

基于多源数据和模型融合的考研成功率预测研究

刘彬¹, 黎俊伟¹, 邓欣¹, 岳曦², 丁晓宇¹, 胡珂¹, 王一雄¹, 王进¹

¹重庆邮电大学人工智能学院, 重庆

²重庆邮电大学集成电路学院, 重庆

收稿日期: 2025年7月30日; 录用日期: 2025年8月28日; 发布日期: 2025年9月10日

摘要

面对考研报考规模激增而录取率持续走低的现实困境, 以及考生普遍存在的自我评估与志愿决策难题, 本研究致力于构建精准的考研成功率预测模型。研究整合了涵盖学业、行为、心理及主观认知等多源数据, 通过特征工程方法提取关键信息, 并构建了基于LightGBM与CatBoost的融合预测模型。模型在真实考研学生数据上展现出优异的预测性能, F1-score和AUC指标分别达到0.78和0.88。该模型为考生提供了客观、个性化的能力评估与志愿决策支持, 有助于缓解因信息不对称和集群效应导致的资源错配问题, 提升研究生报考的科学性与成功率。

关键词

考研预测, 多源数据整合, 特征工程, 模型融合

Research on Predicting Postgraduate Entrance Examination Success Rate Based on Multi-Source Data and Model Fusion

Bin Liu¹, Junwei Li¹, Xin Deng¹, Xi Yue², Xiaoyu Ding¹, Ke Hu¹, Yixiong Wang¹, Jin Wang¹

¹School of Artificial Intelligence, Chongqing University of Posts and Telecommunications, Chongqing

²School of Integrated Circuits, Chongqing University of Posts and Telecommunications, Chongqing

Received: Jul. 30th, 2025; accepted: Aug. 28th, 2025; published: Sep. 10th, 2025

Abstract

Facing the dual challenges of a booming postgraduate entrance examination applicant pool

文章引用: 刘彬, 黎俊伟, 邓欣, 岳曦, 丁晓宇, 胡珂, 王一雄, 王进. 基于多源数据和模型融合的考研成功率预测研究[J]. 教育进展, 2025, 15(9): 584-591. DOI: 10.12677/ae.2025.1591711

alongside persistently declining admission rates, as well as the widespread difficulties candidates encounter in self-assessment and decision-making regarding institution/program choice, this paper aims to construct an accurate predictive model for postgraduate examination success probability. Our research integrates multi-source data encompassing academic performance, behavioral patterns, psychological traits, and subjective cognitive factors. Key predictive information was extracted through feature engineering methodologies, and a fusion prediction model based on LightGBM and CatBoost was developed. Evaluated on real-world data from examination candidates, the model demonstrates superior predictive performance, achieving an F1-score of 0.78 and an AUC of 0.88. This model provides candidates with objective and personalized ability assessment and decision support for institution/program selection. It contributes to mitigating resource misallocation stemming from information asymmetry and herding effects, thereby enhancing the scientific rigor and success rate of postgraduate application strategies.

Keywords

Postgraduate Entrance Examination Prediction, Multi-Source Data Integration, Feature Engineering, Model Fusion

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

高等教育普及化与数字化转型的交互作用正在重塑研究生教育生态格局。《2024 全国研究生招生调查报告》揭示了考研报考规模五年倍增与录取率持续走低的剪刀差现象，如图 1 所示，图中蓝色线条为报考人数，橙色线条为录取人数，橙色线条上方标注着录取比例，五年来，报考人数由 238 万倍增至 474 万，录取率却从 32.6%降低至 24.15% [1]，残酷的数字背后，是百万人的梦想与现实的激烈碰撞。

