑回归模型的准确率：0.95
  - 逻辑回归模型的F1分数：0.94
  - 混淆矩阵图示：展示模型在不同类别上的预测情况。
  - 等
- **无监督学习**：
  - K-Means聚类的标签分布：展示每个样本所属的聚类标签。
  - PCA降维后的可视化图示：展示数据在二维空间上的分布情况。
  - 等

实验时间安排（4学时）

- **第1学时**：了解实验背景、目标和环境配置，对实验题目、数据集及任务等进行详细分析，技术选型及方案规划等准备工作。
- **第2学时**：进行监督学习实验（鸢尾花分类），包括数据预处理、模型训练和评估展示。完成代码编写和结果分析。
- **第3学时**：进行无监督学习实验（客户细分），包括数据处理、降维处理、聚类分析、评估展示等。完成代码编写和结果分析。
- **第4学时**：实验总结与报告撰写，包括实验内容步骤、结果分析、可视化展示和实验报告的撰写。提交实验报告和代码。

注意事项

- 确保实验环境配置正确，包括Python版本和相关库的安装。
- 代码实现过程中注意代码的可读性和注释的清晰度，便于他人理解和复现实验结果。
- 实验结果分析要全面，包括对不同模型参数影响的讨论，以及模型性能的优缺点分析。
- 实验报告要结构清晰，逻辑严谨，表达简明，包括实验背景、目标、方法、结果和结论等部分。