Hacettepe Üniversitesi Eğitim Fakültesi Dergisi

Hacettepe University Journal of Education

e-ISSN: 2536-4758

Investigation of Variables Explaining Mathematics Literacy in PISA 2018 Turkey and China Samples Through C4.5 Decision Tree Algorithm*

Zeynep Begümhan ÖZCAN, Sevda ÇETİN*****

Article Information

Received: 21.03.2024

Accepted: 31.10.2024

Online First: 31.10.2024

Published: 31.10.2024

ABSTRACT

The aim of this study is to identify the factors that most affect students' mathematical literacy in the PISA 2018 Turkey and China samples using the C4.5 algorithm. The analyses were conducted on the datasets of 5959 students from the Turkey sample and 12058 students from the China sample. The independent variables of the study are 22 student questionnaire items that are considered depending on the constructs of self-perception, adaptability, school belonging, family support and school environment. After organising the data set, the consistency ratio for the dependent variable, mathematical literacy, was calculated for both samples to evaluate the performance of the C4.5 classification. The decision tree method was then applied and reported separately for each sample. As a result of the analyses, the items that had the most impact on mathematics performance in the Turkey PISA 2018 application were determined as "I am proud that I have achieved something", "My parents support my educational efforts and achievements", and "I feel that I can manage many things at the same time" under the variables of self-perception and family support. In the Chinese sample, the items that most affected mathematics performance were “My parents support my educational efforts and achievements” and “Students seem to compete with each other” under the variables of family support and school environment.

Keywords: Data Mining, C4.5 Algorithm, PISA 1018, Mathematical Literacy

doi: 10.16986/HUJE.2024.531 Article Type: Research Article

Citation Information: Özcan, Z. B., & Çetin, S. (2024). Investigation of variables explaining mathematics literacy in PISA 2018 Turkey and China samples through C4.5 decision tree algorithm. *Hacettepe University Journal of Education, 39*(4), 378-390. https://doi.org/10.16986/HUJE.2024.531

1. INTRODUCTION

Mathematical literacy can be defined in various ways, such as the ability to use basic procedural and geometric skills in daily life situations, understanding fundamental mathematical concepts, the capacity to develop sophisticated mathematical models, or the ability to comprehend and evaluate the way others use numbers and mathematical models (Jablonka, 2003). In PISA 2012, mathematical literacy is defined as the ability to formulate, use and interpret mathematics in different contexts. Mathematical reasoning involves using mathematical expressions and operations to explain and predict phenomena. Mathematical literacy enables individuals to understand the role of mathematics in their lives and use it consciously to meet their needs (Yılmaz et al., 2011).

By using the data obtained within the scope of PISA, many analyses related to mathematical literacy can be carried out, such as observing the changes in countries over the years, examining the links between mathematical literacy and other field performances, examining the changes in performance of subgroups by using the responses to student questionnaires and identifying the factors affecting performance. Data mining applications are also one of the methods frequently used to identify patterns in the PISA data set, which is quite large.

Data mining is the analysis of large data sets to find meaningful patterns and rules. Linoff and Berry (2011) define data mining as the discovery of meaningful paths, patterns and rules from large amounts of data. Data mining methods are specifically designed to work with large data sets and give the most accurate results in large data. The basic idea of data mining is to efficiently combine the data processing power of the computer with the ability of the human eye to find patterns (Pujari, 2001).

* This study is based on the first author's thesis supervised by the second author.

** Hacettepe University, Faculty of Education, Department of Educational Sciences, Division of Educational Measurement and Evaluation, Ankara-TÜRKİYE. e-mail: begumhanokyay@gmail.com (ORCID: 0009-0006-4396-6325)

*** Assoc. Prof. Dr., Hacettepe University, Faculty of Education, Department of Educational Sciences, Division of Educational Measurement and Evaluation, Ankara-TÜRKİYE. e-mail: tsevda@hacettepe.edu.tr (ORCID: 0000-0001-5483-595X)

In other words, data mining involves methods of converting large amounts of data into useful information. Many different patterns can be discovered with data mining techniques.

The field of educational data mining has emerged with the application of diversified analysis methods to data collected in the field of education and shaped according to the structure of these data. Educational data mining focuses on developing methods to discover certain types of data from information collected in the context of education. This data can be obtained from traditional face-to-face classroom settings, educational software, online courses and large-scale tests. The search for previously undetected causal relationships is of great importance in the development of educational programmes.

With the acceleration of technological developments, the amount of data collected through various sources has reached larger and larger dimensions, leading to the diversification of data analysis and interpretation methods and making this diversification possible. All these changes have led to the development of data mining, which basically aims to identify interesting patterns, models and many other information in large data sets (Han et al., 2001). Decision trees, one of the data mining methods, are tree-shaped structures similar to flow diagrams. Nodes represent the features being tested, branches represent the output of the test and leaves represent classes or class distributions. The decision tree algorithm illustrates a hierarchical method of sample classification, starting from the root node and progressing downwards. At each internal node, a test is conducted to determine the path to traverse, eventually leading to a leaf node where a label class is assigned based on the outcome of the traversal (Yu et al., 2010). The process in the decision tree is to transform the data table into a tree model and then convert the tree model into rules. Decision trees can be easily converted into classification rules. One of the main benefits of decision trees is that knowledge can be extracted and represented as classification rules. Each rule represents a unique path from the root to each leaf. Decision trees are models that allow the use of categorical and numerical variables, and at higher levels they allow to identify the results of the interactions of certain subgroups (Castro & Lizasoain, 2012).

C4.5, an extended version of the ID3 algorithm, is the most frequently used algorithm in decision trees. This algorithm, which can work with continuous data, generates decision trees from a training data using the concept of information entropy (Witten et al., 2011). The C4.5 algorithm selects the feature with the highest information gain to determine the root of the decision tree and then proceeds to repeatedly split the data based on the remaining features (Quinlan, 1986).

