Advances in Education 教育进展, 2024, 14(5), 100-107
Published Online May 2024 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2024.145662

Hans汉斯

基于随机森林和K-Means算法的高校学生评教指标的应用研究

梅　灿[1]，陈　琦[1*]，郝亚兵[2]，刘志鹏[1]

[1]湖北师范大学计算机与信息工程学院，湖北　黄石
[2]湖北师范大学数学与统计学院，湖北　黄石

收稿日期：2024年4月4日；录用日期：2024年5月3日；发布日期：2024年5月10日

摘　要

本文旨在探讨随机森林和K-means算法在高校学生评教体系中的应用及其有效性。首先，通过构建随机森林模型对评教数据进行拟合，分析模型的均方误差和拟合优度，验证其预测能力。进一步利用随机森林的特征重要性评估功能，筛选出对评教结果影响较大的指标，为优化评教体系提供科学依据。同时，对评教指标进行相关性分析，揭示指标间的相互关系。其次，采用K-means算法对评教数据进行聚类分析，通过轮廓系数确定最佳聚类数，并成功将数据划分为三个具有明显差异的聚类。聚类结果揭示了不同教师在教学理念、风格和要求上的多元性，为教学改进和提升提供了参考依据。本文的方法论和结果对优化高校学生评教体系、提升教学质量具有重要意义。

关键词

学生评教，随机森林算法，K-Means聚类算法，评价指标

Research on the Application of Teaching Evaluation Indicators for College Students Based on Random Forest and K-Means Algorithm

Can Mei[1], Qi Chen[1*], Yabing Hao[2], Zhipeng Liu[1]

[1]College of Computer and Information Engineering, Hubei Normal University, Huangshi Hubei
[2]College of Mathematics and Statistics, Hubei Normal University, Huangshi Hubei

Received: Apr. 4th, 2024; accepted: May 3rd, 2024; published: May 10th, 2024

[*]通讯作者。

文章引用: 梅灿, 陈琦, 郝亚兵, 刘志鹏. 基于随机森林和 K-Means 算法的高校学生评教指标的应用研究[J]. 教育进展, 2024, 14(5): 100-107. DOI: 10.12677/ae.2024.145662

Abstract

The purpose of this paper is to explore the application and effectiveness of random forest and K-means algorithm in the evaluation system of college students. Firstly, a random forest model was constructed to fit the evaluation data, and the mean square error and goodness-of-fit of the model were analyzed to verify its prediction ability. Furthermore, the feature importance evaluation function of random forest was used to screen out the indicators that had a great impact on the evaluation results, so as to provide a scientific basis for optimizing the evaluation system. At the same time, the correlation analysis of the evaluation indicators was carried out to reveal the correlation between the indicators. Secondly, the K-means algorithm was used to analyze the clustering of the evaluation data, and the optimal number of clusters was determined by the contour coefficient, and the data were successfully divided into three clusters with obvious differences. The clustering results revealed the diversity of teaching concepts, styles and requirements of different teachers, and provided a reference for teaching improvement and promotion. The methodology and results of this paper are of great significance for optimizing the student evaluation system and improving the teaching quality of colleges and universities.

Keywords

Student Evaluation, Random Forest Algorithm, K-Means Clustering Algorithm, Evaluation Indicators

1. 引言

随着高等教育的蓬勃发展，教学质量的评价与提升已成为教育改革的核心议题。2020 年，《深化新时代教育评价改革总体方案》的出台，为深化教育改革和提升教学质量指明了方向，其中，学生评教作为高校教学质量评价的重要组成部分，受到了广泛的关注[1] [2]。学生评教不仅是衡量教师教学效果的直观反馈，更是推动教师专业化发展、优化教学管理的重要机制。然而，如何确保评教体系的科学性、合理性和公正性，一直是高校教学管理面临的难题。

