Journal of Innovations
in Science, Mathematics and Computing Education
2025, 1(1), pp. 1-14 Research Article

Predictive Analytics of Math Anxiety in Students: A Machine Learning Study on PISA 2022 Turkey Data

Büşra YAĞCI[1*] Murat ŞAHİN [2], Zehra AKBIYIK[3], Yunus DOĞAN[4]
[a*]PhD Candidate, Dokuz Eylül University, https://orcid.org/0000-0002-6553-7760
[*]busra.yagci23@ogr.deu.edu.tr
[b] PhD Candidate, Dokuz Eylül University, https://orcid.org/0000-0002-2866-8796
[c] Master's Student, Dokuz Eylül University, https://orcid.org/0009-0006-3234-5080
[d] Assistant Prof., Dokuz Eylül University, https://orcid.org/0000-0002-0353-5014

Submission: 03.03.2025 Acceptance: 29.03.2025

Abstract

Mathematics anxiety is the worry, fear, and stress individuals experience in mathematics-related situations. Mathematics anxiety is an important problem in the education system and an important factor affecting students' academic success. In this context, studies to prevent or reduce mathematics anxiety are of great importance. Machine learning algorithms significantly contribute to such studies by enabling the extraction of information from large datasets. PISA 2022 dataset focuses on the assessment of student performance in mathematics, reading and science to measure the extent to which students can use what they learned in and out of schools for their full participation in societies. Some 690 000 students took the assessment in 2022, representing about 29 million 15-year-olds in the schools of the 81 participating countries and economies. The primary purpose of this study is to predict mathematics anxiety of students in Turkey using the PISA 2022 dataset. So, the dataset has been filtered based on Turkey. The new dataset includes 7250 instances and 1280 feature attributes. In order to use this dataset, a multi-stage preprocessing is carried out. Two different datasets are developed by selecting different attributes. In Dataset A, there are 26 attributes and 6065 instances. The current study also generated another dataset including attributes containing PISA weighted scores which is called Dataset B. Variables with weighted averages of the PISA 2022 dataset were used in feature selection for Dataset B. Mathematics anxiety values in both datasets are calculated using Decision Tree (DT), Random Forest (RF), Ada Boost (AB), Gaussian Naive Bayes (GaussianNB), K Nearest Neighbors (KNN), Multi-Layer Perceptron Classifier (MLPC), and XGBoost (XGB). These models are compared to calculating Precision, Recall, F1-Score, and Accuracy values.
Keywords: PISA 2022, Mathematics Anxiety, Machine Learning Algorithms, Data Mining

Öğrencilerde Matematik Kaygısının Tahmini Analitiği: PISA 2022 Türkiye Verileri Üzerine Bir Makine Öğrenmesi Çalışması

Öz

Matematik kaygısı, bireylerin matematikle ilgili durumlarda deneyimledikleri endişe, korku ve strestir. Matematik kaygısı, eğitim sisteminde önemli bir sorundur ve öğrencilerin akademik başarısını etkileyen önemli bir faktördür. Bu bağlamda, matematik kaygısını önleme veya azaltma çalışmaları büyük önem taşımaktadır. Makine öğrenimi algoritmaları, büyük veri kümelerinden bilgi çıkarılmasını sağlayarak bu tür çalışmalara önemli ölçüde katkıda bulunmaktadır. PISA 2022 veri seti, öğrencilerin okullarda ve okul dışında öğrendiklerini toplumlara tam katılımları için ne ölçüde kullanabildiklerini ölçmek için matematik, okuma ve fen alanlarındaki öğrenci performansının değerlendirilmesine odaklanmaktadır. 2022'de yaklaşık 690.000 öğrenci değerlendirmeye katıldı ve bu, 81 katılımcı ülke ve ekonominin okullarındaki yaklaşık 29 milyon 15 yaşındaki öğrenciyi temsil ediyor. Bu çalışmanın temel amacı, PISA 2022 veri setini kullanarak Türkiye'deki öğrencilerin matematik kaygısını tahmin etmektir. Bu nedenle, veri seti Türkiye bazında filtrelenmiştir. Yeni veri seti 7250 örnek ve 1280 özellik niteliği içermektedir. Bu veri setini kullanabilmek için çok aşamalı bir ön işleme gerçekleştirilir. Farklı nitelikler seçilerek iki ayrı veri seti oluşturulur. Veri Seti A'da 26 nitelik ve 6065 örnek bulunur. Mevcut çalışmada ayrıca PISA ağırlıklı puanları içeren nitelikler içeren Veri Seti B adı verilen başka bir veri seti de üretilmiştir. PISA 2022 veri setinin ağırlıklı ortalamalarına sahip değişkenler, Veri Seti B için özellik seçiminde kullanılmıştır. Her iki veri setindeki matematik kaygısı değerleri Karar Ağacı (DT), Rastgele Orman (RF), Ada Boost (AB), Gauss Naive Bayes (GaussianNB), K-En Yakın Komşu (KNN), Çok Katmanlı Algılayıcı Sınıflandırıcı (MLPC) ve XGBoost (XGB) kullanılarak hesaplanmıştır. Bu modeller Kesinlik, Duyarlılık, F1 Puanı ve Doğruluk değerlerinin hesaplanmasıyla karşılaştırılmıştır.
Anahtar kelimeler: PISA 2022, Matmatik Kaygısı, Makine Öğrenmesi Algoritmaları, Veri Madenciliği

INTRODUCTION

Anxiety is an abnormal and overwhelming sense of apprehension and fear often marked by physical signs (such as tension, sweating, and increased pulse rate), by doubt concerning the reality and nature of the threat, and by self-doubt about one's capacity to cope with it (Merriam-Webster Dictionary, n.d). Anxiety, which is one of 9 negative emotions are defined by Lazarus, is considered mood disorder. People can encounter with anxiety, whose the most common symptoms are nervous and worried feelings and thoughts and physiological symptoms such as elevated heart rate, muscle tension, and shakiness, in each moment and each field (Boreham & Schutte, 2023). Individuals may experience anxiety about various issues such as being judged by others, being ridiculed, being separated from loved ones, being abandoned, not being able to move comfortably in a crowd, and failing in business or school life. Experiences, especially in childhood, are very important in the formation of anxiety. When examined from an educational perspective, it is common for children or adolescent students to have anxiety about failing classes. The most common example, especially in our country, is mathematics anxiety.

