1.题目：机器学习实践。

2.数据集：从 UCI Machine Learning Repository、Kaggle、Git Hub、Gitee 等数据源，自行下载实际数据集。此外，如果需要自己生成人工数据集。

3. 任务描述：完成如下任务。

3.1 分类任务：对决策树、KNN、逻辑斯蒂回归、SVM 等分类器的分类效果进行比较。

（1）准备实际数据集 1

（2）准备人工数据集 2（可选）

（3）装载数据

（4）必要的预处理与探索

（5）训练集与测试集的划分

（6）建模与调参

（7）模型评价与比较（分类的客观评价指标、分类效果可视化（必要时可以降维））

（8）预测新样本（该样本属于哪个类别）

3.2 聚类任务：对 K-Means 与 GMM 等算法的聚类效果进行比较。

（1）准备实际数据集 3

（2）准备人工数据集 4（可选）

（3）装载数据

（4）必要的预处理与探索

（5）建模与调参

（6）模型评价与比较（聚类的客观评价指标、聚类效果可视化（必要时可以降维））

（7）处理新样本（该样本属于那个类簇）

4. 成果提交：请于截止时间前，把数据与源代码（ipynb）一起压缩为 zip 文件，提交到 obe.ruc.edu.cn 网站。

备注：notebook 的每个代码 Cell，要求加上必要的注释。

5. 参考文献：无

6.评价方法：

（1）本练习为必选练习，根据上述任务完成度评分。
（2）部分完成、晚交适当扣分。
（3）自行完成，不要抄袭，抄袭 0 分。

7.提示

选择有实际意义的题目和数据集进行实验，比如“基于 X 射线胸片的肺炎辅助诊断”、“基于卫星图片的农产品产量预测”等，就很有实际意义。

如下是可选的题目（也可以自行选题）。

（1）客户群划分 customer segmentation
（2）股票聚类 stock clustering

（3）白酒的分类 Wine Classification
（4）信用卡欺诈检测 Credit Card Fraud Detection
（5）确定狗的种类 Identifying Dog Breeds
（6）放贷评分 Loan Eligibility Prediction
（7）客户流失预测 customer churn prediction
（8）乳腺癌检测 breast cancer detection
（9）虚假新闻检测 fake news detection
（10）文本分类 text classification
（11）饭馆、产品评论的情感分析 restaurants/product review sentiment analysis
（12）语音情感分析 Speech Emotion Analyzer
（13）音乐流派分类 Music Genre classification
（14）垃圾邮件分类 Spam Filter

（15）房价预测 House Sale Prices
（16）股票价格预测 Stock Market
（17）故障预测 manufacturing failures
（18）自行车租赁预测 Bike Rentals

（19）新闻、电影、书籍、音乐推荐 news movie book music recommendation