ADBA Computer Science
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8544
RESEARCH ARTICLE
Vol.1 / No.1 / 2024 / pp.8-13
https://doi.org/10.69882/adba.cs.2024072

Evaluating the Effectiveness of Machine Learning Models in Predicting Student Academic Achievement

Emre Deniz *,[1]
*Department of Computer Engineering, Hitit University, 19030, Corum, Turkiye.

ABSTRACT This study evaluates the effectiveness of various machine learning models in predicting student academic achievement using a dataset of 1000 students. The data includes demographic, psychological, social, and institutional factors. Models such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, K-Nearest Neighbors Regressor, Support Vector Regressor (SVR), Gradient Boosting Regressor (GBR), XGBoost Regressor, and Neural Network (MLP) were employed. Results show that test preparation courses significantly enhance student performance, with SVR and Linear Regression models demonstrating the best predictive performance. The study highlights the importance of optimized educational strategies to enhance academic outcomes.

KEYWORDS
Academic achievement
Linear regression
Machine learning
Student performance

INTRODUCTION

Factors influencing student performance are multifaceted and encompass a wide array of determinants. Demographic factors such as gender, ethnicity, and parental education level have been recognized as crucial influencers of academic success (Korantwi-Barimah *et al.* 2017; Cantekin 2020; Jones *et al.* 2012). Gender has been highlighted as a factor impacting student academic performance, with studies indicating variations in success levels between male and female students (Degé *et al.* 2014). Similarly, ethnicity plays a critical role in student academic success, with research emphasizing the influence of ethnic identity and parents' goals on students' academic achievements (Abbasi *et al.* 2019). Additionally, parental education level is linked to student performance, where higher parental education levels are generally associated with better academic outcomes (Tang 2011).

Psychological factors like achievement motivation, locus of control, and academic self-concept are also key determinants of academic success (Wenglinsky 1996; Bizuneh 2021). Motivation levels and beliefs about one's abilities significantly affect academic performance, with high achievement motivation correlating with greater academic success (Hosova and Duchovicova 2019). Locus of control, representing individuals' beliefs about control over their lives, is associated with academic achievement, where internal locus of control is linked to better performance (Iyengar *et al.* 2022). Academic self-concept, reflecting students' perceptions of their academic abilities, plays a pivotal role in shaping their academic outcomes (Corbière *et al.* 2006).

Social factors including parental involvement, peer influence, and socioeconomic status have been shown to influence student performance (Jaiswal and Choudhuri 2017; Marsh and Yeung 1997; Erkman *et al.* 2010). Parental engagement in education is consistently linked to increased academic success, with supportive family environments contributing positively to students' achievements. Peer influence can also impact student performance, with social networks and friendships influencing academic outcomes . Additionally, socioeconomic status is a significant predictor of academic success, with students from higher socioeconomic backgrounds generally achieving better educational outcomes .

In conclusion, student performance is influenced by a complex interplay of factors, encompassing individual characteristics like motivation and self-concept, social influences such as parental involvement and peer relationships, and broader institutional practices and educational environments. Understanding these multifaceted determinants is crucial for developing effective strategies to support student success and enhance academic achievement.

The main research question of this study is: "What are the key factors that influence student performance and how to determine the relative effects of these factors on student academic achievement?"

This research question aims to analyze the effects of demographic, psychological, social and institutional factors on student performance and determine the importance of these factors. In the realm of machine learning, various regression models have been

Manuscript received: 13 June 2024,
Revised: 28 June 2024,
Accepted: 28 June 2024.

[1]emredeniz@hitit.edu.tr (**Corresponding author**)

extensively studied and applied across different domains to predict outcomes and make informed decisions. Among the popular regression models are Linear Regression, Decision Tree Regressor, Random Forest Regressor, K-Nearest Neighbors Regressor, Support Vector Regressor (SVR), Gradient Boosting Regressor (GBR), XGBoost Regressor (XGB), and Neural Network (Multi-Layer Perceptron, MLP). These models have been employed in diverse fields such as healthcare, environmental science, physics, and more to address a wide range of prediction tasks (Hassanzadeh *et al.* 2022).

