目录

第 1 章　绪论

1.1　研究背景与意义

1.1.1　研究背景

数据分析和关联规则是数学统计领域里的一个备受关注的研究方向，是知识发现的一个重要课题。我们的生活中每天都会积累下来大量的数据。我们打电话时会产生通话记录我们购物时会产生消费记录，我们坐飞机出行时会产生出行记录；我们网上冲浪时会产生上网记录；我们每天的活动产生的数据中蕴含着一些未知的信息，合理的利用这些数据从中分析出有用的信息，这些数据将会成为一个宝藏。数据分析和关联规则正是我们用来利用这些数据的一个得力助手。在这样环境里，数据分析和关联规则的研究有了快速的进步，令它被各行业所关注。

数据分析和关联规则应用领域极为广泛。在金融行业，数据分析和关联规则可以用于信贷风险评估和顾客的信誉度的评估，还有洗黑钱行为和其他类型的金融犯罪侦破等。在零售行业，数据分析和关联规则可以用于商品的市场定价分析和客户的购物习惯分析等。在电信行业，数据分析和关联规则可以应用于电信数据的多维分析，以及用户的消费行为分析等。在生物学行业，的数据分析和关联规则可以用于基因变异的预测和蛋白质路径的分析等。

数据分析和关联规则技术，它的应用领域越来越宽，它的应用也渐渐升深入。在教育领域，同样存在海量的数据来等待挖掘。在学生的日常生活中也会产生各种类型的数据，如考试产生的成绩、消费产生的消费记录、社会实践产生的实践报告等。如果能利用数据分析和关联规则从这些数据中挖掘出潜在的、我们所需要的知识，来为学生的学习和生活做出针对性的指引，那么数据分析和关联规则技术就能为教育领域做出重大贡献。

学生成绩是在教育领域的一个重要数据，但是目前对于这一数据的利用还停留在简单的统计分析的初步阶段。本论文利用关联规则技术对学生的成绩书进行深层次的分析，得出的结果能提供给学生、老师以及学校的管理人员一些合理的建议。学生能够依据这些规则了解自己的学习情况，为自己后阶段如何学习提供参考；老师可以依据这些规则，了解学生的学习情况，为学生出谋划策；学校管理人员能够依据这些规则，发现问题，为后阶段的教学安排调整提供依据。

1.1.2　研究意义

近年来我国高等教育事业发展的重点已从规模扩张转移到了质量的提升上来, 质量是高等学校的生命线。学习成绩是反映大学生积极和消极学习心理的重要的综合指标，了解当前大学生在这个问题上的现状，对高校深化教学改革，提高教学质量具有重要的现实意义。[1]

在信息化不断发展的今天，如果我们依然仅仅依靠教师对学生成绩数据的简单人工分析来进行科学教育已然跟不上当前时代的步伐了。我们需要学会利用计算机前沿技术对学生成绩管理进行信息化的指导。通过科学的数据分析找到影响高校课程成绩的潜藏因素，[2] 例如：个人成长家庭的好坏是否会对后续课程产生影响，以及学生的性别、入学成绩等客观因素对成绩的影响等，这些都需要我们对成绩数据做更深层次的分析，从而得出更多的潜藏信息，为教育管理者进行课程的优化设置、提高教学质量提供参考和依据，这就需要我们找到一种创新且更有效的成绩分析方法。

因此为了提高学生各门课程的平时成绩，学生课程成绩的分析预测是解决学生课程成绩的关键因素。[3] 通过对课程成绩数据分析处理，能让我们发现平时不易被察觉的有价值的信息，从而对学生进行课程预警，指导学生自主学习。

由于在影响学生成绩的因素除了课程间的关联性外，还有该学生、教师等因素对成绩的影响，致使目前对学生成绩的分析预警不准确问题，较难找到一种比较可行的方案，为此本文针对上述现状及问题实现一种将关联规则与决策树算法相结合的方法，来完成学生成绩的有效分析，并将该算法应用在学生成绩预警功能上，实现对[4] 学生在校成绩的合理预测，以及根据预测结果做出准确预警，帮助学生及时准确的掌握自己的课程状态，增强学习意识，为良好的教学质量和效果提供保证。

1.2　国内外相关工作研究进展

国际信息化科学技术的快速发展，数据分析和关联规则技术在人工智能等科学技术的不断突破、融合的基础上逐渐应运而生。

数据分析和关联规则的出现是一个逐渐演变的过程。数据分析和关联规则概念最早于八十年代晚期提出的，这个概念刚一提出，国际各大期刊、会议就把他作为核心议题刊登发表，在国

际上引起了非常高的关注。短短几年里，国外发达国家在数据分析和关联规则技术进行全面研究，在数据分析和关联规则领域飞速发展。并且许多国家的各个行业均已得到数据挖掘的技术支持，如把他运用在商业领域[5]，研究消费者消费规律，对物品进行捆绑销售，刺激消费。近年来，数据分析和关联规则技术在高等学校的应用也越来越受到人们的重视，例如预测大学入学比例、预测学生毕业状况、研究学生学习经历、帮助学生选课等等。

我国在 90 年代中期才开始逐步从事数据分析和关联规则方面的研究工作，远远晚于西方等发达国家。因此，国内研究数据分析和关联规则的开端比较良好，但是国内科研基础比较薄弱，底子较差，科研力量与国外相比也比较薄弱，在短时间内难以形成整体的力量，而且国内数据分析和关联规则的研究重点在于其理论知识，不注重其在应用方面的发展。

国内数据分析和关联规则的理论与应用研究存在着很多的局限性，理论研究的局限性的原因在于从事数据分析和关联规则技术的科研工作者们大多数是高等学校的计算机学科老师，从而导致科研工作者学习背景比较单一，缺乏国外科研专家的交叉学科背景，而数据分析和关联规则技术却是一个需要多方面知识来学习的深层次学科。[6] 数据分析和关联规则技术应用研究的局限性表现在数据分析和关联规则技术发展最好的是互联网行业、商业等领域，教育行业领域该技术才刚刚起步。

近年来全国很多科研单位，还有空军第三研究所、清华大学、海军装备论证中心以及中科院计算技术研究所等许多学校竞相开展知识发现的基础理论及其应用研究。其中，许多高校对关联规则开发算法进行了改造和优化处理，如浙江大学、中国科技大学、复旦大学、吉林大学、华中科技大学及中科院数学研究构化数据的知识发现进行了研究和讨论。同时，国内高校对数据分析和关联规则技术在成绩分析方面的研究与应用也渐渐增多，董欢、刘志妩针对学生的期末考试成绩，利用决策树算法建立了分析预测模型，能够通过模型完成对成绩较为准确的预测分析；丁勇、武玉艳通过实际成绩数据进行分类预测实验，证明了决策树有较高的预测准确率；付希、刘美玲等研究了数据分析和关联规则在学生的成绩等级评价中的应用，通过对聚类分析结果的深入分析发现了学生的成绩等级评价方式中存在的问题，根据聚类结果对学生成绩评价进行动态的层次划分的评价方法，使等级评价的结果更客观、可信。申义彩、杨枫等将关联规则 Apriori 算法应用到学生等级考试成绩分析中，得出的结果能够帮助学生提离考试通过率；李昊、周振华通过对不同学生课程的成绩进行关联规则分析，得出了学业的预警信息；设计并实现了基于关联规则算法的成绩预警系统。[7] 邝继红、孙月昊将关联规则数据分析和关联规则方法应用于教务信息管理系统中。

这些研究都是使用数据分析和关联规则中的一种算法从单一的角度对成绩数据进行分析，得到的分析结果不够全面，例如决策树的优势在于依靠信息增益或信息增益率排序节点，找出影响分类的关键因素，还可以清晰的显示出各项的重要程度，但无法得知各项间的关联也不能得出满足规则的数据项占总体的具体比例是多少；而关联规则算法则是可得到各项间的关联和满足规则的样本占总体的比例而不能得到各属性的重要程度排序。