Mathematics anxiety is the worry, fear, and stress individuals experience in mathematics-related situations. This anxiety can negatively impact students' math performance and relationship with math. Mathematics anxiety is an important problem in the education system and an important factor affecting students' academic success. In this context, studies to prevent or reduce mathematics anxiety are of great importance. These studies aim to determine mathematics anxiety, understand its causes, and reduce its effects. Machine learning algorithms significantly contribute to such studies by enabling the extraction of information from large datasets. The primary purpose of this study is to predict mathematics anxiety using the PISA 2022 dataset. Machine learning algorithms can help identify math anxiety by identifying patterns in the dataset. This study may be more effective than traditional methods in determining students' mathematics anxiety and may offer a new perspective on this subject.

The contribution of this study to the literature is that it emphasizes the use of machine learning algorithms in identifying and reducing mathematics anxiety. It offers a data-driven and analytical approach in addition to traditional methods, allowing math anxiety to be addressed more effectively. Additionally, this study may open new research avenues by emphasizing the importance of using machine learning techniques in mathematics education. In summary, this study highlights the importance of using machine learning algorithms in identifying and reducing mathematics anxiety. It offers a new perspective in this field by addressing mathematics anxiety more effectively with a data-oriented and analytical approach.

LITERATURE REVIEW

PISA datasets have been periodically sharing comprehensive self-reported data about students, parents and school administrators since 2000. These datasets are highly valued for many countries globally and are used for many purposes by them. For example, Tzora has proposed a predictor to detect of economic status for Greek high-school students in 2025; Khine et al. have presented a study with its results of effects of economic, social, and cultural status on Australian students' learning in 2024; while Li & Li have studied on disadvantaged Chinese students in 2024, Hernández-Ramos & Martínez-Abad have been researched Spanish teachers' commitment to their professional development in 2023; Bayirli et al. and Bernardo et al. have proposed models for Mathematics in respectively 2023 and 2022, with respectively 12 Asia-Pacific countries' data and Philippines data; while Araújo & Costa have studied on Literature for Portuguese students in 2023, Liu et al. have proposed a regression model for Mathematics, Literature and Science subjects for Chinese students in 2023 (Table 1).

Table 1. Summary of Studies Addressing PISA Data with Various Methods and Aims

REF.	YEAR	REGION	SUBJECT	METHODS	PERFORMANCE METRICS
Tzora	2025	Greece	Detection of economic status for high-school students	Canonical Functions	R^2
Khine et al.	2024	Australia	Effects of economic, social, and cultural status on students' learning	Ridge Linear Regression, K-Nearest Neighbours, Decision Trees, eXtreme Gradient Boosting, Support Vector Machines	Average Absolute Error, Average Absolute Percentage Error, Mean Squared Error, R^2
Li & Li	2024	China	Disadvantaged students	Path Analysis Methodology, Mediation Model	Average, Standard Deviation, Correlations
Bayirli et al.	2023	12 Asia-Pacific countries	Mathematics	Linear Regression, Random Forest, Support Vector Machines	Precision, Recall, F1, Accuracy
Araújo & Costa	2023	Portugal	Literature	Linear Regression	Percentage, Average, The number of observations
Bernardo et al.	2022	Philippines	Mathematics	Logistic Regression, Random Forest, Support Vector Machines Multilayer Perceptron, Decision Tree	Precision, Recall, F1, Accuracy
Liu et al.	2023	China	Mathematics, Science, Literature	Regression	Percentage, Average, The number of observations
Hernández-Ramos & Martínez-Abad	2023	Spain	Teachers' commitment to their professional development	Decision Tree	Precision, Recall, Receiver Operating Characteristic

In addition to official reports on PISA data, these datasets are previously analyzed and reported by researchers using statistical methods and machine learning algorithms (Arpa & Çavur, 2024). It can be said that PISA data is a subject that is mostly focused on in determining and predicting the factors affecting students' reading proficiency and literacy. Accordingly, reading literacy (Dong & Hu, 2019) and reading proficiency (Bernardo et al., 2021), English reading skills (Luo, 2023), reading achievement (Dai et al., 2023), digital reading (Zheng et al., 2024), there are many studies using machine learning algorithms in examining reading self-concept (Ramazan et al., 2023). In addition, there are studies including machine learning approaches to predict students' academic performance (Acıslı Celik & Yesilkanat, 2023; Haw & King, 2023; Lee, 2022; Masci et al., 2018; Puah, 2020; Rebai et al., 2020). Students' attitudes towards ICT (Lezhnina, & Kismihók, 2022) and ICT engagement (Sirganci, 2023) are among the other topics examined. In addition to these studies, it is seen that the tendency and success in mathematics have begun to be examined. Gabriel et al. (2018) analyzed the role of mathematics self-efficacy, mathematics self-concept, mathematics anxiety, motivation, perceived control, subjective norms and attributions of failure and demographics in predicting mathematics literacy with a boosted regression tree. As a result of the analysis made on the PISA 2012 dataset, it was reported that mathematics self-efficacy was a strong predictor. Pejić et al. (2022), neural networks and random forest algorithms were used to predict mathematics performance in three classes: low, mediocre and high, on the PISA 2012 dataset. Manually, background variables were determined as math attitudes, math intentions, student math behavior, math out of school lessons, math experience and math concepts familiarity (Von Lorenz, 2025). In addition, the most influential variables that were automatically selected were possessions - literature, books at home, math efficacy, out-of-school study time, guided homework, out-of-school study time and personal tutor. Test results showed that the accuracy value was

78.39% when recursive feature elimination was used with a neural network. The random forest algorithm together with recursive feature elimination produced an accuracy value of 72.41%.

