Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 9, S. 191-196, 6, 2025

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 9, pp. 191-196, 6, 2025

Research Article

Machine Learning-Based Prediction of LGS Scores from Middle School Exam Results in Kütahya

Zehra Bilici*, Şevval Demiral [2] and Metin Demir [3]

[1]*Department of Computer Engineering, Faculty of Engineering, Gebze Technical University, Kocaeli, Turkey*
[2,3]*Department of Primary Education, Faculty of Education, Kütahya Dumlupinar University, Kütahya,*

(zbilici@gtu.edu.tr)

(Received: 24 May 2025, Accepted: 18 June 2025)

(1st International Conference on Pioneer and Academic Research ICPAR 2025, June 13-14, 2025)

ATIF/REFERENCE: Bilici, Z., Demiral, Ş. & Demir, M. (2025). Machine Learning-Based Prediction of LGS Scores from Middle School Exam Results in Kütahya. *International Journal of Advanced Natural Sciences and Engineering Researches*, 9(6), 191-196.

Abstract – Accurate prediction of student performance is crucial for improving educational outcomes and enabling early interventions. This study examines the predictability of national high school entrance exam (LGS) scores based on in-school exam results across six core subjects from 818 students in grades 6 to 8 at 15 middle schools in Kütahya, Turkey. Fourteen supervised machine learning regression models, including ensemble methods such as Extra Trees, Random Forest, and XGBoost, were applied independently for each subject to forecast LGS net scores. Performance was evaluated using Mean Squared Error (MSE) and Coefficient of Determination (R^2). The results show that ensemble-based models significantly outperform traditional algorithms and achieve high accuracy in all subjects.The findings highlight the effectiveness of these models in capturing complex patterns in educational data and their potential for early identification of at-risk students. This research supports the integration of machine learning techniques into educational assessment systems to foster data-driven, personalized interventions.

Keywords – Student Performance Prediction, Machine Learning, Regression Models, Educational Data Analysis

I. INTRODUCTION

In educational systems, accurately and timely predicting student performance is of critical importance both for improving individual learning processes and for shaping education policies based on data-driven approaches. Traditional assessment methods typically rely on teacher observations, end-of-term grades, and limited-scale performance evaluations, offering only a narrow perspective on a student's developmental trajectory. In this context, Educational Data Mining (EDM) and Machine Learning (ML) techniques hold significant potential to provide more comprehensive and accurate predictions by analyzing large-scale, multidimensional educational data [1].

Particularly, the integration of diverse data sources—such as student absenteeism records, midterm exam results, course grade averages, socio-demographic variables, and digital traces obtained from Learning Management Systems (LMS)—enhances the reliability of individual performance prediction [2]. Commonly used methods for modeling student performance include Naïve Bayes, Decision Trees, Support Vector Machines (SVM), Artificial Neural Networks (ANN), Logistic Regression, and ensemble-based models [3].

These models not only aim to forecast academic outcomes, but also contribute to the early identification of at-risk students, improvements in instructional design, and more effective decision-making regarding resource allocation [4]. Systematic reviews in the field highlight that the success of such modeling efforts depends not only on the choice of algorithm but also on factors such as data diversity, preprocessing techniques, and hyperparameter optimization [5].

This study investigates the relationship between periodic exam scores and scores from the High School Entrance Examination (LGS) for students in grades 6, 7, and 8 across 15 different middle schools in Kütahya, Turkey. The data include students' first and second written exam results from both the first and second semesters in subjects such as Turkish, Mathematics, Science, Social Studies, English, and Religious Culture and Ethics. Using this dataset, various machine learning algorithms—such as Random Forest, Support Vector Machines, k-Nearest Neighbors, and Artificial Neural Networks—were applied to predict students' performance on the LGS.

The aim of the study is to analyze the predictability of central exam performance based on in-school assessment data and to determine the models that yield the highest accuracy. Ultimately, the goal is to enable early detection of academic risk and to support the development of individualized intervention mechanisms.

