**introduction**

在本次实验中，我首先分离了Programme=3的样本，将本次实验的任务从四分类问题转变为三分类问题。接着，按照任务要求使用SVM，随机森林，朴素贝叶斯和集成分类器对学生Programme进行了分类。在每种分类方法中，我都对输入特征的组合与超参数的选择进行了网格搜索，同时使用交叉验证对性能进行评估，在确保不会导致过拟合的前提下选择了最优的情况。此外，由于类别不均衡，本实验也尝试使用欠采样和Synthetic Minority Over-sampling Technique (SMOTE)提高对Programme=3的样本的分类性能。最后，提出了关于超参数与特征组合的选择与分类性能的猜想。为确保实验的可验证性，相关结果与代码保存在以下仓库中：<https://github.com/MushihimePepsi/XJTLU_ICS_Y2S2_Course-notes_23-24/tree/main/INT104-%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD>

**关于特征选择，防止过拟合和类别不平衡的申明**

**-特征选择**

尽管在CW1中使用了PCA方法，通过选取能解释数据集中尽可能大的方差方向（原特征的线性组合）进行降维，从而期待这些新特征的选取可以作为更优秀的分类器输入。但是在实际分类中，这一行为并不会带来更好的分类性能，因此在本报告中，分类器的输入都将会是原特征的组合。类似的，尽管可以利用随机森林对特征的重要性进行排列或使用Recursive Feature Elimination (RFE)选择最重要的特征，但在实际训练中不同的分类器的最佳特征输入往往不同，因此在本实验中每个分类器都会使用最适合的特征组训练。对不同分类器最佳输入特征的网格搜索研究已放入仓库中供您参考。

**-防止过拟合**

在本次实验中，由于数据集规模较小，需要特别注意防止模型的过拟合。在训练模型的过程中使用了k-fold cross validation的方法，设定k=5，将训练集分为5份，每次使用4份作为模型训练，1份用于验证，避免模型仅依赖于单一的训练集，增强泛化能力。此外在分类器的超参数设定上，也进行了正则化或向损失函数中添加惩罚项来限制模型的复杂度。最后，测试集与训练集的划分比例为3:7，在确保有足够训练样本的情况下尽可能测试更多样化的样例。

**-类别不平衡**

本次实验的数据集中的类别分布并不平衡，除去占比极少的Programme=3的样本，Programme=2样本的数量约为Programme=1或Programme=4的样本的一半，导致了类别不平衡问题，部分分类器虽然在整体表现较好（Accuracy≈0.7），但却无法有效分类Programme=2的样本。为解决这个问题，本实验中采取了欠采样和SMOTE操作，降低模型对多数类别的过度依赖，提高泛化能力。

在多分类问题中为了防止过拟合进行的交叉验证和对分类器的正则化，为解决类别不平衡问题从而对训练集进行欠采样或SMOTE操作。

**分离Programme=3的样本**

在CW1中，我曾论证由于Programme=3的学生与Grade=3的特征的高度相关性，可以利用这一关系预先分离Programme=3的学生，从而降低特征维度和对后续其他Programme分类的复杂度。

这一分类的表现相当好，本文后续的操作将着重对Programme={1,2,3}的其余样本进行分离。由于Programme=3的学生共有35名（5.65%），在这步处理后，总样本数从619降为584，同时特征Grade也被去除。

**任务一：决策树和随机森林**

经过对比，本实验使用基尼不纯度而不是信息增益作为衡量特征的重要性程度，并用CART算法构造决策树。通过对'max\_depth'，'min\_samples\_split'和'min\_samples\_split'的限制性的网格搜索对决策树进行正则化和修剪并选择最佳参数。确保决策树在拥有一定泛化能力的同时拥有良好的分类性能。

几个典型输入特征组合的性能：

可以发现，不同输入特征对分类的影响较大，且不符合先前重要性的预测。树的深度越深，越倾向过拟合，对`Programme'=2的分类效果越好。综合考虑性能选择'MCQ', 'Q1', 'Q3', 'Q4'作为输入特征。

最好的混淆矩阵：

树：

随机森林

随机森林的特点，属于集成学习，随机采样样本和特征，不易过拟合。随机森林结合了bagging和随机子空间的做法，同时对训练集的样本和样本特征进行随机采样。这样的做法增强了基学习器之间的差异和独立性，使得集成模型的方差更小，泛化能力更强，具有天然抗过拟合的能力。可以见到，尽管Accuracy类似，但随机森林增强了预测的Precision。

最好的参数：

混淆矩阵

任务二：支持向量机

几个超参数的意义和选择，最后暴力搜索的最好参数。

支持向量机通过核函数将输入数据映射到高维特征空间，使得原本非线性可分的数据在特征空间中变得线性可分，同时避免了维度灾难。使用OVR（One-Versus-Rest）的多分类策略，通过网格搜索不同的核函数，调整决策边界的软硬程度‘C’，核函数的影响范围‘gamma’找到最合适的参数。由于类别不平衡，将正则化的训练集欠采样后作为输入，从而提升对Programme2的分类的性能。

任务三：朴素贝叶斯

朴素贝叶斯的原理和结果

Naïve Bayes的工作原理基于贝叶斯定理和特征条件独立假设，通过统计先验概率和条件概率最终用贝叶斯定理求出后验概率，最终选择后验概率最大的类别作为分类结果。输入特征经过网格搜索后确定为'Gender', 'Total', 'Q1', 'Q2', 'Q3', 'Q4', 'Q5'。

任务四：集成分类

adaboost

投票

分析软投票硬投票结果的差异

在集成学习中，除了先前使用的随机森林和未论述的adaBoost方法，我使用了投票的方法将先前已经训练完毕的模型进行组合，期待用多个弱分类器合成一个强分类器。然而投票产生的模型性能平庸。此外，尽管软投票通常被认为能够提供更准确的预测结果，在实践中却弱于硬投票。这可能是基分类器的同质性和不正确的权重选择导致的。

结论，局限性和猜想

在本次实验中，我对每个类别的分类器进行了交叉验证下的网格搜索，同时对不同的输入特征组合进行研究，提出了特征组合与分类器性能间的猜想，找到适合每个分类器的不同特征组合。但本研究仍有局限性，在分类的最初就将Programme3分离，以期简化后续分类流程。但是Programme3本身依然包含了有别于其他类别的信息，在监督学习中可能有助于分离其他类别。此外，本实验也没有使用经过降维后得到的新特征集，尽管PCA降维后的数据对结果影响不大，但仍可采用Locally Linear Embedding（LLE）等非线性降维方法，提取数据在高维上的关系特征。