DATASET

The dataset used in this paper is “Programme for International Student Assessment (PISA) 2022”. PISA, which is created by the Organisation for Economic Co-operation and Development (OECD), is a large dataset containing many attributes by which students' educational trends and achievements in OECD countries are measured. 81 countries and economies took part in the 2022 assessment, which especially focused on mathematics and creativity, and the data were released by the OECD on 5th December 2023.

PISA focuses on the assessment of student performance in mathematics, reading and science to measure the extent to which students can use what they learned in and out of schools for their full participation in societies. Some 690 000 students took the assessment in 2022, representing about 29 million 15-year-olds in the schools of the 81 participating countries and economies. In mathematics and reading, a multi-stage adaptive approach has been applied in computer-based tests. Students also have answered a background questionnaire, which took about 35 minutes to complete. The questionnaire has sought information about the students themselves, their attitudes, dispositions and beliefs, their homes, and their school and learning experiences. The PISA database contains the full set of responses from individual students, school principals and parents. These files are open sources for statisticians and professional researchers who would like to undertake their own analysis of the PISA data and available at https://www.oecd.org/pisa/data/.

In this study, the set of responses that individual students have given to questionnaires in Turkey, are used. The research data was obtained from 7250 Turkish students participating in Pisa 2022. Since mathematics is being focus point of PISA 2022, the features attributes related with mathematics in the dataset have been chosen to aim predicting mathematic anxiety with machine learning approaches in this study.

PRE-PROCESSING

The original version of PISA 2022 dataset has 660 000 instances and 1280 attributes. This is a very large dataset. So, multistage preprocessing is needed. First of all, the aim of this study is to focus on predicting mathematic anxiety of students in Turkey. So, the dataset has been filtered based on Turkey. The new dataset includes 7250 instances and 1280 feature attributes.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CNT	CNTRYID	CNTSCHID	CNTSTUID	CYC	NatCen	STRATUM	SUBNATIO	REGION	OECD	ADMINMO	LANGTEST	LANGTEST
2	TUR	792	79200148	79200001	08MS	79200	TUR36	7920000	79200	1	2	344	344
3	TUR	792	79200192	79200002	08MS	79200	TUR14	7920000	79200	1	2	344	344
4	TUR	792	79200021	79200003	08MS	79200	TUR31	7920000	79200	1	2	344	344
5	TUR	792	79200007	79200004	08MS	79200	TUR14	7920000	79200	1	2	344	344
6	TUR	792	79200065	79200005	08MS	79200	TUR23	7920000	79200	1	2	344	344
7	TUR	792	79200026	79200006	08MS	79200	TUR12	7920000	79200	1	2	344	344
8	TUR	792	79200182	79200007	08MS	79200	TUR15	7920000	79200	1	2	344	344
9	TUR	792	79200152	79200008	08MS	79200	TUR35	7920000	79200	1	2	344	344
10	TUR	792	79200064	79200009	08MS	79200	TUR07	7920000	79200	1	2	344	344
11	TUR	792	79200100	79200010	08MS	79200	TUR35	7920000	79200	1	2	344	344
12	TUR	792	79200163	79200011	08MS	79200	TUR11	7920000	79200	1	2	344	344
13	TUR	792	79200047	79200012	08MS	79200	TUR34	7920000	79200	1	2	344	344
14	TUR	792	79200089	79200014	08MS	79200	TUR09	7920000	79200	1	2	344	344
15	TUR	792	79200052	79200015	08MS	79200	TUR14	7920000	79200	1	2	344	344
16	TUR	792	79200142	79200016	08MS	79200	TUR29	7920000	79200	1	2	344	344
17	TUR	792	79200170	79200017	08MS	79200	TUR25	7920000	79200	1	2	344	344
18	TUR	792	79200022	79200018	08MS	79200	TUR18	7920000	79200	1	2	344	344
19	TUR	792	79200110	79200019	08MS	79200	TUR08	7920000	79200	1	2	344	344
20	TUR	792	79200138	79200020	08MS	79200	TUR32	7920000	79200	1	2	344	344
21	TUR	792	79200089	79200021	08MS	79200	TUR09	7920000	79200	1	2	344	344
22	TUR	792	79200157	79200022	08MS	79200	TUR18	7920000	79200	1	2	344	344
23	TUR	792	79200146	79200023	08MS	79200	TUR15	7920000	79200	1	2	344	344
24	TUR	792	79200058	79200024	08MS	79200	TUR18	7920000	79200	1	2	344	344
25	TUR	792	79200096	79200025	08MS	79200	TUR35	7920000	79200	1	2	344	344
26	TUR	792	79200086	79200026	08MS	79200	TUR13	7920000	79200	1	2	344	344
27	TUR	792	79200167	79200027	08MS	79200	TUR22	7920000	79200	1	2	344	344
28	TUR	792	79200171	79200028	08MS	79200	TUR26	7920000	79200	1	2	344	344
29	TUR	792	79200136	79200029	08MS	79200	TUR15	7920000	79200	1	2	344	344
30	TUR	792	79200015	79200030	08MS	79200	TUR25	7920000	79200	1	2	344	344

Figure 1. *The sample of dataset.*

In the dataset, some columns represent the option of each question in the questionnaire and the other columns represent the scores of tests that have been made to students. The values in the dataset are numeric and nominal. In Figure 1, an excel table of the first 30 samples in the dataset is given.