1.1. Statement of the Problem

Within the scope of PISA, in addition to questions on knowledge and skills at different cognitive levels, students are also asked about their motivation, views about themselves, psychological characteristics related to learning processes, school environments and families (MoNE, 2019). This application provides a very large and diverse database on the students included in the assessment. The diversity of the data collected through the PISA application enables many analyses to be made with this data. The fact that the data collected from the samples in OECD countries include not only the answers given to the questions prepared to measure knowledge and skills, but also the answers given to the questionnaire questions prepared to understand the student profile opens a space for using data mining methods and searching for cause-effect relationships on the data.

There have been many studies of the published PISA data sets, trying to identify the factors that affect performance (Boman, 2023; Caponera, 2019; Demir & Kılıç, 2010; Dev, 2020; Güzel İş & Berberoğlu, 2010; Güzel İş, 2014; Pinto et al., 2016; Predic et al., 2018). In this study, it is aimed to examine the effects of self-perception, adaptability, school belonging, family support and school environment variables on mathematical literacy.

Self-perception refers to an individual's emotional response to both positive and negative experiences and information encountered throughout their life. It is explained by two sub-dimensions: self-efficacy and self-esteem (Demoulin, 1998). The PISA questionnaire measures adaptability through items such as changing attitudes when faced with new situations, adapting to different situations under stress and pressure, easily adjusting to new cultures, finding a middle ground in solving difficult situations involving different people, and resolving difficulties involving other cultures. The variable of school belonging refers to aspects such as students' sense of social bonding, including feelings of comfort, friendship and lack of loneliness while at school. In a study conducted using PISA 2012 data, it was found that school belonging was a strong predictor of performance (Aydıner & Kalender, 2015). Similarly, another study using PISA 2018 data revealed that school belonging is related to academic achievement (Hofer et al., 2024). The PISA questionnaires measure the family support variable through items that assess parents' support for school-related achievements, assistance with school-related difficulties, and encouragement of student self-confidence. While no study in the literature directly addresses the relationship between student achievement and family support, parental involvement is thought to be associated with achievement (Caponera, 2019). The variable measuring the school environment is assessed through items that focus on the competitive nature of the school environment in PISA student surveys. Research examining the relationship between school environment and academic performance has shown a significant relationship between the two (Gomez & Suarez, 2020; Yıldırım et al., 2017).

The purpose of selecting these variables, which are also mentioned in the literature, is to investigate whether there is a significant difference in mathematics performance based on the variables related to students' school bonding and family support.

1.2. Purpose of the Study

The aim of the study is to is to examine the variables that explain mathematical literacy in PISA 2018 Turkey and China samples with the C4.5 algorithm. For this purpose, in addition to the data from Turkey, the data from China, which has the highest mean score in 2018 PISA mathematics literacy, will also be discussed. Reviewing the PISA mathematics literacy data from 2009, 2012, 2015, and 2018, China participated in PISA applications with various regions, including Beijing, B-S-J-G (Beijing, Shanghai, Jiangsu, and Guangdong), and B-S-J-Z (Beijing, Shanghai, Jiangsu, and Zhejiang) and China's mathematics literacy performance in all regions remained above the OECD mean score.The B-S-J-Z region, which was included in the scope of the research, was included for the first time in the 2018 PISA application and became the participant with the highest average mathematical literacy performance.

The 2018 PISA application, in which current data were collected, was applied to students from 79 countries in total and the data collected together with the success rankings of these countries in different fields were published by the OECD. In the 2018 application, Turkey remained below the PISA average with a mean score of 454 in the field of mathematical literacy (OECD, 2019). Similarly, in the 2015 application, Turkey was below the OECD average (OECD, 2016). Analysing the collected data and identifying causal links in the data is very important as it can play a role in the development of educational policies. Focusing not only on Turkish data, but also on data from a country that has consistently scored above the average in PISA mathematical literacy, provides a different perspective on the meaning of the variables studied.

Mathematical literacy scores and averages are calculated with the achievement tests applied within the scope of PISA. In addition, detailed information about the student profile can be obtained through questionnaires. In comprehensive questionnaires, data on many different variables such as socioeconomic status, school climate and teacher attitudes are collected. There are many studies examining the relationship between socioeconomic status and achievement using PISA data. (Aydın et al., 2012; Dolu, 2020; Jehangir et al., 2015; Karaağaç&Cingöz, 2020; Perry&Mcconey, 2010; Schulz, 2015). Furthermore, although there are few in number, there are also studies that analyse data collected through PISA student and school questionnaires (Aydıner & Kalender, 2015; Hofer et al, 2024; Caponera et al, 2019; Dev, 2020; Gomez & Suarez, 2020; Ötken, 2021; Predic et al. 2018; Yıldırım et al., 2017). Although the relationship between self-perception, adaptability, school belonging, family support, and school environment and performance were examined separately, they were not examined together. In addition, the effect of these variables on performance has been pointed out in indirect studies but not directly addressed. Therefore, within the scope of this study, the relationship between self-perception, adaptability, school belonging, family support and school environment variables and mathematical literacy will be examined.

The C4.5 algorithm was chosen as the model because it allows for the use of continuous variables and is more advanced than other decision tree algorithms that use continuous variables, such as ID3. With the outputs obtained as a result of the model applied separately for two countries, the factors affecting mathematical literacy will be analysed.

1.3. Problem of the Study

The aim of this study is to contribute to the literature by interpreting the effects of independent variables, including self-perception, adaptability, school belonging, family support, and school environment, on mathematical literacy levels in Turkey and China. The variables will be evaluated and compared in these two countries, which have different performance rankings and have maintained this trend.

For this purpose, an answer to the following question will be sought.

What are the effects of independent variables such as self-perception, adjustment, school belonging, family support and school environment on mathematical literacy levels in Turkey and China?

