基于机器学习的创造力影响因素挖掘——以 PISA2022 港澳学生为例

Mining the Factors Influencing Creativity Based on Machine Learning: A Case Study of Hong Kong and Macao Students in PISA 2022

彭天伟[1*]，李艺凡[2]，袁裕添[3]，宿博涵[4]，张缨斌[5]，穆肃[6]
[1,2,3,4]华南师范大学教育信息技术学院
[5,6]华南师范大学教育人工智能研究院
[*]2024020894@m.scnu.edu.cn

【摘要】 创造力对解决诸多自然环境、人类社会和自我发展等问题起到关键作用。本研究基于创造力投资理论，使用机器学习算法分析 PISA2022 港澳数据，挖掘港澳学生创造力的影响因素。结果表明：（1）LightGBM 模型性能在港澳学生数据中优于随机森林、XGBoost 和 Lasso 回归模型；（2）在 83 个影响因素中，数学素养、科学素养、阅读素养、ICT 相关资源、经济文化与社会地位等对学生的创造力有着重要影响。
【关键词】 创造力；PISA；机器学习；影响因素；

Abstract :*Creativity plays a crucial role in addressing a multitude of issues related to the natural environment, human society, and self - development. It also plays an important role in helping students actively cope with uncertainties in the development and changes of the world. Based on the creativity investment theory, this study uses machine - learning algorithms to analyze the PISA 2022 data of Hong Kong and Macao, exploring the influencing factors among students in these regions. The results show that: (1) The performance of the LightGBM model is superior to that of the Random Forest, XGBoost, and Lasso Regression models in the data of Hong Kong and Macao students; (2) Among 83 influencing factors, Mathematical literacy, scientific literacy, reading literacy, ICT - related resources, and economic, cultural and social status have a significant impact on students' creativity.*

Keywords: Creativity, PISA, Machine Learning, Influencing Factors

1. 引言

创造力是国家发展和民族进步的不竭动力，教科文组织也将创造力视为人的核心技能之一，创造力逐渐成为极其重要的能力。创造力是指根据一定目的，运用一切已知信息，产生出某种新颖、独特、有社会价值或个人价值的产品的智力品质，是拔尖创新人才的关键性特征(Sternberg & Karami,2022)，表现为创新思维水平、创新实践能力等多种品质的统一(Kaufmanj & Glaveanu,2019)，在帮助学生积极应对世界发展与变革中的不确定性中扮演着重要角色(胡朗宁 & 吕立杰，2024)。因此，探究学生创造力发展的影响因素至关重要，是助力提升学生创造力发展的首要条件。相关研究表明学生创造力发展受教学模式、教学环境、性格特征、认知风格等因素影响，但以往研究多基于线性回归方程，单个研究中纳入的影响因素有限，且忽视因素对创造力的非线性影响及多因素间的交互作用。鉴于此，本研究基于 PISA2022 的港澳学生数据，采用机器学习算法探究学生创造力的影响因素，以期为学生创造力的提升、拔尖创新人才的培养及未来的教育发展提供有益的优化方向。

2. 文献综述

2.1 创造力发展的影响因素研究

关于创造力发展的影响因素研究，众多心理学及教育学领域学者进行相关探索，形成较为经典的创造力模型及理论，如吉尔福特（Guilford）的智力三维结构模型、阿玛布丽（Amabile）的阿氏创造力结构模型、考夫曼（Kaufman）等人的 4C 理论、格鲁韦努（Glăveanu）的 5A 理论、斯滕伯格（Sternberg）的创造力投资理论等。其中，斯滕伯格在吸收其他创造力理论内核和有益思想基础上提出影响创造力发展的六个核心资源（见表 1），包括智力、知识、思维风格、人格特征、动机、环境（Sternberg,2014），对学生创造力发展的解释力较强（李硕豪，2020）。此外已有研究表明个体创造力发展受到外部及内部双重因素影响，如教育环

