 DOI: https://doi.org/10.12345/jxffcxysj.v6i23.14975

Application of Machine Learning in Predicting Academic Performance—Taking *College Physics* as an Example

Rongyu Zhang Wei Wang Xu Yang Junmei Yang Qin Nie

College of Science, Shenyang Aerospace University, Shenyang, Liaoning, 110135, China

Abstract

This paper investigates the application of machine learning in predicting academic performance, using the course *College Physics* as an example. Traditional education commonly faces issues of standardization and generalization, which cannot meet the individual differences of students and provide personalized feedback and support. Machine learning techniques analyze students' learning data and behaviors to achieve personalized learning and provide teachers with information about students' learning status and needs. This paper comprehensively applies linear regression, random forest, and decision tree algorithms to analyze the data of the *College Physics* course over the past five years, achieving good prediction results. The results show that the decision tree algorithm performs excellently in predicting student performance and has practical significance in predicting and alerting final grades. Additionally, the results of feature contribution extraction indicate that the teaching level of teachers has a greater impact on student performance. This research provides references for teaching reflection and reform, demonstrating the potential and application prospects of machine learning in predicting academic performance.

Keywords

College Physics; machine learning; teaching reform; score prediction

机器学习在学习成绩预测中的应用——以《大学物理》为例

张蓉瑜 王微 杨旭 杨俊梅 聂琴

沈阳航空航天大学理学院，中国·辽宁 沈阳 110135

摘 要

论文以《大学物理》课程为例，探讨机器学习在学习成绩预测中的应用。传统教育普遍存在标准化与泛化的问题，无法满足学生的个体差异和提供个性化反馈和支持。机器学习技术通过分析学生的学习数据和行为，实现个性化学习，并为教师提供学生的学习状态和需求信息。论文综合应用线性回归、随机森林和决策树算法，分析了《大学物理》课程近五年的数据，取得了良好的预测效果。结果显示决策树算法在预测学生成绩方面表现优异，对期末成绩进行预测和预警具有实际意义。此外，特征贡献提取结果显示教师的授课水平对学生成绩的影响更大。本研究为教学反思与改革提供了借鉴，展示了机器学习在学习成绩预测中的潜力和应用前景。

关键词

《大学物理》；机器学习；教学改革；成绩预测

1 背景介绍

传统教育的特点之一是教育的标准化与泛化。传统通常将知识传授给所有学生，忽视了每个学生的独特需求和学习方式。这种泛化式教育无法满足不同学生的个体差异，可能导致部分学生学习困难或者失去兴趣。此外，传统教育缺乏即时的个性化反馈和支持，学生往往无法及时发现自己的问题并加以改进。另外，高校中很多课程采取大班授课，教师需要面对几百人的出勤、作业完成情况等海量数据，难以对学生的学习状态、困难和需求进行有效及时的监控。

近年来，随着科技的发展和进步，机器学习在教育领域发挥着越来越重要的作用[1]。首先，机器学习技术可以通过分析学生的学习数据和行为来实现个性化学习。通过应用机器学习算法，教育可以根据学生的需求和特点，为他们提供个性化的学习资源、反馈和支持。这种个性化学习可以更好地满足学生的学习需求，提高学习效果和学习动力。其次，机器学习可以自动分析和处理大量的数据，并向教师提供学生的学习状态、困难和需求等信息，帮助教师更好地了解学

【课题项目】“八何分析法”在《大学物理》课程中的应用（项目编号：JG2023084）；2022年度辽宁省普通高等教育本科教学改革研究优质教学资源建设与共享项目；基于“立德树人、信息技术、多元评价”的《大学物理》跨校修读课程教学模式研究与实践。

【作者简介】张蓉瑜（1987–），女，中国辽宁盘锦人，博士，讲师，从事物理学研究。

生的学习情况，提供个性化的教学建议和支持，从而改善教学质量和效果。最后，通过对学生学习数据和行为的分析，机器学习还可以预测学生的学习进度、困难和可能遇到的学习障碍。这种学习监测和预测有助于学生及时发现和解决学习问题，向教师提供相关的教学措施，提高学生的学习成绩和自我管理能力。