The application of machine learning techniques in the field of education has been receiving increasing attention, particularly in areas such as predicting student performance, identifying at-risk students, and developing early intervention strategies. Systematic reviews in the domain of Educational Data Mining (EDM) indicate that various machine learning algorithms are effective in forecasting academic outcomes [6].

Among the algorithms commonly used for predicting student performance are Decision Trees, Random Forest, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Naïve Bayes, and Artificial Neural Networks (ANN). The performance of these algorithms often depends on the characteristics of the dataset and the configuration of the predictive models [7]. In addition to academic metrics, several studies have demonstrated that factors such as student motivation levels, learning strategies, and behavioral data significantly influence academic achievement. Incorporating such features into predictive models can enhance accuracy [8].

The success of machine learning applications in education depends not only on algorithmic accuracy but also on the interpretability of the models. Particularly for “black box” models, it is essential to ensure transparency in decision-making processes to promote trust and adoption among educators and administrators. In conclusion, machine learning techniques provide powerful tools for predicting student success and developing proactive educational strategies. The growing body of research in this field contributes significantly to the design of more data-driven and effective education policies.

II. Materials and Method

A. Dataset Description and Preprocessing

This study utilizes subject-specific datasets obtained from 15 public middle schools located in Kütahya, Turkey. The data encompasses 818 students across grades 6, 7, and 8, including their first and second semester written exam scores in six core subjects: Turkish, Mathematics, Science, Social Studies, Religious Culture and Ethics, and Foreign Language. For each subject, a separate dataset was constructed containing 12 in-school exam scores as predictive input features and the corresponding net score obtained by the student in the national high school entrance exam (LGS) as the target output.

Prior to modelling, the datasets were loaded using the pandas library. Column names were sanitized to remove inconsistencies, and non-numeric fields or Boolean artifacts were excluded. Instances with missing target values were dropped, and missing input features were imputed using the column-wise mean. Each dataset was independently split into training (80%) and testing (20%) subsets using a fixed random seed (random state=42) to ensure reproducibility. All numerical inputs were normalized using feature scaling.

B. Machine Learning Models

Fourteen distinct supervised regression algorithms were employed to predict LGS net scores for each subject. A detailed description of each model, including its theoretical basis and practical rationale for this study, is provided in Table 1.

Table 1. Overview of Machine Learning Models Used for Predicting Subject-Specific LGS Net Scores

Model	Description
Linear Regression	Assumes a linear relationship between exam scores and LGS performance. Fast and interpretable but limited in modeling complex relationships.
Ridge Regression	Incorporates L2 regularization to reduce overfitting and multicollinearity. Effective when input features (exam scores) are highly correlated.
Lasso Regression	Uses L1 regularization to shrink less important coefficients to zero, enabling automatic feature selection. Helps with sparse or noisy inputs.
ElasticNet	Combines L1 and L2 penalties to balance sparsity and stability in feature contributions. Useful when some exam variables are redundant.
Decision Tree Regressor	Constructs a hierarchy of rules by partitioning feature space. Interpretable but prone to overfitting when used alone.
Random Forest Regressor	An ensemble of decision trees built on random feature and sample subsets. Demonstrated strong performance across all subjects.
Extra Trees Regressor	Similar to Random Forest but uses fully randomized splits. Enhanced variance reduction, especially beneficial with uniform input distributions.
Gradient Boosting Regressor	Sequentially fits weak learners to minimize prediction error. Performed well on subjects with less linearity or weaker correlations.
AdaBoost Regressor	Adjusts focus on previously mispredicted samples. Improved performance in low-variance subjects with subtle distinctions.
Bagging Regressor	Trains multiple base learners on bootstrap samples. Reduces variance in stable exam-score distributions.
K-Nearest Neighbors (KNN)	Predicts LGS net by averaging outcomes of most similar students. Sensitive to data dimensionality; worked best on well-clustered features.
Support Vector Regression (SVR)	Margin-based regression tolerant to minor prediction errors. Underperformed in some subjects due to poor scalability and complexity.
Multi-Layer Perceptron (MLP)	Fully connected neural network for modeling non-linear patterns. Overfitting observed in smaller or homogeneous subject datasets.
XGBoost Regressor	Optimized gradient boosting algorithm with embedded regularization. Achieved top accuracy in multiple subjects with fast convergence.