1.3 本文主要研究思路

1.3.1 研究思路和框架

论文主要的研究内容是：针对目前在学生成绩分析研究方面的欠缺，及具体应用方面的不足。通过建立相应的数据集实现对学生课程成绩的数据处理和存储，并为后续的成绩分析提供可靠的数据支持。然后针对数据原始属性涵盖信息量不足的问题，通过关联规则算法进行深入的学习研究，提高了分析预测结果的覆盖面，同时在成绩分析时通过个人家庭因素进一步挖掘分析，实现具体学生的课程成绩的准确分析预测，弥补了目前大多数系统在个体学生成绩预警方面不足的问题，从而为高校教育的各个方面提供服务技术和支持。[8]

本论文的主要工作将针对以下三个方面进行研究论述：

（1）在对学生成绩进行数据挖掘分析前数据仓库的设计和建立，以及数据的处理过程。首先要对材工专业的研究生个人成绩和个人家庭背景进行分析和研究，建立面向成绩分析主题需求的数据集。然后，对我们收集的原始数据进行数据的清洗、抽取和转换等处理流程，然后将获得的基础数据加载到新建的主题数据集中。在实际设计处理过程中。。

（2）利用关联规则建立分析模型，对学生课程成绩分析进行了全面分析，既可以有效地挖掘出个人因素的关联性以及隐藏在成绩数据中潜在的有价值的信息，又发现了该方式很难分析到课程先后和学生家庭等个体因素对学生成绩的影响。

（3）利用数据分析的方法对数据集进行描述性分析和相关性分析，学生成绩在性别、家庭状况等先天性条件下的区别，在不同性别、不同家庭情况、不同压力水平下成绩的变化。沉迷于网络游戏的同学，经常锻炼的同学往往学习成绩变化。并且发现平均每天上网时长与除日常消费

意外额外开销之间的关系；平均每周锻炼时间与日常额外开销、对学习生活的满意程度之间的联系；压力感知、心理资本、锻炼、消费都与学习成绩之间的联系。通过回归分析得出学习成绩分别与平均每周锻炼时间、压力感知、心理资本的表达式。[9]

1.3.2 研究方法

第一章绪论。经过相关方面的深入学习研究，对本课题的研究背景、研究意义进行具体阐述，然后对本课题国内外的研究现状做了简单分析，最后提出本论文的研究内容和科研工作。

第二章关联规则技术。先对数据集的相关概念及知识做了简单介绍。然后对数据分析常用的关联规则算法进行了详细学习介绍。

第三章数据分析。对学生成绩数据集的特点进行了分析描述，对数据集进行相关性分析，描述性分析，回归范围内西等数据分析方法。

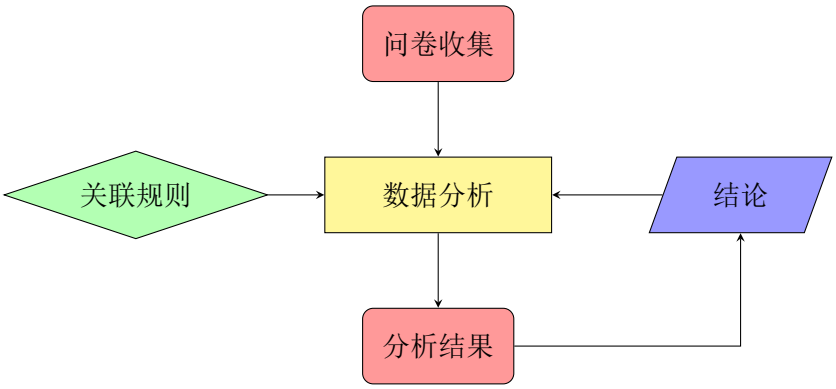
第四章总结出成绩分析的结论并且制定出有效的解决办法，提出具有参考性的建议。

1.3.3 研究路线

问卷调查的研究一般基于两个方面，一是定性分析，另一个是定量分析。定性分析是一种探索性调研方法，目的是对问题定位或启动提供比较深层的理解和认识，或利用定性分析来定义问题或寻找处理问题的途径。问卷定量分析是首先要对问卷数量化，然后利用量化的数据资料进行分析。问卷的定量分析又可以根据分析的难易程度分为简单的定量分析和复杂的定量分析。[10]

简单的定量分析是对问卷结果做出一些简单的分析，比如利用数据的百分比、平均数、频数来进行分析。简单分析常用于单变量和双变量的分析，但问卷调查的项目是多种多样的，单单用两个变量无法满足分析的要求、完成分析的目标，此时就要用到复杂的定量分析。复杂的定量分析有两种形式：多元分析和正交设计分析。这里主要用到多元正交分析，通过对观测数据的分析，由表及里来研究多个变量之间相互依赖的规律性，或者根据实际问题的需要对研究对象做出某种评价、分类和判别。通常有聚类分析、因子分析、主成分分析三种形式。[11]

图 1 研究生成绩分析流程



第 2 章　关联规则

2.1　关联规则理论

2.1.1 关联规则介绍

关联规则挖掘是数据挖掘的研究中一个重要的方面。下面介绍管理规则的一些基本定义。

(1) 项和项集

设存在集合，I 中的元素各不相同，对于集合 I 中的每个元素 i 我们叫做一个项，项目的集合 I 叫做项集。m为项集的长度，长度为 k 的项集称为 k 项集。

(2) 事务

对于项集 I，他的每一个子集我们称之为事务。我们通常用 TID 一个事务, 务都有一个唯一的事务标识。我们一般用 D 来表示事务数据库的所有事务 D 示事务数据库中事务的数量[12]

(3）持度与最小支持度

假设存在项集 X，我们定 count(X) 事务集 D 中 X 事务所出现的次数，那么其支持度的计算方式为:

$$Support(X) = \frac{count(X)}{|D|} \tag{1}$$

最小支持度是由用户自己定义的，它的实际含义是某个事务出现次数在总事务数中所占有的比重，当某个项的支持度不大于它时，代表了这个项不是我们关心的。最小支持度是关联规则的一个重要度量标准，我们把实际的支持度不小于最小支持度的项叫做频繁项。

（4）关联规则

关联规则可以表示成一个蕴含式 $X \Rightarrow Y$，对于 X,Y 有 $X \subset I$,$Y \subset I$，且 X 交 Y 为空。这个蕴含式表达了当 X 在一个事务中出现时，Y 也会以一定概率同时出现在这个事务中。判定一个关联规则是否为我们满足我们的要求时，通常用支持度和置信度这两个标准来衡量

(5) 置信度与最小置信度

置信度可以定义为事务 X 和 Y 的一起出现的次数与事务 X 的独自出现次数的比值。具体公式为:

$$confidence(X \Rightarrow Y) = \frac{support\,(X \Rightarrow Y)}{support\,(X)} \tag{2}$$

置信度反映了在一个记录中，当事务中出现 X 时，一起出现 Y 的概率。最小置信度也是由用户自己定义的，他和最小支持度是用来度量一个关联规则是否准确的标准。通常，只有实际的支持度和置信度分别大于其阀值时，这样的规则才是我们所在意的。[13] 对于关联规则 Y，如果他满足这样的关联规则叫做强关联规则

2.1.2 关联规则分析过程

关联规则挖掘的流程主要分为两步:

第一步是寻找频繁项集。在这个过程中，我们要找出所有项集 X，找出所有的项集之后，需要判断这些项集必须满足 $confidence(X) > SUPmin$，找出的满足要求的项集叫做频繁项集。

