The First International Conference on
Machine Learning and Knowledge Discovery (MLKD 2024)
Amirkabir University of Technology, December 18-19, 2024

Find Effective Features and Optimal Algorithms for Predicting Student Performance in Mathematics and Science

Mahboubeh Molavi-Arabshahi* Narges Ghanbari†

Abstract

In this paper, using data from the 2019 TIMSS (Trends in International Mathematics and Science Study) test in Iran, we aim to find a method to predict students' scores in two subjects: mathematics and science. The major challenge with this database is the large number of features related to schools, teachers, and students, which makes finding and training a model for this classification difficult. Therefore, by applying feature selection and dimensionality reduction techniques, first on each type of feature and then on a dataset composed of all selected features, we attempt to find the best classification algorithm and the set of most impactful features on this performance. Finally, using the RFE (Recursive Feature Elimination) method for dimensionality reduction and three different classification algorithms, we identify 13 features that influence student scores. Additionally, we determine the algorithm that performs best on these selected features, which, in this research, is the gradient boosting algorithm.

Keywords: TIMSS 2019, Educational Data Mining, and Feature Selection

1 Introduction

The data related to student academic performance is constantly increasing, and this can optimize and simplify the process of analyzing and predicting student outcomes. If an instructor can estimate students' performance before final exams based on their behavior and environmental factors, it can greatly aid in optimizing the continuation of education. Additionally, examining which factors have had a positive impact on students' academic results over the course of a school year can significantly influence educational policy-making in different countries.

In this research, we aim to answer two questions:

1. How many parameters have the greatest impact on predicting student performance, and which parameters are they?

2. Which classification methods perform better in predicting student performance?

The implementation details and code used in this research are publicly available at GitHub repository to ensure reproducibility and transparency. `https://github.com/nargesghan/Classifying-students`

2 Related works

In the field of predicting student performance, recent research often seeks to address questions such as which classification algorithms yield the best predictions and which student-related features have the most significant impact on academic outcomes. For example, in [1], the predictive power of various data sources, including institutional data, learning management system (LMS) data, and survey data, is analyzed to predict periodic scores and final GPA scores of students. This study also explores biases in predictions across diverse student groups, such as disadvantaged populations and ethnic minorities.

Numerous studies have investigated factors influencing academic success. For instance, in [2], the effect of prior academic performance on graduation outcomes is examined using data from 1,841 engineering students at Covenant University, Nigeria. The study utilizes six algorithms—Probabilistic Neural Network (PNN), Random Forest, Decision Tree, Naive Bayes, Tree Ensemble, and Logistic Regression—on the KNIME platform, comparing their performance via accuracy metrics and regression models. Similarly, [3] evaluates the influence of demographic factors, such as gender and age, on academic performance, while psychological effects, including interest, stress, and anxiety, are explored in [4].

To enhance interpretability for non-expert users such as educators, studies like [5] leverage associative classification techniques. Using data from 5,000 engineering students at Polytechnic University of Turin, this research generates interpretable rules linking features (e.g., high school grades, LMS usage) with the likelihood of passing or failing exams, comparing these classifiers against models like Decision Trees, Support Vector Machines, and Neural Networks.

Building on this body of work, recent studies have specifically leveraged data mining methods to explore students' academic engagement. For example, Şevgin

*School of Mathematics and Computer Science, Iran University of Science and Technology, `molavi@iust.ac.ir`

†Department of Mathematical Sciences, Sharif University of Technology, `narges.ghanbari81@sharif.edu`

and Eranıl (2023) investigate school engagement (SE) among Turkish eighth-grade students using the TIMSS 2019 dataset[7]. Employing Random Forest models, the study identifies key factors such as bullying, instructional clarity, and students' confidence in science and mathematics, with SE explaining 69.6% and 75.7% of variance in science and mathematics, respectively. These findings underscore the importance of reducing bullying and enhancing instructional practices to foster engagement and improve educational outcomes.