在教学方面，曹梦川等人收集了学生的平时成绩、考勤、性别和期末成绩等多种因素的数据，并使用线性回归模型进行数据建模和预测分析，旨在提前预警学生可能存在的挂科或成绩下降风险，从而帮助学生和教师更好地制订学习计划，提高教学效率 [2]。张峰等人则基于非过程性特征和过程性特征两类数据，采用决策树、支持向量机等方法进行了成绩预测，并取得了良好的效果 [3]。这些研究为教学工作提供了有益的参考。

我们跟踪监测了近五年的《大学物理》课程的平时、期中和期末成绩等数据，利用线性回归、随机森林以及决策树算法进行了学习和预测，取得了良好的效果，为今后的教学反思与改革提供了借鉴。

2 实施步骤

首先，我们收集了学生的历史学习数据，包括作业成绩、出勤情况、随堂测验和视频观看记录等。我们还根据教学大纲给出了学生的平时成绩。同时，我们还获取了近五年各个专业和班级的期中考试成绩、期末考试成绩以及相应的考试分析数据，包括每道题的得分率、每个章节知识点的得分率和授课教师信息。

接下来，我们进行了数据预处理，以优化数据集，确保机器学习的良好效果。预处理包括缺失值处理、异常值处理、数据类型转换以及数据一致性和规范化。对于缺失值处理，我们使用插值方法填充了部分可插值的数据，并删除了无法填充的数据。对于字符串类型的特征值，如授课教师和课程类型（大学物理 A、B 和 C），我们采用了独热编码（One-Hot Encoding）来将其转换为二进制向量。以课程类型为例，利用 Python 中的 sklearn 库进行独热编码分析，部分代码如下：

```
from sklearn.preprocessing import OneHotEncoder
encoder = OneHotEncoder()
class_type = [[’大学物理 A’], [’大学物理 A B’], [’大学物理 A C’]]
ncoded_values = encoder.fit_transform(feature_values).toarray()
```

然后我们将其可视化，得到图 1。

这样做的好处是将每个特征值映射到一个唯一的二进制位，因此不会丢失任何特征值的信息。此外，独热编码不会赋予特征值之间的关系或顺序任何偏好，因此该方法使得特征值可以很好地兼容大多数的机器学习算法。

然后，我们进行了机器学习模型的选择。我们采用线性回归算法、随机森林算法和决策树算法。我们没有选择深度神经网络的原因是数据集的规模较小，导致训练效果不理想，并且模型的可解释性较差。与此相反，随机森林算法和决策树算法都是树形分类器，可以处理离散特征值和连续特征值，对数据集的要求较小，并且具有较强的模型可解释性。

最后我们进行了模型训练与评估。训练过程中，对于线性回归算法，我们采用了最小二乘法进行训练，而对于随机森林和决策树算法，因篇幅关系，不再赘述各种超参数的调优。验证模型准确性时，我们采用了十折验证法，即 90% 的数据用来训练，10% 的数据用来预测。

3 分析与讨论

我们首先使用了线性回归算法对训练集进行训练，并在验证集上进行验证。结果如图 2 所示，纵坐标表示预测的班级期末考试平均成绩，横坐标表示验证集中实际的期末考试平均成绩。两者之间的差距越小，说明预测越准确。在图中可以明显看出，预测成绩和实际成绩存在较大差距，准确率仅为 5.3%。接下来，我们采用了随机森林算法，预测准确率提高到了 60.9%。如图 2（b）所示，与线性回归算法相比，随机森林算法的预测值与测试集中的实际值呈现出线性特征，这表明采用该算法取得了一定的效果。我们将预测

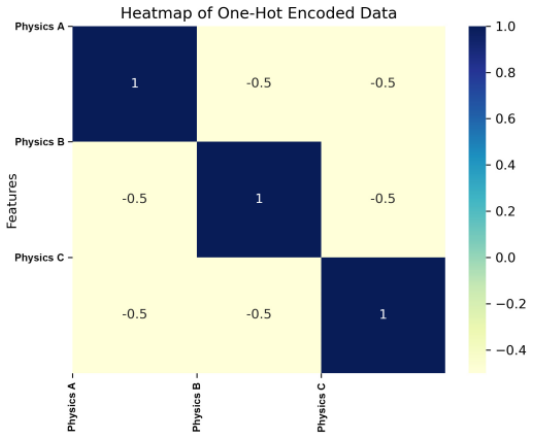
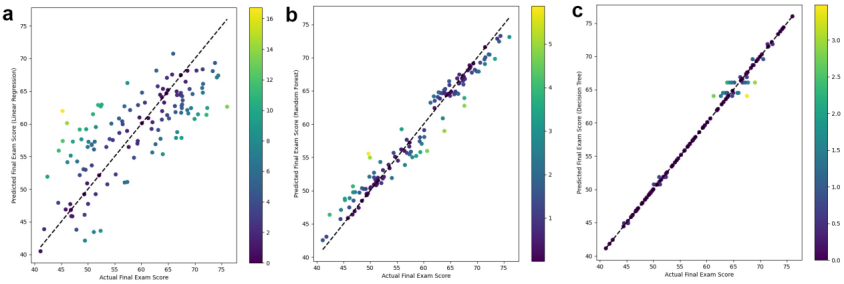