第二步是生成关联规则。找出满足要求的频繁项集后，逐个计算各频繁项的置信度，找出满足的项，这些满足要求的关联规则就是我们需要的强关联规则。

寻找频繁项集所占用的时间在关联规则挖掘的整个过程中占比最重。原因在于生成频繁项集的过程中会产生大量的候选项，我们需要判断这些候选项是否满足要求，在计算这些候选项的支持度时需要对数据库扫描，占用的时间较多。而我们最终得到的满足最小支持度的频繁项会少很多，生成关联规则的时间开销也会小很多。另外，生成关联规则的过程中，不再需要对事务数据库进行遍历，而寻找频繁项时多次遍历数据库会占用大量时间。

2.1.3 关联规则分类

关联规则是受到人们广泛关注的一个研究热点，也是本文研究的一个重要内容。从最早的超市购物篮案例中得出啤酒尿布这一条关联规则以后，关联规则挖掘就吸引了人们的眼球。后来，沃尔玛公司甚至从一个女孩的购物记录中挖掘出一个女孩怀孕的信息，这足以证明关联规则挖掘的强大。[14] 本小节对关联规则进行了分类，从规则维度的角度，我们把它分成下面几类：

（(1) 数值型和布尔型关联规则

依据项目类别的差别，我们可以把关联规则分成数值型和布尔型。例如最经典的超市购物篮案例中得出的尿布二啤酒这条关联规则就是布尔型关联规则。数值型关联规则举个简单的例子：电脑＝价格＜ 1000。我们处理数值型关联规则时，首先要对原数据进行数据转换，使其变成便于算法挖掘的形式，才能得出正确的数值型关联规则。

(2) 多层和单层关联规则

依据关联规则蕴含式的前项与后项所在的层次的不同，我们可以把关联规则分为多层与单层关联规则。联想＞华硕两个项处在同一层次，是一个单层规则。台式〉联想两项处在不同的层次，这样的规则叫做多层关联规则。

(3) 多维和单维关联规则

依据关联规则蕴含式的前项与后项的维度的不同，我们可以把关联规则分为多维和单维关联规则。当蕴含式的前项和后项均都为只包含一个元素时，我们把它叫做单维关联规则。例如，联想＞华硕。当蕴含式的前项和后项均都为只包含一个元素时，我们把它叫做多维关联规则。如苹果，梨＞香蕉，西瓜。[15]

(4) 特殊类型关联规则

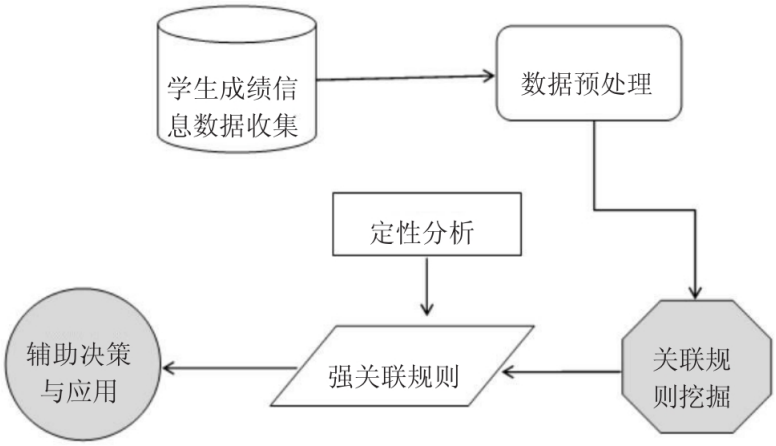


图 2 关联规则分析图

当关联规则蕴含式的前项或者后项中出现了伴随条件，我们把这样的规则叫做特殊关联规则。例如: 苹果＞梨，表示用户不购买苹果时，购买梨的概率会变得更大。

2.2 关联规则应用

2.2.1 寻找频繁项目集

通过问卷的形式接受各类学生的问卷，将问卷以文件的方式储存在数据集中，由于不同的数据和问题具有不同的关联，进行储存之后进行后续分析。

在对学生成绩数据进行关联规则分析时，这里采用了 Apriori 算法来寻找全部的频繁项目集。Apriori 算法是一种重要的关联规则挖掘算法，它使用了一种被称为逐层搜索的迭代算法，k-项集用于搜索（k+1）-项集。首先需要扫描事物数据库，累积每个项的计数，[16] 然后收集满足最小支持度的项，从而找出频繁 1-项目集的集合 L1。L1 用于寻找频繁 2-项目集的集合 L2，而 L2 用于寻找频繁 3-项目集的集合 L3，如此下去，直至不能找到频繁 k-项目集 Lk 为止。

运用频繁 k-项集用于搜索（k+1）-项集是 Apriori 算法的核心，该步骤分为连接步和剪枝步：

(1) 连接步骤：为了寻找 Lk，在 k（k ＞ 1）次扫描数据库时，通过 Lk-1 与自身连接产生候选 k-项集的集合 Ck。

(2) 剪枝步骤：由于 Ck 是 Lk 的超集，即 Ck 的成员可能是也可能不是频繁的。需要扫描全部的事务数据库，确定 Ck 中每个候选的计数，判断是否大于或者等于最小支持度计数，如果是，那么便认为该候选是频繁的。为了压缩 Ck，可以运用 Apriori 性质：任何一个频繁项集的全部非空子集也一定是频繁的，若某个候选的非空子集不是频繁的，那么该候选项集肯定也不是频繁的，从而可以将其从 Ck 中删去。[17]

Apriori 算法描述如下:

2.2.2 生成强关联规则

对于上面得到的每个频繁项目集 L，生成强关联规则的步骤如下：

(1) 生成 L 的所有非空子集；

(2) 对于 L 的每个非空子集 S，令 R=L-S。

如果有

$$\frac{Support\,(S \cup R)}{Support\,(S)} \geq Min \sim Confidence \tag{3}$$

即 $S \Rightarrow R$ 满足最小置信度阈值，那么输出关联规则 $S \Rightarrow R$。又因为这个规则是从频繁项目集 L 中生成的，因此一定满足最小支持度阈值，所以这个规则为强关联规则。根据上面的两个步骤，就可以得出事物数据库 D 的全部强关联规则。[18]

输入：数据库 D；最小支持度 min_Support
输出：D 中的频繁项目集 L
方法：
L_1=find_frequent_1−itemsets(D);;
for(k=2;$L_{k-1}\neq\Phi$;k++){
C_k=apriori_gen(L_{k-1},min_Support)
for each transaction t ∈ D{
C_t=subset(C_k,t);
for each candidate c ∈ C_t
c.count++;
}
L_k={c ∈ C_k|c.count ≥ min_Support}
}
return L=$U_k L_k$

图 3 算法描述及编程流程

2.2.3 挖掘关联规则

选择恰当的关联规则挖掘算法对数据进行分析处理。这里采用关联规则 Apriori 算法对离散化后的学生成绩数据信息进行挖掘。设定最小支持度为 25%、最小置信度为 60%。运行关联规则 Apriori 算法程序后，得到的部分实验结果如表 1 所示。[19]

通过算法的求解可得各问题的关联规则：

表 1 部分问题关联规则

问卷问题编号	关联规则	支持度 (%)	置信度 (%)
7	A>=B	43	46
8	B<=G	56	70
9	C<H	55	75
10	H>D	42	59
…	…	…	…