Similarly, a study using TIMSS 2019 data from Morocco[8] explores machine learning models to predict student performance in mathematics. Logistic Regression, KNN, SVM, Decision Trees, and Random Forests are evaluated, with Decision Tree and Random Forest models achieving the highest accuracy (68%). This research highlights the utility of socio-academic features, such as students' home environment and attitudes toward learning, in guiding educational interventions.

Together, these studies demonstrate the potential of data mining and machine learning approaches to not only predict student outcomes but also provide actionable insights for educators and policymakers. By integrating socio-psychological variables with advanced algorithms, these works contribute to a deeper understanding of the complex dynamics influencing student performance.

3 Materials and methods

3.1 workflow

The TIMSS exam is a test that is conducted every four years in several countries around the world for students in the fourth and eighth grades. The significance of this exam lies in the fact that all participating countries respond to common questions with a specific number of sampled schools. Therefore, it serves as a valuable source for examining and comparing these countries. In this research, we use the results of the 2019 TIMSS exam for Iran. This dataset, in addition to the exam results, contains extensive information about students, teachers, and schools, making the range of analyzable factors quite broad.

In this study, we have four different types of data. The first set includes student-related data, such as the education level of their parents, students' expectations of themselves in mathematics and science, gender, and their access to educational resources like books, tablets, the internet, computers, and more. Another dataset contains teacher-related features, including their personal characteristics, such as years of experience, level of education, and their teaching methods.

A third dataset deals with school-related information, including economic conditions, parental involvement, and more.

Lastly, we have the dataset related to exam results.

Each of these types of data can be considered as impactful features on the test results. Therefore, we run each classification algorithm three times, once for each type of data, to observe the effect of each environmental, teacher, and student-related parameter on the final TIMSS test results. However, each of these datasets contains numerous features, which can still lead to overfitting and may also irrationally affect the prediction of student performance. Thus, we use feature selection and dimensionality reduction methods to limit the data to those features that enhance the model's performance and contribute meaningfully to predicting the final results. The methods employed in this research include Random Forest and Recursive Feature Elimination (RFE).

After reducing the data dimensions, we use classification algorithms such as Random Forest and Logistic Regression to predict student performance. Finally, we evaluate the models. At this stage, in addition to comparing the two selected models on one dataset, we also compare the prediction results using different types of data, examining which data types have the greatest impact on student performance.

In the end, we select the most important features from each dataset and combine them into a new dataset, aiming to predict student scores based on these features, which span all types of data. For this, we use four models: SVM, Logistic Regression, Decision Tree, and Random Forest. Each of these models may use a different number of input features, which we compare based on the number of features, ultimately identifying the most important features and the best algorithm.

3.2 dataset

The data used in this research includes the 2019 TIMSS exam results for Iran. This dataset contains extensive information about students, teachers, and schools, but for simplicity, we examine only a few of them. The student data analyzed in this study includes: The teacher-related data. Each math and science teacher has different characteristics and teaching methods, which are shown in the Table 1.

The school-related features are displayed in Table 2.

Finally, we had a Table 3 that included numerous features related to student performance in the exam, their results, and their scores. Questions from each math and science topic were included in the exam, and for each topic, we calculated an average score based on all related questions. We then recorded the overall average of these columns as the students' final math and science scores. This approach ensured that each topic and each question contributed equally to the target column. Afterward, by calculating the overall average math and science scores for the country of Iran, we divided the students into two categories: above average (1) and below average (0).

