# 设计说明

* 1. **设计背景**

信息技术高速发展，高校在教育信息化推进的过程中，构建了大量的教育教学以及学工管理的相关学习平台或业务管理系统，随着这些的平台系统的推广使用，累计了大量的数据，将这些数据加以分析和利用，必然能为学校的教学和管理工作带来巨大的帮助。数据挖掘是一门新兴交叉学科，能从海量的、不完整的且有噪声的随机样本数据中训练出分析模型，能快速、有效地挖掘出隐藏在数据里的信息和关系。利用数据挖掘技术提取高校学生数据中潜在的规律和信息，为学校的教育教学改革和学生管理水平的提高提供支持，已经成为当前教育信息化研究的热点。

* 1. **设计目的**

本课程设计希望利用数据技术，以分析学生的历史学习数据为着手点，寻找被预警学生的特征和能够反映学生学业成绩下滑状态的属性变量，探究隐藏在被预警学生背后有价值的信息、规律或模式，构建基于数据挖掘的学业预警模型，预测学生的状态是否处于异常并及早地向具有学习风险的学生发出预警信号，使教师及时调整教学策略、优化教学模式，使学生提高学习效果。

* 1. **设计内容**

采用决策树的数据挖掘方法，基于两所葡萄牙中学的学生成绩属性，构建一棵以信息增益划分的决策树，并利用此决策树对学生成绩进行提前预测。同时在此基础上，利用模型评分方式，对模型参数进行优化，降低过拟合对预测的影响，提高模型的泛化能力。

# 数据集说明

**2.1 数据集概述**

这些数据接近两所葡萄牙中学的学生成绩。数据属性包括学生成绩、人口统计、社会和学校相关特征)，并通过学校报告和问卷收集。两个数据集是关于两个不同科目的表现:数学(mat)和葡萄牙语(por)。在[Cortez和Silva, 2008]中，两个数据集在二元/五级分类和回归任务下建模。重要提示:目标属性G3与属性G2、G1具有很强的相关性。这是因为G3是最后一年的职系(在第三期颁发)，而G1和G2对应第一和第二期的职系。在没有G2和G1的情况下预测G3更加困难，但这样的预测更有用(更多细节见论文来源)。

**2.2 数据集来源**

数据集来自于kaggle，数据集的名称为Student Grade Prediction，该数据统计了两所葡萄牙学校的中学学生的学习成绩，数据属性包括学生成绩，人口统计学，社会和与学校相关的特征，通过使用学校报告和调查表进行收集。提供了两个关于两个不同学科表现的数据：数学（mat）和葡萄牙语（por），关于数据集的详细介绍可以参照kaggle的官方说明，数据集链接如下：https://www.kaggle.com/datasets/dipam7/student-grade-prediction。

**2.3 数据集特征**

|  |  |
| --- | --- |
| 数据集特征 | 多元 |
| 实例数量 | 649 |
| 性质 | 社会 |
| 属性特征 | 整数 |
| 属性数量 | 33 |

属性特征如下：

1 school-学生的学校(二进制:“GP”-加布里埃尔·佩雷拉或“MS”-穆萨尼奥·达·西尔韦拉)

2 sex-学生性别(二元:“F”代表女性，“M”代表男性)

3 age-学生年龄(数字:从15岁到22岁)

4 address-学生的家庭地址类型(二进制:'U' -城市或'R' -农村)

5 famsize - 家族规模(二进制:'LE3' -小于或等于3或'GT3' -大于3)

6 Pstatus - 父母的同居状态(二进制:“T”--同居，“A”--分居)

7 Medu - 母亲教育(数字:0 -无，1 -小学教育(四年级)，2 â€" 5 - 9年级，3 â€"中等教育或4 â€"高等教育)

8 Fedu - 父亲教育(数字:0 -无，1 -小学教育(4年级)，2 â€" 5 - 9年级，3 â€"中等教育或4 â€"高等教育)

9 Mjob - 母亲的工作(名义上:“教师”、“与保健有关的”、“公务员”(例如行政或警察)、“在家”或“其他”)

10 Fjob - 父亲的工作(名义上:“教师”、“与保健有关的”、“公务员”(如行政或警察)、“在家”或“其他”)

11 reason - 选择这所学校的理由(名义上:离家近、学校声誉好、课程偏好或其他)

