Proceedings of the 4th International Conference on Educational Innovation and Philosophical Inquiries
DOI: 10.54254/2753-7048/11/20230743

Analysis of the Influence of Parents' Educational Level on Students' Test Scores

— Based on Decision Tree

Peixun Huang[1,a,*]

[1]*School of Safety Science and Engineering, Nanjing Tech University, Nanjing, 211816, China*
a. henry98x@163.com
**corresponding author*

Abstract: Under the background of the "double reduction" policy promulgated in July 2021 and the "People's Republic of China Family Education Promotion Law" promulgated in January 2022, improving family education and promoting home-school cooperation have become the focus of China's education reform and the key to improve the quality of students. The dataset of this paper comes from the Kaggle website. By using data visualization technology and ID3 algorithm of decision tree, the influence of various indicators on the accuracy of the decision tree prediction model is compared, and the relationship between parents' education level and students' test scores is further explored. The final results show that the education level of parents has a certain correlation with the test scores of children, and the students whose parents have higher education level generally have better scores.

Keywords: students' test scores, parents' education level, decision tree, data analysis, influencing factors

1. Introduction

The excessive student burden is a well-known issue that has persisted for a long time and has yet to be resolved in China's basic education stage, which has deep roots in the system, society, and culture [1]. In July 2021, the "double reduction" policy was officially introduced, which further clarified the promotion and implementation of burden reduction work for primary and secondary school students [2]. On January 1, 2022, the "People's Republic of China Family Education Promotion Law" was promulgated and implemented, which also emphasizes the importance of family education in today's society [3]. Under this dual background, the improvement of family education and the promotion of home-school cooperation have become the focus of education reform and the key to enhance students' quality [4].

In this work, a prediction model is constructed by using ID3 algorithm of decision tree, and the importance of the indicators affecting students' test scores is sorted and analyzed [5]. In addition, python data visualization technology is used to further explore whether there is a link between the education level of the parents and their children's test scores. This research can help improve parents' attention to children's education, improve the level of family education, and help promote home-school cooperation and improve the quality of teaching [6].

DOI: 10.54254/2753-7048/11/20230743

2. Methodology

2.1. Theories

2.1.1. Decision Tree and ID3 Algorithm

Decision tree is a kind of tree structure, which is a basic technique for classification and regression. This paper only discusses the decision trees for classification. If-then statements can be used to indicate the classification of instances based on features in classification problems, where each leaf node stands in for a category, each branch for a test output, and each internal node for a test on an attribute [7].

As a traditional decision tree approach, ID3 chooses the best test attributes based on information entropy, chooses the test attributes that have the highest information gain value in the current set of samples, and builds the decision tree recursively. For the ID3 algorithm, we need to understand three important theories, namely Shannon entropy, empirical conditional entropy, and information gain [8]. With these concepts in mind, we can recursively build our decision tree based on information gain.

2.1.2. Shannon Entropy

Shannon entropy, or simply entropy, is a measure of aggregate information.

The expected value of the information is defined as entropy. Shannon entropy is a extent of the random variable's degree of uncertainty in information theory and probability statistics. If there are many categories into which the objects to be classed can be subdivided, the information of c_i is defined as follows:

$$l(c_i) = -\log_2 p(c_i) \tag{1}$$

$p(c_i)$: the likelihood of selecting the category.

We can obtain all kinds of information using the above formula. The following formula can be used to determine the expected value of information (mathematical expectation) included in all potential values for all categories, which is necessary to compute Shannon entropy:

$$H = -\sum_{i=1}^{n} p(c_i) \log_2 p(c_i) \tag{2}$$

2.1.3. Empirical Conditional Entropy

Empirical conditional entropy $H(D|A)$ indicates that in the case of known of random variable A, the random variable D's degree of uncertainty. In the case of given random variable A, empirical conditional entropy $H(D|A)$ of random variable D, defined as when A is given, the conditional probability distribution of D on A's expected mathematical entropy:

$$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|} \tag{3}$$

2.1.4. Information Gain

Making unordered data more ordered is the basic goal of partitioning data sets, but each approach has pros and cons of its own. A field of study known as information theory deals quantitatively with information. Information gain refers to the difference between the information before and after the data collection is divided. You must first understand how to compute the information gain since the feature with the maximum information gain is the best option.

