基于随机森林算法的专业核心课程成绩与毕业结论研究

胡少启[1] 贾梦杰[1] 彭月[2]

1 滁州学院 教务处，安徽滁州，239000；

2 明光市管店中心小学，安徽滁州，239000；

摘要：为实现对学生学业修读情况的早期精准监测与有效干预，提高地方应用型本科高校人才培养质量，已成为当下高校亟待深入探究的重要课题。本研究以软件工程专业 2019-2023 届 398 名毕业生样本为数据源，选取 12 门专业核心课程成绩作为关键分类特征。基于机器学习领域中的随机森林算法，构建了学生专业核心课程成绩与毕业情况的预测模型。实验结果表明，该模型总体预测准确率高达 89.59%，显著优于其他几类常见的机器学习算法模型，表明该模型在学生学业与毕业结论预测具有良好的应用效果。

关键词：决策树；随机森林；专业核心课程；毕业结论

DOI：10.69979/3029-2735.25.07.049

随着教育部《关于深化本科教育教学改革 全面提高人才培养质量的意见》的发布，地方应用型本科高校在优化人才培养方案、提升毕业生培养目标达成度等方面面临着严峻挑战[1]。在日常教学管理与教学评价实践中，专业核心课程成绩不仅直接作用于关联课程的考核通过情况，还间接左右着学生能否按期毕业。因此，根据学生专业核心课程考试成绩，分析并预测学生是否会延期毕业，构建前置学业干预与学情预警机制，对有效降低学生结业率至关重要。

金城等人基于随机森林算法，对学生学情信息与程序设计课程成绩展开相关性分析，研究成果能够有效识别影响不同学生学习成绩的主要相关因素[2]。董建文等人运用决策树模型，剖析影响学生微积分课程成绩的师生特征，得出生源类别、数学成绩、生源地身份等与课程考核成绩紧密相关的结论[3]。然而，上述关于学生成绩的分析研究，大多聚焦于单一课程或相近课程成绩之间的相关性，针对专业核心课程与学生毕业情况的关联研究则相对匮乏。刘丽娟等人以大一新生的修课成绩作为直接影响特征，将与修课成绩相关的 7 个其他特征作为间接影响特征，构建学生学业预警预测模型，并就多维特征对学业预警预测的影响及其重要性进行了量化和对比研究[4]。马昕等人选取学生大学一年级的总学分和 13 门课程成绩作为特征，借助随机森林算法模型，对学生异动情况展开预测分析[5]。这些研究虽建立了学生成绩与学籍异动、学业预警之间的关联，对学生在校期间的学业情况进行了分析预测，但难以通过专业核心课程成绩及通过情况预测学生是否能够按期毕业。

鉴于此，本研究采用基于随机森林算法，深入探究高校教学过程中积累的海量成绩数据，挖掘数据间隐藏的关联信息，运用数据挖掘技术揭示专业核心课程成绩与毕业情况之间的逻辑关系，旨在尽早察觉学生学业问题，并采取相应的学业帮扶措施，进而推动高校教学质量和教学管理水平的提升。

1 随机森林算法

作为一种先进的集成学习算法，随机森林由著名学者 Breiman 开发，适用于处理高维度、大规模数据集，以及在单一决策树模型表现欠佳的场景下使用。该算法本质上是决策树方法的扩展与优化，通过构建多个基础分类器形成组合模型，且各个基分类器在训练过程中保持完整结构，无需进行复杂度控制[6-8]。在模型构建阶段，该算法引入了双重随机机制：其一为样本空间的随机抽样，其二为特征子集的随机选取。这种独特的随机特性使其在模型鲁棒性和预测准确性方面优于传统监督学习方法。此外，该算法在防止模型过拟合方面表现突出，对数据噪声具有良好的容忍度，能够有效处理不完整数据集[9]。具体实施过程可分为以下环节：

训练集抽样：从样本数据集 T={D1,D2,D3,……}中，运用 Bootstrap 抽样方法获取 K 个相互独立且样本量均为 n 的训练集{Tt,t=1,2,…,K}，该抽样过程通过对原始数据集进行有放回抽样，使每个训练集都具有独特性，同时又保留了原始数据集的分布特征。