境、社会支持、教学课程等外部因素（王倩等人，2024；Al-Kumaim et al., 2021），性别及年龄、知识经验、认知风格、个体情感等内部因素（刘芝延等人，2021；徐瑾劼等人，2024；TEMPOE，1998）。与此同时，随着科技飞速发展，生成式人工智能等新兴技术凭借其大数据的特征及个性化、定制化的服务等优势，为学生创造力培养开辟了新的路径。尽管这些研究在一定程度上揭示了创造力的影响因素，但研究大多局限于单一影响因素研究，忽视创造力的复杂动态过程，在探讨创造力的多因素交互作用及其综合效应方面存在局限（Xu, 2024）。此外，现有研究大多集中于西方文化背景，对我国学生创造力影响因素的探讨相对不足。

表 1 斯滕伯格创造力投资理论

核心资源	含义
智力	智力在创造过程中所起的关键作用就是综合、分析与实践，包含综合的、分析的和实践的三种智力形式
知识	知识是智力加工的材料，任何领域的创造行为都必须以知识为基础，包含正式知识和非正式知识
思维风格	思维风格作为心智自我管理方式，指个体在解决问题过程中运用智力能力与知识的倾向性
人格特征	人格在创造性思维转化为创造性行为中扮演着举足轻重的作用，高创造力个体具有五种人格特征：对模糊的容忍、坚持的毅力、超越自我的意愿、敢于冒险的勇气、自我认同的信心
动机	动机是引导个体行为的驱动力量或刺激因素，包含外在动机与内在动机
环境	环境的鼓励与重视，对创造力的发挥是必要的，包含物质、精神、制度环境等

2.2 基于机器学习的学生发展影响因素研究

机器学习（Machine Learning）是一组使计算机能够在没有人为编程干预的情况下进行自我学习的技术（Navamani et al., 2015），其大致可分为监督学习、无监督学习和强化学习（K. M. Lee et al., 2019b）。相较于传统方法，机器学习对数据的基础假设条件较为宽松，它通过预设的训练数据集对实际数据进行深度挖掘，并运用独立的验证数据集来校验模型的普适性（陈涛等人，2019）。同时，机器学习模型并不对响应变量与解释变量之间施加严格框架，从而能够更有效地挖掘非数值型数据中所蕴藏的信息，进而增强预测精度。此外，机器学习通过整合多种模型，提升整体性能，有效规避传统探索策略中因采用单一方法而产生的误差（钱浩祺等人，2021）。近年来，机器学习已被广泛应用于教育领域，有学者基于机器学习技术构建预测模型实现探究学生学习绩效的影响因素（万力勇，2022；Gaftandzhieva et al., 2022；Qazdar et al., 2019b），并依据探究结果实现学习行为诊断，助力教学精准干预（胡航等人，2021）；还有学者关注学生素养达成，利用机器学习技术分析创新素养、数字素养、信息素养、阅读素养等素养的影响因素，助力素养水平提升（王蕊等人，2023; Zhai et al., 2024; J. Li et al., 2024; Kong & Wang, 2024）。然而，现有研究聚焦于学习者学习绩效的影响因素研究，通过构建模型，探究学习者学习绩效影响因素，包括学习行为、学习状态、学习成绩、学习投入度等变量。鲜有研究关注学生创造力影响因素挖掘。因此，本研究基于机器学习算法，从大量因素中识别出学生创造力影响最大的因素，为创造力的培养方向提供证据指引。

3. 研究过程与研究方法

3.1 数据来源

本研究数据来源于国际学生评估项目 PISA（Program for International Student Assessment）2022 年数据库。PISA 由世界经济合作与发展组织 OECD（Organization for Economic Co-operation and Development）于 2000 年发起，每三年进行一次，是目前世界上最具影响力的国际学生学习评价项目之一。PISA2022 测评了 10291 名港澳学生的创造力分数。