DOI: 10.54254/2753-7048/11/20230743

The information gain of feature A on training dataset D $g(D, A)$ refers to the extent to which the information uncertainty of class D is lowered after knowing the information of feature A. It is described as the difference between empirical conditional entropy $H(D|A)$ of D under certain conditions of feature A and Shannon entropy $H(D)$ of D in the dataset.[9]:

$$g(D, A) = H(D) - H(D|A) \tag{4}$$

So how do this paper judge how much each feature in the dataset affects the accuracy of the decision tree prediction model?

First, we calculate the average score of each student in the dataset and classify the average score, as shown in the table 1 below:

Table 1: Average score classification chart.

Average Score	0~60	60~70	70~80	80~90	90~100
Rating	F	D	C	B	A

Then, we quantified the categories of each feature in the dataset, such as feature 'race/ethnicity': 'group A' we use 0 instead, 'group B' we use 1 instead, 'group C' we use 2 instead, etc., and normalized them.

Then get rid of a certain feature of the dataset, respectively, calculate the accuracy of the decision tree model

Finally, the accuracy of each decision tree prediction model is compared. The model with lower accuracy means that the feature removed from the model has greater influence on the prediction, and the feature is more important [10].

2.2. Data Sources

The dataset used in this paper is from Kaggle website, which has 1000 samples (518 girls and 482 boys). The students' scores include math, reading and writing scores. Other information includes 'race/ethnicity', 'lunch', 'parental level of education' and 'test preparation course'. There are two types of 'lunch': 'standard' and 'free/reduced', and 'test preparation course' indicates whether or not the student has finished the test preparation course. We will use decision tree to study and compare these four features.

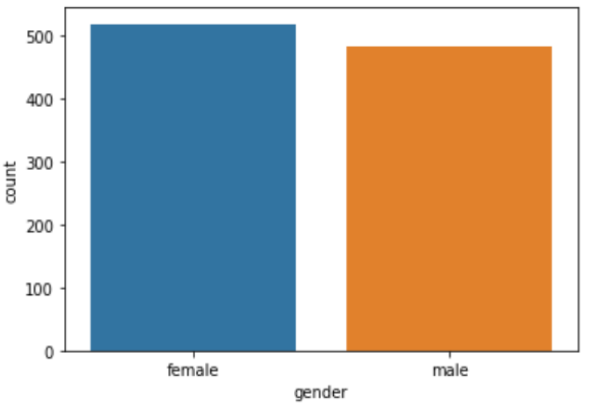

Figure 1: Males and females' samples quantity statistical figure.