	A	B	C
1	CNT	String	Country code 3-character
2	CNTRYID	Numeric	Country Identifier
3	CNTSCHID	Numeric	Intl. School ID
4	CNTSTUID	Numeric	Intl. Student ID
5	CYC	String	PISA Assessment Cycle (2 digits + 2 character Assessment type - MS/FT)
6	NatCen	String	National Centre 6-digit Code
7	STRATUM	String	Stratum ID 5-character (cnt + original stratum ID)
8	SUBNATIO	String	Adjudicated sub-region code 7-digit code (3-digit country code + region ID + stratum ID)
9	REGION	Numeric	REGION
10	OECD	Numeric	OECD country
11	ADMINMODE	Numeric	Mode of Respondent
12	LANGTEST_QQQ	Numeric	Language of Questionnaire
13	LANGTEST_COG	Numeric	Language of Assessment
14	LANGTEST_PAQ	Numeric	Language of Parent Questionnaire
15	Option_CT	Numeric	Creative Thinking Option
16	Option_FL	Numeric	Financial Literacy Option
17	Option_ICTQ	Numeric	ICT Questionnaire Option
18	Option_WBQ	Numeric	Well-Being Questionnaire Option
19	Option_PQ	Numeric	Parent Questionnaire Option
20	Option_TQ	Numeric	Teacher Questionnaire Option
21	Option_UH	Numeric	Une Heure Option
22	BOOKID	Numeric	Form Identifier
23	ST001D01T	Numeric	Student International Grade (Derived)
24	ST003D02T	Numeric	Student (Standardized) Birth - Month
25	ST003D03T	Numeric	Student (Standardized) Birth -Year
26	ST004D01T	Numeric	Student (Standardized) Gender
27	ST250Q01JA	Numeric	Which of the following are in your [home]: A room of your own
28	ST250Q02JA	Numeric	Which of the following are in your [home]: A computer (laptop, desktop, or tablet) that you can use for school work
29	ST250Q03JA	Numeric	Which of the following are in your [home]: Educational Software or Apps
30	ST250Q04JA	Numeric	Which of the following are in your [home]: Your own [cell phone] with Internet access (e.g. smartphone)

Figure 2. *The descriptions of attributes.*

The descriptions (label) of the first 30 of 1280 feature attributes are also given in Figure 2. The column A includes the names of feature attributes in the dataset. The column B represents the kinds of values, and the column C contains the descriptions of feature attributes in the dataset.

Table 2. The Descriptions and Input Varieties of The First 10 Attributes

ATTRIBUTES	DESCRIPTIONS	INPUT VARIETIES
ST001D01T	Grade	7,8,9,10,11,12
ST003D02T	Date of birth (Year)	2006
ST003D02T	Date of birth (Month)	1,2,3,4,5,6,7,8,9,10,11,12
ST004D01T	Gender	1: Girl, 2: Boy
ST230Q01JA	Number of siblings	1,2,3,4
ST295Q05JA	Exercise or practise a sport (e.g. running, cycling, aerobics, soccer, skating)	1: 0 days, 2: 1 days, 3: 2 days, 4: 3 days, 5: 4 days, 6: 5 or more days
ST295Q01JA	Eat dinner (after school)	1: 0 days, 2: 1 days, 3: 2 days, 4: 3 days, 5: 4 days, 6: 5 or more days
ST268Q06JA	<Science> is easy for me.	1: Strongly disagree, 2: Disagree, 3: Agree, 4: Strongly agree
ST297Q06JA	Small group study or practice (2 to 7 students)	1: Participate, 0: Not Participate
ST283Q02JA	The teacher showed us how mathematics can be useful in our everyday lives.	1: Never or almost never, 2: Less than half of the lessons, 3: About half of the lessons, 4: More than half of the lessons, 5: Every lesson or almost every lesson

In order to understand this complex dataset, firstly the questionnaire that used to create this dataset was examined. The questions that are related to the basic demographics and mathematic anxiety of students were determined. Then, the attributes that are not related with these questions have been removed. The dataset in the last version has 7250 instances and 131 attributes. The descriptions and input varieties of the first 10 features attributes are in Table 2 as sample.

Column names have been replaced with their descriptions to make the dataset more understandable. For the same purpose, the nominal input varieties have been represented with their labels instead of numbers. The new version of the dataset is given in Figure 3.

	A	B	C	D	E	F	G	H	I
1	Gender	BirthYear	NumOfSiblings	Grade	BirthMonth	Exercise or practise a sport (e	Eat dinner	<Science> is easy for me.	Small group study or practice
2	2	2006	4	10	6	0 days	5 or more days	Agree	Not Participate
3	1	2006	4	10	6	0 days	5 or more days	Disagree	Not Participate
4	1	2006	4	10	5	2 days	5 or more days	Agree	Not Participate
5	1	2006	2	10	7	5 or more days	5 or more days	Agree	Not Participate
6	1	2006	4	10	11	1 days	5 or more days	Disagree	Not Participate
7	2	2006	2	10	9	0 days	5 or more days	Strongly disagree	Not Participate
8	2	2006	4	10	1	0 days	5 or more days	Agree	Not Participate
9	2	2006	2	10	3	5 or more days	5 or more days	Strongly disagree	Not Participate
10	2	2006	4	9	10	5 or more days	5 or more days	Strongly disagree	Not Participate
11	2	2006	3	10	7	5 or more days	5 or more days	Disagree	Not Participate
12	1	2006	4	10	8	3 days	5 or more days	Agree	Not Participate
13	2	2006	3	10	3	3 days	5 or more days	Agree	Participate
14	1	2006	4	10	2	3 days	5 or more days	Strongly agree	Not Participate
15	1	2006	2	10	11	1 days	5 or more days	Agree	Not Participate
16	2	2006	2	10	6	2 days	5 or more days	Disagree	Not Participate
17	1	2006	4	10	3	3 days	5 or more days	Agree	Not Participate
18	1	2006	2	10	7	4 days	5 or more days	Agree	Participate
19	1	2006	4	10	4	4 days	5 or more days	Agree	Not Participate
20	1	2006	4	10	5	0 days	2 days	Disagree	Not Participate
21	2	2006	2	10	7	2 days	4 days	Strongly agree	Not Participate
22	2	2006	4	10	12	4 days	5 or more days	Agree	Participate
23	2	2006	4	10	7	5 or more days	5 or more days	Agree	Not Participate
24	2	2006	4	10	2	4 days	5 or more days	Agree	Participate
25	2	2006	1	10	9	3 days	5 or more days	Agree	Not Participate
26	1	2006	4	10	6	5 or more days	5 or more days	Agree	Not Participate
27	2	2006	2	10	7	5 or more days	5 or more days	Strongly agree	Not Participate
28	2	2006	4	10	10	2 days	5 or more days	Agree	Participate
29	1	2006	2	10	8	3 days	5 or more days	Agree	Participate
30	2	2006	4	10	6	3 days	999	Strongly agree	Participate

