Proceedings of the 4th International Conference on Educational Innovation and Philosophical Inquiries
DOI: 10.54254/2753-7048/11/20230750

Analysis of the Influence of Some External Factors on Students' Mathematics Achievement Based on Decision Tree Algorithm

Yichao Wang[1,a,*]

[1]*School of Software, North University of China, Taiyuan City, Shanxi Province, 030000, China*
a. 1220448818@qq.com
**corresponding author*

Abstract: With the continuous progress of the educational system in various countries, people attach more and more importance to education. In today's school education, the math score of students is an important indicator for the assessment of student learning outcomes, on the one hand, can be a real, objective reflect students' actual learning and teachers' teaching level, on the other hand can also choose for students after learning methods, teachers to play a good role in guiding the teaching plan. This study analyzes the student performance data of math students in kaggle, explores the factors that affect the students' performance of this course from many aspects, and then puts forward reasonable suggestions and applies them to practical teaching so as to improve the teaching quality. All of the data the research has been getting is from Kaggle math students. The data set contains nearly 400 pieces of relevant educational data about students, and there are dozens of factors that affect the performance of each student, such as parental cohabitation, educational support and the desire for higher education. The topic of our study is to analyze the impact of students' family situations and educational support on their math performance by using the C4.5 algorithm in the decision tree algorithm as an auxiliary tool. And this paper finally came to this interesting conclusion: motivated students with poor parental relationships and little educational support from school and family tend to do better in math.

Keywords: decision tree algorithm, parental relationship, educational support, student intention students' math scores

1. Introduction

With the continuous improvement and promotion of education informatization in the global field, both domestic and foreign, the methods and auxiliary tools used for education are also increasing. At the same time, people's research degrees and educational cognition are deepening. And the factors that affect student achievement are more and more diverse [1]. The vast majority of social concerns have been directed toward education in society and student achievement. In education, the achievement of students can not only be a rough customization of students' future, but also can be used for schools and countries to formulate some corresponding policies and methods. Therefore, this study explores the influencing factors of student achievement based on some student achievement and some influencing factors on kaggle website. The study wanted to look at math

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/11/20230750

achievement by looking at certain factors affecting math achievement among hundreds of students in certain schools. Our goal was to identify the most important factors among several factors, such as parental cohabitation, family and parental educational support, and whether students aspire to higher education.

2. Methodology

2.1. Decision Tree Algorithm

Decision tree is a classification algorithm based on tree structure to make decisions [2]. We hope to learn a model (namely decision tree) from a given training data set and use this model to classify new samples. The decision tree can intuitively show the classification process and results. Once the model is successfully built, the classification efficiency of new samples is also quite high. The most classical decision tree algorithms include ID3, C4.5 and CART, among which ID3 is the first proposed algorithm, which can deal with the classification of discrete attribute samples, while C4.5 and CART algorithms can deal with more complex classification problems [3]. C4.5 algorithm is used in this study. Classification tree (decision tree) is a very common classification method. It is a kind of supervised learning, the so-called supervised learning is given a bunch of samples, each sample has a set of attributes and a category, these categories are determined in advance, then through learning to get a classifier, this classifier can give the correct classification of the newly emerged objects [4]. This kind of machine learning is called supervised learning.

2.2. Data Collection

The data set named math students on kaggle website was used as the data information for this study. Among them, the research uses four factors that have a large impact on students' math scores as variables in this study, namely parental cohabitation, educational support, and the desire for higher education.

2.3. Data Preprocessing

After sorting out the data, there are 1980 pieces of data in the data table.

Table 1: Related data affecting student achievement.

	Pstatus	schoolsup	famsup	higher	G1	G2	G3
1	T	no	no	yes	19	19	20
2	A	no	no	yes	18	19	19
3	T	no	no	yes	18	19	19
4	T	no	no	yes	19	18	19
5	T	no	yes	yes	19	18	18
6	T	no	yes	yes	18	18	19
7	T	no	no	yes	18	18	18
8	T	no	yes	yes	18	18	18
9	no	no	yes	18	18	18	54
10	no	yes	yes	18	18	18	54
11	no	yes	yes	17	18	18	53
12	no	no	yes	17	18	18	53

DOI: 10.54254/2753-7048/11/20230750

Table 1: (continued).