图 1 大学物理课程类型的独热编码转化与可视化

值和实际值的偏差映射到散点的颜色上，颜色越深表示差距越小。总体来看，预测与实际值的偏差在 2~3 分之间。然后，我们又采用了决策树算法进行训练和验证，准确率达到了 97.2%。如图 2（c）所示，除少量散点误差较大之外，绝大多数点的预测值与实际值的偏差都在 1 分以内。决策树模型的拟合效果优于随机森林模型，一方面是因为数据集较小，决策树算法在处理小数据集时更占优势，另一方面原因是随机森林算法具有较强的泛化能力和抗过拟合能力，而决策树算法可能存在过拟合的风险。随着今后数据集的不断增加，拟合效果有望进一步提高。



（a）线性回归算法　（b）随机森林算法　（c）决策树算法

图 2 利用不同方法预测期末成绩和实际期末成绩的关系

除此之外，我们在教学研讨中常遇到一些问题，比如班级成绩的优劣究竟更受教师之间教学水平的差距影响大，还是学生所在学院和专业的学生生源影响大。此外，针对不同学院和专业的需求，我们开设了三种不同类型的课程：《大学物理 A》《大学物理 B》和《大学物理 C》。其中，《大学物理 A》的课时较多，对电磁学和热学部分的学习内容更加详实。《大学物理 B》和《大学物理 C》的课时稍少，前者更注重电磁学，后者则更注重热学。课程类型对学生成绩的影响究竟有多大？实际上，影响成绩的因素远不止这些。归结为数学模型，就是多个因素对一个结果的影响问题。凭借教学经验和人的直觉，很难将这些影响量化。

针对以上问题，我们采用机器学习的方法做了进一步的研究。决策树算法通过构建一个树状结构来进行决策。每个决策树的节点表示一个特征，树的分支表示该特征的取值，而叶节点表示最终的决策结果。决策树的构建过程包括选择最佳的特征来进行分裂，并递归地构建子树。通过分析决策树，可以提取特征贡献，帮我们理解数据特征对决策的影响程度。

我们对授课教师和专业班级对成绩的影响进行了量化分析。发现有些教师对成绩的影响更大，而有些班级对成绩的影响更大。在专业相近的班级中，录取分数接近，我们认为学生生源差距较小，此时教师的水平将发挥更为重要的作用。而对于授课水平相近的教师，不同的生源将对成绩产生更大的影响。这也是我们在讨论时总是无法确定教师影响还是生源影响更大的原因。综合而言，就近五年的考试成绩分析来看，教师对成绩的影响略高于班级对成绩的影响。这也给我们提供了一些启示，任课教师应更多地夯实基础，提升个人教学能力，通过精心教授能够使基础较差的班级取得相对较好的成绩。

4 结论

综上所述，我们利用机器学习分析了近五年《大学物理》课程相关数据，利用线性回归、随机森林和决策树等算法，进行了模型训练和预测。决策树算法取得了良好的预测效果，在今后的工作中，可以根据平时成绩和期中成绩，对期末成绩进行预测和预警。此外特征贡献提取结果显示，任课老师的授课水平比生源的影响要更大一些。

参考文献

[1] Ceren Korkmaz, Ana-Paula Correia. A review of research on machine learning in educational technology[J]. Educational Media International,2019(56):250-267.

[2] 曹梦川,欧阳仪,伍丹,等.基于机器学习的学生学情预警方法研究[J].现代信息科技,2023(7):142-150.

[3] 张峰,陈静静.适用于学生成绩预测的学习数据特征综述[J].软件工程,2023,26(10):1-4.