All models were trained for each subject independently using the same pipeline and hyperparameters unless stated otherwise. Model training and evaluation were performed using the scikit-learn and xgboost libraries.

C. Evaluation Metrics

Model performance was assessed using two widely adopted regression metrics: Mean Squared Error (MSE) and Coefficient of Determination (R^2). The mathematical formulations of these metrics are presented in Equation (1) and Equation (2), respectively.

Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n the number of observations. A lower MSE indicates better predictive accuracy and model fit.

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad (2)$$

where $\bar{y}$ is the mean of actual values. R^2 quantifies the proportion of variance in the dependent variable that is explained by the model. Values approaching 1 indicate a strong explanatory capacity.

D. *Subject-Specific Modeling*

To preserve subject-level granularity and accuracy, model training and evaluation were conducted independently for each subject. That is, a separate training pipeline and performance assessment was implemented for Turkish, Mathematics, Science, Social Studies, Religious Culture and Ethics, and Foreign Language.

This subject-wise modeling approach allowed the investigation of how each model's predictive power varies across disciplines, reflecting potential differences in how student achievement in each domain translates to LGS performance. The results of these experiments are reported separately in the Results section, along with comparative performance rankings.

III. RESULTS

In this study, fourteen supervised regression models were independently trained and evaluated for each of the six core subjects—Mathematics, Turkish, Science, Social Studies, Religious Culture and Moral Knowledge, and Foreign Language—to predict students' LGS net scores based on their in-school exam performances. Model performance was assessed using Mean Squared Error (MSE) and Coefficient of Determination (R^2), with detailed results presented in Table 2.

Table 2.Comparative Performance Analysis of Student Achievement Prediction Models (MSE & R^2 Values)

	Mathematics		**Turkish**		**Science**		**Social Studies**		**Religious Culture and Moral Knowledge**		**Foreign Language (English)**	
	MSE	**R^2**	**MSE**	**R^2**	**MSE**	**R^2**	**MSE**	**R^2**	**MSE**	**R^2**	**MSE**	**R^2**
LinearRegression	1.9329	0.9228	2.1920	0.9340	2.2605	0.9341	0.6636	0.9339	1.1954	0.8715	1.5782	0.8855
Ridge	1.9328	0.9228	2.1919	0.9340	2.2606	0.9341	0.6636	0.9339	1.1954	0.8715	1.5782	0.8855
Lasso	2.2052	0.9120	2.3087	0.9305	2.3615	0.9312	0.9139	0.9089	1.4333	0.8460	1.8155	0.8683
ElasticNet	2.1096	0.9158	2.2488	0.9323	2.3061	0.9328	0.8044	0.9191	1.3683	0.8529	1.7053	0.8763
DecisionTree	0.0369	0.9985	0.1353	0.9959	0.0251	0.9993	0.0084	0.9992	**0.0011**	**0.9998**	0.0064	0.9995
RandomForest	0.0217	0.9991	0.0417	0.9987	0.0206	0.9994	0.0086	0.9991	0.0025	0.9997	0.0058	0.9996
GradientBoosting	0.0340	0.9986	**0.0070**	**0.9998**	0.0102	0.9997	0.0053	0.9995	0.0033	0.9996	0.0120	0.9991
AdaBoost	0.5845	0.9767	0.3791	0.9886	0.3760	0.9890	0.1648	0.9836	0.1397	0.9849	0.1165	0.9915
Bagging	0.0211	0.9992	0.0454	0.9986	0.0212	0.9994	0.0086	0.9991	0.0019	0.9998	0.0060	0.9996
ExtraTrees	**0.0147**	**0.9994**	0.0106	0.9997	**0.0020**	**0.9999**	**0.0004**	**0.9999**	0.0018	0.9998	**0.0004**	**0.9999**
KNN	6.1679	0.7538	8.1587	0.7544	8.0687	0.7648	3.3782	0.6633	3.3497	0.6400	5.1833	0.6239
SVR	4.8238	0.8075	7.6843	0.7687	17.712	0.4838	7.4945	0.2531	7.1152	0.2353	11.052	0.1980
MLP	3.0185	0.8795	2.0242	0.9391	2.1783	0.9365	1.3447	0.8660	1.5747	0.8308	1.8169	0.8682
XGBoost	0.0159	0.9994	0.0125	0.9996	0.0055	0.9998	0.0024	0.9998	0.0094	0.9990	0.0022	0.9998