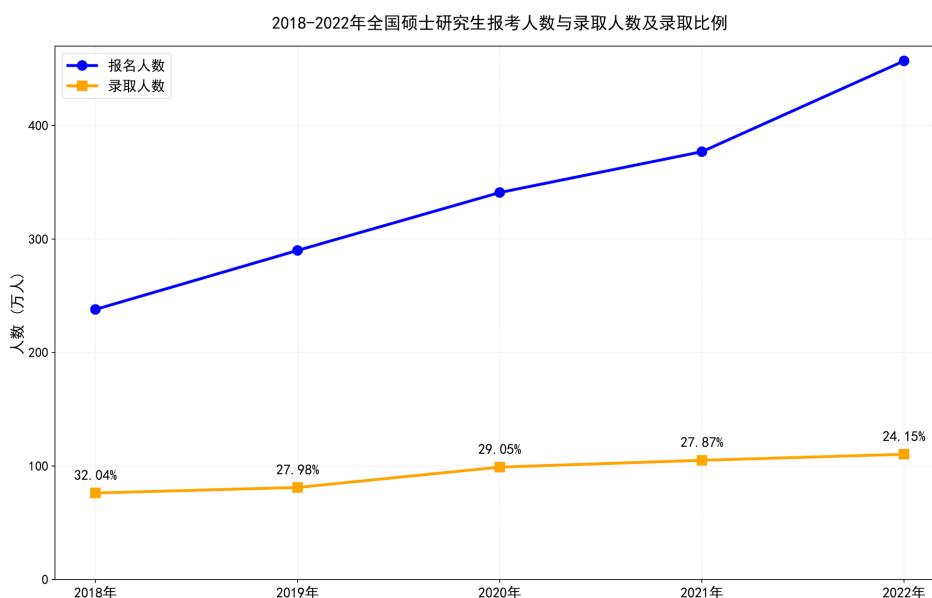


Figure 1. Postgraduate entrance exam stats: applicants, admissions, rates (2018~2022)

图 1. 2018~2022 年研究生入学考生报考人数和录取人数及比例

研究生教育生态中的矛盾不仅体现在规模与录取率的失衡上,更深入到志愿分布、考核模式以及导学关系等微观层面。志愿填报呈现典型的非对称分布特征,61%考生的目标院校高度集中于仅占高校总量12%的双一流院校,这种空间集聚效应导致优质教育资源的竞争烈度呈几何级数增长[1]。

传统备考模式下,学生难以客观自我评估,多依赖离散的模拟考试和有限的题目演练来判断复习效果。在有限的检索能力下,学生对考研的目标设定困难,缺乏有效的决策支持依据。面对激烈的竞争环境,学生在报考决策中需要平衡自身实力、学术兴趣和职业规划等多重因素,但现有的指导体系往往局限于经验分享和历年分数线比较,缺乏个性化的量化分析[2]。部分考生在志愿填报时存在“集群效应”,盲目追随热门院校和专业,导致部分学生高估自己实力盲目冲刺名校最终落榜,另一部分则因保守心理错失更好机会。这种决策偏差不仅降低了整体录取效率,也带来了个人发展的机会成本。

2. 考研预测研究现状

在考研预测领域,国内研究团队已经存在一些先验的探索。李梦莹等人[3]针对工程硕士与学术型硕士选拔机制差异,采用双路注意力机制分别构建模型,预测准确度因此提高约15%,实现了考研预测的分类精细化处理。在此基础上,郑宝乐等人[4]着眼于历史和多维度的数据,整合近五年考研分数、院校分档线及导师评价指标,结合机器学习与并行计算技术开发预测系统,将误差率控制在约3%,使预测结果具备实际参考意义。张懿等人[5]则发现传统线性模型难以准确描述考研复杂场景,通过在回归分析中引入纠偏项,结合考生本科表现与复试情况进行量化分析,改进了录取概率评估方法。这些研究分别从分类模型构建、历史数据挖掘和非线性关系处理等角度推动了考研预测技术的演进。