3.2 变量

3.2.1 因变量

响应变量（即因变量）是来在 PISA2022 学生问卷的创造力分数，创造力分数范围为 0-60，

代表着对学生创造力水平的总体评价，创造力分数共有 10 个合理值，本研究使用第十个合理值作为因变量。

3.2.2 自变量

特征（即自变量）是从 PISA2022 学生问卷中提取，为了全面分析创造力影响因素，因此保留所有自变量。自变量个数为 83 个，智力方面含有 1 个特征变量，为 PROBSELF(自主学习问题)。知识方面含有 5 个特征变量，例如 PV1MATH (数学素养)。思维风格方面含有 9 个特征变量，例如 COOPAGR(合作)。个性方面含有 26 个特征变量，例如 PERSEVAGR(坚持不懈)与 CURIOAGR(好奇心)。动机方面含有 11 个特征变量，例如 CREATEFF(创造性思维自我效能感)，环境方面含有 31 个特征变量，例如 FAMSUP(家庭支持)与 TEACHSUP(教师支持)。

3.3 研究方法

3.3.1 随机森林

随机森林（Random Forest）是一种集成学习算法，广泛应用于分类和回归问题。其核心思想是通过构建多个决策树并综合它们的预测结果来提高模型的准确性和稳定性。随机森林通过自助抽样（bootstrap）方法从原始数据集中生成多个子数据集，每个子数据集用于训练一棵决策树。在构建每棵决策树时，不仅随机选择样本，还会在每个节点分裂时随机选择一部分特征进行分割。每棵树独立生长，不进行剪枝，最终通过投票或平均的方式整合所有树的预测结果来形成最终输出。

3.3.2 XGBoost

XGBoost（Extreme Gradient Boosting）是一种基于梯度提升树的高效机器学习算法，并在 Kaggle 等多项比赛中表现出色，广受学术界和工业界的认可。它是一种集成学习方法，通过结合多个基础分类器（通常是决策树）来实现 1+1>1 的效果。XGBoost 提供以下关键特性：卓越的预测性能、并行处理能力、对各种模型形式的支持、内置的模型验证和提前停止机制、强大的正则化功能以及处理稀疏数据和缺失值的能力。

3.3.3 LightGBM

LightGBM（Light Gradient Boosting Machine）是一种由微软开发的高效梯度提升（Gradient Boosting）框架，基于决策树算法。它旨在解决传统梯度提升框架（如 XGBoost）在处理大规模数据时的效率问题，通过多项优化技术显著提高了训练速度和内存使用效率。

3.3.4 Lasso

Lasso（Least Absolute Shrinkage and Selection Operator）是一种在统计学和机器学习中广泛使用的回归分析方法，其核心在于通过对系数进行压缩，以达到变量选择和复杂度调整的目的，从而提高模型的预测精度和解释能力。

3.4 数据准备

数据准备过程使用 SPSS（版本 27）与 jupyter notebook 进行，首先从 PISA2022 数据库下载学生问卷数据文件，其次从 SPSS 文件中提取港澳样本的数据。第三，对数据进行预处理，预处理规则如下：（1）删除创造力分数缺失的样本（2）对行进行处理，如果该行控制占比超过百分之 50 则删除该行（3）对列进行处理，如果该列空值占比超过百分之 20 则删除该列（4）使用随机森林进行缺失值的插补（插补时未使用因变量）。经过数据预处理后，剩余样本量为 9954。

3.5 模型选择

随机森林与 Lasso 模型使用 Python 中的 scikit-learn 包(Pedregosa et al.,2011)实现，XGBoost 与 LightGBM 则导入相对应 XGBoost 库与 LightGBM 库的库实现。

3.6 性能指标

均方根误差（RMSE）、平均绝对误差（MAE）和决定系数（R^2）被用作性能指标。

3.7 超参数调整

超参数调整是确定定义模型架构的参数的最优值的过程。在本研究中，使用随机搜索（Random Search）作为超参数调整方法。随机搜索是一种优化算法，用于在给定的搜索空间中寻找最优解。它通过随机选择候选解，并评估这些解的性能来工作。随机搜索不依赖于梯

度信息，因此适用于那些难以或无法使用梯度下降等基于梯度的方法来优化的问题。

3.8 交叉验证

交叉验证是 ML 领域中使用的一种增强技术，用于以更稳健的方式评估模型的性能，特别是在涉及超参数调整时。交叉验证确保性能评估不会因模型在调整过程中接触测试数据而产生偏差，交叉验证通过将数据集分割成若干个子集，然后使用不同的子集组合进行训练和验证，以评估模型在未知数据上的泛化能力。