DOI: 10.54254/2753-7048/11/20230743

3. Experiment Results

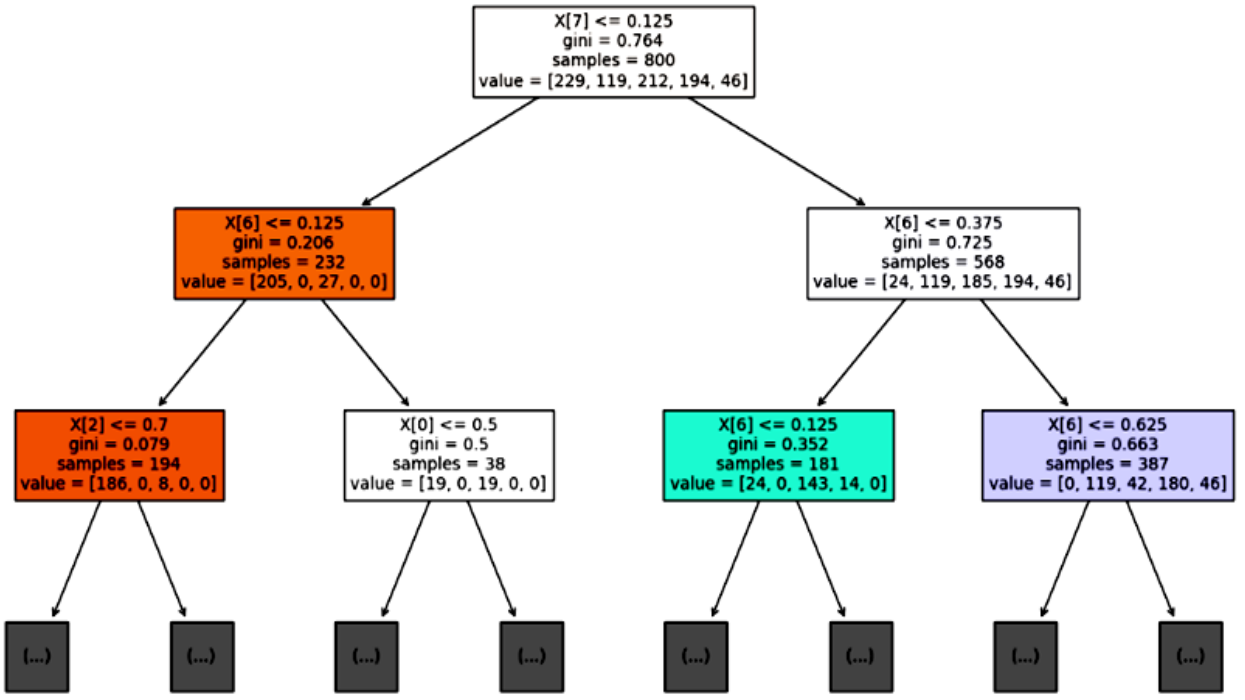

Figure 2: Student test score prediction - decision tree model.

Based on a 4:1 ratio, we separated the data set into the training set and the test set, and used python visualization tool to obtain the decision tree model structure of this paper (Fig. 2). This model is a partial decision tree model with max_depth = 3, including all features, and the prediction accuracy is 86%.

Table 2: Comparison of decision tree prediction accuracy.

Remove	race/ethnicity	lunch	parental level of education	test preparation course
Accuracy	83%	84%	82.50%	79.50%
Decline	↓3%	↓2%	↓3.5%	↓6.5%

By using the methods mentioned above, the experiment's outcomes obtained are displayed in Table 2. It is clear from the figure that whether students participate in test preparation course has the greatest impact on the prediction accuracy (decreased by 6.5%), followed by the education level of parents (decreased by 3.5%).

So there is a more obvious relationship between examination results and completion of test preparation course. However, the influence of parents' education level cannot be ignored, as it is the second most influential factor, and to prove this, we calculated the average scores of each subject and the final scores of children of parents with different education levels, and used the python visualization tool to obtain the experimental results as shown in Fig.3:

DOI: 10.54254/2753-7048/11/20230743

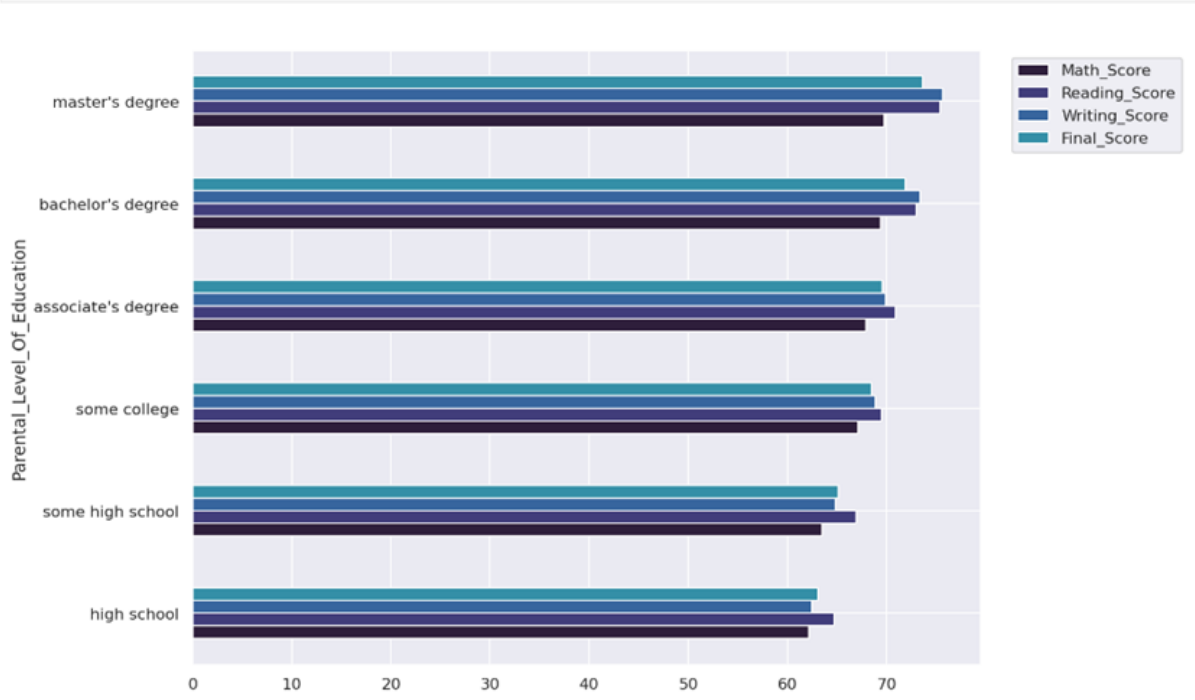

Figure 3: Comparison of children's test scores of parents with different educational levels.