13	no	yes	yes	16	18	19	53
14	no	yes	yes	16	18	18	52
15	no	yes	yes	16	18	18	52
16	no	yes	yes	17	17	18	52
17	no	yes	yes	17	17	17	51
18	no	no	yes	16	17	18	51
19	no	no	yes	18	16	16	50
...							
389	no	yes	yes	7	0	0	7
390	no	no	yes	7	0	0	7
391	no	no	yes	6	0	0	6
392	no	yes	no	6	0	0	6
393	no	yes	yes	5	0	0	5
394	no	yes	no	5	0	0	5
395	no	yes	yes	4	0	0	4
396	no	yes	yes	4	0	0	4

2.4. Data Classification Mining

Combined with the research object of this study, that is, students' achievement data contains both continuous data and discrete data, C4.5 algorithm of the decision tree algorithm is chosen as the method of this study. As an improved algorithm of the ID3 algorithm, the C4.5 algorithm uses information gain rate selection properties to avoid the ID3 algorithm when selecting attributes. It tends to choose the maximum attribute values, but also can deal with incomplete data, in terms of decision tree pruning C4.5 algorithm has made the corresponding improvement, can be in the process of pruning [5].

2.5. Build the Decision Tree Model

The model building process of C4.5 algorithm is shown in the figure below [6] :

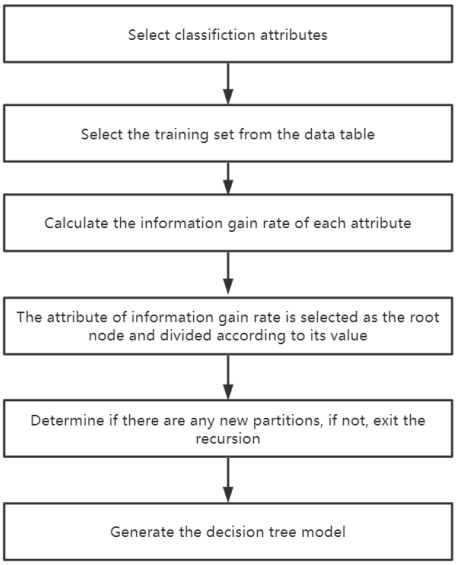

Figure 2: The model building process of C4.5 algorithm.

DOI: 10.54254/2753-7048/11/20230750

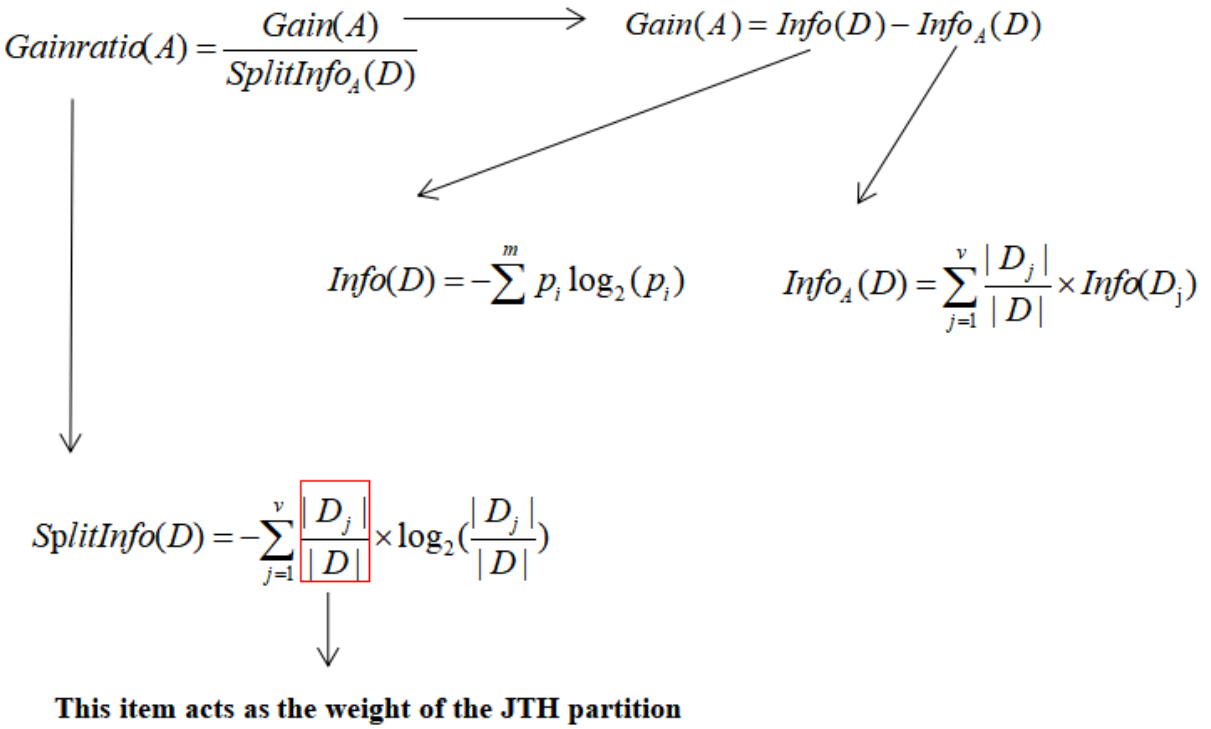

Figure 3: Information gain rate.