1.3.1. Sub-problems of the study

The general purpose of the study is to examine the variables that explain mathematical literacy in PISA 2018 Turkey and China samples with the C4.5 algorithm. In line with the determined purposes, answers to the following questions are sought within the scope of this study:

Which variables respectively explain mathematical literacy levels in the PISA 2018 Turkey sample?
Which variables respectively explain mathematical literacy levels in the PISA 2018 China sample?

2. METHODOLOGY

Within the scope of the present study, the variables explaining mathematical literacy in PISA 2018 Turkey and China samples will be analysed using the C4.5 decision tree algorithm. In this sense, this research on understanding and defining the existing situation will be carried out using descriptive analysis with a quantitative approach (Büyüköztürk et al., 2018).

2.1. Participants

A total of 12,058 students from 362 schools in China, representing 992,302 students aged 15 in B-S-J-Z (China), were included in the study (OECD, 2019). In the Turkey sample, a total of 5959 students from 186 schools representing 884,971 15-year-old students were included in the study (OECD, 2019). Table 1 shows demographic characteristics of the samples.

Table1.
Demographic Characteristics of the Samples

China Sample	
N	12058
Female	51 %
Male	49 %
10th Grade	56 %
11th Grade	29 %
12th Grade	15 %
Turkey Sample	
N	6890
Female	49.6 %
Male	50.4 %
9th Grade	17.7 %
10th Grade	78.8 %
11th Grade	2.9 %
Other Grade	Below 1 %

As seen in table 1 in Turkey sample, 49.6% of the sample consisted of female students and 50.4% of the sample consisted of male students. In terms of gender groups, there is a fairly balanced distribution in the Turkish sample. When the grade level distribution of the students in the sample is analysed, it is determined that 78.8% of the students in the sample continue their education in 10th grade, 17.7% in 9th grade and 2.9% in 11th grade. The total proportion of students in other grade levels is below 1% (MoNE, 2019). In the B-S-J-Z (China) sample, 49% of the sample consisted of female students and 51% of male students. Of these students, 56% were 10th grade students, 29% were 11th grade students, and 15% were 12th grade students (OECD, 2019).

2.2. Data Collection Process

The data used in the study were obtained from the OECD database in SPSS data format (https://www.oecd.org/pisa/data/2018database/). The data include the questionnaires and test results of the students who participated in the PISA 2018 assessment in China and Turkey. After the data for all countries were downloaded, they were transferred to the R Studio programme and the relevant variables (self-perception, adaptability, school belonging, family support and school environment) and the data except for the countries to be examined within the scope of the research were deleted from the data set.

2.3. Data Collection Tools

In the study, student questionnaires and mathematical literacy test used in the PISA 2018 application were used. The data used in the research were downloaded from the OECD database. The data set used includes data on self-perception, adaptability, school belonging, family support and school environment variables, modules on mathematics performance and country codes.

Within the scope of the study, the extent to which the items belonging to the variables of self-perception, adaptability, school belonging, family support and school environment affect mathematics performance was investigated by using the questionnaire and data collected through the PISA 2018 application. Table 2 shows all variable codes, related headings, item text in the decision trees created with the C4.5 algorithm for Turkey and China samples:

Table 2.
Variables Included in Decision Trees

Heading	**Code**	**Item**
School Belonging	ST034Q01TA	"I feel like an outsider (or left out of things) at school"
	ST034Q02TA	"I make friends easily at school"
	ST034Q03TA	"I feel like I belong at school"
	ST034Q04TA	"I feel awkward and out of place in my school"
	ST034Q05TA	"Other students seem to like me"
	ST034Q06TA	"I feel lonely at school"
Family Support	ST123Q02NA	"My parents support my educational efforts and achievements"
	ST123Q03NA	"My parents support me when I am facing difficulties at school"
	ST123Q04NA	"My parents encourage me to be confident.

Self-perception	ST188Q01HA	"I usually manage one way or another"
	ST188Q02HA	"I feel proud that I have accomplished things"
	ST188Q03HA	"I feel that I can handle many things at a time"
	ST188Q06HA	"My belief in myself gets me through hard times"
	ST188Q07HA	"When I'm in a difficult situation, I can usually find my way out of it"
School Environment	ST205Q01HA	"Students seem to value competition"
	ST205Q02HA	"It seems that students are competing with each other"
	ST205Q03HA	"Students seem to share the feeling that competing with each other is important"
	ST205Q04HA	"Students feel that they are being compared with others"
Adaptability	ST216Q01HA	"I can deal with unusual situations"
	ST216Q03HA	"I can adapt to different situations even when under stress or pressure"
	ST216Q04HA	"I can adapt easily to a new culture"
	ST216Q05HA	"When encountering difficult situations with other people, I can think of a way to resolve the situation"

2.4. Analysing the Data

The two sub-problems of the research are identical in terms of the analysis methods and variables used. The dependent variable used within the scope of the research is mathematical literacy performance and to obtain this value, the values of the module scores PV1MATH-PV10MATH were averaged. The independent variables were self-perception, adaptability, school belonging, family support and school environment. For the analysis of the data, firstly, the data that would not be used were removed from the downloaded data set. The raw data downloaded from the OECD database is the SPSS data named "Student questionnaire data", which consists of 1120 columns and 612,004 rows and contains all the questionnaire and performance data collected in the PISA 2018 application, country codes and all other data that can be used in related research. After the editing process, a new data set consisting of 18017 rows and 26 columns was obtained. Prior to running the analyses, missing data analysis was performed to ensure that the algorithms worked more efficiently and since the missing data was less than 5%, all rows with missing data were deleted and a new dataset was created.

