数据挖掘在普通高校大学生体质健康数据中的应用

鹿瀚升
青岛黄海学院　山东青岛　266427

摘　要：本研究通过数据挖掘技术对 H 大学 2020-2023 年学生体质健康数据进行分析，揭示影响大学生体质健康的关键因素。本文采用 Python 进行数据清洗、预处理及数理统计分析，并利用 matplotlib 实现数据的可视化，结果表明通过合理的数据挖掘方法可以有效识别和分析体质健康数据中的潜在模式和关联，为提高大学生体质健康水平和改进高校体育教学提供了科学依据。

关键词：数据挖掘；体质健康；Python；matplotlib；大学生

引言 : 高校学生体质健康问题备受关注，体质健康数据的积累为研究大学生体质健康提供了宝贵的资源。然而，仅依靠传统的数据分析方法难以从海量数据中挖掘出有价值的信息，因此，数据挖掘技术的应用显得尤为重要。

1 数据挖掘概述

1.1 数据挖掘的定义与发展

数据挖掘是一种从大量数据中提取出潜在、有用信息和知识的技术，通过分析和处理海量数据来揭示数据之间的隐含模式和规律，从而为决策提供支持。近年来，数据挖掘技术涌现出一系列新的方法和工具，如：神经网络、决策树、支持向量机、聚类分析等，这些技术极大地提升了数据分析的效率和精度。

在教育领域，高校学生体质健康数据的分析中数据挖掘技术具有广泛的应用前景。数据挖掘过程通常包括数据收集、数据清洗、数据转换、数据建模和结果评估五个步骤。通过对学生体质健康数据的挖掘可以识别出影响体质健康的关键因素（如运动频率、饮食习惯、生活作息等）[1]。

1.2 数据挖掘的主要方法

数据挖掘主要方法包括分类、聚类、关联规则挖掘、回归分析、序列模式挖掘、异常检测等，表 1 是数据挖掘方法的特点和应用场景。

表 1 数据挖掘方法的特点和应用场景

数据挖掘方法	特点	应用场景
分类	将数据项分配到预定义类别	健康风险预测，疾病倾向分析
聚类	划分数据项为多个相似组	识别相似健康特征的学生群体
关联规则挖掘	发现数据项间的关联模式	分析健康行为和健康状况之间的关系
回归分析	研究因变量与自变量间的关系	健康指标预测，如体重、血压等
序列模式挖掘	发现时间序列中的模式	研究健康行为随时间变化的规律
异常检测	识别显著不同的异常数据	发现和干预潜在的健康问题

本研究采用 Python 作为主要的编程工具，通过其丰富的数据处理库（如 pandas、numpy）和机器学习库（如 scikit-learn、tensorflow），实现数据清洗、预处理、建模和分析。利用 matplotlib 和 seaborn 等可视化库能够直观展示数据分析结果，帮助识别和理解数据中的模式和规律。

2 数据处理与预处理

2.1 数据收集与选择

本研究的数据来源于 H 大学 2018-2021 年间学生体质健康测试数据，包括体重、身高、BMI（体质指数）、肺活量、坐位体前屈、50 米跑、800/1000 米跑等多个体质健康指标。数据共涉及 10，000 名学生，涵盖男女生各 50%，其中男生 5，000 名，女生 5，000 名，数据记录完整、格式统一，具备良好的分析基础。

在数据选择过程中需考虑数据的完整性和一致性。对缺失数据和异常数据进行严格筛查和处理，确保所选数据具备良好的分析基础；对于缺失值较少的数据，采用插值法进行补全；对于异常值，通过对比历史数据和现场复核来确认其合理性，必要时进行剔除。