在本研究中，三种模型均采用五折交叉验证对模型进行性能优化。

4. 研究结果

4.1 模型指标

本研究的三种模型预测性能如表 2 所示。

表 2 模型预测性能

模型名称	RMSE(标准差)	MAE (标准差)	决定系数 R^2(标准差)
XGBoost	7.61(1.31)	6.20(0.09)	0.46(0.01)
随机森林	7.83(1.35)	6.41(0.11)	0.42(0.02)
LightGBM	7.55(1.47)	6.14(0.12)	0.47(0.01)
Lasso	7.66(1.44)	6.23(0.10)	0.45(0.02)

从模型预测性能来看，LightGBM 在港澳学生数据集的表现优于随机森林回归模型、XGBoost 模型和 Lasso 回归模型。同时，各模型的评价指标在 5 重交叉验证中的标准差较小，说明模型稳定性较好。

4.2 特征重要性

由于 LightGBM 模型性能最优，因此选用该模型的特征重要性方法来对数据集中最为重要的 20 个特征变量进行筛选，结果如图 1 所示，同时，本研究对前十个变量对创造力的正负影响进行了探究，变量含义解释与正负相关性如表 3 所示。从表 3 可以看出，分类变量中，香港学生的平均数高于澳门学生，女生的平均数分数高于男生，从未留级学生的平均数高于留级学生。在连续变量中，只有 ICTWKEND(信息与通信技术活动的频率(周末))呈负相关关系，其余变量均与创造力呈正相关关系。