Across all subjects, ensemble-based learning methods such as Extra Trees, XGBoost, Bagging, and Random Forest consistently delivered the most accurate results, with R^2 scores approaching 1.00 and very low MSE values, indicating almost perfect model performance. For example, Extra Trees achieved an R^2 of 0.9999 in both Science and English, while XGBoost yielded MSE as low as 0.0022 in English and 0.0055 in Science. These results highlight the robustness, generalizability, and consistency of ensemble methods across all academic disciplines.

Among traditional linear models, Linear Regression, Ridge, and ElasticNet demonstrated relatively strong but lower performance, with R^2 values between 0.85 and 0.93, and notably higher MSE values compared to ensemble models. While these models are more interpretable, they were limited in capturing the non-linear and high-dimensional patterns present in student performance data. Decision Tree, Gradient Boosting, and Bagging regressors also achieved very high R^2 values (≥ 0.9985) with minimal prediction error, further reinforcing the utility of tree-based approaches in educational prediction tasks. These models were particularly effective in subjects like Social Studies, Religious Culture, and Science,

which may have more standardized assessment patterns and lower variance. The AdaBoost Regressor showed comparatively lower performance than other ensemble models but still significantly outperformed traditional regressors in most subjects, especially in Foreign Language ($R^2 = 0.9915$). It demonstrated effective learning in subjects with lower variance but lagged slightly in more complex or nonlinear domains. On the other hand, K-Nearest Neighbors (KNN) produced moderate predictive power (R^2 between 0.62 and 0.75) but exhibited high MSE values, indicating weaker predictive accuracy, likely due to its sensitivity to feature dimensionality and data sparsity. Support Vector Regression (SVR) consistently underperformed across all subjects, particularly in Science, Social Studies, and Foreign Language, where R^2 scores dropped below 0.50, suggesting poor generalization and limited suitability for this dataset. Lastly, the Multi-Layer Perceptron (MLP) model yielded mixed results: it performed relatively well in Turkish and Science ($R^2 = 0.9391$ and 0.9365, respectively), but showed lower accuracy in Religious Culture and English, potentially due to overfitting or distributional issues within the input data.

In summary, the findings strongly emphasize the superior performance of ensemble learning algorithms, particularly Extra Trees and XGBoost, in modeling student achievement across multiple academic domains. They also reveal subject-dependent variations in model effectiveness, underlining the importance of contextual model selection in educational data mining applications.

IV. DISCUSSION

The findings of this study reaffirm that ensemble-based learning methods offer a distinct advantage in predicting subject-specific LGS scores based on in-school exam data. Models such as Extra Trees, XGBoost, Bagging, and Random Forest consistently achieved extremely low error rates (MSE) and R^2 values approaching 1.00, confirming their ability to capture the complex and non-linear relationships inherent in educational performance data. Among them, Extra Trees and XGBoost particularly stood out, yielding near-perfect accuracy across all subjects.

In contrast, traditional regression models such as Linear Regression, Ridge, Lasso, and ElasticNet showed only moderate predictive power. These models were useful for modeling simpler patterns but lacked the flexibility needed to handle high-dimensional inputs and interactions between variables. Their relatively higher MSE values and slightly lower R^2 scores illustrate their limited capacity in this context.