Linear Regression, a fundamental and widely used regression model, forms the basis for many predictive analytics tasks. It establishes a linear relationship between the input variables and the target variable, making it a simple yet effective tool for prediction. Decision Tree Regressor operates by recursively partitioning the data into subsets based on certain features, creating a tree-like structure to make predictions. Random Forest Regressor, an ensemble method built on decision trees, combines multiple trees to improve prediction accuracy and reduce overfitting (Azar *et al.* 2022). K-Nearest Neighbors Regressor predicts the target variable by considering the 'k' nearest data points in the feature space.

Support Vector Regressor (SVR) utilizes support vectors to find the optimal hyperplane that best separates the data points in a high-dimensional space. Gradient Boosting Regressor sequentially builds multiple weak learners to create a strong predictive model, minimizing errors at each step. XGBoost Regressor, known for its efficiency and performance, implements gradient-boosted decision trees and is considered a state-of-the-art model for structured data, often outperforming deep learning models in regression tasks (Ferreira *et al.* 2024).

In the context of specific applications, these regression models have been leveraged for various predictive tasks. For instance, in the medical field, machine learning models like Random Forest, Support Vector Machine (SVM), and XGBoost have been utilized for survival prediction in diseases such as ovarian cancer. These models play a crucial role in analyzing patient data and making informed decisions regarding treatment and prognosis (Fei *et al.* 2019). Moreover, in environmental science, regression models like Gradient Boosting Regressor, Linear Regression, K-Nearest Neighbors Regressor, Random Forest Regressor, and XGBoost have been employed to predict outcomes related to solar energy harvesting and air pollution forecasting. These models aid in optimizing processes, enhancing efficiency, and making data-driven decisions in environmental research (Gonçalves *et al.* 2023). Furthermore, in physics and material science, regression models such as XGBoost, Random Forest, and Support Vector Regression have been utilized for tasks like predicting reduction potentials for complexes and conducting single-molecule conductance measurements.

In the domain of public health, machine learning models have been instrumental in predicting outcomes related to pandemics like COVID-19. Regression models such as Linear Regression, Support Vector Machine Regressor, Random Forest Regressor, and XGBoost Regressor have been employed to forecast disease outbreaks, analyze mortality rates, and guide public health interventions. These models provide valuable insights for policymakers and healthcare professionals to make informed decisions and mitigate the impact of health crises (Belho and Rawat 2023). Overall, the diverse applications of regression models in various fields underscore their significance in predictive analytics, decision-making, and knowledge discovery. By leveraging the strengths of different regression algorithms, researchers and practitioners can extract valuable insights from data, optimize processes, and drive innovation across a wide range of domains.

MATERIALS AND METHODS

In this study, a data set containing performance data of 1000 students was used (SPSScientist 2018). The variables included in the data set are:

Gender: Female and male

Ethnicity: Group A, Group B, Group C, Group D, Group E

Parental Level of Education: Some high school, high school, some college, associate's degree, bachelor's degree, master's degree

Lunch Type: Free/reduced and standard

Test Preparation Course: None and completed

Course Scores: Mathematics, reading and writing scores (between 0-100)

In the data preprocessing stage, categorical variables were converted to numerical values and numerical variables were normalized. This study was carried out using exploratory data analysis, correlation analysis and various machine learning models (linear regression, decision trees, random forest, K-Nearest Neighbors, Support Vector Regressor, Gradient Boosting Regressor, XGBoost Regressor, Multi-Layer Perceptron).

Data Preprocessing

First of all, missing and incorrect data in the data set were checked. Categorical variables were converted to numerical values and numerical variables were normalized. These operations are important to make the data set suitable for machine learning models.

Figure 1 shows the detailed exploratory anaylsis of the data. The distributions of the variables in the data set were examined using histograms. Additionally, a correlation matrix was created to understand the linear relationships between variables.

Gender: The number of male and female students in the data set is almost equally distributed.

Ethnicity: Although there is no significant difference between ethnicity groups, Group C and Group D seem to have the most students.

