Journal of Artificial Intelligence and Metaheuristic (JAIM)

♦ ISSN: 2833-5597 Vol: 07, No. 02, 2024 ♦ PP. 38–42

Article DOI: https://doi.org/10.54216/JAIM.070206



FULL LENGTH ARTICLE

Predictive Modeling of Global Educational Outcomes: A Comparative Analysis Using Machine Learning Regression Techniques

Abdelhameed Ibrahim[1,*] **Abdelaziz A. Abdelhamid**[2] **Ehab M. Almetwally**[3]

[1] Computer Engineering and Control Systems Department, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt
[2] Department of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt
[3] Department of Mathematics and Statistics, Faculty of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

Emails: afai79@mans.edu.eg · abdelaziz@cis.asu.edu.eg · emalmetwally@imamu.edu.sa

Received: July 05, 2023 Revised: September 29, 2023 Accepted: March 02, 2024 ⋆ Corresponding author

ABSTRACT

Education is central to social and economic development, making school enrollment and the quality of learning important global priorities. This study applies machine-learning regression to the World Educational Data dataset from Kaggle. After data exploration and preprocessing, eleven models are compared: Support Vector Regression (SVR), CatBoost, Random Forest, Extra Trees, XGBoost, a multilayer perceptron (MLP), Gradient Boosting, Decision Tree, K-Nearest Neighbors, Linear Regression, and a pipeline model. Performance is evaluated using MSE, RMSE, MAE, MBE, the correlation coefficient, R^2, NSE, and WI. CatBoost reports the strongest correlation and coefficient of determination (both approximately 0.99999) with an MSE of 0.04024. The results illustrate the potential of machine-learning methods to support analysis and prediction of educational outcomes and to inform educational planning.

Keywords: Educational Data Analysis ▪ Regression Models ▪ Machine Learning ▪ Predictive Modeling ▪ Global Education Outcomes

1. INTRODUCTION

Educational quality is an important determinant of a country's social and economic development, affecting employment opportunities and national competitiveness. As countries seek to improve their education systems, identifying the factors that influence educational outcomes becomes essential. Regression analysis is a useful statistical approach for examining relationships among these factors and informing decisions by policymakers and educators [1, 2, 3].

Regression analysis can estimate associations between an outcome, such as educational quality, and one or more explanatory variables. In cross-national analyses, it can help identify predictors of differences in educational performance, including socioeconomic conditions, teacher qualifications, school financing, and education policy [4, 5, 6].

This study examines global educational outcomes using predictors that include national education expenditure, student–teacher ratios, and educational attainment. It aims to identify factors associated with education quality and compare their predictive contribution across regression models. Findings of this kind can help guide resource allocation and policy development, including decisions about teacher training and support for communities facing socioeconomic disadvantage [7, 8, 9].

© 2024 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

Regression-based analysis can therefore contribute to broader efforts to improve fairness and effectiveness in education systems. The remainder of the paper reviews related work, describes the dataset and methods, presents the model comparison, and discusses the implications and conclusions [10, 11, 12].

2. RELATED WORK

Regression methods have been used to study educational quality by measuring relationships between educational outcomes and features of education systems. Linear regression remains a common approach because it is relatively simple and its results are interpretable. It can, for example, estimate associations between student–teacher ratios, education funding, parental involvement, and student performance, although it may not capture nonlinear relationships or interactions [13].

Polynomial regression extends linear models to represent nonlinear relationships between predictors and outcomes, including possible higher-order effects of educational resources and socioeconomic conditions [14]. Generalized linear models, including logistic and Poisson regression, support outcomes with distributions other than the normal distribution, including categorical and count data [15].

Machine-learning methods have also been applied to educational data. Decision trees and random forests can represent nonlinear relationships and interactions among variables, while remaining useful for exploring which factors may influence educational outcomes [16, 17]. Support Vector Machines can model nonlinear relations through kernel functions and have been used to distinguish levels of educational quality [18].

Neural networks can learn complex patterns from educational datasets, and ensemble approaches combine multiple models to improve prediction or reduce variance. Boosting methods update models sequentially, whereas bagging combines predictions from models trained on different samples [19, 20]. Factor analysis and Principal Component Analysis (PCA) have also been used to identify latent dimensions and reduce the number of measured attributes while retaining much of the data variation [21]. Integrating regression with machine-learning approaches has been proposed as a way to combine interpretable baseline relationships with stronger nonlinear prediction [22].