Science Teacher Features	**Math Teacher Features**
Years_Teaching	Years_Teaching
Education_Level_Completed	Education_Level_Completed
Major_Mathematics	Major_Mathematics
Major_Biology	Major_Biology
Major_Physics	Major_Physics
Major_Chemistry	Major_Chemistry
Major_Earth_Science	Major_Earth_Science
Major_Edu_Mathematics	Major_Edu_Mathematics
Major_Edu_Science	Major_Edu_Science
Major_Edu_General	Major_Edu_General
Ask_New_Content	Ask_New_Content
Ask_Other_Plan	Ask_Other_Plan
Ask_Different_Experiment	Ask_How_to_Solve
Ask_Plan_Experiments	Ask_Review_Procedures
Ask_Conduct_Experiments	Ask_Apply_Learned_Concepts
Ask_Conduct_Experiment	Ask_Model_Already
Ask_Plan_Procedures	Ask_Same_Ability
Ask_Interpret_Data	Ask_Use_Evidence
Ask_Use_Evidence	Homework_Assigned_Frequency
Ask_Read_Textbooks	Time_Spent_on_Homework
Ask_Mentor_Facts	Computer_Tablet
Ask_Mentor_Results	Number_of_Students_in_Class
Ask_Same_Ability_Groups	School_Emphasis_Academic_Success
Ask_Same_Ability	Teaching_Limited_by_Student_Needs
Ask_Same_Ask	
Frequency_Set_Homework	
Textbook_Used	
Teacher_Availability_Class	
School_Emphasis_Academic	
Teachers_Engages_Science	

Table 1: Science and Math Teacher Features

School Features
School_ID
Economic_Affluence
Existing_School_Library
Parental_Involvement
Emphasis_on_Academic_Success

Table 2: School Features

3.3 data preprocessing and adjusting feature selection algorithms

For better performance of the selected algorithms, some preprocessing steps must be performed on the data before training the algorithms. In this study, the following steps were taken:

1. **Converting Categorical Features to Numeric:**

   Several columns in the dataset contained non-numeric values, rendering them incompatible with classification algorithms. However, as these columns represented ordinal data, they could be effectively converted to numerical values by mapping the ordinal categories to corresponding numerical codes.

Student Features
Gender
Books_at_Home
Home_Computer_Tablet
Home_Own_Room
Home_Internet_Connection
Parent_A_Education
Parent_B_Education
Education_Expectation
Enjoy_Math
Math_is_Boring
Do_Well_in_Math
Math_is_Difficult
Enjoy_Science
Science_is_Boring
Do_Well_in_Science
Science_is_Difficult
Absent
Belong_to_School

Table 3: Student Features

2. **Data Normalization and Dimensionality Reduction:**

   Feature selection is a crucial step in data preprocessing, particularly when working with high-dimensional datasets. The objective is to identify a subset of relevant features that have the greatest impact on the model's predictive performance. In this study, we applied Recursive Feature Elimination (RFE) to reduce the dataset's dimensionality. RFE was implemented with several models, including Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting.

   RFE is a wrapper-based feature selection method that recursively eliminates the least important features, building a model with the remaining ones at each step. This process continues until a specified number of features remain. Feature importance is determined based on the model's performance.

   In the first step, we train a model (e.g., Support Vector Machine, Random Forest) on the entire dataset. Then, we rank all features according to their importance (e.g., in decision trees, feature importance can be measured by the impurity index, while in logistic regression, it can be determined by the feature coefficients). Next, the least important feature(s) are removed, and the process is repeated with the reduced set of features until the desired number of features is reached.

   To determine the number of output features for each algorithm, we first calculated the covariance of the initial features with the target column. The number of features was set to match the number of columns whose absolute covariance was greater than 0.1.

3.4 classifier selection and evaluation metrics

We applied the algorithms Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting on datasets created from student, teacher, and school data, using the selected features from the previous step. These models were trained twice—once for math and once for science—to classify students into two categories: above average and below average.