Let the data partition D be the training set of labeled class tuples. Assume that the class label attribute has m distinct values. Info(D) represents the average amount of information needed to identify the class labels of tuples in sample set D. Info(D) is also called the entropy of D. In total, there are v data with different values and attribute A in the training set.

2.6. Generate Classification Rules

The decision tree algorithm uses rules in the form of if-then. The expression of the if-then rule is if is the condition and then is the conclusion. The method of extracting if-then rules is as follows: each path from the root node to the leaf node in the decision tree generates a classification rule, the test conditions in the path constitute the conjunction item of the rule's precursor, namely the if part of the rule, and the class label of the leaf node is assigned to the rule's later part, namely the then part of the rule.

3. The Concrete Application of Decision Tree Algorithm

We define students whose total score is above 39 as excellent, 20-39 as good, and below 20 as poor.By using the table tool, we can know that there are a total of 396 students. There are 103 samples of excellent students, 238 samples of good students and 55 samples of poor students. The information entropy of 'excellent' can be calculated by the formula.

It can be calculated from the data that the entropy of the final score integral is 0.82689, so we take the final score integral as the root node of the decision tree.

$$\inf\ o(D) = -\frac{103}{396} \times \log_2 \frac{103}{396} - \frac{293}{396} \times \log_2 \frac{293}{396} = 0.82689 \quad (1)$$

According to the decision tree model, the final score integral is taken as the root attribute, and then the final root node is established, as shown in the figure below:

DOI: 10.54254/2753-7048/11/20230750

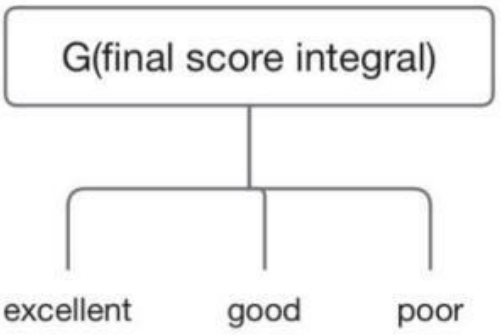

Figure 4: The final root node.

Table 2: Related data and score chart affecting students' scores.

	Pstatus	schoolsup	famsup	higher	G1	G2	G3	G	level
1	T	no	no	yes	19	19	20	58	excellent
2	A	no	no	yes	18	19	19	56	excellent
3	T	no	no	yes	18	19	19	56	excellent
4	T	no	no	yes	19	18	19	56	excellent
...									
102	T	no	no	yes	13	13	14	40	excellent
103	T	no	yes	yes	12	14	14	40	excellent
104	T	no	no	yes	13	13	14	40	excellent
105	T	no	no	yes	15	12	12	39	good
106	T	no	yes	yes	13	13	13	39	good
107	T	no	yes	yes	14	12	13	39	good
...									
339	T	no	no	no	5	8	7	20	good
340	T	no	no	yes	7	6	7	20	good
341	T	no	yes	yes	6	6	8	20	good
342	T	no	no	yes	10	9	0	19	poor
343	T	no	yes	yes	9	10	0	19	poor
344	T	no	yes	yes	10	9	0	19	poor
...									
394	T	no	yes	no	5	0	0	5	poor
395	A	no	yes	yes	4	0	0	4	poor
396	T	no	yes	yes	4	0	0	4	poor

Taking 'whether parents cohabit' as the test attribute, this attribute has two attribute values of 'yes' and' no ', the corresponding number of samples are 355 and 41 respectively, among which 93 are excellent, 210 are good, and 52 are poor. In terms of parental separation, 10 were excellent, 27 were good, and 4 were poor.