3. METHODOLOGY

3.1 Data Source and Description

3.1.1 Data Source

The study uses the *World Educational Data* dataset obtained from Kaggle. The dataset covers multiple aspects of education, including enrollment, literacy, teacher–student ratios, government expenditure on education, and socioeconomic factors related to education.

3.1.2 Key Variables and Their Importance

The dataset includes the following groups of variables:

- **Enrollment rates:** Enrollment in primary, secondary, and tertiary education, reflecting access to educational institutions.
- **Literacy rates:** The proportion of people in an age group who can read and write, providing an indicator of educational attainment.
- **Teacher–student ratios:** The number of students per teacher, a measure related to the learning environment and the attention available to students.
- **Government expenditure on education:** The share of gross domestic product or total government expenditure allocated to education.
- **Socioeconomic factors:** Economic growth, purchasing power, consumer and poverty indices, and employment status, which provide context for educational outcomes.
- **Test scores and assessments:** Achievement scores and averages from international assessments such as PISA and TIMSS.
- **Infrastructure and resources:** The availability of classrooms, books, technology, and other resources used in teaching and learning.

Taken together, these variables provide a broad basis for examining educational performance and constructing models to predict educational outcomes.

3.2 Data Exploration and Preprocessing

3.2.1 Preprocessing and Cleaning

The source study describes the following data preparation steps:

- **Initial inspection:** Review the dataset structure, variable types, missing values, descriptive statistics, and distributions.
- **Missing values:** Identify missing observations; use mean or median imputation for numerical variables, mode imputation for categorical variables, or advanced methods such as K-Nearest Neighbors or model-based imputation. Variables with extensive missing data may be excluded when they are not important.
- **Outliers:** Detect outliers using measures such as Z-scores and the interquartile range, along with graphical tools. Possible treatments include winsorization, transformations such as a logarithm, or removal of erroneous or unrepresentative observations.
- **Normalization:** Standardize numerical variables to mean zero and standard deviation one, or use Min–Max scaling to map features to a range such as $[0, 1]$. These transformations are relevant to scale-sensitive models such as SVR and neural networks.
- **Feature engineering:** Derive interaction or polynomial features and convert categorical variables to numerical representations using one-hot or label encoding.
- **Data splitting:** Separate training and test data, typically using either a 70/30 or 80/20 split, and apply cross-validation to assess model validity and reduce overfitting.
- **Exploratory analysis:** Use histograms, scatter plots, correlation matrices, and statistical tests to examine patterns and associations among variables.

The feature-importance plot in Figure 1 shows the relative importance assigned to predictors of educational outcomes and can help identify variables that contribute to model predictions.

Figure 2 presents the distribution of demographic indicators in the dataset. Exploring these distributions provides context

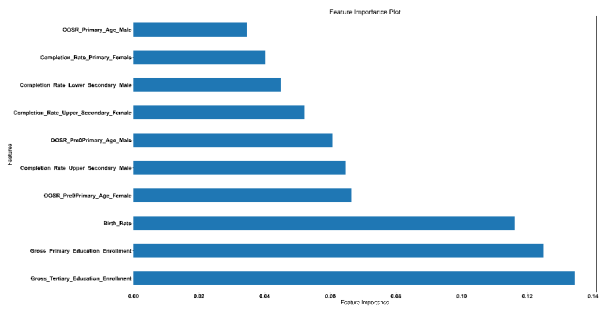


Figure 1. Feature-importance plot.

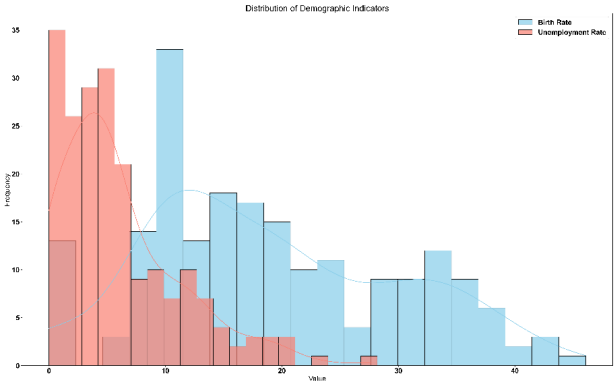


Figure 2. Distribution of demographic indicators.

for interpreting the regression results and the relationships among variables.