Parental Level of Education: The majority of parents have received education up to undergraduate level.

Lunch Type: The majority of students receive standard lunch.

Test Preparation Course: The majority of students have not completed the test preparation course.

Mathematics, Reading and Writing Scores: These scores show a wide distribution and the density is concentrated around the average score.

The correlation matrix which showed in Figure 2 evaluates linear relationships between variables. The findings obtained in the correlation analysis are as follows:

Test Prep Course: Mathematics correlates positively with reading and writing scores. This shows that students who completed the test preparation course received higher scores.Math, Reading, and Writing Scores: There are strong positive correlations between these three scores. Students who perform well in one subject often perform well in other courses.

Machine Learning Models

In this study, various machine learning models were used to predict student performance. Training and performance evaluation of the models were performed by hyperparameter optimization using GridSearchCV. The models and hyperparameter settings used are:

Linear Regression Model: LinearRegression Hyperparameters: No hyperparameter tuning is done.

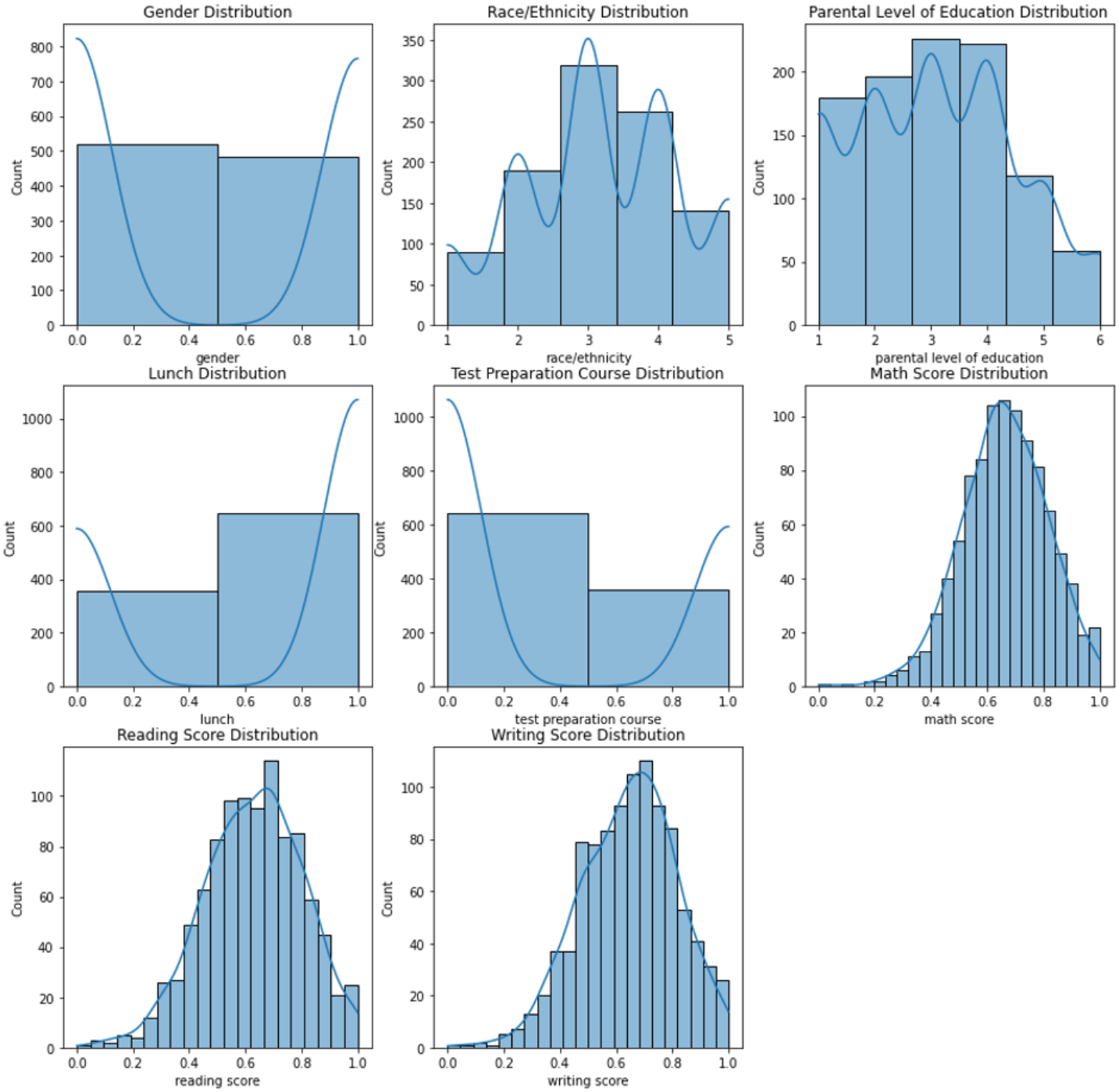

Figure 1 Exploratory Data Analysis

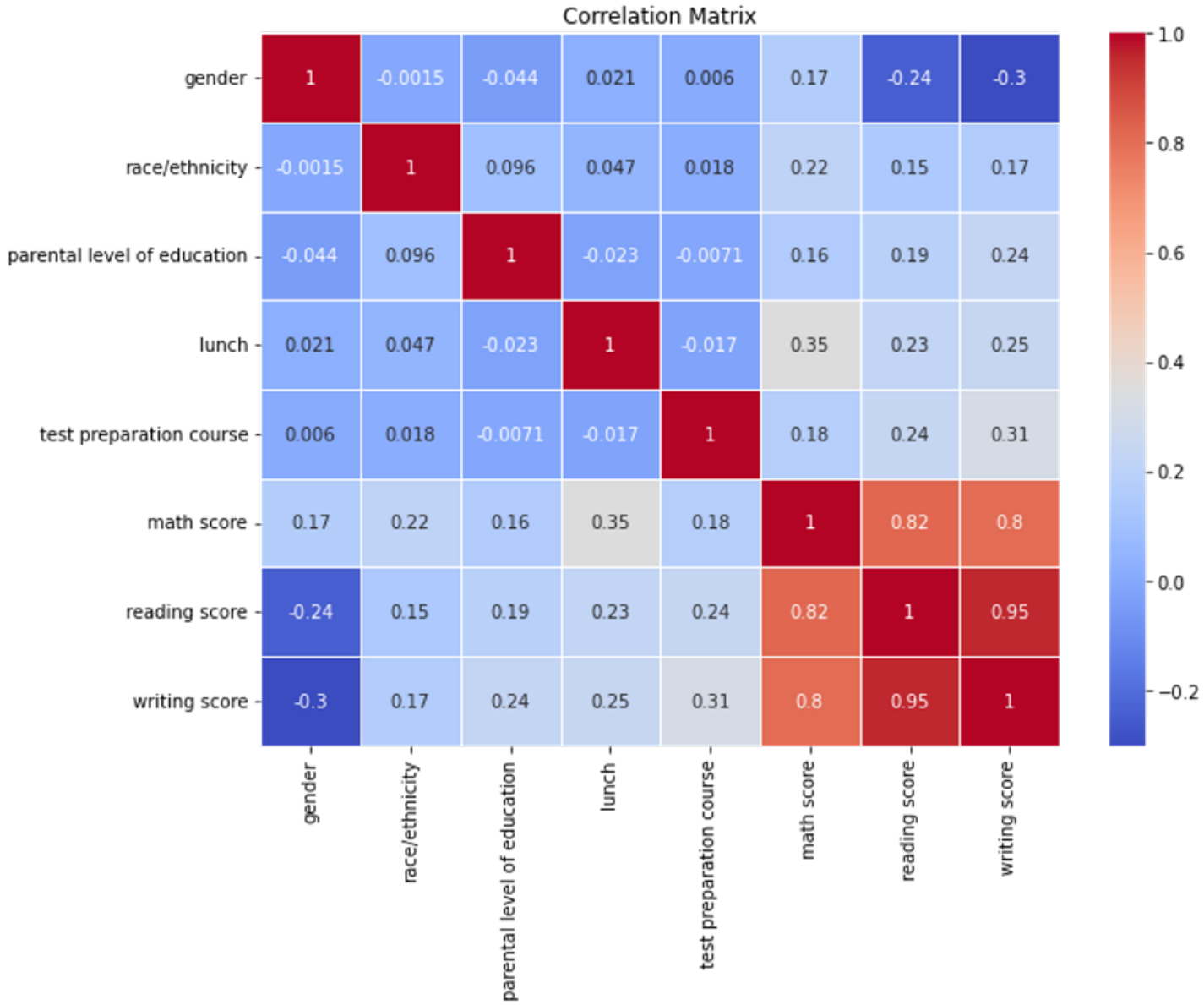

Figure 2 Correlation Matrix

Decision Tree Model: DecisionTreeRegressor Hyperparameters: 'max-depth': 10, 'min-samples-leaf': 4, 'min-samples-split': 10

Random Forest Model: RandomForestRegressor Hyperparameters: 'max-depth': 10, 'min-samples-leaf': 2, 'min-samples-split': 10, 'n-estimators': 200