Figure 3. *The new version of the dataset.*

In order to determine the predictive variables in understanding the antecedents of mathematics anxiety, first, the dataset was cleaned, and the missing values were determined. According to the output there is no missing value in any attributes. However, the missing values have been labeled as 95, 96, 97, 99, 999 in the original dataset. These labels have been standardized as 999. So, 999 represents the missing values in the dataset. The missing values of the first 100 instances are given in a matrix in Figure 4.

Figure 4. *The missing values of the first 100 instances.*

Attributes containing approximately more than 22% missing values or have single type value, and instances include missing values for more than one features in the dataset were removed from the analysis. Thus, the number of attributes and the number of instances were equal to 37 and 6065, respectively. The missing values in each column were filled with their mode values, which is one of the methods to fill missing data.In the next step, the values in the dataset have been transformed from categoric to numeric using label encoding. Hierarchy between variables has not been ignored while label encoding has been made on the attributes which include variables that have hierarchical ordering.

In order to predict mathematics anxiety using this dataset, an attribute related to mathematic anxiety is needed. However, in Pisa 2022, mathematics anxiety is measured with six items such that

- I often worry that it will be difficult for me in mathematics classes,
- I get very tense when I have to do mathematics homework,
- I get very nervous doing mathematics problems,
- I feel helpless when doing a mathematics problem,
- I worry that I will get poor marks in mathematics,
- I feel anxious about failing in mathematics.

All these attributes are also the questions that measure mathematics anxiety, and their responses can be “Strongly disagree”, “Disagree”, “Agree”, or “Strongly agree”. In other words, this is a Likert scale. These attributes were combined by averaging the answers hierarchically labeled from 0 to 3 and a single attribute was obtained. The median of this attribute, obtained from 6 attributes, is 1.17. Instances in the dataset are labeled as anxious or non-anxious depending on whether their math anxiety attribute values are less or greater than this value. Thus, a dataset suitable for making a binary classification model was obtained.

In the last version of the dataset, there are 26 attributes and 6065 instances. These attributes and their short descriptions are the followings;

1. Gender: Student's gender.

2. Grade: Current school grade level.

3. BirthMonth: Month of birth.

4. <Science> is easy for me: Student's perception of science difficulty.

5. Small group study or practice (2 to 7 students): Participation in small study groups.

6. <Science> is one of my favorite subjects.: Student's preference for science.

7. I want to do well in my <science> class.: Student's motivation in science.

8. Mathematics is easy for me.: Student's perception of math difficulty.

9. <Test language> homework: Time spent on language homework.

10. One-on-one tutoring with a person: Private tutoring sessions.

11. Mathematics is one of my favorite subjects.: Student's preference for math.

12. Internet or computer tutoring with a programme or application: Use of digital tools for learning.

13. How many hours per week do you usually need to attend the following courses? (For all subjects including mathematics in one week): Weekly class hours for all subjects.

14. Large group study or practice (8 or more students): Participation in large study groups.

15. Total time for all homework in all subjects, including subjects not listed above: Weekly homework duration.

16. <Science> homework: Time spent on science homework.

17. I want to do well in my <test language> class.: Student's motivation in language studies.

18. How many hours per week do you usually need to attend the following courses? (In one week for math class): Weekly math class hours.

19. <Test language> is easy for me.: Student's perception of language difficulty.

20. How would you rate the quality of your mathematics course during the school year?: Student's evaluation of math course quality.

21. I do not participate in <additional mathematics instruction>.: No extra math lessons attended.

22. <Test language> is one of my favourite subjects.: Student's preference for the test language.

23. I want to do well in my mathematics class.: Student's motivation in math.

24. Mathematics homework: Time spent on math homework.

25. Video-recorded instruction by a person: Learning from pre-recorded lessons.

26. Math anxiety: Student's level of anxiety about math.

This dataset which is obtained using the attributes related with basic demographics and mathematic anxiety of students in the original dataset is called Dataset A. In Figure 5, the first 30 samples in Dataset A are given.