国际研究上对于此方面的研究更着眼于对学生的数据画像或者特征工程的构建。ALJASMI 等人[6]首先关注多维度指标的整合问题,构建了融合英语语言能力、累积平均绩点、大学排名和推荐信等参数的录取预测模型,通过机器学习技术实现各指标的动态权重调优,该模型在跨校验证中展现出良好的普适性。FEDYNICH 等人[7]的研究将学生数据画像的多样性推向深入,通过对考试成绩与实际学术成就关联性的系统分析,发现单一考试分数难以全面反映学生潜力,特别是对具备跨学科背景、丰富实践经验或创新思维但考试表现不突出的学生。

综合国内外研究成果可见,考研预测正逐步向多模型融合、多维度数据整合和精细化学生画像构建方向发展,但与理想的预测系统相比,当前研究仍存在一些局限。最显著的局限是对动态行为数据关注不足,缺乏对考生复习过程中行为变化和心理波动的实时捕捉,使得预测模型难以反映考生状态的动态演变;其次,数据完整性问题普遍存在,院校数据,尤其是日志化的数据,普遍存在缺失或不准确情况,如果不针对性进行处理,会影响模型训练和预测的质量。最后也是最需要对考研过程中学生能力和状态的高速改变进行适配的问题是,现有预测模型主要依赖静态指标,如历史分数线和考生背景特征等,难以动态反映考生学习过程中的能力提升和知识结构变化。这些挑战不仅揭示了当前考研预测研究的技术瓶颈,也为未来研究指明了突破方向,即需要构建能够捕捉学习过程动态变化、整合多源异构数据并适应不同院校录取特性的新一代预测模型。

3. 数据集和特征工程

3.1. 多源数据集成

本研究构建了一个贯穿2018至2023学年的综合性教育数据集,整合了2532名在校生完整的学习与发展轨迹。数据采集采用了多方协同的策略,通过教务处、信息化办公室与辅导员团队的紧密合作,建立了一个高维度、多层次的学生数据体系。教务处提供的课程成绩与学业规划数据构成了学术评估的基础层;信息化办公室采集的在线行为与资源使用数据形成了行为分析层;辅导员团队记录的课外活动与

心理健康评估补充了社会情感层；结构化问卷与深度访谈获取的自评信息则完善了主观认知层。这四层数据交织融合，构建了包含认知能力、学习行为模式、心理特质与社会参与度的立体化学情画像。

数据集内容结构涵盖四大类特征群，系统性地描绘了学生的全维度状态(如图 2 所示)。学业数据构成了学生能力评估的客观基础，来源于学校教务系统的结构化记录。行为数据通过时序日志形式捕捉了学生的日常学习习惯与活动轨迹，数据主要来自校园信息化设备的自动采集。心理数据整合了学生的心理健康状况与情绪变化轨迹，主要来源于学校心理健康中心的专业评测与记录。主观问卷数据弥补了客观指标难以直接测量的内在因素，通过结构化问卷获取，该问卷通过对数十位不同专业的同学进行深度访谈获得。调查内容涵盖学习满意度的多维评分(学习环境、资源可及性等)，学习动机的内外驱动力评估与目标定向分析，学习策略的选择偏好、执行效果与元认知监控能力。