3.3 Regression Models Applied

The study compares linear and nonlinear models, ensemble methods, a neural network, and a pipeline model. The set of models is intended to cover different modelling assumptions and improve the breadth of the comparison.

- **Support Vector Regression (SVR):** Supports linear and nonlinear regression using an epsilon-insensitive margin. It is included for its performance in high-dimensional spaces and its ability to represent nonlinear relations through kernels.
- **CatBoost:** A gradient-boosting method designed to handle categorical variables and limit overfitting. It is included for its performance with categorical features.
- **Random Forest Regressor:** Combines predictions from decision trees trained on different data subsets. Averaging across trees can reduce overfitting and support high-dimensional data.
- **Extra Trees Regressor:** Similar to Random Forest, with additional randomness in tree construction, intended to reduce variance and improve speed on large datasets.
- **XGBoost:** A flexible and efficient gradient-boosting library included for its speed and suitability for large-scale and sparse data.
- **MLP Regressor:** A multilayer perceptron that can learn nonlinear dependencies and interactions among variables.
- **Gradient Boosting Regressor:** Trains models sequentially, with each new model addressing errors made by earlier models.
- **Decision Tree Regressor:** A non-parametric model that provides a straightforward, interpretable point of comparison.
- **K-Neighbors Regressor:** Predicts an outcome by averaging the values of nearby observations and is included as a simple method for smaller datasets.
- **Linear Regression:** Provides a simple and interpretable baseline by modelling the outcome as a linear function of one or more predictors.
- **Pipeline:** Combines preprocessing steps with a final estimator to support consistent transformations during model fitting and evaluation.

3.4 Model Validation and Evaluation Metrics

Model performance is assessed with the following metrics:

- **Mean Squared Error (MSE):** The mean of the squared differences between observed and estimated outcomes; lower values indicate smaller errors.
- **Root Mean Squared Error (RMSE):** The square root of MSE, expressed in the units of the outcome variable.
- **Mean Absolute Error (MAE):** The mean absolute difference between actual and predicted values; it is less sensitive to outliers than MSE.
- **Mean Bias Error (MBE):** The average signed prediction error, indicating whether predictions tend to be above or below actual values.
- **Correlation coefficient (r):** Measures the strength and direction of the linear relationship between predicted and actual outcomes.
- **Coefficient of determination (R^2):** Indicates the proportion of variance in the dependent variable explained by the predictors.
- **Nash–Sutcliffe Efficiency (NSE):** A measure of predictive skill that ranges from negative infinity to 1, with 1 indicating a perfect fit.
- **Willmott Index (WI):** Assesses prediction accuracy on a scale from 0 to 1, with 1 indicating a perfect fit.

The validation techniques described are K-fold cross-validation, which rotates the test fold to use each data point for testing once, and grid search with cross-validation, which evaluates parameter combinations to select model hyperparameters.

4. RESULTS AND DISCUSSION

4.1 Comparison of Regression Models

Table 1 summarizes the reported performance metrics for the eleven regression models. The table reports CatBoost with the highest correlation and R^2 values, while SVR and CatBoost share the lowest reported MSE.

The discussion in the source study attributes the performance of ensemble methods such as CatBoost, Random Forest, and Extra Trees to combining multiple trees and reducing variance. CatBoost and XGBoost use boosting to improve weak learners iteratively. The MLP Regressor can learn complex patterns, while SVR can model nonlinear relationships through kernel functions. Linear Regression serves as a simple and interpretable baseline.

Figure 3 reproduces the reported comparison of mean squared error across models.

Table 1. Performance metrics for the regression models.

Model	MSE	RMSE	MAE	MBE	r	R^2	NSE	WI	Fitted time (s)
SVR	0.04024	0.20060	0.14844	0.03410	0.95740	0.91662	0.09460	0.56054	0.31393
CatBoost	0.04024	0.20060	0.14813	0.02646	0.99999	0.99999	0.09458	0.56145	45.01141
RandomForestRegressor	0.04261	0.20641	0.15770	0.02047	0.99998	0.99996	0.04133	0.53313	0.68596
ExtraTreesRegressor	0.04262	0.20644	0.15515	0.02737	0.99966	0.99933	0.04113	0.54067	0.00399
XGBoost	0.04269	0.20661	0.15800	0.03294	0.99522	0.99046	0.03947	0.53222	26.82379
MLPRegressor	0.04299	0.20734	0.14794	0.05397	0.99240	0.98485	0.03271	0.56202	4.43813
GradientBoostingRegressor	0.04329	0.20807	0.15688	0.03757	0.99121	0.98249	0.02590	0.53554	0.00450
DecisionTreeRegressor	0.04700	0.21679	0.16149	0.01497	0.98986	0.97981	0.02590	0.52189	0.22202
KNeighborsRegressor	0.04750	0.21796	0.16215	0.03281	0.98650	0.97317	0.06886	0.51994	0.02729
LinearRegression	0.05103	0.22589	0.17029	0.04412	0.96248	0.92636	0.02729	0.49585	0.01876
pipeline	0.13915	0.37303	0.23620	0.03873	0.95740	0.91662	0.52189	0.30070	0.24061