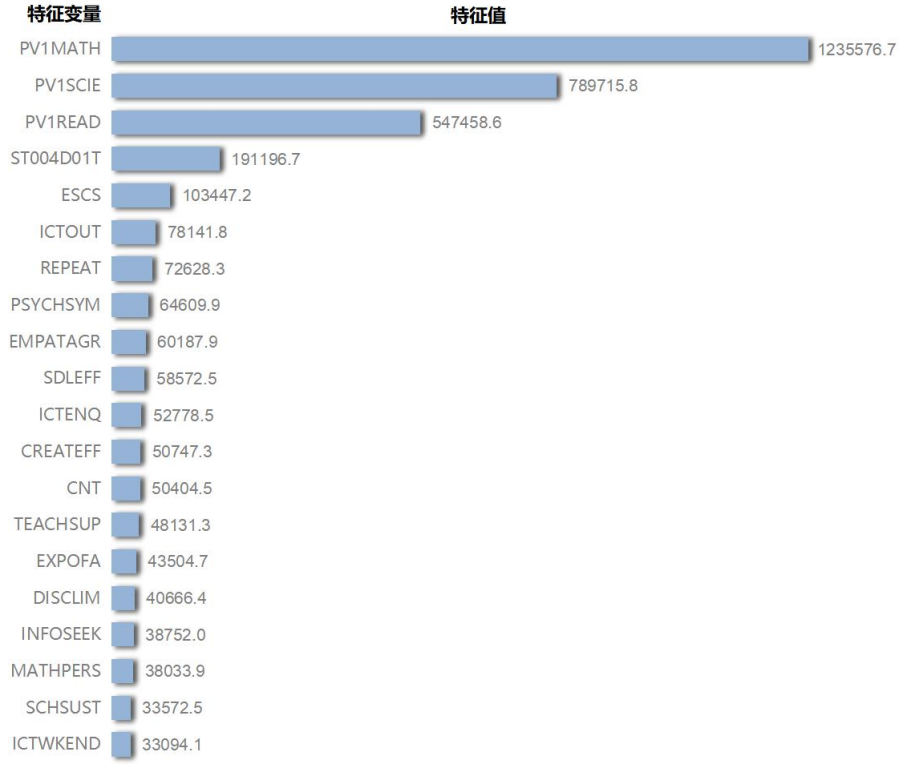

图 1 特征重要性直方图

表 3 特征变量含义与正负相关性

特征变量	特征内涵	正负相关性
PV1MATH	数学素养	正相关
PV1SCIE	科学素养	正相关
PV1READ	阅读素养	正相关
ST004D01T	性别	女生平均数高于男生平均数
ESCS	经济、社会和文化地位指数	正相关
ICTOUT	将信息与通信技术（ICT）用于课堂外的学校活动	正相关
REPEAT	留级	从未留级学生平均数高于留级学生平均数
PSYCHSYM	身心情况	正相关
EMPATAGR	共情能力	正相关
SDLEFF	自主学习自我效能感	正相关
ICTENQ	信息与通信技术在探究式学习活动中的使用	正相关
CREATEFF	创新自我效能感	正相关
CNT	城市	香港学生平均数高于澳门学生平均数
TEACHSUP	数学教师支持	正相关
EXPOFA	对符号和应用数学任务的接触	正相关
DISCLIM	数学学科的学科氛围	正相关
INFOSEEK	关于未来职业的信息搜索	正相关
MATHPERS	数学中的努力和坚持	正相关
SCHSUST	促进学习的学校活动	正相关
ICTWKEND	信息与通信技术活动的频率（周末）	负相关

5. 讨论

本研究从理论和实践的角度出发，对创造力影响因素研究提供一定参考。本研究主要使用 ML 算法探索港澳学生创造力影响因素，并基于创造力投资理论对影响因素进行归类，进一步对其进行特征重要性排序与正负相关性探究。研究发现，数学素养、科学素养、阅读素养、性别、经济、社会和文化地位指数与信息通信技术的应用等对创造力的解释作用占比较高，这些影响因素也对应着创造力投资理论中的各个核心资源，已有研究表明个体创造力发展受到外部及内部双重因素影响，如教育环境、社会支持、教学课程等外部因素，性别及年龄、知识经验、认知风格、个体情感等内部因素(王倩等人，2024；Al-Kumaim et al.,2021)，这与研究结果相一致。而在本研究中，数学素养、阅读素养与科学素养是对创造力最具有影响的三个特征，三者在创造力投资理论中均属于知识核心资源，在一定程度上表明创新要以知识为基础(初庆春等人，1999)，只有广阔的知识储备才能进一步进行创新创造活动。

已有研究认为数学天赋与创造力存在关联（Bicer et al.,2023），而在本研究中则发现环境核心资源中的数学教师支持、数学学科的学科氛围、对符号和应用数学任务的接触与个性核心资源中的数学中的努力和坚持也对创造力发展有着重要影响。因此，加强教育环境中数学教师的核心素养、营造更浓厚的数学氛围并提升数学任务的质量是有效提升学生创造力的重要途径。

课堂外的信息技术使用作为环境核心资源之一，主要为学生主动浏览与搜索相关兴趣信息，从而激发个体对未知探索欲望，成为创意构思起点，促使其在相关或跨界领域尝试创新实践。其次与知识核心资源相结合来看，网络中丰富多样的资源为学生提供了夯实的知识库，学生足不出户便可获得大量知识资源，为学生的知识拓展提供了良好的途径。

6. 研究局限性与未来研究方向

本研究只对港澳学生进行创造力影响因素探究，研究结果可能并不适用于其他地区的所有人群，结果的泛化性可能会受到影响。此外，本研究的影响因素主要针对学生个体层面，未涉及班级和学校层面。未来研究可以更加侧重于多样化的地区、国家，考虑纳入人口统计数据，例如民族、城乡等，同时考虑班级和学校方面对于学生创造力的影响，以此提高研究结果的深度和适用性。

参考文献

[1]陈涛,王鹏翀,林轩,裴欢昌,邢怡伦,罗捷... & 王亚. (2019). 机器学习在绘画测验预测 青少年依赖型人格偏离中的应用. 中国心理卫生杂志 (10), 769-773.

[2]初庆春，刘荣，汪克夷. (1999). 知识、创新和创造力. 大连理工大学学报(社会科学版) (02), 53-56.

[3]胡航,杜爽,梁佳柔 & 康忠琳.(2021).学习绩效预测模型构建:源于学习行为大数据分析.中国远程教育,(04),8-20+76.