K-Nearest Neighbors (KNN) Model: KNeighborsRegressor Hyperparameters: 'algorithm': 'brute', 'n-neighbors': 9, 'weights': 'distance'

Support Vector Regressor (SVR) Model: SVR Hyperparameters: 'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'

Gradient Boosting Regressor (GBR) Model: GradientBoostingRegressor Hyperparameters: 'learning-rate': 0.1, 'max-depth': 3, 'n-estimators': 100, 'subsample': 1.0

XGBoost Regressor (XGB) Model: XGBRegressor Hyperparameters: 'learning-rate': 0.1, 'max-depth': 3, 'n-estimators': 100

Neural Network (MLP) Model: MLPRegressor Hyperparameters: 'activation': 'tanh', 'alpha': 0.05, 'hidden-layer-sizes': (50, 100, 50), 'learning-rate': 'constant', 'solver': 'adam'

RESULTS

The performance of various machine learning models was evaluated using Mean Squared Error (MSE) and R-squared (R^2) metrics, as shown in Table 1. The Support Vector Regressor (SVR) and Linear Regression models demonstrated the best performance, with MSE values of 0.0028 and 0.0028 and R^2 values of 0.8866 and 0.8854 respectively. These models showed superior ability in predicting student performance accurately.

The strong performance of these models suggests that linear relationships among variables play a significant role in predicting academic achievement. Additionally, models like Random Forest and Gradient Boosting also showed high accuracy, indicating their robustness in handling complex data interactions. The results highlight the importance of optimizing test preparation strategies and suggest that focusing on interrelated academic subjects can lead to enhanced student performance.