Finally, we evaluated the models using the following classification metrics:

Notation

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives
- **Accuracy**: The ratio of correctly predicted samples to the total samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- **Precision**: The ratio of correctly predicted positive observations to the total predicted positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **F1-Score**: The harmonic mean of Precision and Recall, providing a balance between the two:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

We then analyzed the models' performance on each dataset to determine which type of data (student, teacher, or school) is most effective in predicting student performance. The best classification methods for each dataset were identified. We selected a number of features from each of the three separate dataframes (student, teacher, and school) based on their correlation with the target column, which was the class of math or science scores for the students. We then combined these features into a new dataframe that included the selected features and the main target variables. We applied Support Vector Machine (SVM), Random Forest, Logistic Regression, and Decision Tree models on this new dataframe. Additionally, we reduced the number of features step-by-step and retrained the models to observe changes in accuracy.

4 Results

Each of the different datasets related to school, student, and teacher was evaluated using four algorithms: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. These evaluations were based on accuracy, precision, recall, and F1 score. Among the algorithms, Gradient Boosting achieved the highest accuracy, 76%, for predicting math and science scores using student-related data. For each dataset, the algorithm with the best performance was used as a benchmark for feature selection. For instance, Random Forest, with its high accuracy, was identified as the optimal choice for feature selection in school-related datasets for predicting science scores (Figure 2). A similar approach was applied to other datasets and target variables to select the best algorithms for dimensionality reduction.

Two new datasets were then created by combining selected features—a total of 18 features—for math and science. Classification and dimensionality reduction algorithms were applied to these combined datasets. As shown in Figures 3 and 4, features for each dataset were selected based on wrapper algorithms combined with Recursive Feature Elimination (RFE). For the school dataset, Random Forest and Decision Tree were chosen for science and math, respectively. For teacher-related features, Gradient Boosting was selected for science, while Random Forest was used for math. Student-related features had the highest correlation with the target variables, contributing the most features (10 for both math and science) due to their significant impact on predictions.

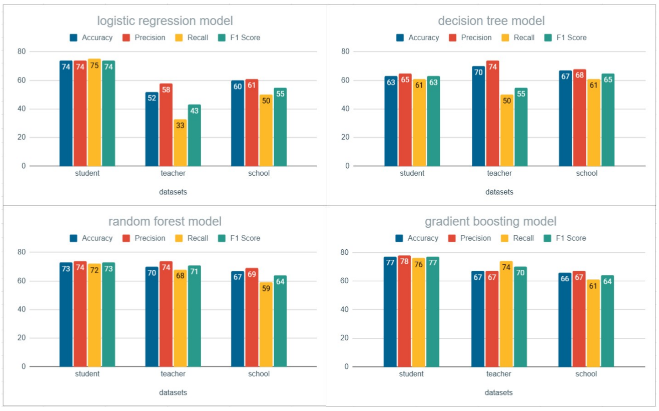


Figure 1: Prediction of Students' Math Scores with Different Datasets Using Four Algorithms: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting.

The performance of the four algorithms was also analyzed based on the number of input features (Figure 5). Decision Tree performed well with fewer than three features, but all algorithms underperformed with five fea-

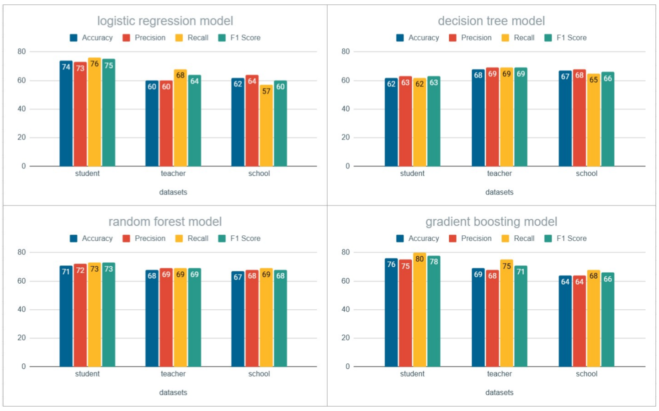


Figure 2: Prediction of Students' Science Scores with Different Datasets Using Four Algorithms: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting.