2.2 数据清洗与预处理

在本研究中，对 H 大学 2018–2021 年间的学生体质健康数据进行详细的清洗与预处理，保证后续数据挖掘的顺利进行。数据清洗与预处理包括数据去重、缺失值处理、异常值处理、数据转换及标准化等环节。本研究为后续的数据挖掘提供了高质量的数据基础，确保了分析结果的科学性和有效性 [2]。通过这些步骤，可以挖掘大学生体质健康数据中的潜在模式，为提升学生体质健康水平和改进高校体育教学提供科学依据。

2.3 数据转换与标准化

数据转换是对原始数据进行格式变换、特征工程的操作，而标准化是将数据进行归一化处理，消除不同变量之间的量纲差异，从而提高模型的性能和分析结果的可靠性。

数据转换是指将原始数据转换为更适合分析的格式。例如，对于身高和体重这两个变量，通过计算其比例可以得到一个新的变量 BMI（体质指数），其公式为：BMI = BMI= 体重（公斤）÷ 身高（米）^2。通过这种转换可以将两个相关变量合并为一个新的特征，从而简化分析过程。

特征工程是通过对原始数据进行加工生成更具代表性和预测力的特征。特征工程包括但不限于特征选择、特征组合和特征提取等操作。在本研究中，特征选择的过程包括从大量的体质健康指标中筛选出与体质健康密切相关的变量，如体重、身高、肺活量、坐位体前屈、50 米跑和 800/1000 米跑等。特征组合将多个变量组合生成新的特征，例如，通过组合不同的运动指标可以生成反映学生综合运动能力的综合指标。数据转换与标准化还包括对时间序列数据的处理，在本研究中，学生体质健康数据的收集具有时间连续性，通过对时间序列数据进行差分处理和移动平均处理可以消除时间序列中的趋势和周期性，从而更准确地捕捉数据中的模式和规律 [3]。这些处理方法在时间序列数据分析中具有重要意义，有助于揭示学生体质健康随时间变化的趋势和影响因素。

3 数据挖掘技术应用与分析

3.1 使用 Python 进行数理统计分析

本研究采用 Python 进行数理统计分析，利用其强大的数据处理和分析库，如：pandas、numpy 和 scipy 能够确保分析过程的高效性和准确性。数据分析包括描述性统计分析、假设检验和相关分析等内容。为了直观展示数据的统计特征和分布情况，采用 matplotlib 和 seaborn 等可视化工具绘制统计图表。表 3 是主要体质健康指标的描述性统计量和相关系数，数理统计分析不仅揭示了数据中的显著模式和规律，还为模型构建和预测提供了坚实的理论基础。在本研究中，通过使用 Python 进行高效的数理统计分析能够准确捕捉数据中的关键特征和模式，为提高大学生体质健康水平和改进高校体育教学提供科学依据。

表 2 主要体质健康指标的描述性统计量和相关系数

指标　名称	均值	标准差	中位数	最小值	最大值	BMI 相关系数	肺活量相关系数	50 米跑相关系数
体重（kg）	65.2	12.5	63.0	40.0	110.0	0.87	0.32	–0.15
身高（cm）	172.5	8.7	171.0	150.0	200.0	0.85	0.28	–0.12
BMI	22.0	21.5	16.0	35.0	1.00	0.30	–0.18	–0.18
肺活量（ml）	3500	800	3400	2000	6000	0.30	1.00	–0.45
坐位体前屈（cm）	5.2	6.3	5.0	–10.0	25.0	–0.18	–0.15	–0.45
50 米跑（s）	8.5	1.2	8.4	6.0	12.0	–0.15	–0.45	1.00
800/1000 米跑（s）	230.5	45.6	45.6	225.0	150.0	360.0	–0.12	–0.4

BMI（身体质量指数）是通过测量身高和体重后，使用公式体重（kg）除以身高（m^2）计算得出的，为了了解学生的 BMI 对其他身体素质的影响，假设当 BMI 处于正常范围内时，相比其他 BMI 状态，学生在其他身体素质项目中的表现最佳 [4]。基于这一假设，本研究将 BMI 作为自变量，以其他身体素质项目的成绩为因变量，并按性别对学生的 BMI 与其他身体素质成绩之间的关系进行探讨。