The information entropy of "whether parents live together" is:

DOI: 10.54254/2753-7048/11/20230750

$$\text{info(parents' cohabitation)} = \frac{103}{396}\times(-\frac{93}{103}\times\log_2\frac{93}{103}-\frac{10}{103}\times\log_2\frac{10}{103})+\frac{237}{396}\times(-\frac{210}{237}\times\log_2\frac{210}{237}-\frac{27}{237}\times\log_2\frac{27}{237})+\frac{56}{396}\times(-\frac{52}{56}\times\log_2\frac{52}{56}-\frac{4}{56}\times\log_2\frac{4}{56})=0.4783 \quad (2)$$

The split information of "whether parents live together" is:

$$split\inf o(parents'cohabitation) = -(\frac{103}{396}\times\log_2\frac{103}{396}+\frac{237}{396}\times\log_2\frac{237}{396}+\frac{56}{396}\times\log_2\frac{56}{396})=0.65141 \quad (3)$$

The information gain rate of "whether the parents live together" is:

$$gainratio(parents'cohabitation) = \frac{0.82689-0.4783}{0.65141}=0.5351 \quad (4)$$

Then the subtree of decision tree is established after the root node, and the four factors affecting students' mathematics achievement are taken as the subattributes.As shown below(Here we use the cohabitation of parents as a case study):

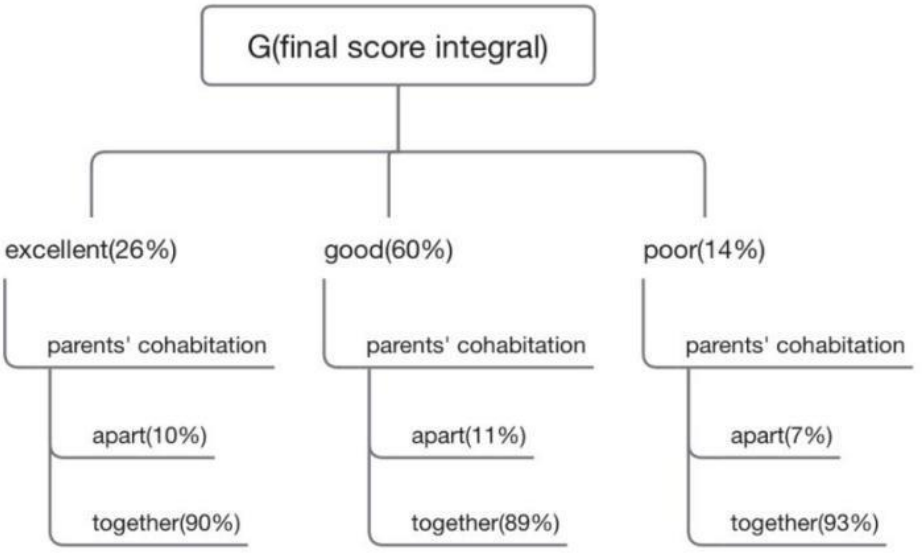

Figure 5: The subtree of decision tree.

Using the above if-then classification rule of C4.5, the following categories can be obtained:

The first one is that most students scored higher in math if their parents were separated. The second one is that if the same math score is good, the average score of the students whose parents live apart is higher than that of the students whose parents live together. Based on the exploration and discovery of the above questions, we can also use this method to find the conclusions drawn from other factors that affect students' math achievement: The first one is that most of the students have family support but those who don't have performed better in Math exam. The second one is that students who do not have educational support performed better in Math exam and especially female candidates. The third one is that those who do not have family support and not looking for higher education are scoring least in the Math exam. The forth one is that.those who are not having family support but looking for higher education have the highest scores in Math exam. The fifth one is that those who are looking for higher edu. are performing much better that those who are not looking for higher edu. Then, according to the data calculated from the decision tree model constructed above, we obtained the following four table graphs and conclusions by using python [7].

DOI: 10.54254/2753-7048/11/20230750

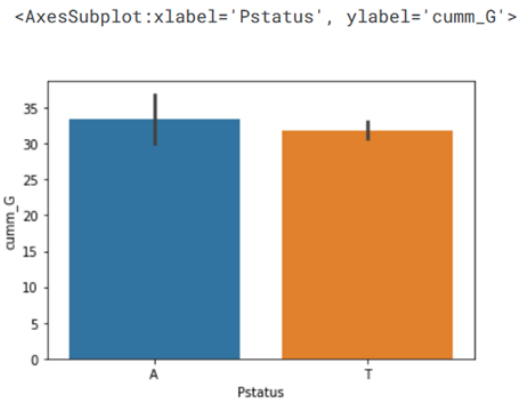

Figure 6: The influence of parents' cohabitation on grades.

From the fig 6 , we can find that students whose parents live apart do better in school.

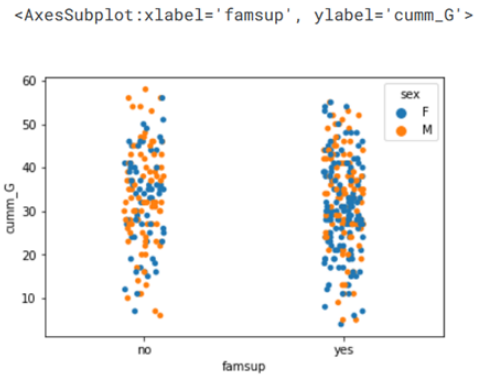

Figure 7: The influence of family support on grades.