12 guardian - 学生的监护人(名义上:“母亲”、“父亲”或“其他”)

13 traveltime - 从家到学校的旅行时间(数字:1 - <15分钟，2 - 15 - 30分钟，3 - 30分钟到1小时，或4 - >1小时)

14 studytime - 每周学习时间(数字:1 - <2小时，2 - 2 - 5小时，3 - 5 - 10小时，或4 - >10小时)

15 failures - 过去班级失败的次数(数值:如果1<=n<3，则为n，否则为4)

16 schoolsup - 额外教育支持(二进制:是或否)

17 famsup - 家庭教育支持(二进制:是或否)

18 paid - 课程科目内的额外付费课程(数学或葡萄牙语)(二进制:是或否)

19 activities - 课外活动(二进制:是或否)

20 nursery - 托儿所(二进制:是或否)

21 higher - 想接受高等教育(二进制:是或否)

22 internet - 在家上网(二进制:是或否)

23 romantic - 有浪漫的关系(二进制:是或否)

24 famrel - 家庭关系质量(数字:1 -非常差到5 -极好)

25 freetime - 放学后的空闲时间(数字:从1 -非常少到5 -非常多)

26 goout - 和朋友出去(数字:从1 -非常低到5 -非常高)

27 Dalc - 工作日酒精消耗量(数字:1 -极低至5 -极高)

28 Walc - 周末饮酒(数字:1 -极低至5 -极高)

29 health - 当前健康状况(数字:从1-非常差到5-非常好)

30 absences - 学校缺勤次数(数字:从0到93)

#这些分数与课程科目相关，数学或葡萄牙语:

G1 -第一阶段等级(数字:从0到20)

G2 -第二阶段等级(数字:从0到20)

32 G3 -最终等级(数字:从0到20，输出目标)

# 决策树算法

分类作为数据挖掘的一个重要领域备受关注，相应的算法也很多，使用最频繁的是决策树方法。决策树方法采用自顶向下的递归方式来构造决策树36l，通过有目的地对大量数据进行分类，进而挖掘隐藏在数据中的有价值的信息。描述简单、分类速度快.特别适合大规模数据处理是其主要优点。基于信息嫡的ID3方法是决策树算法中最为经典的一种。

**3.1 决策树算法概述**

以样本属性为节点、属性取值为分支的树状结构称为决策树。根节点取所有样本中信息量最大的属性,根节点子树样本信息量最大的属性作为树的中间节点,样本的类别值作为叶节点。从根到叶节点的一条路径就对应着一条分类规则,整个决策树则对应着一组析取表达式规则。

决策树算法的分类包括构造树(Tree Building)和树剪枝(Tree Pruning)两个阶段。

(1)构造树阶段:从根节点开始按设定标准计算每个节点的值，根据值大小选择测试属性，并按照相应测试属性的可能值依次向下来建立分支，删除测试属性后对新划分的训练样本继续以相同方式递归执行。一个节点上的所有数据都被归类或某节点中的样本数据的数量为空或低于设定值时结束，决策树生成。在树的节点上怎样选择最佳测试属性将训练样本进行最佳划分是构造阶段的关键，用作选择测试属性的标准很多，常用的有:信息增益、信息增益比、基尼指数等。

(2)树剪枝阶段:为提高数据集分类的科学性与准确性，就需要把决策树中因为测试数据过度拟合或因噪声、孤立点造成的不合理的分支剪掉，使决策树得到优化，称为树剪枝。树剪枝即通过检测来去掉一些分支，常用的树剪枝方法有三类，分别是先剪枝法、后剪枝法、两者结合的方法。

**3.2 ID3算法**

ID3算法属于贪心算法，以信息论为理论基础，使用信息增益为属性选择标准。基本原理为:若N为集合T中的元组，选择信息增益最高的属性为T的分类属性，可使结果分类中对T分类所需要的信息最小。