3.2 关联规则挖掘与频繁模式分析

关联规则挖掘与频繁模式是从大量数据中发现数据项之间的有趣关系和潜在模式。本研究中，通过对 H 大学 2020–2023 年学生体质健康数据进行关联规则挖掘和频繁模式分析，揭示不同体质健康指标之间的关联关系，为提高大

学生体质健康水平提供科学依据。频繁项集是指在数据集中频繁出现的项集，而关联规则是描述一个频繁项集中项之间的条件依赖关系。常用的算法包括 Apriori 算法和 FP-Growth 算法。在本研究中，采用 Apriori 算法进行频繁项集挖掘和关联规则生成。为了进行关联规则挖掘，首先需要将连续变量离散化，例如，将 BMI 分为”偏瘦”、”正常”、”超重”三个类别，将肺活量分为”低”、”中”、”高”三个类别。通过这种离散化处理，可以更直观地揭示不同体质健康指标之间的关联关系 [5]。

这些关联规则和频繁模式能够指导高校在制定体育教学和健康干预措施时，重点关注这些关键因素。例如，通过加强体重管理和肺活量训练来改善学生的体质健康水平；通过对偏瘦学生进行营养指导和适当的体能训练也可以提高他们的肺活量和整体健康水平。在本研究中，应用关联规则挖掘技术发现体重、肺活量和跑步成绩等指标之间的显著关联关系，为改进高校体育教学和学生健康管理提供了重要参考。

3.3 数据可视化与结果展示

本研究采用 Python 的 matplotlib 和 seaborn 库对 H 大学 2020-2023 年学生体质健康数据进行可视化分析，揭示出数据中的关键信息和潜在关联。使用 matplotlib 和 seaborn 可以生成多种类型的图表，如：箱线图、散点图、热图、柱状图等 [6]。通过对不同类型图表的综合分析可以更全面地理解学生体质健康状况及其影响因素。通过数据可视化，不仅可以揭示数据中的模式和规律，还能有效地传达分析结果，便于相关人员理解和利用。本研究通过对学生体质健康数据的可视化分析发现许多重要的关联关系和潜在问题，为进一步的健康干预和教育政策制定提供了科学依据。数据可视化在数据分析中发挥着关键作用，通过将抽象的数据转换为直观的图表和图形，帮助识别和理解数据中的关键模式和关系。

结论

通过合理应用数据挖掘技术可以从大量的体质健康数据中提取出有价值的信息，本研究的研究结果对提高大学生体质健康水平和改进高校体育教学具有重要意义。未来的研究可以进一步扩展数据样本，采用更多的数据挖掘技术，如：机器学习和深度学习可以获得更全面和深入的分析结果。结合实际的健康干预措施和教育政策也可以更有效地提升学生的体质健康水平，实现健康教育的目标。

参考文献：

[1] 陈裕臻 . 基于聚类分析的大学生体质健康测试成绩分析与研究 [J]. 西安文理学院学报（自然科学版）， 2023，26（02）: 1-6.

[2] 张坤 . 基于数据挖掘的大学生体质健康评价优化算法及平台设计研究 [D]. 山东大学， 2022.

[3] 马丽亚 . 基于 Apriori 算法的大学生体质健康指标联动与对策研究 [J]. 福建体育科技， 2021， 40（05）: 24-29.

[4] 江婷 . 基于数据挖掘技术的大学生体质健康分析 [D]. 安徽师范大学， 2021.

[5] 张崇林，王世香，王卉，等 . 基于关联规则数据挖掘的大学生体育锻炼行为阶段体质健康知识发现 [J]. 井冈山大学学报（自然科学版）， 2020， 41（03）: 80-84.

[6] 张满意 . 健康中国背景下苏北普通高校大学生体质健康促进路径研究 [D]. 中国矿业大学，2021.