While selecting the data mining method to be used within the scope of the research, the studies in the literature were examined in terms of the methods used and the selected algorithm was selected according to the characteristics of the variables used. C.4.5 algorithm, which is one of the decision tree methods, was used in the research with its ability to analyse complex data in an understandable and easy way. The selected C4.5 algorithm is one of the applications used in the literature to analyse PISA data. C4.5 algorithm, one of the most frequently used algorithms in decision trees, can work with continuous variables and creates decision trees using training data (Osmanbegović & Suljić, 2012). The C4.5 algorithm, which is the continuation of the ID3 Algorithm, is the same as the ID3 algorithm in terms of process steps, but while the ID3 algorithm cannot analyse numerical data, the C4.5 algorithm can also analyse numerical data (Karaboga et al., 2022).

In order to evaluate the performance of the C4.5 classification, the data consistency ratio, which is obtained by dividing the number of correctly classified data predictions by the number of all prediction data, was calculated. The formula used for the consistency calculation (Karaboğa et al., 2022) is given below:

$$\text{Consistency} = \frac{\text{Number of Correctly Classified Predictions}}{\text{Total Number of Predictions}}$$

Data analysis was carried out using R programming language and RStudio IDE. R is a programming language accessible through CRAN (Comprehensive R Archive Network). RStudio is frequently used in research in the field of education thanks to its open source library and different function packages, allowing R applications to be made

3. FINDINGS

Within the scope of the study, using the questionnaire and test data collected through the PISA 2018 application, the extent to which the items belonging to the variables of self-perception, adaptability, school belonging, family support and school environment affect mathematics performance was investigated.

3.1. Findings related to the first sub-problem-Turkey Sample

Table 3.
Classification Table for Success Status – Turkey Sample

Observed	Predicted		
	Successful	Unsuccessful	Success Percentage
Successful	1717	1129	0.60
Unsuccessful	1141	1972	0.63
Total	3689	2270	0.62

As seen in Table 3, it is seen that 1717 (60%) of 2846 successful students were correctly classified in the model, but 1129 (40%) students were classified as unsuccessful despite being successful. Likewise, it is seen that 1972 (63%) of the 3113 unsuccessful students were correctly classified, while 1141 (37%) students were classified as successful despite being unsuccessful. The overall success of the algorithm in classifying successful and unsuccessful students is 62%. The margin of error of the algorithm is 38%.

A decision tree was created with the C4.5 algorithm and the properties of the created decision tree were analysed. This decision tree is given at https://github.com/svdctn/decision-tree-/blob/1b99a81b0b4da39c78e4f270c3c69111aae0d574/T%C3%BCrkiye_c45.pdf. Table 4 shows which independent variables are most frequently used in the decision tree and the percentages of their use, together with the explanation of the relevant independent variable:

Table 4.
Percentage of Use-Turkey Sample

Item Code	Relevant Independent Variable	Item	Percentage of Use
ST188Q02HA	Self-perception	"I feel proud that I have accomplished things"	100.00
ST123Q02NA	Family Support	"My parents support my educational efforts and achievements"	97.62
ST205Q04HA	School Environment	"Students feel that they are being compared with others"	71.45
ST188Q03HA	Self-perception	"I feel that I can handle many things at a time"	55.38
ST216Q01HA	Adaptability	"I can deal with unusual situations"	46.15
ST034Q05TA	School Belonging	"Other students seem to like me"	45.24
ST205Q02HA	School Environment	"It seems that students are competing with each other"	42.24
ST216Q03HA	Adaptability	"I can adapt to different situations even when under stress or pressure"	35.09
ST188Q01HA	Self-perception	"I usually manage one way or another"	23.26
ST205Q01HA	School Environment	"Students seem to value competition"	12.23
ST034Q03TA	School Belonging	"I feel like I belong at school"	10.15
ST216Q04HA	Adaptability	"I can adapt easily to a new culture"	9.23
ST205Q03HA	School Environment	"Students seem to share the feeling that competing with each other is important"	2.82
ST188Q06HA	Self-perception	"My belief in myself gets me through hard times"	1.49

In Table 4, it is seen that the self-perception variable with the item "I am proud of achieving something" is the most frequently used item in the decision trees created with the C4.5 algorithm. ST188Q02HA item was used in all decision trees created to create the final decision tree. The fact that it was found in all of the repetitions made to finalise the model means that the answers given to this item have an important effect on the classification of students' performence. Another of the most frequently used variables in decision trees is the item "My parents support my educational efforts and achievements" under the family support heading, which was used in 97.62% of the decision trees. It is seen that this item, which forms nodes in almost all of the decision trees, has a significant effect on the classification of mathematical literacy performance. The variables that were used more than 50% frequently in the decision tree and had a significant effect on categorisation were the item "Students feel that they are being compared with others" (71.45%) related to the school environment variable and the item "I feel that I can handle many things at a time" (55.38%) related to the self-perception variable.

Table 5.
Classification error percentages by node

Node	Classification	n	Percentage of Error
38. Node	1	8	12.5
11. Node	0	17	17.6
8. Node	0	51	21.6
2. Node	0	142	23.9
41. Node	0	54	24.1
20. Node	1	12	25
19. Node	0	26	26.9
21. Node	1	472	31.6
10. Node	1	151	33.8
13. Node	0	45	35.6
40. Node	0	59	35.6
32. Node	0	176	36.4
26. Node	0	441	37.2
36. Node	1	557	37.2
22. Node	0	40	37.5
14. Node	1	284	37.7
6. Node	0	1969	38
39. Node	0	81	40.7
34. Node	1	365	41.1
27. Node	1	164	45.1
33. Node	1	845	47.2

*0- unsuccessful, 1- successful

As seen in Table 5, at Node 38, a total of 8 students were classified as successful with an error rate of 12.5%. In Node 38, the responses to the statement "I feel proud that I have accomplished things" were strongly disagree, strongly agree to the statement "My parents support my educational efforts and achievements", positive (agree or strongly agree) to the statement "I feel that I can handle many things at a same time", positive (over-defines me, defines me well or defines me) to the statement "I can deal with unusual situations". The students who answered "I can adapt to different situations even when under stress or pressure " as "Does not describe me at all", "Students feel that they are being compared with others" as "True", "Other students seem to like me" as "Strongly disagree", " My belief in myself gets me through hard times" as "Strongly disagree and disagree" were classified as successful - 1.

