# Ensemble Learning Lab Report / 集成学习实验报告

## 1. Introduction / 实验介绍

This lab explores ensemble learning using the Pima Indians Diabetes dataset. Tasks include data preprocessing, hard and soft voting, pipeline evaluation, and hyperparameter tuning using RandomizedSearchCV.  
本实验使用 Pima 印第安人糖尿病数据集进行集成学习探索，包括数据预处理、硬投票与软投票、Pipeline 评估与超参数调优等任务。

## 2. Data Loading and Exploration / 数据加载与探索

Dataset: pima-indians-diabetes.csv

✔ Column Names and Data Types / 列名与数据类型: [截图位置 2]

✔ Missing Value Check / 缺失值检查: [截图位置 3]

✔ Descriptive Statistics / 描述性统计信息: [截图位置 4]

✔ Class Balance Check / 类别分布情况: [截图位置 5]

We observed some features like 'Insulin', 'SkinThickness' contain many zeros, which may imply missing values. The dataset is slightly imbalanced (500:267).  
我们观察到如 Insulin、SkinThickness 等特征存在许多 0 值，可能表示缺失。类别分布略不平衡（500:267）。

## 3. Data Preprocessing / 数据预处理

Used StandardScaler to normalize all numerical features.  
使用标准化对所有数值型特征进行缩放处理。

✔ Data Split (70/30) / 数据划分: [截图位置 1]

## 4. Hard Voting Ensemble / 硬投票集成

✔ Individual Classifier Predictions / 各分类器预测: [截图位置 6]

✔ Hard Voting Results: [截图位置 6]

Conclusion: Most classifiers predicted [0, 0, 0], except DT which misclassified 1st instance.  
结论：大部分分类器预测一致为 0，只有决策树错误预测第一个样本为 1。

## 5. Soft Voting Ensemble / 软投票集成

✔ Soft Voting Predictions / 输出结果: [截图位置 7]

Soft voting slightly improved consistency by using probability averaging.  
软投票通过概率平均，提升了一致性。

## 6. Pipeline Evaluation / Pipeline 评估

✔ ExtraTrees Pipeline Accuracy: [截图位置 8]

✔ DecisionTree Pipeline Accuracy: [截图位置 8]

Pipeline with ExtraTrees performed better overall.  
整体上 ExtraTrees 的 Pipeline 准确率更高。

## 7. Metrics & Evaluation / 模型评估

✔ Confusion Matrix, Precision, Recall, Accuracy [截图位置 9]

ET Precision: ~60%, Recall: ~57%, DT Recall higher (~68%).  
ExtraTrees 精确率约 60%，召回率 57%；决策树召回率略高（~68%）。

## 8. Hyperparameter Tuning / 超参数调优

✔ RandomizedSearch Best Params: [截图位置 10]

✔ Fine-tuned Accuracy & Metrics: [截图位置 11]

Using the best estimator improved accuracy from 0.76 to 0.78.  
调参后模型精度由 0.76 提升至 0.78，说明调参有效。

## 9. Conclusion / 实验总结

Ensemble learning with hard/soft voting and tuned models can significantly improve prediction performance. RandomizedSearchCV is effective in finding better parameters.   
本实验展示了集成学习在提升分类性能方面的价值，特别是配合网格搜索后精度明显提升。