决策树构建：对每个通过 Bootstrap 抽样方法生成的训练子集，采用完全扩展策略建立决策树模型，过程中保持模型的自然生长状态，不实施任何形式的复杂度控制。在构建决策树的节点分裂过程中，随机从 M 个总特征中选取 m 个特征（m≤M），并从这 m 个特征中筛选出最优特征属性，以实现决策树的生长。通过上述建模过程，最终生成由 K 棵决策树组成的“森林”。每棵决策树均在独特的训练子集和特征子集下进行训练，这种差异性为模型带来了必要的多样性，有助于提升整体泛化性能。

分类决策：将训练得到的 K 个基分类器类比为多领域的决策专家，共同参与新样本的类别判定。具体实现过程中，通过整合所有基分类器的判别结果，采用多数表决机制来确定样本的最终类别归属。对于待分类样本 X，设随机森林中各决策树集合模型为{md1(x),md2(x),md3(x),……,mdk(x)}，则该随机森林的分类决策函数可表示为：

$$f_c(x) = \max \sum_{i=1}^{K} I(md_i(x) = Y)$$

其中，fc(x)表示随机森林模型的最终分类结果，mdi(x)表示第 i 棵决策树对样本 X 的分类结果，Y 为类别标签空间中的任意类别，I 为布尔判别函数。

当 mdi(x)=Y 时，I（mdi(x)=Y）=1，否则 I（mdi(x)=Y）=0。通过这种投票机制，随机森林综合了多棵决策树的判断，选取获得票数最多的分类选项作为最终输出，从而提升了分类的准确性与稳定性。

2 实验模型构建分析

2.1 模型数据来源

本研究数据来源于某应用型本科高校计算机学院软件工程专业 2019-2023 年的毕业生数据，包含五届毕业生的 12 门专业核心课程成绩及毕业结论情况。近五届软件工程专业共有 398 名毕业生，其中 326 人当年按期毕业，72 人结业，具体分布详见表 1。

表 1 各年份学生毕业情况统计表

年份	学生总数	毕业生数	结业生数
2019	68	62	6
2020	81	62	19
2021	80	63	17
2022	88	73	15
2023	81	66	15

表 2 学生专业课程成绩表

StuID	DSP	JOOP	PODS	MAP	OPS	……	BYJL
1	80	87	82	75	77	……	是
2	91	86	88	84	78	……	是
3	75	72	69	68	63	……	是
4	60	78	70	77	65	……	否
……	……	……	……	……	……	……	……
398	60	86	90	77	86	……	是

该专业 398 名毕业生就读期间修读同一版本的人才培养方案，学生成绩及毕业结论数据库包含学生学号（StuID）等关键信息，数据示例见表 2。12 门专业核心课程分别为：数据结构（DSP）、Java 程序设计（JOOP）、数据库原理与应用（PODS）、操作系统（OPS）、动态网页设计（WPBD）、计算机网络（CNET）、软件工程（CSE）、移动应用开发（MAP）、算法设计与分析（DAA）、JavaEE 应用开发（J2EE）、面向对象程序设计（C++）（COOP）以及软件测试（STD）。这些课程构成了该专业学生专业知识体系的核心部分，对其毕业情况有着重要影响。

2.2 模型数据预处理

为进一步优化数据挖掘的运算效率，显著提升随机森林算法的计算性能，对表 2 中的数据进行预处理操作。预处理严格依据学校学分制学籍管理办法的规定，将课程成绩由百分制转换为绩点制。具体转换规则如下：成绩小于 60 分，对应绩点为 0；成绩大于等于 60 分且小于 70 分，绩点为 1；成绩大于等于 70 分且小于 80 分，绩点为 2；成绩大于等于 80 分且小于 90 分，绩点为 3；成绩大于等于 90 分，绩点为 4。通过量化处理，将原始实验数据转化提升随机森林算法的运行效率，减少计算复杂度，使数据分析结果能够以更清晰的方式呈现，为深入探究专业核心课程成绩与毕业结论之间的内在关联奠定了坚实的数据基础。

2.3 数据分析模型构造

在构建学生专业核心课程成绩与毕业结论的随机森林模型训练过程中。首先，运用 Bootstrap 抽样法对包含 398 条数据的数据集进行处理，将其划分为训练样本和测试样本。其中，训练集数据占比设定为 80%，测试集数据占比为 20%，确保模型在训练阶段能够充分学习数据特征，同时在测试阶段能够准确评估模型性能。