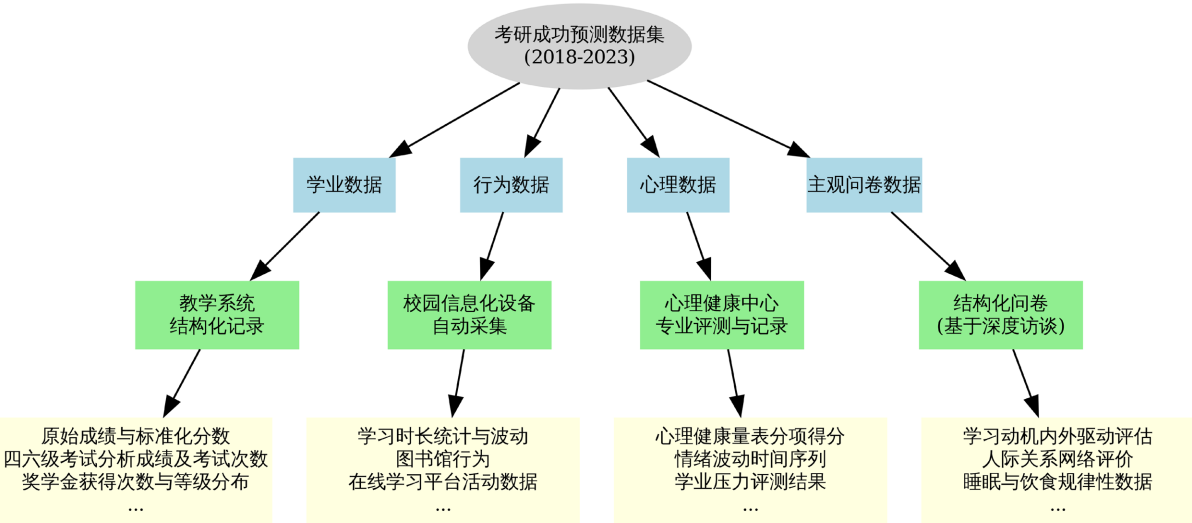


Figure 2. Data source architecture
图 2. 数据来源架构

3.2. 数据安全和隐私保护

为严格保障数据合规使用与隐私安全，本项目实施了全面的多层防护体系。在基础设施层面，研究团队搭建了隔离于校园网主网的专用虚拟研究环境，采用双重认证的跳板机作为唯一访问入口，实现了物理与逻辑隔离。本系统部署了细粒度的角色权限分级矩阵，根据研究人员职责精确分配最小必要权限，确保敏感数据不出安全域。此外，还实施了全程日志记录与异常行为监测机制，对敏感操作进行实时监控与预警。

数据处理层面采取了严格的隐私保护措施。针对学号、姓名等直接标识符，应用基于 ϵ -差分隐私的哈希加盐算法进行不可逆脱敏，设置合理的隐私预算 ϵ 值平衡可用性与安全性。对间接标识符采用 k -匿名化与 t -接近度技术进行泛化处理，确保即使在多维度交叉查询情况下也无法重建个体信息。同时，通过数据字段映射表维护数据间的引用完整性，保留关键研究价值的关联性，满足分析建模需求。

在管理流程层面，建立了三级审批的数据访问机制，包括项目负责人初审、数据安全委员会复审及伦理委员会最终审批，严格控制数据使用范围与时限。本系统集成高精度的操作留痕功能，记录数据访问的 5W1H 信息(何人、何时、何地、何因、何物及如何操作)，实现全链路可追溯。定期进行权限清理与数据访问审计，确保授权状态与实际需求保持一致。

在法律约束层面，研究团队成员全员签署了符合《网络安全法》与《个人信息保护法》要求的保密

协议与数据使用承诺书，详细规定了数据使用边界、禁止行为、责任划分与退出机制。协议明确了数据脱密、匿名化处理与结果发布的标准流程，并针对违规行为设定了对应的法律后果，从法律层面筑牢数据保护屏障。这套完整的数据治理机制不仅为本研究提供了坚实的合规基础，也为教育大数据在尊重隐私前提下的深度应用提供了可复制的实践范式。

3.3. 特征工程

本研究针对考研成功预测任务，设计了一套系统化的特征工程框架，旨在全面捕捉学生学业表现、学习行为与心理状态等多维信息。

基于滑动窗口的统计特征：采用四周滑动窗口，动态提取学生行为时序数据的统计特征，捕捉备考短期变化。计算窗口内的均值评估行为稳定性基准；方差衡量波动程度，识别学习一致性；趋势斜率量化行为增减方向，预警状态变化；稳健极值比反映极端波动幅度。此类特征相比静态指标，更能揭示学生投入的持续性、稳定性及趋势性，为预测模型提供关键时序信息，并辅助识别风险学生进行精准干预。