[4]胡朗宁，吕立杰. (2024). 学生创造力评估：国际经验及未来展望. 外国教育研究, 51(4), 63-80.

[5]李硕豪. (2020). “拔尖计划”学生创造力发展影响因素实证研究. 中国高教研究 (04), 51-58.

[6]刘芝延 & 徐瑾劼. (2021).学生创造力培养的学校路径及挑战——基于 OECD 社会情感能力调查的发现与启示.人民教育,(23),75-78.

[7]万力勇.(2022).算法时代的教育预测及其研究范式转变.远程教育杂志,(03),35-44.

[8]王蕊,王捷,楚天舒.(2023).中国学生创新素养的相关因素及政策建议——基于大数据分析模型的实证研究. 全球教育展望, 52 (09): 3-21.

[9]王倩 & 唐一鹏. (2024). 家庭与学校支持如何影响中小学生的创造力?——基于社会与情感能力测评数据的国际比较分析. 华中师范大学学报(人文社会科学版) (02), 179-188.

[10]孙雍君.(2000).斯腾伯格创造力理论述评. 自然辩证法通讯(01),29-37+46.

[11]钱浩祺,龚嫣然 & 吴力波.(2021).更精确的因果效应识别:基于机器学习的视角.计量经济学报,(04),867-891.

[12]徐瑾劼 & 安德烈亚斯 •施莱歇尔. (2024). 全球 15 岁学生创造力发展水平及培养路径——基于 PISA2022 学生创造性思维表现结果的洞察. 人民教育 (12), 74-78.

[13]Al-Kumaim, N. H., Alhazmi, A. K., Mohammed, F., Gazem, N. A., Shabbir, M. S., & Fazea, Y. (2021). Exploring the impact of the COVID-19 pandemic on university students' learning life: An integrated conceptual motivational model for sustainable and healthy online learning. Sustainability, 13(5), 2546.

[14]Bicer, A., Bicer, A., Capraro, M., & Lee, Y. (2023). Mathematical Connection is at the Heart of Mathematical Creativity. Creativity. Theories–Research-Applications, 10(1-2), 17-40.

[15]Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140.

[16]Gaftandzhieva, S., Talukder, A., Gohain, N., Hussain, S., Theodorou, P., Salal, Y. K., & Doneva, R. (2022). Exploring online activities to predict the final grade of student. Mathematics, 10(20), 3758.

[17]Kong, H., & Wang, X. (2024). Exploring the influential factors and improvement strategies for digital information literacy among the elderly: An analysis based on integrated learning algorithms. Digital Health, 10, 20552076241286635.

[18]Lee, K. M., Yoo, J., Kim, S. W., Lee, J. H., & Hong, J. (2019). Autonomic machine learning platform. International Journal of Information Management, 49, 491-501.

[19]Li, J., Wang, J., & Xue, E. (2024). Applying a Support Vector Machine (SVM-RFE) Learning Approach to Investigate Students' Scientific Literacy Development: Evidence from Asia, Europe, and South America. Journal of Intelligence, 12(11), 111.

[20]Navamani, J. M. A., & Kannammal, A. (2015). Predicting performance of schools by applying data mining techniques on public examination results. Research Journal of Applied Sciences, Engineering and Technology, 9(4), 262-271.

[21]Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830.
[22]Qazdar, A., Er-Raha, B., Cherkaoui, C., & Mammass, D. (2019b). A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco. Education and Information Technologies, 24(6), 3577–3589.
[23]Sternberg, R. J., & Karami, S. (2022). An 8P theoretical framework for understanding creativity and theories of creativity. The Journal of Creative Behavior, 56(1), 55-78.
[24]Xu, J. (2024). Enhancing Student Creativity in Chinese Universities: The Role of Teachers' Spiritual Leadership and the Mediating Effects of Positive Psychological Capital and Sense of Self-Esteem. Thinking Skills and Creativity, 101567.
[25]Zhai, X., Yuan, W., Liu, T., & Wang, Q. (2024). Machine learning investigation of optimal psychoemotional well-being factors for students' reading literacy. Education and Information Technologies, 29(14), 18257–18285.