In Node 8, the students who answered "I feel proud because I have accomplished something" with options other than disagree, agree or strongly agree, "My parents support my educational efforts and achievements" with options other than strongly agree, "It seems that students are competing with each other" with options other than not true at all, "Students feel that they are being compared with others" with extremely true, and "Students seem to value competition" with not true at all were classified as unsuccessful - 0.

In Node 41, a total of 54 students were classified as unsuccessful with an error rate of 24.1%. In Node 41, the students who gave answers other than strongly disagree for the statement "I feel proud that I have accomplished something.", strongly agree for the statement "My parents support my educational endeavours and achievements", positive (agree or strongly agree) for the statement "I feel that I can handle many things at a time", negative (does not describe me very much or does not describe me at all) for the statement "I can deal with unusual situations" were classified as failing- 0.

In the Turkey sample, the nodes that explain mathematical literacy the most are ST188Q02HA, ST123Q02NA and ST188Q03HA. ST188Q02HA item is the item "I feel proud tahat I have accomplished something" related to the variable of self-perception. ST188Q02HA item is located at the 1st node in the decision tree and is found in 100% of the decision trees that form the decision tree. When these statistics are evaluated, it is concluded that the item "I feel proud tahat I have accomplished something" related to the variable of self-perception is the most important factor explaining mathematical literacy in the Turkey sample.

ST123Q02NA item is the item "My parents support my educational endeavours and achievements" related to family support variable. ST123Q02NA item is located at the 3rd node in the decision tree and is found in 97.62% of the decision trees that form the decision tree. In line with these statistics, the item "My parents support my educational endeavours and achievements." is considered as the second most important factor explain mathematical literacy in Turkey sample.
ST188Q03HA item is the item "I feel that I can handle many things at a time" related to the variable of self-perception. ST188Q03HA item is located at Node 15 in the decision tree and is found in 55.38% of the decision trees that form the decision tree. The item "I feel that I can handle many things at a time", which was determined as the last intersecting variable in the decision trees, was evaluated as the third most important factor explaining mathematical literacy in the Turkey sample.

3.2. Findings related to the first sub-problem-China Sample

The data from China were classified using the C4.5 algorithm. Table 6 displays the results of this classification.

Table 6.
Classification Table for Success Status – China Sample

Observed	Predicted		
	Successful	Unsuccessful	Success Percentage
Successful	4892	1609	0.75
Unsuccessful	2992	2563	0.46
Total	7455	4601	0.62

In Table 6, it is seen that the model correctly classified 75% (4892) of the 6501 successful students, but 25% (1609) of successful students were incorrectly classified as unsuccessful. Similarly, 46% (2563) of the 5555 unsuccessful students were correctly classified, while 54% (2992) of unsuccessful students were incorrectly classified as successful. The algorithm's overall success rate in classifying successful and unsuccessful students is 62%, with a margin of error of 38%.

A decision tree was created with the C4.5 algorithm and the properties of the created decision tree were analysed. This decision tree is given at https://github.com/svdctn/decision-tree-/blob/1b99a81b0b4da39c78e4f270c3c69111aae0d574/China_c45.pdf. Table 7 shows which independent variables are most frequently used in the decision tree and the percentages of their use, together with the explanation of the relevant independent variable:

Table 7.
Percentage of Use-China Sample

Item Code	Relevant Independent Variable	Item	Percentage of Use
ST123Q02NA	Family Support	"My parents support my educational efforts and achievements"	99.24
ST205Q02HA	School Environment	"It seems that students are competing with each other"	54.88
ST034Q05TA	School Belonging	"Other students seem to like me"	44.39
ST123Q04NA	Family Support	"My parents encourage me to be confident"	30.20
ST205Q04HA	School Environment	"Students feel that they are being compared with others"	25.92
ST123Q03NA	Family Support	"My parents support me when I am facing difficulties at school"	24.85
ST188Q02HA	Self-perception	"I feel proud that I have accomplished things"	20.48
ST188Q06HA	Self-perception	"My belief in myself gets me through hard times"	18.21
ST188Q03HA	Self-perception	"I feel that I can handle many things at a time"	14.69
ST188Q01HA	Self-perception	"I usually manage one way or another."	9.73
ST034Q01TA	School Belonging	"I feel like an outsider (or left out of things) at school"	1.82

In Table 7 it is seen that that the variable "Family support", specifically the item "My parents support my educational efforts and achievements", is the most frequently used variable in the decision trees generated by the C4.5 algorithm. The other variable that was used more than 50% frequently in the decision trees and had a significant effect on classification was the item "Students seem to compete with each other" (54.88%) associated with the school environment variable. All the remaining items were included in the decision trees with frequencies below 50%.

Table 8.
Classification error percentages by node

Node	Classification	n	Percentage of Error
31. Node	0	76	22.4
21. Node	1	41	26.8
24. Node	0	30	33.3
14. Node	0	64	35.9
28. Node	1	5157	36.1
4. Node	0	2816	38.2
25. Node	1	645	38.3
26. Node	1	518	39.2
22. Node	1	277	39.7
19. Node	0	1115	40
20. Node	1	47	40.4

16. Node	0	87	43.7
17. Node	1	401	43.9
5. Node	1	199	44.7
10. Node	1	422	45
30. Node	1	163	46

*0- unsuccessful, 1- successful

In the analysis of the decision tree for the China sample, Node 31 had the lowest classification error rate. Students who strongly agree with the statement “My parents support my educational efforts and achievements”, strongly disagree with the statement “Other students seem to like me”, and strongly disagree with the statement “I feel excluded at school” are classified as failing (0). In Node 31, 76 students were classified as successful, with an error rate of 22.4%.