周期性行为模式分解与行为量化编码：采用时间序列分解，分离出行为数据的趋势项、季节项(日/周/月多尺度)和残差项，并计算异常聚集度与恢复时间以量化中断与恢复能力。同时，构建多维度行为编码：划分晨/午/夜认知效能时段，计算学习效率加权学习时长；利用多尺度自相关函数评估总体及日内/周内/月内规律性。此类特征深入挖掘了行为的周期性、规律性及认知效能分布，形成高阶行为模式刻画，有效适配树模型处理需求。

多维度交互特征衍生，主要设计了四类核心交互特征：

1) 学习时间利用效率(任务数/时长)，区分投入产出差异：

$$E_t = \frac{C_c}{T_h}$$

其中， C_c 表示完成的备考任务数量， T_h 表示总学习时长。该指标直接反映了单位时间内的学习产出，有助于区分“高投入低产出”和“低投入高产出”的学习模式。

2) 心理韧性指数，衡量高压下学习保持能力：

$$R_p = \frac{T_s}{T_a} \times \left(1 - \frac{S_s}{100}\right)^{-1}$$

其中， T_s 表示学生在高压阶段的日均学习时长， T_a 表示学生在常态压力阶段的日均学习时长， S_s 表示学生在高压阶段的压力自评。该指数实质上衡量了学生在压力增加时保持学习投入的能力，并对压力水平进行了加权。

自我认知偏差指数，量化主客观学习时长差异：

$$D_s = 1 - \frac{\sum_{i=1}^n \min(A_i, S_i)}{\sum_{i=1}^n S_i}$$

其中， A_i 表示第*i*天系统客观记录的实际学习时长， S_i 表示学生在当天报告的主观估计学习时长， n 表示观测天数。该指数衡量了学生自我报告与客观行为之间的偏差程度。

学习深度指数，评估专注度、复习频率与连贯性。

$$I_d = w_1 \frac{T_c}{T_t} + w_2 \frac{N_r}{N_a} + w_3 \left(1 - \frac{N_s}{N_m}\right)$$

其中， T_c 表示持续 90 分钟以上的学习节数， T_t 表示总学习节数， N_r 表示知识点平均复习次数， N_a 表示

样本平均复习次数, N_s 表示单位时间内的学习主题切换频率, N_m 表示样本最大切换频率, w_1 、 w_2 、 w_3 为权重系数。该指数有效整合了专注度、复习频率和学习连贯性三个方面。

4. 考研预测模型

4.1. 单机器学习模型

考研预测任务涉及学生背景、学习行为和考试成绩等多源异构数据, 这些数据间存在复杂的非线性关系与交互模式, 而现有研究中对此类特定场景下模型选择的系统性评估仍然不足。本文从模型结构的角度选择了五种代表性算法进行系统性对比。这些算法代表了当前机器学习领域的三大类模型结构: 线性结构、核方法结构和树集成结构。

线性结构以逻辑回归(LR)为代表, 其特点是参数空间简单、决策边界为超平面, 每个特征通过权重直接贡献预测结果。线性结构的理论优势在于可解释性强(权重代表对应特征贡献)、训练稳定性高, 但其固有限制是无法自动捕获特征间的交互效应和非线性关系, 需要人工构建特征交叉项。

核方法结构以 RBF 核支持向量机(SVM)为代表, 通过核函数将数据隐式映射到高维空间, 理论上能构建任意复杂的非线性决策边界。RBF 核实质上将样本点映射到无限维空间, 使原本线性不可分的数据在高维空间中变得线性可分。核方法的优势在于无需显式构建高维特征空间, 能够自动捕获一定程度的非线性关系; 但其隐式映射方式使模型难以解释, 且在高维特征空间可能面临优化困难。