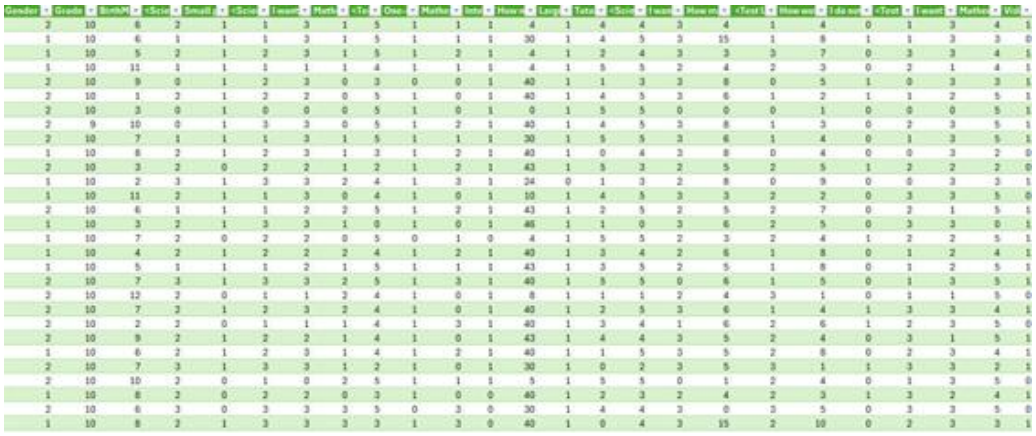

Figure 5. *Dataset A.*

The current study also generated another dataset including attributes containing PISA weighted scores which is called Dataset B. Variables with weighted averages of the PISA 2022 dataset were used in feature selection for Dataset B. By examining the correlations of these variables with the dependent variable, a dataset was created with attributes that gave a significant correlation over 0.10. These attributes are given below with their explanations in Table 3.

Table 3. The Descriptions of The Attributes in Dataset B

ATTRIBUTES	DESCRIPTIONS
MATHPREF	Preference of Math over other core subjects
MATHEASE	Perception of Mathematics as easier than other subjects
MATHMOT	Motivation to do well in mathematics
BULLIED	Being bullied (WLE)
PERSEVAGR	Perseverance (agreement) (WLE)
STRESAGR	Stress resistance (agreement) (WLE)
EMOCOAGR	Emotional control (agreement) (WLE)
DISCLIM	Disciplinary climate in mathematics (WLE)
COGACRCO	Cognitive activation in mathematics: Foster reasoning Version B (WLE)
COGACMCO	Cognitive activation in mathematics: Encourage mathematical thinking Version B (WLE)
MATHEFF	Mathematics self-efficacy: formal and applied mathematics - response options reversed in 2022 (WLE)
MATHEF21	Mathematics self-efficacy: mathematical reasoning and 21st century skills (WLE)
FAMCON	Subjective familiarity with mathematics concepts (WLE)
MATHPERS	Effort and Persistence in Mathematics (WLE)
ICTQUAL	Quality of access to ICT (WLE)

Math anxiety scores were discretized as the outcome variable. Discretization was carried out according to whether the weighted scores were below or above zero.

METHODS

Decision trees are a popular method for both classification and regression tasks because they are intuitive and easy to interpret. A decision tree algorithm seeks to create a model that predicts the value of a target variable based on several input variables. Each internal node of the tree corresponds to an input variable; and each leaf node corresponds to a predicted output value. The Random Forest algorithm improves upon the decision tree by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. This helps to reduce overfitting and improve the prediction accuracy (Breiman, 2001).

AdaBoost, short for Adaptive Boosting, is another ensemble technique that combines multiple weak learners into a strong learner. In AdaBoost, each subsequent model is tweaked in favor of those instances misclassified by the previous models, and it focuses on achieving high accuracy by assigning a higher weight to the more difficult cases during the training process (Freund & Schapire, 1996). Gaussian Naive Bayes applies the Naive Bayes principles with an assumption of Gaussian (normal) distribution of the input variables. It works well in cases where the assumption about the distribution holds relatively true, making it efficient, especially for high-dimensional datasets (Bi, Han, Huang, & Wang, 2019). K-Nearest Neighbors is a non-parametric, instance-based learning algorithm that is often used for classification and regression. In KNN, the output is a class membership: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (Guo, Wang, Bell, Bi, & Greer, 2003). The Multi-layer Perceptron Classifier is a type of feedforward artificial neural network. MLPC consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. Unlike other techniques, each node uses a nonlinear activation function. This allows MLPC to capture complex relationships in data. XGBOOST, or Extreme Gradient Boosting, is an advanced implementation of gradient boosting that is more efficient and flexible. It uses a gradient boosting framework and excels in handling sparse data. XGBOOST has gained popularity due to its effectiveness in numerous machine learning competitions (Chen, & Guestrin, 2016). Although each of these algorithms has its unique characteristics and applications, they share the common goal of analyzing complex data to make predictions or decisions, distinguishing them in how they construct models from the data. These methodologies are crucial in advanced analytics, enabling insights that guide strategic decision-making across various domains.

ETHICAL APPROVAL

It is declared that during the preparation process of this study, scientific and ethical principles were followed, and all the studies benefitted from are stated in the bibliography.

RESULTS

The prediction results for Dataset A and Dataset B are given in Table 4 and Figure 6. As seen in Table 4, the accuracy values obtained for Dataset A vary between 59% and 66% (DT: 59%, RF:67%, AB: 68%, NB:66%, KNN:62%, MLP: 68%, XGB:66%). Among these results, the highest prediction accuracy belongs to the Ada Boost and Multi-Layer Perceptron algorithms. Precision values for these algorithms are calculated as 69% for AdaBoost and 68% for Multi-layer Perceptron. Recall and F1 score values vary between 66% and 68%. It can be stated that multiple metric results indicate consistent results in terms of showing the performance of these two algorithms. Accuracy values obtained for Dataset B vary between 65% and 74% (DT: 65%, RF: 74%, AB: 74%, NB: 70%, KNN: 70%, MLP: 73%, XGB: 72%) Accordingly, it is seen that accuracy values indicate higher results than Dataset A.