Table 1 Results of Machine Learning Models

Model	Mean Squared Error (MSE)	R-squared (R^2)
Linear Regression	0.0028	0.8854
Decision Tree	0.0046	0.8116
Random Forest	0.0036	0.8533
K-Nearest Neighbors	0.0054	0.7798
Support Vector Regressor	0.0028	0.8866
Gradient Boosting Regressor	0.0030	0.8755
XGBoost Regressor	0.0032	0.8685
Neural Network (MLP)	0.0036	0.8512

CONCLUSION

This study has demonstrated the effectiveness of various machine learning models in predicting student academic achievement and highlighted the significant impact of test preparation courses on student performance. The findings indicate that demographic factors such as gender and ethnicity are not direct determinants of academic success, suggesting that educational policies should focus on enhancing educational experiences and preparation.

Educational institutions are recommended to prioritize test preparation courses and integrate data-driven approaches to identify and support students at risk of underperforming. Policies should be designed to foster interconnected learning across subjects to maximize student achievement. Future research should aim to validate these findings using larger and more diverse datasets and explore the long-term effects of different educational strategies.

Availability of data and material

Not applicable.

Conflicts of interest

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

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Abbasi, H., V. Mehdinezhad, and M. Shirazi, 2019 Impact of jigsaw technique on improving university students' self-concept. Educational Research in Medical Sciences **8**.