变量合并：

1. 将问卷中问题 1,3-6,9 的数据作处理作为“社交活动”变量，该变量数值越大，表示社交活动参与度越高；

2. 将问卷中问题 10-16 的数据作处理合并为“网络依赖”变量，该变量数值越大，表示对网络的依赖度越高；

3. 将问卷中问题 19-22,26 的数据作为“学习热情”变量，该变量数值越大，表示学习热情越高；

4. 将问卷中问题 28-32,34 的数据作为“心理资本”变量，该变量数值越大，表示心理状态越积极；

5. 将问卷中问题 39-41 的数据作为“成功渴望”变量，该变量数值越大，表示追求成功的动机越高。

对于问卷调查来说，对收集之后的数据进行处理是非常必要的。在处理中，对一些反映同一方面的指标进行合成，将 49 个指标采用相加的方式分类合成 5 个指标。通过关联规则得到各指标之间的关系。[21]

表 2 基于关联规则的合成表

分类指标	合成指标	指标描述
1,3-6,9	社交活动	变量数值越大，表示社交活动参与度越高
10-16	网络依赖	变量数值越大，表示对网络的依赖度越高
19-22,26	学习热情	变量数值越大，表示学习热情越高
28-32,34	心理资本	变量数值越大，表示心理状态越积极
39-41	成功渴望	数值越大，表示追求成功的动机越高

第 3 章　数据分析

3.1　多指标线性回归分析

3.1.1 描述性统计分析

本文采用武汉理工大学 18 级专业三个班 186 个学生样本进行分析。问卷共分为 4 类问题，包括个人信息的填写、家庭背景的问题、兴趣爱好问题、心理健康问题（主要分为以下 6 个子问题：学业倦怠问题、压力感知问题、心理资本问题、成就动机问题、职业预期结果问题、职业决策自我效能问题）等，并结合学生的期末《统计计算》成绩形成问卷反馈表。本文重点关注心理健康测试的问题，并结合家庭背景与兴趣爱好问题，探究学生学习、生活现状。其中，兴趣爱好与心理健康问题为离散数值型数据，分为 4 挡，数值越高表明学生在该测试上表现越好，属于效益型指标。家庭背景问题多为二类离散数值型问题，填写方式主要为是或否。[22]

数据的描述性分析既是从数据出发概括数据特征，主要包括数据的位置特征、分散性、关联性等数字特征和反映数据整体结构的分布特征，它是数据分析的第一步，也是对数据更进一步分析的基础。

通常意义上的描述性分析，可分为两个方面：一维和多维。一维的数据常常从表示位置的数字特征、表示分散性的数字特征和表示分布形状的数字特征三个方向考虑。常见于峰度和偏度的描述，偏度为正数据左偏；偏度为负数据右偏；峰度小于 0，分布状况为中高细尾；峰度大于 0，分布状况为中低粗尾。对于数据的分布通常采用直方图、茎叶图等，图中数据点的分布可以更为直观的反映数据之间的相关关系。多维数据主要是从数据相关性的角度进行分析。

以下是一些重要统计量的计算公式：

均值：

$$\overline{x}=\frac{1}{\mathrm{n}}\sum_{i=1}^{n}x_i \tag{4}$$

标准差:

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2} \tag{5}$$

偏度:

$$g_1=\frac{n}{(n-1)(n-2)}\frac{1}{s^3}\sum_{i=1}^{n}(x_i-\overline{x})^3 \tag{6}$$

峰度:

$$g_2=\frac{n(n+1)}{(n-1)(n-2)(n-3)}\frac{1}{s^4}\sum_{i=1}^{n}(x_i-\overline{x})^4-3\frac{(n-1)^2}{(n-2)(n-3)} \tag{7}$$

变异系数:

$$CV=\frac{s}{\overline{x}}\times 100\% \tag{8}$$

首先对典型的离散指标性别进行分析，得出下列表格

表 3 男女对应指标的数值

	社交活动	网络依赖	学习热情	心理资本	成功渴望
男	0.60	0.52	0.66	0.58	0.59
女	0.40	0.35	0.70	0.49	0.52

可以看出女生在社交活动积极性上高于男生，女生的社交表现更好；而男女生对网络的依赖度基本相同，可推测男生热爱网络游戏的程度与女生热爱网络追剧的程度近乎相等。在学习热情方面，女生略高于男生，对比于男生的心理资本与追求的成功的动机略高于女生，我们可以推测女生步入大学后仍保持了较高的学习热情，但因生活的丰富而使得心态较为波动；而男生虽然在大学中受游戏的影响较多，会在一定程度上耽误学习，但人处青年，胸怀大志，渴望实现自己的人生抱负，心态更为积极与有目标感。

接着计算其他的离散指标，家庭状况，独生子女，父母婚姻状况，助学贷款如下表所示，

表 4 18 级材料工程专业学生属性统计表

指标	属性	占比
家庭状况	农村	42%
	乡镇或县城	28.%
	地级及以上城市	29%
独生子女	是	48%
	不是	52%
父母婚姻状况	丧偶	3%
	离异	4%
	正常	93%
助学贷款或他人资助	有	30%
	没有	70%

通过上述表格，我们可以对家庭状况、是否为独生子女、父母婚姻状况和是否有助学贷款等与学生个人基本情况相关且离散的指标进行了频率分析，从问卷数据可以看出，18 级材料工程专业性别分布方面，男女比例基本持平；家庭状况方面，来自农村、乡镇县城和大城市的同学占比相近；非独生子女的同学略多于独生子女同学；大部分同学没有接受助学贷款或他人资助；部分同学的父母婚姻状况正常。

在计算完连续性变量之后，接着计算了剩余的连续性变量

表 5 制药 17 级学生状态统计表

指标	最小值	最大值	均值	均值标准误差	方差	偏度	峰度
《统计分析》成绩	1	4	2.6	0.1	0.7	-0.4	0.1
社交活动	6	10	9	0.2	1.6	-0.3	-0.6
网络依赖	5	11	8	0.4	1.8	-0.1	0.4
学习热情	5	25	16	0.8	14.5	0.4	-0.5
心理资本	6	8	4	0.3	2.9	0.8	0.7
成功渴望	5	19	8	0.3	2.7	0.6	0.5

对《统计分析》成绩、社交活动网络依赖学习热情心理资本成功渴望等连续型变量进行分析，以及反映学生个人能力、状态的指标进行了描述性分析。从表 5 的相关结果可以看出，18 级

研究生《统计分析》成绩平均等级在 2.6 左右，且大多数同学的《统计分析》成绩优于平均成绩，峰度正向接近于 0，表示《统计分析》成绩分布比较均匀。从偏度的方面还可以看出，大多数同学的成就动机较低、心理资本较少、职业预期较差，在职业决策自我技能方面各层次分布较为对称。从峰度方面可以看出，网络依赖学习热情情况比较集中，而心理资本成功渴望中极端数值分布范围较广。从均值标准误差和方差来看，这组数据平均值可以很好地反映大部分同学的实际情况。