树集成结构选择了三种主流模型: XGBoost、LightGBM 和 CatBoost [8]。这类结构通过梯度提升决策树(GBDT)框架, 集成模型通过不断添加新的决策树来拟合前一阶段模型的残差, 每棵树实质上学习了一个函数映射, 用于纠正当前模型的预测误差。GBDT 算法通过损失函数的负梯度来近似残差, 然后训练决策树拟合这些残差, 从而逐步提升模型性能。树集成结构的理论优势在于: 可自动发现特征交互、处理混合类型特征的能力强、能通过集成减少方差。三种树集成算法在树生成策略上存在差异: XGBoost 采用层级生长策略, 注重平衡性; LightGBM 采用叶级生长策略, 强调效率和深度特征交互; CatBoost 采用对称树结构, 专注于类别特征处理和降低预测偏移。

4.2. 模型融合

不同结构的模型倾向于捕获数据的不同方面, 通过组合这些互补视角可以降低总体方差并减少过拟合风险, 为了提高预测的可信度和准确性, 模型融合策略是不可或缺的关键因素。通过对单模型预测结果的 Pearson 相关系数分析, CatBoost 与 LightGBM 的相关系数为 0.78, 处于一个能够提供互补信息的理想相关性区间。从模型结构来看, 这种互补性可能源于两个模型在处理不同特征类型时的算法特点, LightGBM 的叶级生长策略和直方图算法在处理大量连续型特征时计算效率高且能够构建更深的决策路径, 而 CatBoost 的有序目标统计编码和对称树结构在处理分类特征时会存在明显优势, 并且 CatBoost 模型能够实现特征的浅层交互, 这使得两种模型对样本的观测视角存在明显差异, 最终导致了预测结果的差异。两者算法差异最终转化为两个模型在不同决策边界区域的互补预测能力, 为后续融合系统提供了理想的组件基础。因此, 融合模型采用了 LightGBM 与 CatBoost 的双模型加权融合方案通过交叉验证确定最优权重为 0.55:0.45。

5. 实验结果与分析

为评估模型对未来数据的泛化能力, 研究团队采用了时间序列切分策略, 将数据集划分为训练集(2018~2022 年, 占比 60%)、验证集(2022 年, 占比 20%)和测试集(2023 年, 占比 20%)。这种基于时间的前向划分方法, 模拟了模型在实际应用中面对的预测场景, 确保了评估结果的实践参考价值, 同时也考

验了模型对教育环境动态变化的适应能力。实验采用分层五折交叉验证设计，并保留独立测试集评估最终性能。考虑到样本类别不平衡(考研成功样本约占 33%)，对所有模型应用一致的类别权重调整策略。评估采用 F1-score 作为主要指标，同时参考 AUC-ROC 提供全面评估。

表 1 展示了各模型在独立测试集上的性能结果。实验数据呈现出明显的性能梯度：从线性结构(F1 = 0.66)到核方法结构(F1 = 0.69)再到树集成结构(F1 = 0.73~0.75)，性能逐级提升。这表明考研预测任务确实存在非线性决策边界，线性模型的表达能力不足以完全捕获特征间的复杂关系，例如“高 GPA 但学习行为不稳定”这类复合模式。线性结构与核方法结构之间的性能差异可以发现非线性建模确实带来了预测能力的提升，但提升幅度有限。核方法结构与树集成结构之间存在显著的性能差距，这表明不同结构的模型对特征组合确实具有不同的表达效率。三种树集成模型之间的性能差异虽然相对较小，但存在着预测分布的差异。

相较单一树模型的结果，LightGBM + CatBoost 组合在各项评估指标上均有显著提升：F1-score 提升了 3 个百分点(0.75→0.78)，AUC-ROC 提升了 4 个百分点(0.84→0.88)。融合模型在精确率和召回率的平衡上取得了显著提升，两项指标分别提升了 3 个和 4 个百分点，表明融合模型不仅扩大了对多样化样本的包容能力，也在评分阈值调整时保持了较好的稳定性。LightGBM + CatBoost 加权融合方案在预测性能、计算效率和应用稳健性之间取得了最佳平衡。该方案配合合理的特征工程与参数调优，可以在大规模实际应用场景中高效部署，为考研预测提供更精准的技术支持。