From the fig 7, we can find that the math scores of students without family support are slightly better than those with family support.

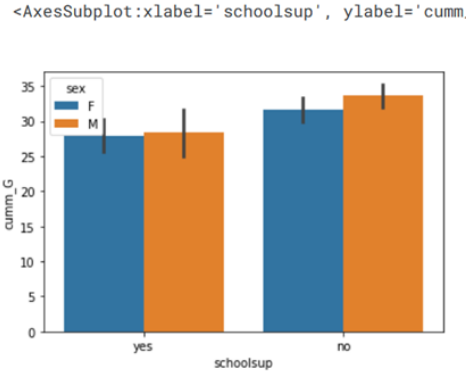

Figure 8: The influence of school support on grades.

From the fig 8, we can find that students without educational support perform better in math tests.

DOI: 10.54254/2753-7048/11/20230750

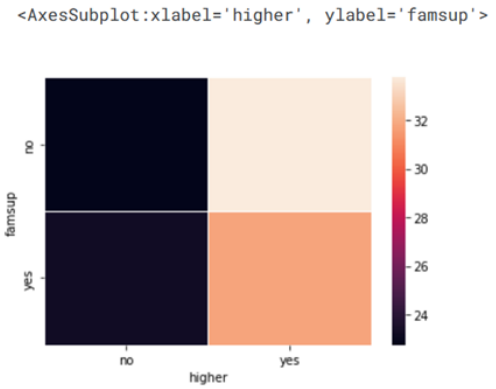

Figure 9: Chart of the influence of students' desire for higher education on their academic performance.

From the fig 9, we can see that those who seek higher education without family support perform better.

4. Conclusion

The topic of the study is to analyze the impact of students' family situations and educational support on their math performance by using the C4.5 algorithm in the decision tree algorithm as an auxiliary tool. And the project finally came to this interesting conclusion: motivated students with poor parental relationships and little educational support from school and family tend to do better in math.

The research still have lots of questions. Due to the limited ability of individuals, this study also has certain limitations. First and foremost, this study analyzed the entire group of students' math scores without taking individual differences into account; the data cannot cover all cases; the decision tree model has some empty branches; and the study's results are not comprehensive; additionally, the previous study of presuppositions was not comprehensive, considering only the default five factors; and there may still be important factors that have not been considered.

And in the future, this topic still has much work to do. This study only used one data mining method and was conducted using course result analysis. Many methods, such as association rules in data mining and clustering, can be combined with actual teaching to improve students' abilities. It is also more important to improve teaching quality; this is also one of the most important directions for future research.

Acknowledgement

First of all, I would like to thank Professor Shlomotaasan for teaching me both in mathematics and computer knowledge, which has benefited me a lot. Let me have a new cognition and further understanding and learning of the field of data science.

Secondly, I would like to thank my teaching assistant for their help. His patient answers helped me, a novice, adapt to the course of data science more quickly, whether it was about academic questions or about the data and format in the paper.

Next, I would like to thank my classmates. As a novice in data science, I have consulted other students for a lot of knowledge that I do not know, and they are very enthusiastic , which make me feel pleasant.

Finally, I would like to thank the teachers and parents who helped me to complete my learning tasks.

DOI: 10.54254/2753-7048/11/20230750

References

[1] *P. Cortez and A. Silva. April, 2008.Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12. https://www.kaggle.com/datasets/janiobachmann/math-students*

[2] *IT23131, October, 2021, The watermelon of the decision tree, https://blog.csdn.net/IT23131/article/details/121068259*

[3] *Love_YourSelf, Novermber, 2022, Classification based on decision tree, https://blog.csdn.net/qq_48068259/article/details/127640315*

[4] *Nine door data analysis research center, in June, 2021, the advantage of decision tree, http://www.jiudaomen.com.cn/question/newsdetail_68974687.html*

[5] *Hu Mingming.,2008, Research on the Application of Decision Tree Algorithm in the Analysis of Students' course scores,22-27. Master's Thesis, Harbin Normal University.*

[6] *Hu Mingming.,2008, Research on the Application of Decision Tree Algorithm in the Analysis of Students' course scores,17-18. Master's Thesis, Harbin Normal University.*

[7] *Yogesh Sachdeva, Ayushi Arora and Kriti Suri，January,2022,Data Exploration & Visualization Project on Student-Mat dataset using Python. https://www.kaggle.com/code/yogeshsachdeva223/student-mat-exploration-and-visualisation/notebook*