实验任务 – 2025

一、熟悉 scikit-learn 数据挖掘包（https://scikit-learn.org/stable/），特别是如下分类与聚类算法的使用：

1. 分类算法：

(1) 集成学习中的 Adaboost

(2) 朴素贝叶斯（Naive Bayes）

(3) 决策树 C4.5（Decision Trees）

(4) 集成学习中的 Gradient Tree Boosting

(5) 支持向量机（Support Vector Machine）

(6) 最近邻分类器（Nearest Neighbors）

(7) 集成学习（Ensemble Methods）中的随机森林（Random Forest）

(8) 分类与回归树 CART

2. 聚类算法（若算法需要输入聚类数目，则指定数据集中的类数目；对于层次聚类，则指定该 k，即该数据集的类数目）：

(1). Affinity Propagation

(2). BIRCH

(3). DBSCAN

(4). Hierarchical clustering

(5). K-means

(6). Mean Shift

(7). OPTICS

(8). Spectral clustering

二、熟悉深度学习平台 Pytorch（https://pytorch.org/）或 keras（https://keras.io/）。

注：Keras 是深度学习平台 Tensorflow 的进一步 API 封装，简单易用。

三、实际操作：

1) 针对 UCI 机器学习数据库（http://archive.ics.uci.edu/ml/index.php）的分类任务（Classification Task），截止到 2025-5-12 日，共有 517 个数据集。聚类任务（Clustering Task），也采用分类任务的数据集。
2) 使用的数据集：数据集的样本数要大于等于 5000。
3) 分类或聚类任务的确定：奇数学号为分类任务，偶数学号为聚类任务。
4) 每个人使用的算法以及数据集按如下方法确定：
   (1). 使用算法与数据集的确定：使用中文 Word Embedding 选取与自己姓名最接近的（使用 2-范数）三个算法以及数据集。
   (2). 除上述算法外，每个人还**必做深度学习**（用 keras 或 Pytorch 平台）。

四、注意事项及相关说明：

1. 语言为 Python，以前没有使用过的正好通过这次实验进行熟悉（python 为机器学习的最主流语言，具有最大的开源社区）。
2. 评价方法：对于分类任务—采用 10-折交叉验证（10-fold cross validation）。对于聚类任务— ARI 或 NMI。
3. 提交内容：实验报告、源代码（包括确定算法名以及数据集名的代码）的电子版。其中实验报告部分，要将 4 个算法的结果列在一张表中进行对比。以学号+姓名+数据集名+算法命名文件，以班级为单位由班长统一提交（电子邮件 jbwang@scut.edu.cn)或网盘形式提交，若文件超过 50M，学校邮箱可能拒收）。
4. 提交时间：第 15 周周日晚上 12：00 以前。实验占总评成绩的 30%，逾期未交，此部分成绩为 0 分。
5. 若发现抄袭，抄袭者与被抄者均计 0 分。