Table 4. Prediction Findings of Dataset A and B

Dataset	ML Model	Precision	Recall	F1-Score	Accuracy
Dataset A	Decision Tree	0.58	0.58	0.58	0.59
	Random Forest	0.67	0.66	0.66	0.67
	Ada Boost	0.69	0.69	0.67	0.68
	Gaussian NB	0.68	0.67	0.66	0.66
	K Nearest Neighbors	0.63	0.62	0.61	0.62
	Multi Layer Perceptron	0.68	0.67	0.68	0.68
	XGBoost	0.66	0.65	0.65	0.66
Dataset B	Decision Tree	0.42	0.44	0.42	0.65
	Random Forest	0.61	0.36	0.45	0.74
	Ada Boost	0.62	0.38	0.47	0.74
	Gaussian NB	0.50	0.59	0.54	0.70
	K Nearest Neighbors	0.51	0.33	0.40	0.70
	Multi Layer Perceptron	0.58	0.40	0.46	0.73
	XGBoost	0.55	0.39	0.45	0.72

As seen in Figure 6, accuracy values calculated for Dataset B are above the values calculated for Dataset A for all algorithms. Still, it is important to consider other metrics when interpreting these findings. As a matter of fact, the precision values calculated for Dataset B vary between 42% and 66%. For Dataset B, all recall and precision values are below 50%, except for the Gaussian Naive Bayes algorithm. The reason for this inconsistency between the metrics may be due to the discretization process being performed based on positive and negative weighted averages for Dataset B. In this case, considering that different attributes and weighted averages have different roles for both datasets, these findings can be used in the prediction process.

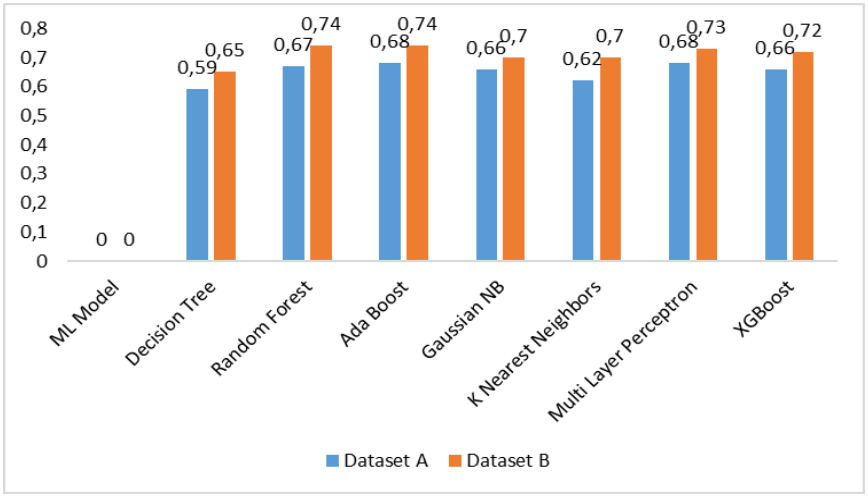

Figure 6. *Comparison of accuracy values between Dataset A and B.*

DISCUSSION AND CONCLUSION

This study conducted using the PISA 2022 dataset released in December 2023, aims to predict mathematics anxiety in Turkey, model its contributing factors, and demonstrate the model's effectiveness. Although previous studies have used datasets from different years of PISA results, the selection of this dataset is crucial due to its real, up-to-date nature and its inclusion of specifically math-focused questions. Additionally, the dataset's substantial amount of missing, repeated values, and the extensive efforts required for labeling make it a study rich in preprocessing procedures.

All models used in the study contain valuable information. However, among the seven models developed, the success of two models is more pronounced. As detailed in the results section, and considering F1-Score, recall, and precision, the methods with the best accuracy are observed to be AdaBoost and Multi-Layer Perceptron. This study holds potential in gaining a deeper understanding of the factors contributing to mathematics anxiety, identifying students with mathematics anxiety, and providing solutions for mathematics anxiety in Turkey.

Future research can advance by experimenting with additional features in the dataset, improving preprocessing methods, and enhancing model parameters to increase predictive accuracy. Moreover, validation through the integration of this model into real-world applications can accelerate scientific steps toward addressing mathematics anxiety in Turkey and provide effective interventions.

Conflict of Interest
The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors Contributions
Yazar 1: Design, Literature Review, Data Analysis, Writing, Critical Review of Content
Yazar 2: Literature Review, Data Analysis, Writing, Critical Review of Content
Yazar 3: Design, Literature Review, Writing, Critical Review of Content
Yazar 4: Design, Literature Review, Data Analysis, Writing, Critical Review of Content

REFERENCES

Acıslı Celik, S., & Yesilkanat, C. M. (2023). Predicting science achievement scores with machine learning algorithms: a case study of OECD PISA 2015–2018 data. *Neural Computing and Applications*, *35*(28), 21201-21228.

Araújo, L., & Costa, P. (2023). Reading to Young Children: Higher Home Frequency Associated with Higher Educational Achievement in PIRLS and PISA. *Education Sciences*, *13*(12), 1240. https://doi.org/10.3390/educsci13121240

Arpa, T., & Çavur, M. (2024). A Comparative Analysis of Machine Learning Techniques to Explore Factors Affecting Mathematics Success in Developing Countries: Turkey, Mexico, Thailand, And Bulgaria Case Studies (Doctoral dissertation, M. Hanefi CALP).

Bayirli, E. G., Kaygun, A., & Öz, E. (2023). An Analysis of PISA 2018 Mathematics Assessment for Asia-Pacific Countries Using Educational Data Mining. *Mathematics*, *11*(6), 1318. https://doi.org/10.3390/math11061318

Bernardo, A. B., Cordel, M. O., Lucas, R. I. G., Teves, J. M. M., Yap, S. A., & Chua, U. C. (2021). Using machine learning approaches to explore non-cognitive variables influencing reading proficiency in English among Filipino learners. *Education Sciences*, *11*(10), 628.