在这一背景下，数据挖掘和机器学习技术的兴起为高校学生评教体系的研究与实践提供了新的视角。随机森林算法，以其强大的预测能力和特征选择功能，为评教结果的预测和评教指标的筛选提供了有力支持。同时，K-means 算法作为一种无监督学习方法，能够深入挖掘评教数据的内在结构，揭示教师的不同教学风格和特点。这两种算法的结合应用，有望为高校学生评教体系的优化提供新的思路和方法。

本文旨在探讨随机森林和 K-means 算法在高校学生评教体系中的应用及其有效性。通过构建随机森林模型分析评教数据，筛选关键评教指标，揭示指标间的相关性；同时，运用 K-means 算法对评教数据进行聚类分析，划分不同的教师群体，以揭示教学风格的多样性。期望本研究的结果能够为高校学生评教体系的完善、教学质量的提升以及教师的专业化发展提供有益的参考和借鉴。通过对学生评教数据的深度挖掘和有效利用，期待能够为深化教育改革、提升教学质量贡献一份力量。

2. 数据来源及方法

2.1. 数据集描述

本文采用的数据集来源于 H 高校的学生评教系统，选取 2022 年 12 月份的学生评教数据，涵盖了各学院、不同课程的学生评教数据。数据集包括了学生基本信息、课程信息、教师信息以及学生对教学的各项评价指标。首先对数据进行预处理，包括数据清洗、数据转换和数据集成等步骤，去除了重复、缺失和异常值，将字符串类型转换为数字类型，以及将各二级指标的得分整合到对应的一级指标下面。这些预处理操作有助于消除数据中的噪声和异常值，使得后续的数据分析和挖掘更加准确和可靠。预处理后的数据集包含：教师编号、教学态度、教学内容、教学用书、教学方法、教学结果及总分 7 的字段，用于后续的分析。预处理后的部分数据见表 1。

Table 1. Some of the data are evaluated by students
表 1. 学生评教部分数据

教师编号	教学态度	教学内容	教学用书	教学方法	教学效果	总分
1	19.9	30	10	10	28.3	98.13
2	20	30	10	10	30	100
3	19.5	29	9.7	9.7	28.9	96.82
4	19.5	29.1	9.7	9.7	28.8	96.71
5	17.6	27.1	9.2	9.1	27.1	90
6	20	30	10	10	30	100
7	20	30	10	10	30	100
8	19.5	29.4	9.8	9.8	29.6	97.63
9	20	30	10	10	30	100
10	19.6	27.8	9.3	9.6	29.8	95.75

2.2. 随机森林算法

随机森林算法是一种强大的集成学习算法，它基于决策树的概念，并通过结合多个弱分类器来形成一个更为准确和稳健的分类模型。该算法在处理分类、回归乃至降维问题时表现出色，同时对异常值和噪声数据具有高度的容忍性。相较于单一决策树模型，随机森林在预测和分类方面展现了显著的优势。作为 Bagging 类型的一种集成算法，随机森林通过集成各个弱分类器的结果，采用投票或平均值的方式来提升整体模型的精确度和泛化能力。这种卓越表现主要归功于算法中的两大核心要素：随机性和森林结构。随机性通过引入数据采样和特征选择的随机性，有效地增强了模型的抗过拟合能力；而森林结构，即多个决策树的集合，则通过集成各个树的预测结果来提高预测的准确度。在随机森林算法中，每个决策树都是基于原始数据集的一个有放回抽样子集进行训练的，这样可以确保各个树之间的独立性。每个树都会产生一个分类结果，而最终的预测结果则是通过取所有树分类结果的众数来确定的。此外，随机森林在每个节点上还采用了随机特征选择机制，进一步增强了模型的泛化能力，使其在各种复杂数据集上都能取得优异的性能。