其次，在随机森林构建决策树的过程中，每次可选取的特征变量个数依据总特征数量的平方根确定。实验

涉及的构造数据分析模型共有 12 个特征变量，为确定最佳特征子集，从 12 个候选特征中随机抽取 3 到 12 个特征进行多组对比实验。通过大量实证分析发现，当每个节点分裂时随机选择 3 个特征进行评估，能够实现模型性能与计算效率的最优平衡，从而构建出高效的分类模型。

进一步对随机森林模型的性能进行评估，随着决策树数量的增加，集成模型的分类性能呈规律性变化。当决策树数量大于 20 个时，随机森林的误分率逐渐趋于平稳，其泛化误差的波动范围显著减小。当决策树数量达到 25 个时，模型的预测精度进入稳定区间，继续增加决策树数量对性能提升的边际效应显著降低。基于此实验结果，为使模型达到最佳性能，将模型中的决策树数目设置为 25 个。该参数配置有效保证了随机森林模型在分析学业表现与毕业结果关联性数据时的稳定性与可靠性，为后续的数据分析和预测提供了可靠的模型基础。

2.4 数据模型验证

为验证随机森林模型训练的准确度，采用 K 折交叉验证法对模型样本数据开展交叉验证。通过将样本数据划分为 K 个互不重叠的子集，在 K 次迭代中，每次选取其中一个子集作为测试集，其余 K-1 个子集作为训练集，以此全面评估模型的泛化能力[10]。实验结果表明，该随机森林模型在训练集上的平均分类准确率达到 89.59%。表明该模型具备良好的识别度，能够较为准确地对学生专业核心课程成绩与毕业结论之间的关系进行分类判断。

此外，为进一步凸显随机森林模型的优势，实验选取其他多种训练模型进行对比。各训练模型的得分情况如下表 3 所示。通过对不同模型分类准确率的对比发现，在处理相同样本数据时，随机森林建模所获得的分类准确率最高，其分类效果在众多模型中表现最为出色[11-12]。

表 3 随机森林和对比模型得分表

序号	训练模型	模型得分
1	决策树 DecisionTree	0.78125
2	引导聚集算法 Bagging	0.86458
3	随机森林 RandomForest	0.89583
4	迭代算法 AdaBoost	0.86458
5	梯度提升决策树 GBDT	0.84375

随机森林模型具备对模型中特征变量的重要性评估功能。实验采用变量重要性度量方法，从训练好的随机森林模型中提取各特征的贡献度指标，绘制出学生毕业结论影响因素权重分布的散点图，从而清晰地展示不同变量对预测结果的相对重要性排序。其中，IncNodePurity 指数作为一种解释变量重要性的评估方法，其本质为残差的平方和，该值恒为非负。在特征重要性评估过程中，节点不纯度减少量的数值大小直接反映了相应特征的重要性水平，该指标值越大，说明该特征对响应变量（即学生毕业状态）的预测贡献度越高。需要说明的是，节点不纯度减少量与基尼系数平均减少量在特征重要性度量上具有等价性，两者均可用于评估特征对模型预测能力的贡献程度。

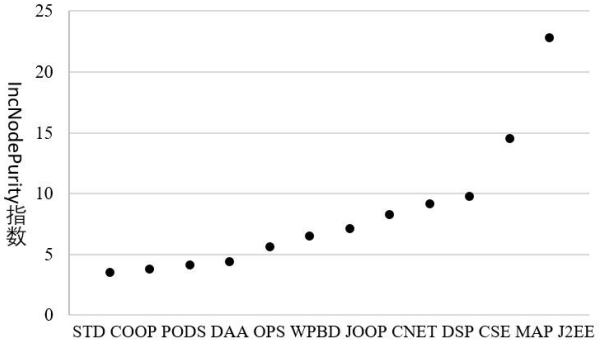

图 1 专业核心课程成绩对学生毕业影响程度散点图

专业核心课程成绩对学生毕业结论影响如图 1 所示，J2EE 课程成绩对学生是否能够正常毕业的影响程度最为显著。在其后的依次排序中，MAP、CSE、DSP、CNET、JOOP、WPBD、OPS 等课程的修读情况，对学生毕业结论也有着不同程度的影响。这些结果为深入理解学生毕业相关影响因素提供了量化依据，有助于教育工作者精准聚焦关键课程，优化教学资源配置，提升人才培养质量。

2.5 结果分析与评价

本研究运用随机森林算法，紧密结合高校学生人才培养目标，针对地方应用型本科高校学生的毕业情况开展分类建模工作。通过该模型成功建立起学生专业核心课程与毕业结论之间的映射关系，并依据各课程对毕业结论的影响程度进行了相应排序。结果表明，此模型具有较高的准确度和稳定性，研究成果为应用型本科高校的人才培养提供了极具价值的决策参考，具体内容如下：