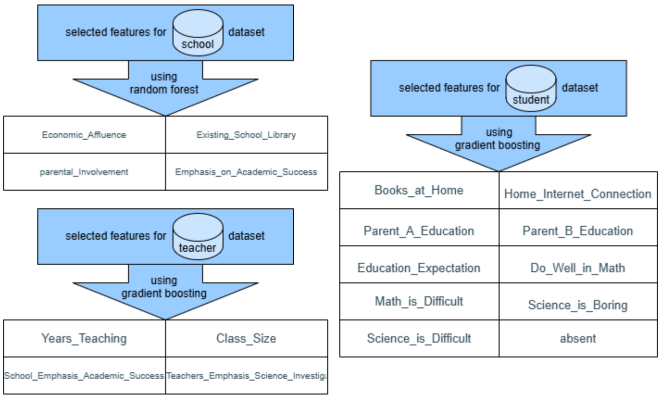


Figure 3: Selected features for science score prediction, grouped by dataset.

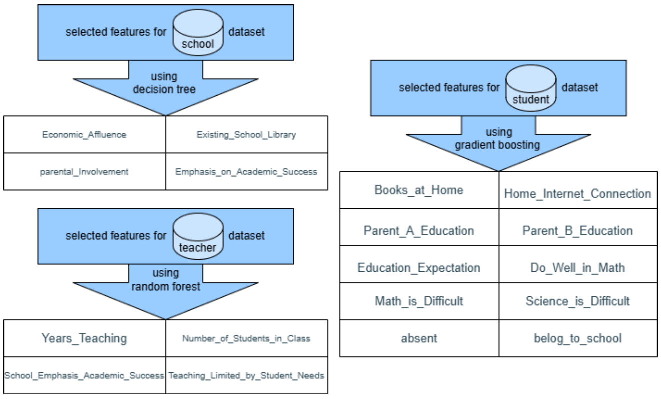


Figure 4: Selected features for math score prediction, grouped by dataset.

tures. The Gradient Boosting algorithm achieved the best overall performance when trained on 13 features for both math and science. These 13 features, identified as the most impactful, are ranked by priority in Figure 6.

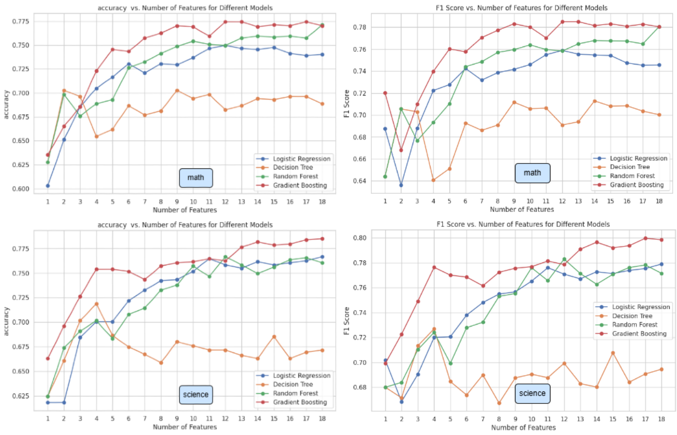
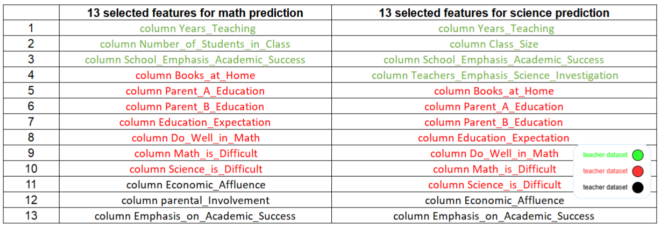


Figure 5: Comparison of feature counts, F1-score, and accuracy across algorithms.