The node with the second lowest error rate is node 21. "My parents support my educational endeavors and achievements" is strongly disagree, disagree, agree, "Students seem to compete with each other" is very true or extremely true, "My parents encourage me to be self-confident" is strongly disagree, disagree and agree, "Students feel that they are being compared to each other" as not true at all, somewhat true or very true, "I feel proud of accomplishing something." as strongly disagree, disagree or agree, "I feel that I can manage many things at the same time" as strongly disagree, disagree or agree, and "I feel that I can manage many things at the same time" as strongly agree. In Node 21, 41 students were classified as successful, resulting in an error rate of 26.8%. This node represents the second most accurate prediction in the China sample within the C4.5 decision tree.

In the China sample, ST123Q02NA and ST205Q02HA seem to have the largest effects on mathematical literacy. ST123Q02NA, which relates to family support, specifically the statement “My parents support my educational endeavors and achievements”, is located at Node 1 in the C4.5 decision tree and is present in 99.24% of the decision trees that comprise the decision tree. Upon evaluation of the statistics, it is concluded that the item “My parents support my educational efforts and achievements” is the most significant factor affecting mathematical literacy in the China sample, in relation to the family support variable. The item “Students seem to be competing with each other”, related to the school environment variable, is located at Node 2 in the C4.5 decision tree and is found in 54.88% of the decision trees that form the decision tree. According to these statistics, the statement “Students seem to compete with each other” is considered the second most significant factor affecting mathematical literacy in the China sample.

4. RESULTS, DISCUSSION AND RECOMMENDATIONS

The aim of this study is to determine the factors affecting mathematical literacy by using the C4.5 algorithm, one of the decision tree methods, using Turkey and China samples. Within the scope of the research, the independent variables of self-perception, adaptability, sense of belonging, family support and school environment were considered and the items related to these variables were analyzed. In assessing the research findings, the factors that have the greatest impact on mathematical literacy in each country were identified by using the most influential values in the decision tree created for each country. In Turkey's PISA 2018, the most influential items on mathematics performance were identified as "I feel proud that I have accomplished something", "My parents support my educational efforts and achievements" and "I feel that I can manage many things at the same time". These items are associated with self-perception and family support. In the China sample, the items that affected math performance the most were "My parents support my educational efforts and achievements" and "Students seem to compete with each other", which were considered under the variables of family support and school environment.

In the Turkey sample, the top three factors affecting mathematics performance are related to self-perception. The first factor, “I feel proud that I have achieved something”, and the third factor, “I feel that I can manage many things at the same time”, measure students' self-belief and high self-evaluation of their capacity. It can be said that students who are aware of and proud of their achievements are more likely to be interested in mathematics and that this positive feeling can have a positive impact on their performance. Similarly, it can also be inferred that feeling proud of one's achievements can also help to motivate. It can be concluded that students who are confident in handling more than one situation will transfer this self-confidence to the field of mathematics and thus be able to perform better by experiencing less anxiety when solving new problems.

Studies examining the relationship between self-perception and mathematics performance also support this finding. For instance, Anigala (2015) conducted a study where a self-perception scale was applied to 3000 middle school students and concluded that there was a positive correlation between students' mathematics performance and their self-perception. Ajayi et al. (2011) examined the relationship between self-perception and mathematics achievement. They administered 2000 middle school students an attitude towards mathematics scale, a mathematics achievement test and a self-perception scale, and concluded that attitude towards mathematics and self-perception have a significant composite effect on mathematics achievement. Akarsu (2009) examined the factors affecting mathematics achievement using data from PISA 2003 and concluded that self-efficacy is a strong predictor of mathematics achievement. Within the scope of the study, self-efficacy, intrinsic motivation, extrinsic motivation and mathematics achievement variables were examined. Considering the results of the research, the significant effect of self-perception is seen.

The second item, "My parents support my educational efforts and achievements", which is one of the factors that affect mathematical literacy performance the most, overlaps with the factor of pride in achievement that we encountered in the first node. It can be interpreted that students who are appreciated by their parents will be more satisfied with their academic achievements and therefore will strive for greater success. Karaağaç (2015), who examined the relationship between mathematics achievement and family, applied the Mathematics Achievement Test and the Family Rating Scale to 319 7th and 8th grade students and examined the effect of family functioning on mathematics achievement. The study found that family functioning, specifically problem solving, roles, emotional response, and general functions, had a significant effect on mathematics performance.

The most important factors affecting mathematical literacy in the China sample are family support and school environment. The first factor was identified as family support with the item "My parents support my educational endeavours and achievements". Lee (2022) examined the factors that most affect academic achievement at the school level using data from the PISA 2018 China application. As a result of the analysis, it was concluded that school and student characteristics positively affect learning outcomes and thus performance. It can be interpreted that family support not only contributes to students' motivation and self-confidence, but also affects mathematics performance by reducing students' performance anxiety.

The second factor that most affects mathematical literacy performance in China is the school environment, with the item "Students seem to be competing with each other". Teng (2020) examines the factors affecting the mathematics performance of students in the China sample in PISA 2012. The study found that certain dimensions of the school environment moderated the impact of family background on mathematics achievement, in other words, the school environment functioned as a safeguard against the negative effects of family background on mathematics performance. In addition, the effect of school environment was found to affect average-performing schools and low-achieving students more than high-performing schools and students. According to the study, the perception of a negative school environment is the main factor explaining low performance. In another study, Satıcı (2008) examined the factors affecting mathematical literacy using the data set of PISA 2003 China sample and concluded that students' competitive thinking is the most important factor affecting mathematics achievement. It should be noted that China is a particularly successful country in mathematics, above the PISA average. In this sample, where performance is generally high, it can be interpreted that competitive school environment motivates students and high expectations positively affect students' performance.