Azar, A. S., S. B. Rikan, A. Naemi, J. B. Mohasefi, H. Pirnejad, *et al.*, 2022 Application of machine learning techniques for predicting survival in ovarian cancer. BMC Medical Informatics and Decision Making **22**.

Belho, K. and M. S. Rawat, 2023 Response of hydro-meteorological hazards to environmental degradation in kohima district of nagaland, north east india. International Journal of Scientific Research in Science, Engineering and Technology pp. 339–349.

Bizuneh, S. M., 2021 Belief in counselling effectiveness, academic self-concept as correlates of academic help-seeking behavior among college students. Journal of Education and Practice .

Cantekin, Ö. F., 2020 The effects of academic self-concept and organizational factors on academic achievement. Bartın University Journal of Faculty of Education **9**: 26–35.

Corbière, M., F. Fraccaroli, V. Mbékou, and J. Perron, 2006 Academic self-concept and academic interest measurement: A multisample european study. European Journal of Psychology of Education **21**: 3–15.

Degé, F., S. Wehrum, R. Stark, and G. Schwarzer, 2014 Music lessons and academic self-concept in 12- to 14-year-old children. Musicae Scientiae **18**: 203–215.

Erkman, F., A. Caner, H. Sakız, B. Börkan, and K. Şahan, 2010 Influence of perceived teacher acceptance, self-concept, and school attitude on the academic achievement of school-age children in turkey. Cross-Cultural Research **44**: 295–309.

Fei, X., Q. Zhang, and Q. Ling, 2019 Vehicle exhaust concentration estimation based on an improved stacking model. IEEE Access **7**: 179454–179463.

Ferreira, R. A. S., S. F. H. Correia, L. Fu, P. Georgieva, M. Antunes, *et al.*, 2024 Predicting the efficiency of luminescent solar concentrators for solar energy harvesting using machine learning. Scientific Reports **14**.

Gonçalves, D. M., R. Henriques, and R. S. Costa, 2023 Predicting metabolic fluxes from omics data via machine learning: Moving from knowledge-driven towards data-driven approaches. Computational and Structural Biotechnology Journal **21**: 4960–4973.

Hassanzadeh, H., J. Boyle, S. Khanna, B. Biki, and F. Syed, 2022 Daily surgery caseload prediction: Towards improving operating theatre efficiency. BMC Medical Informatics and Decision Making **22**.

Hosova, D. and J. Duchovicova, 2019 Gender differences in self-concept of gifted pupils. In *CBU International Conference Proceedings*, volume 7.

Iyengar, R. N., G. Gouri, M. Kumar, and Y. Yanjana, 2022 Academic self concept and academic achievement of indian cbse school students. National Journal of Community Medicine **12**: 405–410.

Jaiswal, S. K. and R. Choudhuri, 2017 Academic self concept and academic achievement of secondary school students. American Journal of Educational Research **5**: 1108–1113.

Jones, M. H., S. Audley, and S. M. Kiefer, 2012 Relationships among adolescents' perceptions of friends' behaviors, academic self-concept, and math performance. Journal of Educational Psychology **104**: 19–31.

Korantwi-Barimah, J. S., A. Ofori, E. Nsiah-Gyabaah, and A. M. Sekyere, 2017 Relationship between motivation, academic self-concept and academic achievement amongst students at a ghanaian technical university. International Journal of Human Resource Studies **7**.

Marsh, H. W. and A. S. Yeung, 1997 Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. Journal of Educational Psychology **89**: 41–54.

SPSScientist, 2018 Students performance in exams [data set].

Tang, S., 2011 The relationships of self-concept, academic achievement and future pathway of first year business studies diploma students. International Journal of Psychological Studies **3**.

Wenglinsky, H., 1996 Measuring self-concept and relating it to academic achievement: Statistical analyses of the marsh self-description questionnaire. ETS Research Report Series **1996**.

How to cite this article: Deniz, E. Evaluating the Effectiveness of Machine Learning Models in Predicting Student Academic Achievement. *ADBA Computer Science*, 1(1), 8-13, 2024.