2.3. K-Means 聚类算法

K-means 算法是一种基于簇聚类的无监督算法，可根据不同因素指标给出较为客观合理的聚类结果，

且算法的复杂度低，适合处理高维数据集。针对给定的多维数据集，通过按照样本数据值之间的欧氏距离大小，将其划分为 K 个簇，使存在同一个簇类之间的样本点尽量靠近，不同簇之间的距离尽量大[3]。假定划分簇类集为$(C_1, C_2, \ldots, C_K)$，则簇内欧氏距离 D 如下：

$$D = \sum_{i=1}^{k} \sum_{x \in C_i} \left\| x - u_i \right\|^2 \tag{1}$$

其中，u_i 为簇 C_i 内的质心，表达式如下：

$$u_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{2}$$

通过在数据集中随机选取质心，计算数据集样本点到质心的距离，根据样本点距质心的距离将其分簇，重复 K 轮，使 D 值达到最小后停止，得到最优的聚类结果。

3. 随机森林算法的应用

3.1. 随机森林模型的构建与结果分析

在构建随机森林模型时，首先对原始数据集进行了划分，将其分为训练集和测试集。训练集用于模型的训练和学习，而测试集则用于评估模型的性能和泛化能力[4]。其次，选择了适当的参数来构建随机森林。

将样本容量按 7:3 的比例划分训练集和测试，选取测试集进行随机森林拟合，拟合得到测试集样本的实测值和预测值，如图 1 所示[5]。

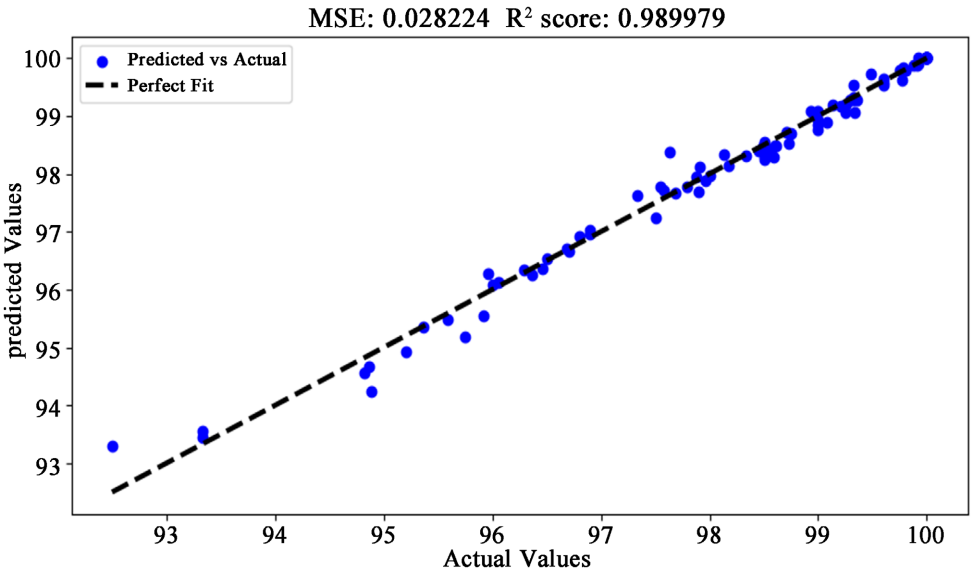

Figure 1. Random forest fitting plot
图 1. 随机森林拟合图

从随机森林模型的拟合结果来看，本模型的均方误差为 0.0282，拟合优度为 0.9900，拟合效果和实际数据基本拟合。这一结果表明，通过本模型所得到的实际综合评分能够精准地反映出学生对教师教学质量评价的真实情况，具有很高的可靠性和有效性。

3.2. 随机森林模型在评教指标筛选和权重确定中的有效性分析

随机森林模型的特征重要性评估功能能够直观地了解各个评教指标在模型中的重要性程度。根据特

征重要性的排序，可以筛选出对评教结果影响较大的指标，作为优化评教体系的重要依据。同时，特征重要性的大小也可以作为确定各个指标权重的参考依据，从而构建更加科学、合理的评教指标体系。

随机森林提供了两种特征选择的方法：平均不纯度减少和平均精确度减少[6]。本文采用平均不纯度减少(Mean Decrease Impurity)方法测量每个特征在树的不纯度减少(如基尼不纯度或信息增益)方面的贡献。当一个特征被用于分割节点时，它通常会减少该节点的不纯度。通过在森林中所有树上平均每个特征的不纯度减少量，我们可以得到每个特征的重要性得分。