	13 selected features for math prediction	13 selected features for science prediction
1	column Years_Teaching	column Years_Teaching
2	column Number_of_Students_in_Class	column Class_Size
3	column School_Emphasis_Academic_Success	column School_Emphasis_Academic_Success
4	column Books_at_Home	column Teachers_Emphasis_Science_Investigation
5	column Parent_A_Education	column Books_at_Home
6	column Parent_B_Education	column Parent_A_Education
7	column Education_Expectation	column Parent_B_Education
8	column Do_Well_in_Math	column Education_Expectation
9	column Math_is_Difficult	column Do_Well_in_Math
10	column Science_is_Difficult	column Math_is_Difficult
11	column Economic_Affluence	column Science_is_Difficult
12	column parental_Involvement	column Economic_Affluence
13	column Emphasis_on_Academic_Success	column Emphasis_on_Academic_Success

teacher dataset 🟢
teacher dataset 🔴
teacher dataset ⚫

Figure 6: Final selected features prioritized by Gradient Boosting.

The 13 selected features included teacher-related factors such as teaching experience, class size, and the school's emphasis on student success. Student-related features, such as the number of books available at home, parents' education levels, students' self-assessment, their expected educational level, and perceived difficulty of math and science, were also highly influential. Among school-related features, school wealth, family involvement, and the emphasis on academic success were notable contributors. Since math and science scores were highly correlated, the selected features for both subjects were largely similar, as shown in Figure 6.

References

[1] Yu, Renzhe; Li, Qiujie; Fischer, Christian; Doroudi, Shayan; Xu, Di. Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data. *ERIC Document Reproduction Service*, Report No. ED608066, 2020. `https://eric.ed.gov/?q=Towards+Accurate+and+Fair+Prediction+of+College+Success%3a+Evaluating+Different+Sources+of+Student+Data&id=ED608066`.

[2] Aderibigbe Israel Adekitan, Odunayo Salau The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, 5(2):e01250, 2019. https://doi.org/10.1016/j.heliyon.2019.e01250.

[3] Waziha Kabir, M. Omair Ahmad, M. N. Shanmukha Swamy. A novel normalization technique for multimodal biometric systems. In *2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 1–4, 2015. https://api.semanticscholar.org/CorpusID:8561917.

[4] Rajni Garg. PREDICTING STUDENT PERFORMANCE OF DIFFERENT REGIONS OF PUNJAB USING CLASSIFICATION TECHNIQUES. *International Journal of Advanced Research in Computer Science*, 9:236–240, 2018. https://api.semanticscholar.org/CorpusID:189438117.

[5] Cagliero L, Canale L, Farinetti L, Baralis E, Venuto E. Predicting Student Academic Performance by Means of Associative Classification. *Appl. Sci.*. 2021; *11*(4):1420. https://doi.org/10.3390/app11041420.

[6] Huang, Anna Y. Q., Owen H. T. Lu, Jeff C. H. Huang, C. J. Yin, and Stephen J. H. Yang. "Predicting Students' Academic Performance by Using Educational Big Data and Learning Analytics: Evaluation of Classification Methods and Learning Logs." *Computers Education*, **129** (2019): 69–87. https://doi.org/10.1016/j.compedu.2018.10.017.

[7] Şevgin, Hikmet, and Anıl Kadir Eranıl. "Investigation of Turkish Students' School Engagement through Random Forest Methods Applied to TIMSS 2019: A Problem of School Psychology." *International Journal of Psychology and Educational Studies*, **10**(4) (2023): 896–909. https://doi.org/10.52380/ijpes.2023.10.4.1260.

[8] Boudad, Saida, and Fatiha Essaaidi. "Using Machine Learning to Predict Mathematics Performance of Moroccan Students: An Analysis of the TIMSS 2019 Dataset." *International Journal of Advanced Computer Science and Applications*, **14**(7) (2023): 493–501. https://doi.org/10.14569/IJACSA.2023.0140758.