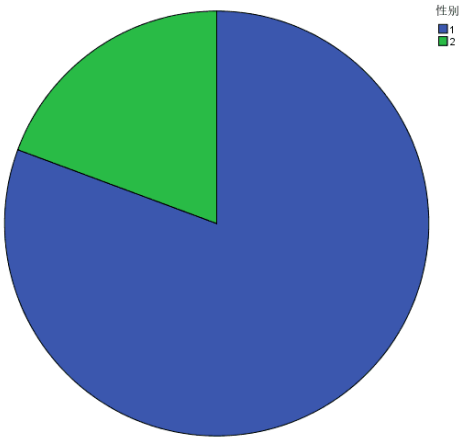


图 4 性别占比图

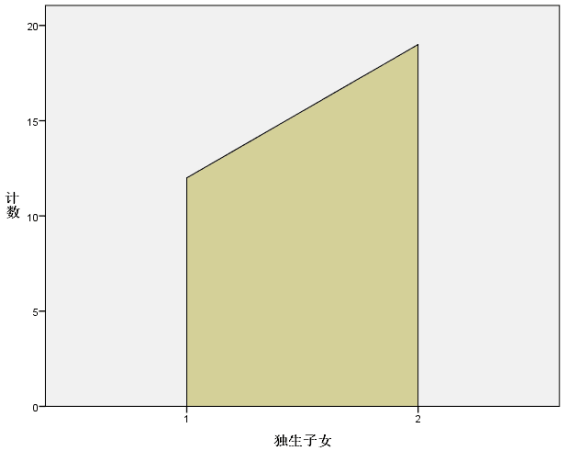


图 5 独生子女计数图

接着以学习好坏的平均值为主要指标，分析了学习好坏分别在性别、独生子女、家庭状况、社会资助等方面的分布情况。通过分析可以得出，男学生的学习好坏程度大于女学生，独生子女与非独生子女学习好坏程度相差不大，来自农村、乡镇或县城的学生的学习好坏程度略大于来自地级或以上城市的学生，接受过助学贷款或他人资助的学生比未接受过的学生的学习好坏程度要大一些。各个层面学习好坏程度略有差异，但区别不大，可以见得学习好坏存在于各个层面之中，不论性别、不论出身都会出现学习好坏的情况，我们可以针对其他方面进行进一步的分析。如下图所示，在学习数学程度相同的情况下，助学贷款和家庭情况之间的关系非常密切，表明越喜欢数学的学生，家庭条件和助学贷款的就越好，说明喜欢数学的学生是需要资本的。

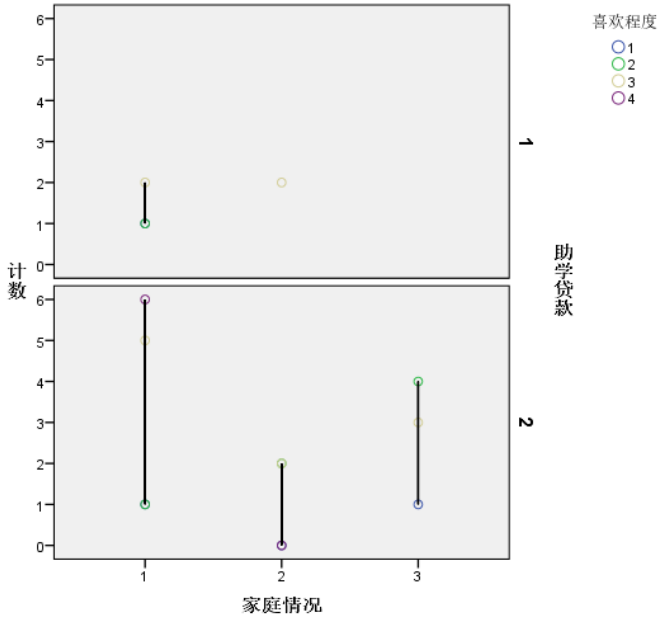


图 6 喜欢数学程度对应的箱图

3.1.2 变量相关分析

相关分析就是对总体中确实具有联系的标志进行分析，其主体是对总体中具有因果关系标志的分析。它是描述客观事物相互间关系的密切程度并用适当的统计指标表示出来的过程，是研究两个或两个以上处于同等地位的随机变量间的相关关系的统计分析方法。

相关系数主要包括 pearson 相关系数和 spearman 相关系数。其中 person 相关系数用来衡量两个数据集合是否在一条线上面，它用来衡量定距变量间的线性关系。可用以下公式表示：

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}} \tag{9}$$

spearman 相关系数是一种秩相关系数，它是衡量两个变量的依赖性的非参数指标。它利用单调方程评价两个统计变量的相关性。如果数据中没有重复值，并且当两个变量完全单调相关时，斯皮尔曼相关系数则为 +1 或 −1。可以用以下公式表示

$$q_{xy} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2 \tag{10}$$

相关系数的绝对值越大，相关性越强：相关系数越接近于 1 或-1，相关度越强，相关系数越接近于 0，相关度越弱。通常情况下通过以下取值范围判断变量的相关强度：

表 6 相关系数强度表

相关系数	强度
0.8-1.0	极强相关
0.6-0.8	强相关
0.4-0.6	中等强度相关
0.2-0.4	弱相关
0.0-0.2	极弱相关或无相关

通过计算六类的相关系数，学习成绩，社交活动，网络依赖，学习热情，学习资本，成功渴望之间的关系，由于使用的 python 软件编程，对于编程的文字，主要使用的是英文编码，所以在画图过程中，使用的是英语，分别表示为 Academic record ，Social affair，Network dependence，Enthusiasm for learning，Learning capital，Desire for success。这六类指标之间的关系如下图所示

分析六类指标的相关性之后，接着又分析学习成绩和个人因素之间的相关性，主要有家庭状况，网络沉迷，锻炼程度，花费开销相关系数的英文表示为 Family status，internet addiction，Exercise level，Cost expenditure。关系图如下所示

3.1.3 回归分析

线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。

回归分析即是利用 Y 与 $X_1, X_2, \cdots, X_{p-1}$ 的观测数据，并在误差项 ε 的某些假定下确定 $f(X_1, X_2, \cdots, X_{p-1})$ 。利用统计推断方法对所确定的函数的合理性以及由此关系所揭示的 Y 与各 $X_1, X_2, \cdots, X_{p-1}$ 的关系做分析，进一步应用于预测、控制等问题。当 f 是 X 的线性函数时，有

$$Y = \beta_0 + \beta_1 X_1 + \cdots \beta_{p-1} X_{p-1} + \varepsilon \tag{11}$$

称此模型为线性回归模型。其中 $\beta_0, \beta_1, \cdots, \beta_{p-1}$ 是未知常数，称为回归参数或回归系数；Y 称为因变量或响应变量；$X_1, X_2, \cdots, X_{p-1}$ 称为自变量或回归变量；ε 称为随机误差项并假定 $E(\varepsilon) = 0$,ε 是不可观测的随机变量，而 Y 和 $X_1, X_2, \cdots, X_{p-1}$ 是可观测的变量。

建立回归模型的第一步就是利用观测数据对回归参数向量 β 做出估计。另外，为了了解误差的分散性并对有关问题作统计推断研究，需要对误差方差 σ^2 做估计

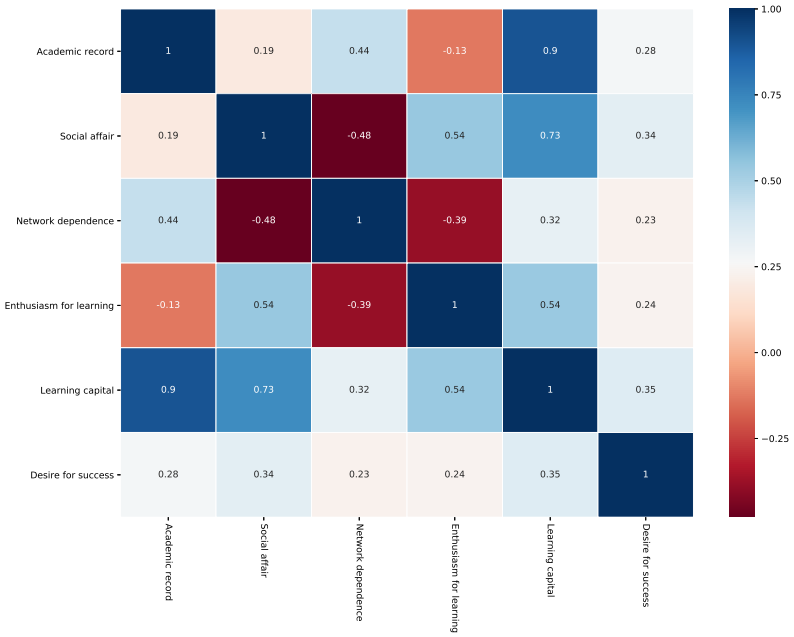


图 7 六类指标的相关系数

β 的最小二乘估计即选择 β 使误差项的平方和

$$S(\beta) = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \sum_{j=0}^{p-1} \beta_j x_{ij})^2 \tag{12}$$