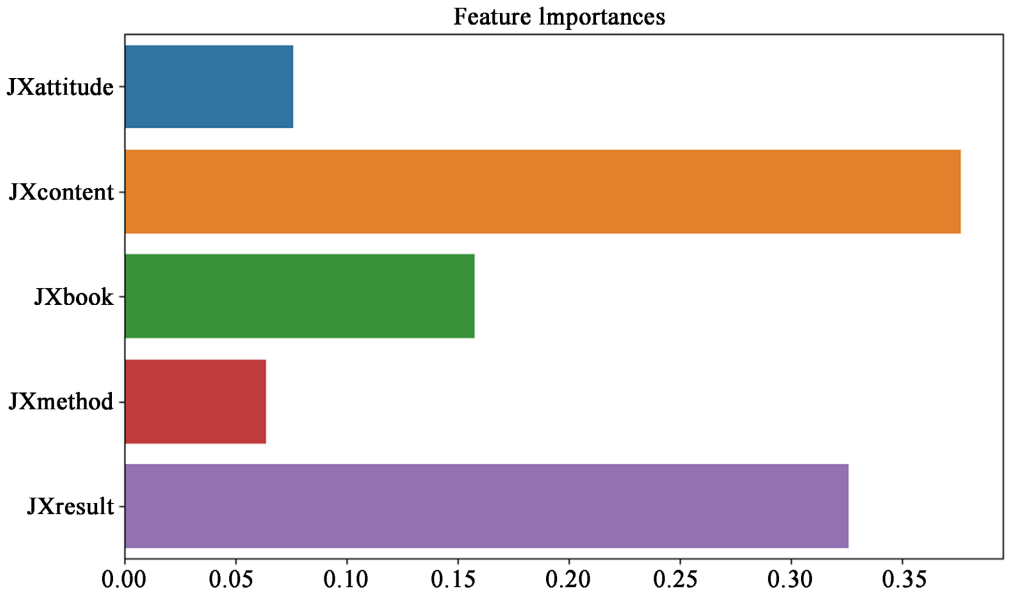

Figure 2. Feature importance visualization
图 2. 特征重要性可视化图

研究结果如图 2 所示，在评教过程中，教学内容的质量和深度对学生最重要，特征重要性值达0.37679。教学内容的充实性、前沿性及满足学生需求的程度显著影响学生的整体评价。其次，教学结果的重要性为 0.32583，反映学生对学习收获的看重。教学用书以 0.15767 排在第三，其选用和内容相关性对学习效果有影响。教学态度虽重要性较低(0.07586)，但教师的热情、耐心等仍受学生关注。教学方法重要性最低(0.06384)，可能因学生对其多样性需求低，但其在实际教学中仍对提高效果和激发兴趣重要。

3.3. 学生评教指标的相关性分析

在高校学生评教体系中，优化评教体系、提升评教质量需深入了解评教指标间的相关性。本文基于随机森林算法，对高校学生评教指标进行相关性分析，并通过热力图直观展示。热力图中的颜色深浅代表指标间的相关系数大小，揭示其相关性强弱：深色表示强相关性，浅色表示弱相关性，从而有助于优化评教体系。

如图 3 所示，五个一级评教指标之间的相关系数均高于 0.65，呈现出较强的正相关性。具体来看，教学内容与教学结果、教学用书与教学方法、教学内容与教学用书之间的相关系数分别达到了 0.81、0.81 和 0.80，均为高度相关。这些高相关系数不仅表明这些指标在评价教学质量时具有较高的一致性，即它们在很大程度上共同反映了教学质量的某些方面；同时也揭示了它们在教学过程中相互关联、相互影响的紧密关系。