优化专业核心课程设置：从市场需求、学生能力匹配以及职业定位等维度出发，对人才培养方案进行全面修订。针对当前专业课程普遍存在的理论知识过多、内容陈旧、开设时间集中且难以激发学生学习兴趣等问题加以改进。对于毕业率普遍偏低的理工类专业，开课学

院应严格贯彻 OBE 理念，深入剖析不及格率较高的专业核心课程在人才培养方案中的作用。从教学大纲、教学设计、教学模式、过程性考核与结果考核等方面，展开全面的教学质量分析，并针对突出问题制定切实可行的改进措施。

调整教育教学方式方法：变革专业核心课程的教学模式，加强对教师教学方式方法和技能技巧的培训，鼓励教师创新教学手段。同时，组建专业的教学质量考评团队，致力于全面提升学校的管理水平和教师的教学水平，推动教学质量稳步提高。此外，大力强化校园学风建设，通过开展“学风建设月”等主题活动，营造良好的校园学习氛围，形成学校组织、学生积极参与、教师共同培育的浓厚学习环境。

完善学业预警及帮扶制度：修订学业预警管理制度，加强学业过程性学习记录的采集与管理，深度挖掘前置学情数据，对可能无法正常毕业的学生进行精准预测，尽早发现学生的学业问题，及时扭转学生的学业困境。充分发挥校内教师、校外家长等多方教育主体的协同作用，凝聚预警与帮扶的合力。调动参与人员的主观能动性，制定学业帮扶人员绩效奖励制度，确保帮扶团队成员的稳定性，构建多方协同、全员参与的学业预警帮扶体系。

3 结论

研究构建了针对学生专业核心课程修读情况与毕业结论的预测模型。该模型以随机森林分类算法为基础，将学生在校期间修读的专业核心课程的考核成绩作为特征变量，训练并预测学生是否能够按期毕业。实验结果表明，该模型能够精准且有效地预测学生的按期毕业情况。利用该模型对低年级学生专业核心课程成绩的分析，提前预测学生毕业情况，有针对性地开展学业管理和学业帮扶工作，全方位提升学校的教育教学管理水平，为培养高质量人才奠定坚实基础。

参考文献

[1]毕瑶家,刘国柱,王华东等.改进随机森林算法在人才培养质量评价中的应用[J].计算机系统与应用.2020,29(7):212-216.

[2]金城,崔荣一,赵亚慧.基于机器学习的高考信息与大学程序设计课程成绩相关性分析研究[J].延边大学学报（自然科学版），2020,46(4):366-370.

[3]董建文,张一春,胡燕.基于决策树算法的学习结果预测模型设计与应用[J].广播电视大学学报，2022,(1):39-46.

[4]刘丽娟,林雨衡,王晓琪等.多维特征对高校学生学业预警预测的影响[J].厦门理工学院学报，2020,28(4):54-61.

[5]马昕,王雪,杨洋.基于随机森林算法的大学生异动情况的预测[J].江苏科技大学学报（自然科学版）.2012,1（26）:86-90.

[6]刘志妩.基于决策树算法的学生成绩的预测分析[J].计算机应用与软件,2012,11（29）:312-314.

[7]QuinlanJR.Induction of decisiontrees[J].MachLearn,1986,（1）:81.

[8]BreimanL.Randomforests［M］.MachLearn,2001,45,(1):5.

[9]胡永培,张琛.基于 AP 聚类与随机森林的客户流失预测研究[J].计算机技术与发展.2021,2(31):86-90.

[10]Wolpert DH,Macready WG.An efficient method to estimate baggings generalization error[J].MachLearn,1999,35(1):41.

[11]班文静,姜强,赵蔚.基于多算法融合的在线学习成绩精准预测研究[J].现代远距离教育，2022,(3):37-45.

[12]罗杨洋,韩锡斌.基于增量学习算法的混合课程学生成绩预测模型研究[J].课程与教学，2021,(7):83-90.

作者简介：胡少启（1990-），男，安徽滁州人，助理实验师，硕士研究生。主要研究方向：教育管理、教育教学改革、教学质量与评价研究。

基金项目：滁州学院教学改革研究项目“‘三全育人’视阈下高校学业预警帮扶工作存在问题和对策研究”（2022jyzb001）