From the figure above, we can clearly see that the children from families with higher education have higher scores. Based on the above experimental results, we can conclude that nowadays, students' test scores have a more direct relationship with whether students prepare for exams carefully. However, the education level of parents also has a certain impact on the examination results of students; the children from families with higher education have higher scores. The first reason for these phenomena may be that well-educated parents pay more attention to the education of their children, and the second may be that well-educated parents know more about how to raise their kids.

According to Figure 3, we can also judge that there is still a problem of class solidification in education. Parents with low education level cannot guide their children's growth correctly. So the government needs to pay attention to the problem of educational inequality, so that every child can get a comprehensive training,

4. Conclusion

In this work, we used the decision tree ID3 algorithm and considered the data normalization principle to explore the relationship between parents' education level and students' test scores. We use the removal of a feature to observe its impact on the prediction accuracy of the whole model, determine the importance of this feature, and draw the following conclusion: Parents' education level has a certain impact on students' test scores, and the children of well-educated parents tend to do better in exams. As for the limitations of the paper, the machine learning algorithm is not only a decision tree, but only a decision tree method is used in this paper. We can also use other algorithms, such as KNN or Bayes algorithm, to explore this kind of problem, and see if there are different results. Besides, for the future, the influencing factors of family education include not only the education level of parents, but also other aspects, such as the degree of harmony between parents and the economic conditions of the family, we hope to find or collect more comprehensive data to conduct research and analysis on the impact of family education.

DOI: 10.54254/2753-7048/11/20230743

References

[1] *She Yu, Que Mingkun, Yang Kaiyong & Shan Dasheng.(2022). Student burden management in basic education in China: "Double Reduction" policy and long-term mechanism construction. Management World (07),163-170. doi:10.19744/j.cnki.11-1235/f.2022.0090.*

[2] *Zhao Keli & Wu Xiaohong.(2022). Review and improvement path of homework burden reduction in primary and secondary schools under the background of "double reduction". Education for Creative Talents (04),61-66.*

[3] *(2022-07-04). Promote the construction of family tutoring and family style to improve the family education level under the "double reduction" policy. Siping Daily, 006.*

[4] *Long Baoxin & Li Haiying.(2022). The transformation and implementation of Home-school Co-education thinking under the background of "double reduction". Journal of suzhou university (education science edition) (03), 29 to 37, doi: 10.19563 / j.carol carroll nki SDJK. 2022.03.002.*

[5] *Yu Shuyun. (2021). Research on online teaching learning effect prediction model based on ID3 algorithm Journal of Lanzhou University of Arts and Sciences (Natural Science Edition) (03), 110-114. doi: 10.13,804/j.cnki.2095-6991.2021.03.022.*

[6] *Li Longzhen. (2021). Online learning performance prediction based on decision tree algorithm Information Technology and Informatization (01), 130-133.*

[7] *Zhu Yao. (2021). License plate number recognition based on decision tree model Practical Technology of Automobile (22), 222-225. doi: 10.16638/j.cnki.1671-7988.2021.057.*

[8] *Wang Huiqing, Chen Junjie, Hou Xiaojing&Guo Kai. (2011). Research on attribute selection method of decision tree classification Journal of Taiyuan University of Technology (04), 346-348+352. doi: 10.16355/j.cnki.issn1007-9432tyut.2011.04.021.*

[9] *Wang Jingxiang. (2022). Principle research and practical application of decision tree algorithm Computer Programming Skills and Maintenance (08), 54-56+72. doi: 10.16184/j.cnki.comprg.2022.08.043.*

[10] *Sun Xiaoxue, Zhong Hui&Chen Haipeng. (2021). A statistical analysis system for students' test scores based on decision tree classification technology Journal of Jilin University (Engineering Edition) (05), 1866-1872. doi: 10.13229/j.cnki.jdxbgxb20210455.*