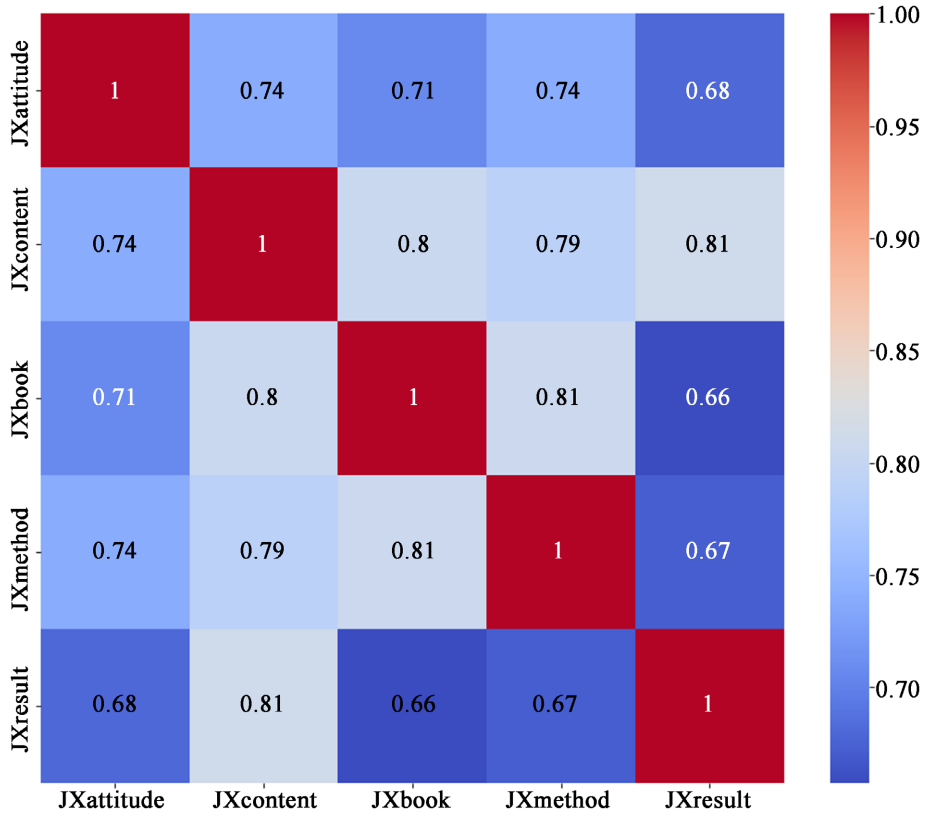

Figure 3. Correlation plot between indicators
图 3. 各指标之间的相关性图

4. K-Means 算法的应用与结果分析

尽管随机森林算法在评教中有其优势，但为了更深入挖掘数据结构和分类特点，本文还采用了 K-means 算法进行聚类分析。作为一种无监督学习方法，K-means 能发现数据的自然分组，有助于更全面地理解评教数据的分布和特征。

本文运用 K-means 算法对五个学生评教指标进行聚类，并利用轮廓系数方法(Silhouette Coefficient)确定最佳聚类数。轮廓系数越接近 1，表明聚类效果越佳，在实际应用中，通常认为轮廓系数大于 0.5 时，聚类效果相对较好。为确保分析的有效性和可视化的便捷性，将聚类数的选择范围限制在 3 至 6 个之间，进一步运用主成分分析法将数据降维至两个主成分进行可视化展示。经过分析比较，当 k 值设定为 3 时，轮廓系数达到最高值 0.563，表明聚类效果相对较好，轮廓系数图如图 4 所示。因此，选择 k = 3 进行最终聚类分析，以获得更具意义的评教结果解读。

通过轮廓系数法得到了最终的聚类数目，并将 k 值设为 3，成功地将评教数据划分为 3 个聚类，聚类结果用散点图展示，如图 5 所示。每个聚类内部的数据点在评教指标上具有高度的相似性，而不同聚类之间则呈现出明显的差异。