达到最小。求出最佳 $\hat{\beta}= (\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_{p-1})^T$ 代入线性回归模型，并略去误差项，则称

$$\hat{\varepsilon} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \cdots, \hat{\varepsilon}_n)^T \tag{13}$$

$$SSE = \sum_{i=1}^{n} (\hat{\varepsilon}_i - \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i)^2 \tag{14}$$

称 SSE 为残差平方和

$$\hat{\sigma}^2 = \frac{SSE}{n-p} \tag{15}$$

为 σ^2 的无偏估计。

通过对五类指标社交活动，网络依赖，学习热情，学习资本，成功渴望进行线性回归分析得，其中学习资本与成功渴望以及学习热情呈线性相关模型，通过 SPSS 求出其中的参数的值分别为 2.366，0.499，2.855，其中 R 方为 0.62 设学习资本为 y, 成功渴望为 x_1，学习热情为 x_2：

$$y = 2.366 + 0.499x_1 + 2.855x_1 \tag{16}$$

3.2 成绩评价

3.2.1 主成分分析

主成分分析的主要目的就是对原变量加以改造，在不损失原变量太多信息的条件下尽可能地降低原变量的维数，即用维数较少的新变量代替原来的各变量。主成分分析，是考察多个变量间相关性的一种多元统计方法，研究如何通过少数几个主成分来揭示多个变量间的内部结构，即从原始变量中导出少数几个主成分，使他们尽可能多地保留原始变量的信息，且彼此间互不相关。[?]

总体主成分的定义：设 $X = (X_1, X_2, \cdots, X_p)^T$ 为 p 维随机向量，其协方差矩阵为

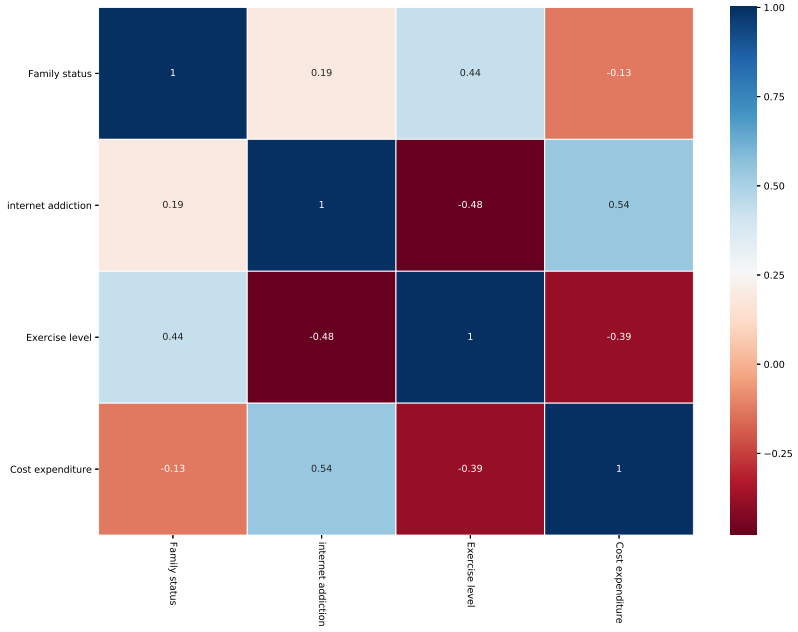


图 8 成绩与个人因素的相关系数

$$Cov(X) = E\left[(X - E(X))(X - E(X))^T\right] \tag{17}$$

这是一个 p 阶非负定矩阵。按照主成分分析的思想，首先构造 $X_1, X_2, \cdots, X_p$ 的线性组合

$$Y_1 = \alpha_1{}^T X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \tag{18}$$

确定 $\alpha_1 = (\alpha_{11}, \alpha_{12}, \cdots, \alpha_{1p})^T$，在约束条件 $a_1^T a_1 = 1$ 之下，使得

$$Var(Y_1) = Var(a_1^T X) = \alpha_{11}\Sigma a_1 \tag{19}$$

达到最大。如果第一主成分 Y_1 在 a_1 方向上的分散性还不足以反映原变量的分散性，则再构造 $X_1, X_2, \cdots, X_p$ 的线性组合

$$Y_2 = \alpha_2{}^T X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \tag{20}$$

为使 Y_1 和 Y_2 所反映的原变量的信息不相重叠，要求 Y_1 与 Y_2 不相关，即

$$Cov(Y_2, Y_1) = Cov(a_2^T X, a_1^T X) = a_2^T \Sigma a_1 = 0 \tag{21}$$

按主成分分析思想，问题转化为在约束条件 $a_2^T a_1 = 1$ 及 $a_2^T \Sigma a_1 = 0$ 之下，求 a_2 使得 $Var(Y_2) = a_2^T \Sigma a_2$ 达到最大。由此 a_2 所确定的随机变量 $Y_2 = a_2^T X$ 称为 X 的第二主成分。由此方法，可以构造出 p 个方差大于零的主成分。

记 $Y = (Y_1, Y_2, \cdots, Y_p)^T$ 为 p 个主成分构成的随机向量，则 $Y = P^T X$，其中 $P = (e_1, e_2, \cdots, e_p)$ 为 Σ 的的 p 个正交单位化特征向量构成的正交矩阵。Y 的协方差矩阵为

$$Cov(Y) = Cov(P^T X) = P^T \Sigma P = Diag(\lambda_1, \lambda_2, \cdots, \lambda_p) \tag{22}$$

各主成分的总方差为

$$\sum_{k=1}^{p} Var(Y_k) = \sum_{k=1}^{p} \lambda_k = tr(\Sigma) = \sum_{k=1}^{p} Var(X_k) \tag{23}$$

由上式可知，$\lambda_k \Big/ \sum_{i=1}^{p} \lambda_i = Var(Y_k) \Big/ \sum_{i=1}^{p} Var(X_i)$ 描述了第 k 个主成分提取的 $X_1, X_2, \cdots, X_p$ 的总（分散性）信息的份额，称此为第 k 个主成分 Y_k 的贡献率。第一主成分的贡献率最大，表明 Y_1 综合原始变量信息的能力最强，其他主成分综合原始变量信息的能力依次减弱。前 m 个

主成分的贡献率之和 $\sum_{k=1}^{m}\lambda_k \Big/ \sum_{k=1}^{p}\lambda_k$ 称为 $Y_1, Y_2, \cdots, Y_m$ 的累计贡献率，它表明前 m 个主成分综合 $X_1, X_2, \cdots, X_p$ 的信息的能力，通常选取使得累计贡献率到达 80% ~ 90% 的 m 个主成分，这样不但可以使原变量的维数降低，而且也不至于损失原变量中的太多信息。

3.2.2 KMO 检验和巴特利特球形度检验

因子分析前，首先进行 KMO 检验和巴特利球体检验。KMO 检验用于检查变量间的相关性和偏相关性，取值在 0 1 之前。KMO 统计量越接近于 1，变量间的相关性越强，偏相关性越弱，因子分析的效果越好。实际分析中，KMO 统计量在 0.7 以上时效果比较好；当 KMO 统计量在 0.5 以下，此时不适合应用因子分析法，应考虑重新设计变量结构或者采用其他统计分析方法。在 spss 中的因素分析时有关于 bartlet 球形检验的选项，如果 sig 值小于 0.05，则数据呈球形分布。

巴特利特球形检验法是以相关系数矩阵为基础的. 它的零假设相关系数矩阵是一个单位阵, 即相关系数矩阵对角线的所有元素均为 1, 所有非对角线上的元素均为零. 巴特利特球形检验法的统计量是根据相关系数矩阵的行列式得到的. 如果该值较大, 且其对应的相伴概率值小于指定的显著水平时, 拒绝零假设, 表明相关系数矩阵不是单位阵, 原有变量之间存在相关性, 适合进行主成分分析；反之, 零假设成立, 原有变量之间不存在相关性, 数据不适合进行主成分分析。[?]