SVR and KNN were among the weakest performers across all subjects. Their predictive power was hampered by sensitivity to feature scaling, dimensionality, and data sparsity, especially in subjects like Science, Social Studies, and Foreign Language. Their R^2 scores dropped significantly in these areas, falling below 0.50 in several cases. Additionally, model effectiveness varied depending on the subject area. Subjects with more structured, standardized content (such as Religious Culture and Social Studies) yielded higher model accuracy, while subjects with greater performance variability across students showed slightly less predictability. This indicates that the nature of the subject content influences the modeling success and should be considered in future applications.

In summary, ensemble methods emerged as highly reliable tools for educational prediction tasks. However, choosing the right model also requires balancing accuracy, interpretability, data complexity, and deployment feasibility in real-world educational contexts.

V. CONCLUSION

This study demonstrates that in-school exam results can be effectively used to predict students' LGS performance using machine learning techniques. Among the 14 regression models evaluated, ensemble methods—especially Extra Trees, XGBoost, Bagging, and Random Forest—consistently outperformed other models, achieving the highest predictive accuracy and explained variance across all six core subjects. These results underline the practical value of ensemble models for early detection of students at academic risk and for supporting personalized educational interventions. Furthermore, the analysis revealed that model performance is not uniform across subjects, emphasizing the role of discipline-specific characteristics in prediction success. For future research, it is recommended to explore methods that enhance model interpretability, such as SHAP or LIME, and to incorporate additional features like

student motivation, attendance, or socioeconomic factors. Such additions could further improve prediction robustness and facilitate real-world educational policy integration.

ACKNOWLEDGMENT

The authors would like to thank the participating schools, teachers, and students in Kütahya for providing the data essential to this study. We also appreciate the support from our colleagues and institutions that contributed to the successful completion of this research.

REFERENCES

[1] M. S. N. Al-Din and H. A. Al Abdulqader, 'Students' Academic Performance Prediction Using Educational Data Mining and Machine Learning: A Systematic Review', International journal of research and innovation in social science, vol. VIII, no. VIII, pp. 1264–1291, Jan. 2024, doi: 10.47772/IJRISS.2024.808095.

[2] V. Nakhipova, L. Suleymenova, and E. Adylbekova, 'Determination of students' academic performance using machine learning methods', ILIM, vol. 41, no. 3, pp. 5–20, Sep. 2024, doi: 10.47751/SKPU.1937.V41I3.1.

[3] M. Bellaj, A. Ben Dahmane, and L. Sefian, 'Educational Data Mining: Employing Machine Learning Techniques and Hyperparameter Optimization to Improve Students' Academic Performance', International Journal of Online Engineering (ijoe), vol. 20, no. 03, pp. 55–74, Feb. 2024, doi: 10.3991/IJOE.V20I03.46287.

[4] T. T. Tin, L. S. Hock, and O. M. Ikumapayi, 'Educational Big Data Mining: Comparison of Multiple Machine Learning Algorithms in Predictive Modelling of Student Academic Performance', International Journal of Advanced Computer Science and Applications, vol. 15, no. 6, pp. 633–645, Jan. 2024, doi: 10.14569/IJACSA.2024.0150664.

[5] H. Nagarajan, Z. Alsalami, S. Dhareshwar, K. Sandhya, and P. Palanisamy, 'Predicting Academic Performance of Students Using Modified Decision Tree based Genetic Algorithm', 2nd IEEE International Conference on Data Science and Information System, ICDSIS 2024, May 2024, doi: 10.1109/ICDSIS61070.2024.10594426.

[6] B. Albreiki, N. Zaki, and H. Alashwal, 'A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques', Education Sciences 2021, Vol. 11, Page 552, vol. 11, no. 9, p. 552, Sep. 2021, doi: 10.3390/EDUCSCI11090552.

[7] M. Yağcı, 'Educational data mining: prediction of students' academic performance using machine learning algorithms', Smart Learning Environments, vol. 9, no. 1, pp. 1–19, Dec. 2022, doi: 10.1186/S40561-022-00192-Z/TABLES/14.

[8] F. A. Orji and J. Vassileva, 'Machine Learning Approach for Predicting Students Academic Performance and Study Strategies based on their Motivation', Oct. 2022, Accessed: Jun. 12, 2025. [Online]. Available: https://arxiv.org/pdf/2210.08186