Bernardo, A. B. I., Cordel, M. O., II, Lapinid, M. R. C., Teves, J. M. M., Yap, S. A., & Chua, U. C.

(2022). Contrasting Profiles of Low-Performing Mathematics Students in Public and Private Schools in the Philippines: Insights from Machine Learning. *Journal of Intelligence*, *10*(3), 61. https://doi.org/10.3390/jintelligence10030061

Bi, Z. J., Han, Y. Q., Huang, C. Q., & Wang, M. (2019). Gaussian naive Bayesian data classification model based on clustering algorithm. In *2019 International Conference on Modeling, Analysis, Simulation Technologies and Applications (MASTA 2019)*, 396-400.

Boreham, I. D., & Schutte, N. S. (2023). The relationship between purpose in life and depression and anxiety: A meta-analysis. *Journal of Clinical Psychology*, *79*(12), 2736-2767.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794.

Dai, S., Hao, T., Ardasheva, Y., Ramazan, O., Danielson, R. W., & Austin, B. (2023). PISA reading achievement: Identifying predictors and examining model generalizability for multilingual students. *Reading and Writing*, *36*(10), 2763-2795.

Dong, X., & Hu, J. (2019). An exploration of impact factors influencing students' reading literacy in Singapore with machine learning approaches. *International Journal of English Linguistics*, *9*(5), 52-65.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm, 96, 148-156.

Gabriel, F., Signolet, J., & Westwell, M. (2018). A machine learning approach to investigating the effects of mathematics dispositions on mathematical literacy. *International Journal of Research & Method in Education*, *41*(3), 306-327.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003*, 986-996.

Haw, J. Y., & King, R. B. (2023). Understanding Filipino students' achievement in PISA: The roles of personal characteristics, proximal processes, and social contexts. *Social Psychology of Education*, *26*(4), 1089-1126.

Hernández-Ramos, J. P., & Martínez-Abad, F. (2023). Professional Development among Secondary Teachers in Spain: Key Associated Factors as of PISA 2018. *Journal of Intelligence*, *11*(5), 93. https://doi.org/10.3390/jintelligence11050093

Khine, M. S., Liu, Y., Pallipuram, V. K., & Afari, E. (2024). A Machine-Learning Approach to Predicting the Achievement of Australian Students Using School Climate; Learner Characteristics; and Economic, Social, and Cultural Status. *Education Sciences*, *14*(12), 1350. https://doi.org/10.3390/educsci14121350

Lee, H. (2022). What drives the performance of Chinese urban and rural secondary schools: A machine learning approach using PISA 2018. *Cities*, *123*, 103609.

Lezhnina, O., & Kismihók, G. (2022). Combining statistical and machine learning methods to explore German students' attitudes towards ICT in PISA. *International Journal of Research & Method in Education*, *45*(2), 180-199.

Li, Z., & Li, Q. (2024). How Social Support Affects Resilience in Disadvantaged Students: The Chain-Mediating Roles of School Belonging and Emotional Experience. *Behavioral Sciences*, *14*(2), 114. https://doi.org/10.3390/bs14020114

Liu, A., Wei, Y., Xiu, Q., Yao, H., & Liu, J. (2023). How Learning Time Allocation Make Sense on Secondary School Students' Academic Performance: A Chinese Evidence Based on PISA 2018. *Behavioral Sciences*, *13*(3), 237. https://doi.org/10.3390/bs13030237

Luo, S. (2023). *Factors Affecting English Reading in Macao, Hong Kong, and Singapore: Combining Machine Learning Methods and Hierarchical Linear Regressions Using PISA 2018 Data* (Doctoral dissertation, University of Macau).

Masci, C., Johnes, G., & Agasisti, T. (2018). Student and school performance across countries: A machine learning approach. *European Journal of Operational Research*, *269*(3), 1072-1085.

Merriam-Webster Dictionary. (n.d). Anxiety. Retrieved May 1, 2024, from https://www.merriam-webster.com/dictionary/anxiety

Pejić, A., Molcer, P. S., & Gulači, K. (2021). Math proficiency prediction in computer-based international large-scale assessments using a multi-class machine learning model. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)* (pp. 49-54). IEEE.

Puah, S. (2020). Predicting students' academic performance: a comparison between traditional MLR and machine learning methods with PISA 2015.

Ramazan, O., Dai, S., Danielson, R. W., Ardasheva, Y., Hao, T., & Austin, B. W. (2023). Students' 2018 PISA reading self-concept: Identifying predictors and examining model generalizability for emergent bilinguals. *Journal of School Psychology*, *101*, 101254.

Rebai, S., Yahia, F. B., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Economic Planning Sciences*, *70*, 100724.

Sirganci, G. (2023). A Machine Learning Approach to Assess Differential Item Functioning of PISA 2018 ICT Engagement Questionnaire: Item Functioning of PISA 2018 ICT Engagement Questionnaire. *International Journal of Curriculum and Instruction*, *15*(3), 2079-2093.

Tzora, V. A. (2025). Defining the Predictors of Financial Literacy for High-School Students. *Journal of Risk and Financial Management*, *18*(2), 45. https://doi.org/10.3390/jrfm18020045

Von Lorenz, A. C. (2025). Exploring Latent Class Profiles of Mathematics Performance: Insights from PISA 2022 Using Growth Mindset Indicators and Group Comparison Analysis. *Journal Evaluation in Education (JEE)*, *6*(1), 150-158.

Zheng, J. Q., Cheung, K. C., & Sit, P. S. (2024). Identifying key features of resilient students in digital reading: Insights from a machine learning approach. *Education and Information Technologies*, *29*(2), 2277-2301.