球形检验主要是用于检验数据的分布，以及各个变量间的独立情况。按照理想情况，如果我们有一个变量，那么所有的数据都在一条线上。如果有两个完全独立的变量，则所有的数据在两条垂直的线上。如果有三条完全独立的变量，则所有的数据在三条相互垂直的线上如果不对数据分布进行球形检验，在做因素分析的时候就会违背因素分析的假设——各个变量在一定程度上相互独立。

3.2.3 基于五大合成指标的主成分分析

对社交活动，网络依赖，学习热情，学习资本，成功渴望等五个合成指标进行进一步的降维，针对这些指标进行了主成分分析，并且根据成分得分矩阵对每个同学进行了打分。得到各成分之间的得分

表 7 成分得分矩阵表

变量	组件		
	1	2	3
社交活动	0.582	0.469	0.090
网络依赖	0.611	0.663	-0.279
学习热情	0.807	-0.252	-0.332
学习资本	0.836	-0.009	1.171
成功渴望	0.616	-0.531	-0.514

根据成分矩阵，可以知道两个主成分的表达式分别为：

$$Y_1 = 0.372X_1 + 0.391X_2 + 0.516X_3 + 0.535X_4 + 0.243X_5 \tag{24}$$

$$Y_2 = 0.467X_1 - 0.660X_2 - 0.251X_3 + 0.009X_4 + 0.532X_5 \tag{25}$$

由主成分表达式可以知道，Y_1 是 5 个指标的加权和，它反映了学生的消极程度，Y_1 值越大，表示学生的消极程度越高。Y_2，Y_2 值越大，表示在消极的同时心理承受能力也很强。将指标返回原数据，发现 Y_1，Y_2 值越大，最后期末成绩分数越高。说明这几个指标在一定程度上影响着期末成绩。

3.3 聚类分析

3.3.1 聚类分析原理

聚类分析与判别分析都是研究分类的，但他们有所区别。聚类分析一般寻求客观的分类方法，在聚类分析以前，对总体到底有几种类型并不知道，一般有两种类型：快速聚类法与谱系聚类法。

快速聚类法也称动态聚类法，该方法首先将样品粗糙地分类，然后再依据样品间的距离按一定规则逐步调整，直至不能再调整为止。快速聚类法适合于样品数目较大的数据集的聚类分析，但需要事先指定分类的数目，此数目对最终分类结果有较大影响，因此在实际中一般要对多个分类的数目进行尝试，以找出合理的分类结果。[?]

设 $X=(x_1,x_2,\cdots,x_p)^T$ 为所关心的 p 个指标，对此指标作 n 次观测能得到 n 组观测值

$$x_i=(x_{i1},x_{i2},\cdots,x_{ip})^T, i=1,2,\cdots,n \tag{26}$$

称这 n 组观测数据为 n 个样品。这时每个样品可看成维空间的一个点，n 个样品组成 p 维空间的 n 个点，此时可以用各点之间的距离 $d(x_i,x_j)$ 来衡量各样品之间的靠近长度。

在快速聚类法中，可以先规定聚类中采用的距离是欧式距离，即

$$d(x_i,x_j)=||x_i-x_j||=[(x_i-x_j)^T(x_i-x_j)]^{\frac{1}{2}} \tag{27}$$

a) 设 k 个初始聚类的集合是

$$L^{(0)}=\{x_1^{(0)},x_2^{(0)},\cdots,x_k^{(0)}\} \tag{28}$$

用以下原则实现初始分类

$$G_i^{(0)}=\{x:d(x,x_i^{(0)})\le d(x,x_j^{(0)}),j=1,2,\cdots,k,j\ne i\},i=1,2,\cdots,k \tag{29}$$

这样，将样本分类成不相交的 k 类，以上初始分类的原则是每个样品以最靠近的初始点聚类，这样的到一个初始分类

$$G^{(0)}=\{G_1^{(0)},G_2^{(0)},\cdots,G_k^{(0)}\} \tag{30}$$

b) 从 $G^{(0)}$ 出发，计算新的聚类点集合 $L^{(1)}$，以 $G_i^{(0)}$ 的中心作为新的聚点：

$$x_i^{(1)}=\frac{1}{n}\sum_{x_l\in G_i^{(0)}}x_l,i=1,2,\cdots,k \tag{31}$$

其中 n_i 是类 $G_i^{(0)}$ 中的样品数。这样得到新的聚点集合

$$L^{(1)}=\{x_1^{(1)},x_2^{(1)},\cdots,x_k^{(1)}\} \tag{32}$$

同上得到新的分类

$$G^{(1)}=\{G_1^{(1)},G_2^{(1)},\cdots,G_k^{(1)}\} \tag{33}$$

c) 依次计算下去，直到第 m 次，$G^{(m+1)}$ 与 $G^{(m)}$ 完全相同。

快速聚类与系统聚类主要区别：快速聚类是先指定分类数，通过调整聚点使得聚类完成；系统聚类则是首先视各样品自成一类，然后把最相近（距离最小）的样品聚为一类，再将已聚合的小类按其相近性再聚合，最后将一切子类聚合成一个大类，从而得到一个按相近性大小聚合起来的普系统，再根据实际情况确定合适的分类。

3.3.2 针对学习好坏程度的聚类分析

首先单一地根据学业倦怠程度进行系统聚类分析，得到如下树状图。根据树状图，可以以不同的分类数进行分类。从分类效果来看，系统聚类对于不同程度的学习成绩的分类效果良好。

3.3.3 针对其他指标的聚类分析

接着进一步以社交活动，网络依赖，学习热情，学习资本，成功渴望为变量，以学习成绩为观测指标进行 K-均值聚类。从的聚类结果来看，这五个综合指标的聚类结果与学习的成绩的好坏程度有很大的关联度，说明这些指标对于学习成绩指标有很强的相关性，在之后的研究中可以通过这些指标对于学生学习好坏进行大概的预测和评判。

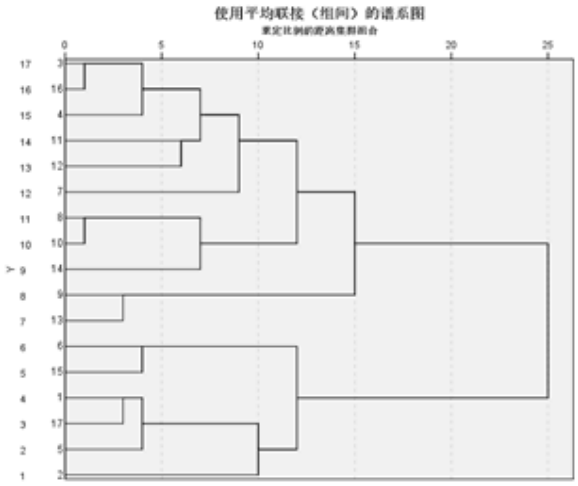


图 9 学习成绩聚类树状图

第 4 章　研究结论及建议

4.1　研究结论

本次主要从两方面进行了研究：学生等个体属性因素对成绩的影响，使分析预测结果全面。对学生个体进行差异性预测，避免普遍性导致的问题。根据以上对大学生学习状况的研究，我们可以得出以下结论：