Considering the two items that stand out in the China sample within the scope of the study, "My parents support my educational efforts and achievements" and "Students seem to compete with each other", it is seen that family support not only causes a decrease in the student's mathematics anxiety, but also that families with a supportive profile can offer more learning opportunities. It can be said that the competitive environment of students in China, a country with a high average success rate, in terms of motivation and future plans, positively affects their mathematics performance.

In this study, the factors affecting mathematical literacy in PISA 2018 Turkey and China samples were examined using the C4.5 decision tree algorithm. The same study can be conducted using different classification methods. Similarly, the consistency of the algorithms can be compared with more data by applying the decision tree methods used to the data of different countries.

The independent variables of self-perception, adaptability, school belonging, family support and school environment were used in the study and their effects on mathematics literacy were examined. Other variables included in the PISA student, teacher and school surveys can be determined as independent variables and their effects on mathematical literacy can be examined with the C4.5 algorithm.

Within the scope of the research, the factors affecting mathematical literacy in the samples determined using PISA 2018 data were examined. The same analysis methods, samples and independent variables can be applied to other PISA applications and the results obtained from different years can be compared.

Using the selected independent variables (self-perception, adaptability, school belonging, family support and school environment) and the analysis method used (C4.5 algorithm), the effects of these variables on science literacy and reading comprehension competence can be examined and the effects on different areas can be compared.

As a result of the studies, it was seen that the most influential factors on mathematical literacy in the Turkey sample were self-perception and family support. These two concepts can be strengthened by organizing workshops for students and parents to raise their awareness.

In terms of self-perception, activities that include relatively complex problems and challenges that students need to solve on their own can support students' self-confidence and belief in their ability to solve problems. For family support, through family counseling during the education process, families can be made aware of the right approach and supportive profile for students and students' belief in the support of their families can be strengthened.

Within the scope of the research, a single data mining classification method was used. For future studies data mining studies on education, evidence can be presented regarding the validity of the results obtained by using more methods. Analyses can be

repeated with different decision tree methods such as C5.0, Artificial Neural Networks, ID3. Furthermore, the consistency of the algorithms can be compared with more data by applying the decision tree methods used to the data of different countries.

Research and Publication Ethics Statement

The authors affirm having followed professional, ethical guidelines in preparing this work. Secondary data were used in this study. Therefore, ethical approval is not required.

Contribution Rates of Authors to the Article

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [ZBO], [SC]. The first draft of the manuscript was written by [ZBO] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Statement of Interest

The authors declare that they have no conflict of interest.

5. REFERENCES

Akarsu, S. (2009). Öz-Yeterlik, Motivasyon ve PISA 2003 Matematik Okuryazarlığı Üzerine Uluslararası Bir Karşılaştırma: Türkiye ve Finlandiya. (Yüksek Lisans Tezi) Abant İzzet Baysal Üniversitesi, Bolu.

Ajayi, K., Lawani, A., & Adeyanju, H. (2011). Effects of Students' Attitude and Self-Concept on Achievement in Senior Secondary School Mathematics in Ogun State, Nigeria. *Journal of Research in National Development.*

Anigala, A. (2015). Self-Concept as a Correlate of Secondary School Students Academic Performance in Mathematics. *International Journal of Education.*

Aydın, A., Sarıer, Y., & Uysal, Ş. (2012). Sosyoekonomik ve Sosyokültürel Değişkenler Açısından PISA Matematik Sonuçlarının Karşılaştırılması. *Eğitim ve Bilim*, 1300-1337.

Aydiner, A., & Kalender, İ. (2015). Student Segments Based on the Factors Related to Sense of Belonging Across Disadvantaged and Resilient Groups in PISA 2012, *Procedia - Social and Behavioral Sciences, Volume 174*, 2015, ISSN 1877 0428, https://doi.org/10.1016/j.sbspro.2015.01.997

Büyüköztürk, Ş., Kılıç, E. K., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2009). *Bilimsel araştırma yöntemleri* (4. Basım) Ankara: Pegem A Yayıncılık.

Boman, B. (2023). Is the SES and academic achievement relationship mediated by cognitive ability? Evidence from PISA 2018 using data from 77 countries. Frontiers in Psychology, 14, 1045568.

Caponera, E., Di Chiacchio, C., Greco, S., & Palmerio, L. (2019). ¿Pueden Los Padres De Estudiantes De 15 Años Influir En Su Rendimiento En Ciencias?. *Journal of Supranational Policies of Education*, (9), 156–176. https://doi.org/10.15366/jospoe2019.9.00

Castro, M. & Lizasoain, L. (2012). Las técnicas de modelización estadística en la investigación educativa: Minería dedatos, modelos de ecuaciones estructurales y modelos jerárquicos lineales [Statistical modeling techniques ineducational research: Data mining, structural equation modeling. *Revista Española de Pedagogía, 131*-148.

Demir, İ. & Kılıç, S. (2010). Using PISA 2003: Examining the factors affecting students' mathematics achievement. *H. U. Journal of Education*, *38*, 44-54.

Demoulin, D. F. (1998). Giving Kids A Good Emotional Start-What Head Start Parents and Teachers Should Know to Ensure Emotionally Healthy Children. *Children and Families*, Fall. https://www.ilikeme.org/article1.html.

Dev, Ş. (2020). PISA Matematik Okuryazarlığını Etkileyen Duyuşsal Faktörlerin İncelenmesi: Sistematik Derleme Çalışması. *Necmettin Erbakan Üniversitesi Eğitim Bilimleri Enstitüsü*.

Dolu, A. (2020). Sosyoekonomik Faktörlerin Eğitim Performansı Üzerine Etkisi: PISA 2015 Türkiye Örneği. *Yönetim ve Ekonomi Araştırmaları Dergisi*, 41-58.