聚类 0：这一类别的教师展现出了高度的教学热忱和专业性，在教学态度、教学用书选择以及教学方法上都得到了很高的评价，教学结果也相应地获得了较高的评分。这类教师在教学内容的组织上表现平稳，整体而言，具备了多方面的出色教学能力。

聚类 1：在这个聚类中，教师在教学态度上显得较为缺乏热情和投入，教学内容的组织也显得不够充实和有价值。尽管他们选用了高质量的教材，但由于教学方法不佳，无法有效地利用这些资源。这些

因素共同导致了教学结果的较差表现，因此，这些教师需要在教学态度、教学方法以及教学内容的组织上进行全面提升。

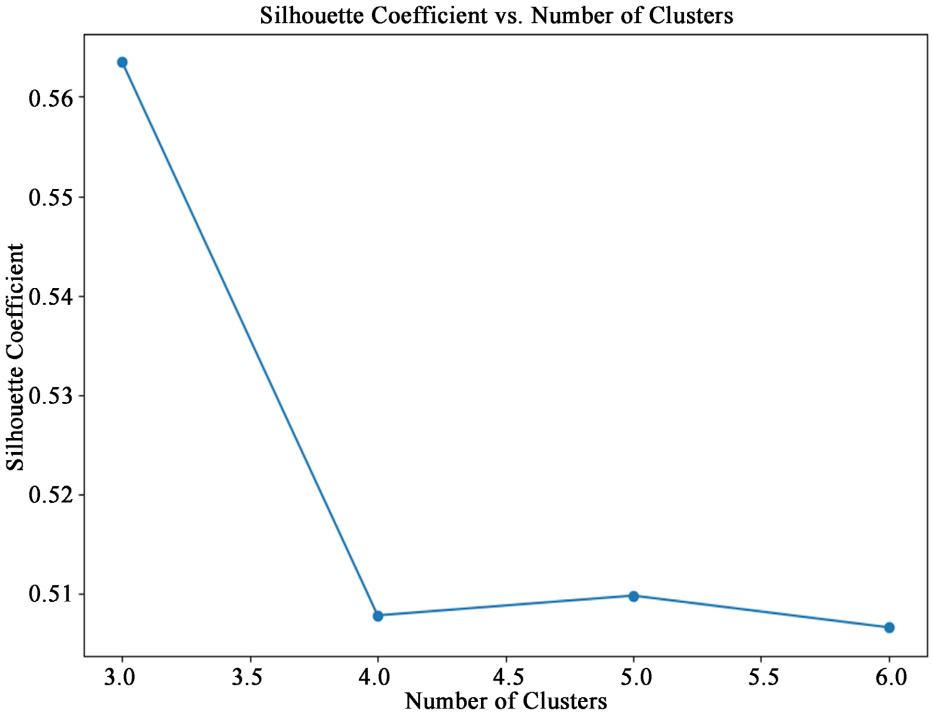

Figure 4. Contour coefficient diagram
图 4. 轮廓系数图

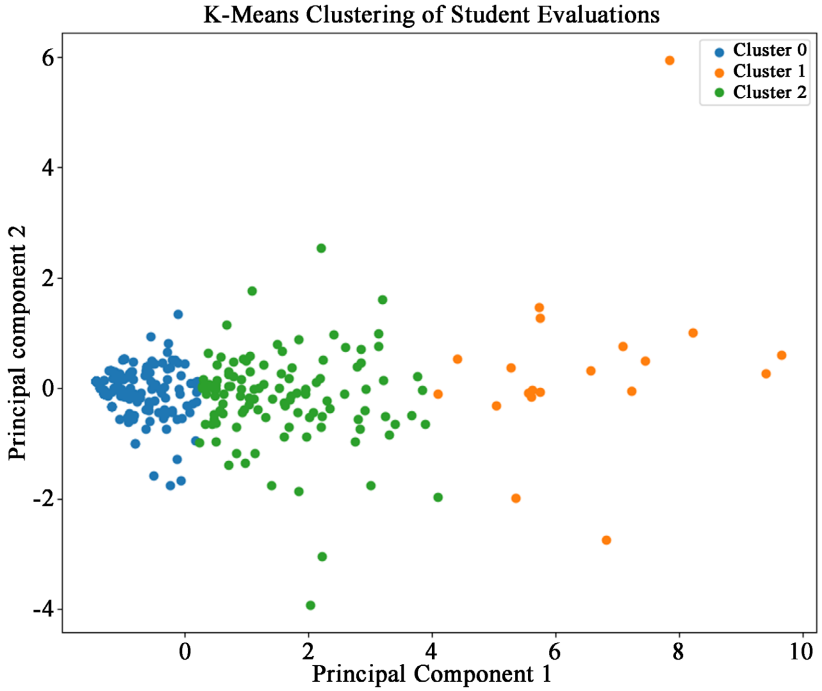

Figure 5. Clustering effect diagram
图 5. 聚类效果图

聚类 2：在这个聚类中，教师在教学态度和教学内容的组织上表现良好，能够为学生提供必要的指导和知识技能。虽然在教学用书的选择上存在不足，但他们采用的教学方法能够有效地促进学生的学习，最终也取得了不错的教学成绩。为了进一步提升教学质量，这类应更加关注教学用书的选择，确保教材与教学内容的高度契合。

不同聚类之间在评教指标上的显著差异，揭示了教师们在教学理念、风格和要求上的多元性，这为教学改进和提升提供了宝贵的参考依据。通过深入分析这些差异，我们可以更有针对性地支持教师发展，优化课程设置，从而提升整体教学质量。

5. 结论与展望

本文运用随机森林算法和 K-means 算法对高校学生评教指标进行深入研究。通过构建随机森林模型，成功地对评教数据进行了拟合，并准确评估了各个评教指标的重要性，为筛选关键指标和确定指标权提供了有力的支持。同时，K-means 算法的应用进一步揭示了评教数据的内在结构和分类特点，为全面理解评教数据的分布和特征提供了新的视角。这两种方法的结合使用，为我们提供了更全面、更深入地理解评教数据的途径。