（1）当代大学生的学习状况个体差异性明显，且总体上存在很多学习问题，如抑郁、焦虑、人际关系不良、自卑、有学习障碍等。究其原因，主要在于大学生这个阶段的人生观和价值观刚刚形成，不够成熟，且为未来比较迷茫；另一方面，大学生刚刚经历中学时代“疯狂式”的封闭学习生活，在大学这个与社会接轨又较为宽松的过渡期中，容易产生不适应感而没有规划好自己的生活，导致了学习上的不自信与自责，而这种消极的学习又阻碍了他们珍惜时间干实事，这是十分可怕的恶性循环。

（2）大学生的学习状况与大学生的心理状态及日常生活状态息息相关。学习状态越积极的学生，对学习的热情越高，他们更加自律，目标感更强，追求成功的动机也更显著。因此，对于一个大学生来说，努力保持一个积极的心态尤为重要，这就需要我们刻意地培养与训练，积极的学习暗示与踏实的实际行动，能让我们的生活越过越好。

（3）为了培养一个积极的学习状态，我们可以从控制用网时间、保证日常睡眠和加强体育锻炼三个方面采取行动。“控制用网时间”要求我们日均用网时间应在 4 小时以内，伏案工作者应注意休息，而网络娱乐也应加以控制并减少；“保证日常睡眠”要求我们日均睡眠 7-8 小时，可以每天晚上睡 7 小时，再配上中午的 30 分钟午休；“加强体育锻炼”要求我们周运动时长要达到 2-3 小时，如每周进行 3-4 次，每次 40 分钟的长跑。

（4）提高学习习惯更高层次上，要求我们对社会的动态也要有所关注，不能“两耳不闻窗外事，一心只读圣贤书”。关注社会引导我们要心胸开阔，胸怀天下，将个人对社会的贡献作为自己人生的价值，这样一来，我们就不会太为个人的蝇头小利、琐碎小事而烦恼了。

4.2　研究的建议

针对以上分析结果，我们可以从以下几个方面对学生进行教育。一方面，加强他们自身探索的实力，这需要老师和父母们正确的引导，让他们早日发现自己的兴趣爱好，从而能够主动的学习，有目标的学习。同样的，也可以邀请优秀的学长学姐以自身实际经历让学生受到启发。但不得不承认，每个人都是不一样的，学生在学习他人的优秀方法时，也应当结合自己的实际情况，多一些对自己的认识与探索。

另一方面，无论是否探索到适合自己的学习方法或感兴趣的方向，都应当保持学习的状态，学习目标则是保持学习状态的动力之源，学生在学习之前，应当首先明确一下自己的学习目标。在制定学习目标时，学生应当根据自身实际情况，为自己制定合适的短期的学习目标，短期的学

习目标越详细越具体，越有利于目标的顺利实行，之后在短期目标的基础上，再制定长期的学习目标，同时也可以先制定长期学习目标，再细化成短期目标。学生在制定目标时，亦可向自己的老师和同学求助，通过教师的帮助与指导，把自身能力条件与现有学习环境相结合，为自己制定切实可行的短期学习目标，并且可以根据实施过程中自身不断变化的学习情况，适当地做出相应的调整。当自己制定的学习目标的实现的时候，都会给学生带来成功的喜悦感以及成就感，长此以往坚持制定目标，并且完成计划，有利于增强学生在学习上的自信心和动力，可以大大提高学生的求知兴趣，从而激发出学生自主学习的积极性，进一步减轻学习倦怠。

同时，也可以适当进行教学改革，教学改革主要用于激发学生学习的热情，老师的目的并不能局限在教授知识，还应当多一些对学生的引导，引导他们更深层次的了解专业所涉及的相关应用，带领他们实际的去解决问题，从而激发他们学习的兴趣，发散学生的思维，让他自己思考。

参考文献

[1] 陈喜华, 黄海宁, 黄沛杰. 基于 Apriori 算法的学生成绩分析在课程关联性的应用研究 [J]. 北京城市学院学报,2018(04):60-65+84.

[2] 栾若星. 基于 R 语言的成绩分析方法 [J]. 智能计算机与应用,2018,8(04):136-139.

[3] 曾兴. 基于关联规则挖掘的学生成绩分析研究 [D]. 海南大学,2018.

[4] 张濠天, 张文卿, 王元元, 施月霞, 曾南焱. 关联规则挖掘在成绩分析中的应用 [J]. 中国高新区,2018(10):47.

[5] 张甜, 尹长川, 潘林, 安杰. 基于改进的聚类和关联规则挖掘的学生成绩分析 [J]. 北京邮电大学学报 (社会科学版),2018,20(02):91-96.

[6] 吴文玲. 基于数据挖掘技术的课程相关性分析及其应用研究 [D]. 四川师范大学,2018.

[7] 王成勇. 基于关联规则 Apriori 算法的学生成绩分析 [J]. 价值工程,2018,37(05):171-173.

[8] 张甜. 基于数据挖掘的高校学生成绩关联分析研究 [D]. 北京邮电大学,2018.

[9] 吴珊珊. 关联规则在专业课教学评价中的应用研究 [D]. 武汉工程大学,2017.

[10] 敖希琴, 费久龙, 陈家丽. 基于关联规则的高校学生成绩分析研究 [J]. 教育现代化,2017,4(45):240-244.

[11] 和铁行, 王伟. 数据挖掘在计算机课程成绩分析中的应用 [J]. 浙江医学教育,2017,16(05):4-6+42.

[12] Li Li,Xin Li,Yuan Yang,Jia Dong. Indoor tracking trajectory data similarity analysis with a deep convolutional autoencoder[J]. Sustainable Cities and Society,2019,45.

[13] Xuesong Wang,Minming Yang,David Hurwitz. Analysis of cut-in behavior based on naturalistic driving data[J]. Accident Analysis and Prevention,2019,124.

[14] 邵峰, 曾普华, 曾光, 郜文辉, 贺佐梅, 夏帅帅, 李静, 黄惠勇. 基于临床数据分析原发性肝癌的证治规律 [J/OL]. 湖南中医药大学学报,2019(01):40-44[2019-01-17].http://kns.cnki.net/kcms/detail/43.1472.R.20190117.1527.020.html.

[15] 王琳. 科学大数据云分析服务的性能优化技术 [J/OL]. 电子技术与软件工程,2019(01):136[2019-01-17].http://kns.cnki.net/kcms/detail/10.1108.TP.20190115.1536.220.html.

[16] 冯锐, 周洋. 基于多元线性回归分析法的数控机床热误差补偿的研究 [J]. 内燃机与配件,2018(13):129-130.

[17] 刘华. 非线性回归模型在边坡变形监测中的应用 [J]. 测绘与空间地理信息,2018,41(04):221-224.

[18] 范继农. 一元回归分析法在工程财务分析中的应用 [J]. 黑龙江交通科技,2017,40(04):198-200.

[19] Weijie Li,Jun Liang,Jingran Ge. Novel designs of charring composites based on pore structure control and evaluation of their thermal protection performance[J]. International Journal of Heat and Mass Transfer,2019,129.

[20] John Z. Wu,Christopher S. Pan,Bryan M. Wimer. Evaluation of the shock absorption performance of construction helmets under repeated top impacts[J]. Engineering Failure Analysis,2019,96.

[21] Rongli Sang,Yuzhu Zhang,Jun Shao,Chunliang Yan,Kai Zhao. Preparation of ZnO/ASC by electrochemical deposition and evaluation of its desulphurisation performance[J]. Journal of Alloys and Compounds,2019,777.

[22] Jian Sun,Cheng Liang,Xianliang Tong,Yafei Guo,Weiling Li,Chuanwen Zhao,Jubing Zhang,Ping Lu. Evaluation of high-temperature CO 2 capture performance of cellulose-templated CaO-based pellets[J]. Fuel,2019,239.