Güzel İş, Ç. (2014). The impact of student and school characteristics and their interaction on Turkish students' mathematical literacy skills in the programme for international student Assessment (PISA) 2003. *Mediterranean Journal of Educational Research*, *15*, 11-30.

Güzel İş, Ç. & Berberoğlu, G. (2010). Students' affective characteristics and their relation to mathematical literacy measures in the programme for international student assessment (PISA) 2003. *Eurasion Journal of Educational Research, 40*, 93-113.

Gómez, R.L. & Suárez, A.M. (2020). *Do inquiry-based teaching and school climate influence science achievement and critical thinking? Evidence from PISA 2015*. IJ STEM Ed 7, 43. https://doi.org/10.1186/s40594-020-00240-5

Han, J., Kamber, M., & Pei, J. (2001). *Data Mining: Concepts and Techniques.* San Francisco, Calif.: Morgan Kaufmann Publishers.

Hofer, S.I., Heine, JH., Besharati, S. Yip, J.C., Reinhold, F., Brummelman, E.(2024). Self-perceptions as mechanisms of achievement inequality: evidence across 70 countries. *npj Sci. Learn.* 9, 2 https://doi.org/10.1038/s41539-023-00211-9

Jablonka, E. (2003). *Mathematical literacy. Second international handbook of mathematics education*, 75-102.

Jehangir, K., Glas, C. A., & van den Berg, S. (2015). Exploring the relation between socio-economic status and reading achievement in PISA 2009 through an intercepts-and-slopes-as-outcomes paradigm. *International Journal of Educational Research, 1*-15.

Karaağaç, M., & Erbay, H. (2015). Aile İşlevselliğinin Matematik Başarısıyla İlişkisi. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 12*(31), 21-33. https://dergipark.org.tr/en/pub/mkusbed/issue/19577/208892

Karaağaç Cingöz, Z. (2020). Ekonomik, Sosyal ve Kültürel Statünün Akademik Başarıya Etkisi PISA 2015 ve TEOG 2017 Sonuçlarının Karşılaştırması. *The Journal of Humanity and Society*, 247-288.

Karaboğa, H. A., Akogul, S., & Demir, I. (2022). Classification of Students' Mathematical Literacy Score Using Educational Data Mining: PISA 2015 Turkey Application. *Cumhuriyet Science Journal, 43*(3), 543-549.

Lee, H. (2022). What drives the performance of Chinese urban and rural secondary schools: A machine learning approach using PISA 2018. *Cities, 123*, 103609. https://doi.org/10.1016/j.cities.2022.103609

Linoff , G., & Berry , M. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer*

MEB. (2019). *PISA 2018 Türkiye Ön Raporu .* Ankara: TC. Milli Eğitim Bakanlığı.

OECD. (2016). *PISA 2015 Results in Focus.* Paris: OECD Publishing.

OECD. (2019). *PISA 2018 Assessment and Analytical Framework.* Paris: OECD Publishing.

Osmanbegović, E., & Suljic, M. (2012). ATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE. *Journal of Economics & Business/Economic Review*, 3-12.

Ötken, Ş. (2021). PISA 2012'de Öğrencilerin Matematik Başarısını Sınıflayan Değişkenlerin Belirlenmesi . *The Journal of Social Science , 5* (9) , 241-249 . DOI: 10.30520/tjsosci.871481

Perry , L. B., & Mcconey, A. (2010). Socioeconomic Status and Student Achievement Using PISA 2003. *Teachers College Record: The Voice of Scholarship in Education, 1137*-1162.

Predić, B., Dimić, G., Rančić, D., Štrbac, P., Maček, N., & Spalević, P. (2018). Improving final grade prediction accuracy in blended learning environment using voting ensembles. *Computer Applications in Engineering Education*, 26(6), 2294-2306.

Pinto, J., Silva, J. C., & Bixirão Neto, T. (2016). Fatores influenciadores dos resultados de matemática de estudantes portugueses e brasileiros no PISA: revisão integrativa. Ciência & Educação (Bauru), 22, 837-853.

Pujari, A. (2001). *Data Mining Techniques*. Hyderguda: Universities Press (India) Private Limited. Relationship Management. Indianapolis: Wiley Publishing.

Satıcı, K. (2008) *PİSA 2003 sonuçlarına göre matematik okuryazarlığını belirleyen faktörler: Türkiye ve Hong Kong - Çin*. (Yüksek Lisans Tezi) Balıkesir Üniversitesi.

Schulz, W. (2015). *Measuring the Socio-Economic Background of Students and Its Effect on Achievement on PISA 2000 and PISA 2003*. San Francisco: American Educational Research Association.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81-106. https://doi.org/10.1007/bf00116251

Yıldırım, Y. , Şahin, M. G., & Sezer, E. (2017). PISA 2012 Türkiye Örnekleminde Okul Özelliklerinin Matematik Okuryazarlığına Etkisi. *İlköğretim Online , 16* (3) , 1092-1100 . DOI: 10.17051/ilkonline.2017.330244.

Yılmaz, H. B., Aztekin, S., Umurhan, H., Aydın, H., Akıncı, B., Yılmaz Fındık, L., & Eser, G. (2011). *PISA Türkiye.* Ankara: Milli Eğitim Bakanlığı Yenilik ve Eğitim Teknolojileri.

Witten, I. H., Frank, E., & Hall, M. (2011). *Data mining: practical machine learning tools and techniques*. Amsterdam, The Netherlands: Elsevier.

Yu, H., Huang, X., Hu, X., & Cai, H. (2010, October). *A comparative study on data mining algorithms for individual credit risk evaluation.* In 2010 International Conference on Management of e-Commerce and e-Government (pp. 35-38). IEEE

Yuan T. (2020) The relationship between school climate and students' mathematics achievement gaps in Shanghai China: Evidence from PISA 2012, *Asia Pacific Journal of Education, 40*,3, 356-372, DOI: 10.1080/02188791.2019.1682516