3.2.1基本概念

定义1:若N为集合T中的元组，对T中元组进行分类需要的期望信息称为T的信息期望，也称为T的信息嫡。

其中pi是T中任意元组属于类Ci的非零概率，用计算。

定义2:信息增益是基于某属性（例如用属性X)对集合T进行分类后，原来的信息嫡与新的信息嫡(对X划分后）之间的差值。

其中的InfoA(T)是依据A对元组T分类所需要的信息嫡。

3.2.2 算法步骤

算法:由给定样本产生决策树ID3\_DecsionTree

输入:训练数据集 Dataset;候选属性集 AttributeList

输出:决策树ID3\_DecsionTree。

方法:

(1)创建一个节点为N;

(2)若样本数据均属于类C，返回N为叶节点并标为类C;

(3)若AttributeList空，返回N为一个叶节点，记为Dataset中一个最普通的类;

(4)在AttributeList选择信息增益最大的属性 Test\_A;

(5)标记节点N为AttributeTest;

(6)对于测试属性AttributeTest中所有己知值a1 ;

(7)以 AttributeTest=ai为条件，从节点N分裂一个分支;

(8)若Dataset中符合AttributeTest=ai的集合si为空,追加一树叶并记为Dataset中的最普遍的类;

(9)否则加上由ID3\_DecsionTree( si,AttributeList\_AttributeTest)返回的节点;

分析算法步骤可知，一个节点上的所有数据都属于同一类或没有属性可再用于数据分割时算法递归终止。条件二发生时应将当前节点通过一定的方式标记为叶节点，否则创建一个叶节点并标记为当前节点所含样本集中类别个数最多的类别。

# 结果分析

本文以数据挖掘为基本方法对高校学业预警进行研究,运用经典数据挖掘算法对学生学业成绩相关属性数据进行了分析。数据挖掘结果显示,在学生日常表现方面,学生的课余时间长短、出勤情况等对学生的预警有较大的影响;学生个人情况方面,学生健康状况、过去班级失败次数、家庭经济状况，家庭教育支持情况，家庭关系质量等对学生的学业预警也有较为明显的影响。根据上述结果，建议在教学和管理中要注意以下几个方面：

（1）学工部门要加强对学生的基础管理，严把学生请假关，并与任课教师共同做好课堂考勤工作。

（2）可以通过大学生导师制、班主任工作、新生研讨等多种方法做好学生指导，增强高考成绩较低学生的信心，做好学生的学习方法、学习习惯的养成工作。

（3）学校各相关部门、班主任、辅导员、任课教师要备加关心和爱护单亲、孤儿学生和来自贫困家庭的学生，充分利用高校的各种奖助政策，减轻学生经济压力，切实将学生的主要精力转到专业学习上来。

（4）学校和家长之间要加强沟通，同时家长与孩子要及时沟通，积极解决可能存在或已存在的矛盾，构建一个家校学生沟通环境。

（5）学校在教学安排上要注意劳逸结合，同时尽量避免过重的学业压力所导致的负面情绪和严重的心理压力。

# 结束语

将数据挖掘技术应用于学生“学业预警”，不仅能提高该项工作的针对性，也能够对高校的学生管理、教学管理决策提供数据支持，也是对高校保存的大量历史数据的科学使用。同样地，数据挖掘技术也可以对高校保存的教师科研数据、学生评价数据、教师业绩数据等深入的挖掘分析，挖掘数据背后的有用的规律来服务高校的人才培养、教学管理等工作。

# 参考文献

[1]宫锋. 数据挖掘在高校学业预警中的应用研究[D].中国石油大学(华东),2017.

[2]宫锋.数据挖掘在高校学生学业预警中的应用[J].电子技术与软件工程,2017(04):202-203.

[3]马丹妮. 基于机器学习的学生学业预警模型研究[D].沈阳理工大学,2020.DOI:10.27323/d.cnki.gsgyc.2020.000193.

[4]林秀科,沈良忠.基于决策树的学生成绩对毕业影响分析[J].电脑知识与技术,2017,13(35):15-16.DOI:10.14004/j.cnki.ckt.2017.4044.

[5]张军,王芬芬.决策树在高校学生学业预警中的应用研究[J].无线互联科技,2020,17(20):171-172.