本研究虽取得一定成果，但仍存在局限，如数据集大小和代表性的问题可能影响结果的稳定性和泛化性。未来，应致力于收集更全面、更具代表性的评教数据以增强研究的可靠性。同时，算法参数的选择也需进一步优化，通过交叉验证、网格搜索等方法提升模型性能，从而深入研究评教指标。

随着技术的不断进步，评教指标研究有望通过引入更先进的算法和技术实现更高的准确性和有效性。深度学习、集成学习等方法的探索应用，将为评教研究开辟新的路径。此外，更全面、详细的评教数据收集将有助于更深入地分析教学质量和学习体验。进一步地，研究评教指标与其他教育因素(如学习成绩、课程满意度等)的关联，将有助于揭示评教在教育过程中的更深层次影响和作用机制，为优化教育资源配置和提升教学质量提供更有力的支持。

参考文献

[1] 新华社. 中共中央国务院印发《深化新时代教育评价改革总体方案》[EB/OL]. http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202010/t20201013_494381.html, 2024-03-13.

[2] 燕姣云, 安俊丽, 孙国红. 课堂教学质量评价指标体系重构[J]. 中国大学教学, 2023(12): 74-78+91.

[3] 戴兴雨, 王卫民, 梅家俊. 基于深度学习的手语识别算法研究[J]. 现代计算机, 2021, 27(29): 63-69.

[4] 许盛彬. 大面积化工园区浅层地下水污染风险预测研究[J]. 低碳世界, 2023, 13(7): 13-15.

[5] 蒋明池, 胡圣波, 孟欣. 基于随机森林算法的高校学生评教指标研究——以程序设计基础课程为例[J]. 凯里学院学报, 2023, 41(3): 74-81.

[6] 段中满, 贾亮亮, 蒋明光, 等. 基于不同特征选择方法和随机森林法的滑坡易发性评价——以湖南中西部地区为例[J]. 华南地震, 2023, 43(2): 115-124.