Table 1. Experimental results of the postgraduate entrance examination prediction models
表 1. 考研预测模型实验结果

模型	F1-score	AUC-ROC
逻辑回归	0.66 ± 0.02	0.75 ± 0.03
SVM (RBF 核)	0.69 ± 0.01	0.77 ± 0.02
XGBoost	0.73 ± 0.01	0.82 ± 0.01
LightGBM	0.75 ± 0.02	0.84 ± 0.01
CatBoost	0.74 ± 0.02	0.83 ± 0.01
LGB + Cat (融合)	0.78 ± 0.01	0.88 ± 0.01

6. 结语

本研究聚焦考研成功率预测问题，从数据整合、特征挖掘与算法优化三个维度构建了基于机器学习的考研预测模型。通过整合 2018~2023 学年 2500 余名学生的学业记录、行为轨迹、心理测评及主观问卷等多源数据，科学构建研究样本库；创新性地采用动态行为分析与周期性规律挖掘方法，结合多维特征交互编码，深度刻画备考行为模式；在算法验证中，发现融合型智能算法在高维数据场景下具有显著优势。实证表明，该模型可精准预测学生考研是否上线，进而有效帮助学生客观评估自身水平与目标院校匹配度，降低学生跟风报考名校的盲目性，提升学生考研成功率。本文的研究为精准考研预测提供了一条可操作的技术路径，对提升人才培养质量具有实践价值。

本研究构建的模型主要基于本校单一学院和单一学科的历史数据样本，其预测效能跨院校、跨学科推广时可能存在一定局限。不同院校的生源结构、培养模式、学习氛围及考研政策等外部变量存在差异，这些因素可能影响模型特征的有效性和泛化能力。未来研究将致力于通过跨校合作，引入更广泛的

外部变量(如不同院校特征、学科学习路径等)进行模型验证与优化,以进一步提升模型的普适性和预测精度。

基金项目

本文得到重庆市高等教育教学改革项目“数字化赋能高校学生高质量升学途径研究”(244057)、“‘学练赛思’四维模式下新工科大数据专业实践型人才培养体系建设”(233210),重庆邮电大学教育教学改革研究项目(XJG24107、XJG24108、XWTJG2106、XJG23229、XJG24230、XJG23107、XJG23108、XJG23230)的资助。

参考文献

- [1] 中国教育在线. 2024 全国研究生招生调查报告[R]. 北京: 中国教育在线, 2024.
- [2] 李东洁, 袁善良, 黄林萱, 等. 本科成绩对考研成绩的影响及成功考研的共性调研与分析——以哈尔滨理工大学自动化学院为例[J]. 高教学刊, 2022, 8(19): 53-55, 59.
- [3] 李梦莹, 王晓东, 阮书岚, 等. 基于双路注意力机制的学生成绩预测模型[J]. 计算机研究与发展, 2020, 57(8): 12-34.
- [4] 郑宝乐, 李思成, 唐庭龙. 基于机器学习的考研结果预测算法设计[J]. 信息通信, 2020(10): 17-46.
- [5] 张懿, 胡春美. 基于线性回归的考研成绩分析与预测[J]. 中国宽带, 2020(5): 21-25.
- [6] Aljasmī, S., Nassif, A.B., Shahin, I., et al. (2020) Graduate Admission Prediction Using Machine Learning. *International Journal of Computers and Communications*, **14**, 79-83. <https://doi.org/10.46300/91013.2020.14.13>
- [7] Fedynich, L.V. (2017) The Grand Question: Do Entrance Examinations Determine Graduate Student Academic Success? *Journal of Academic and Business Ethics*, **11**, 12-29.
- [8] 周建华, 单正义, 覃红霞. 新高考是否促进了学生自主选科?——基于 CatBoost 回归树模型的实证分析[J]. 华东师范大学学报(教育科学版), 2024, 42(3): 12-25.