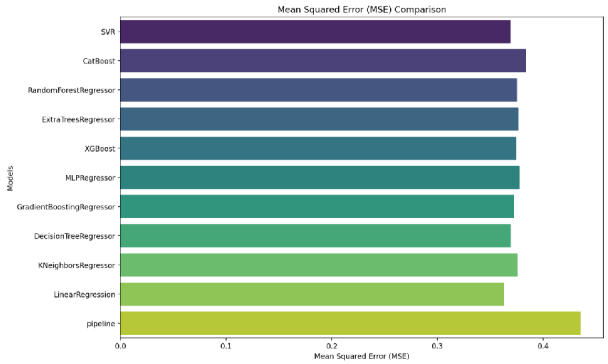


Figure 3. Mean squared error of the regression results.

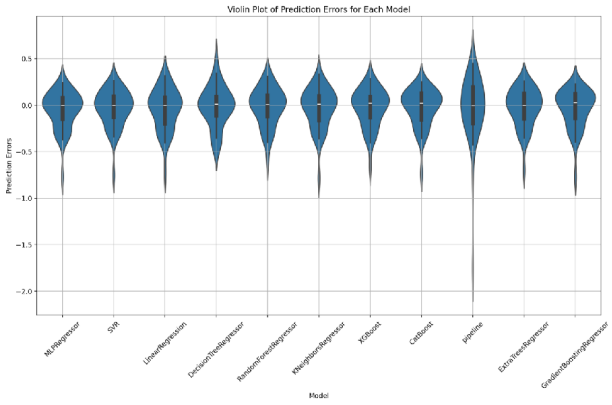


Figure 4. Violin plots of prediction errors for the regression models.

Figure 4 shows the reported violin plot, which summarizes the distribution of model prediction errors and helps illustrate their spread and potential extreme values.

4.2 Implications

The analysis illustrates how machine-learning methods, particularly ensemble techniques and models able to represent nonlinear patterns, can support prediction of educational outcomes. Comparing the models also provides information about their relative strengths and limitations, which can inform the selection of forecasting methods for education data.

5. CONCLUSION

This study compared regression models for predicting educational outcomes using the World Educational Data dataset from Kaggle. The reported results identify CatBoost and Random Forest as strong performers, with CatBoost reporting a correlation of approximately 0.999996 and an MSE of about 0.0402. SVR and the MLP Regressor also performed well, supporting the use of methods that can represent nonlinear patterns. Linear Regression provides a simple baseline, while ensemble models can reduce prediction variance by combining multiple estimators.

Future work can incorporate additional data sources, assess more advanced models, and use longitudinal data to study educational outcomes over time. Improving data collection to reduce missing values and outliers, alongside further feature engineering, may also strengthen predictive models.

REFERENCES

[1] M. Ahmad, Z. Ahmed, X. Yang, N. Hussain, and A. Sinha, "Financial development and environmental degradation: Do human capital and institutional quality make a difference?" *Gondwana Research*, vol. 105, pp. 299–310, 2022.

[2] M. K. Anser, M. Ahmad, M. A. Khan, K. Zaman, A. A. Nassani, S. E. Askar, M. M. Q. Abro, and A. Kabbani, "The role of information and communication technologies in mitigating carbon emissions: Evidence from panel quantile regression," *Environmental Science and Pollution Research*, vol. 28, no. 17, pp. 21 065–21 084, 2021.

[3] M. Azam, L. Liu, and N. Ahmad, "Impact of institutional quality on environment and energy consumption: Evidence from developing world," *Environment, Development and Sustainability*, vol. 23, no. 2, pp. 1646–1667, 2021.

[4] R. Banerjee, V. Mishra, and A. A. Maruta, "Energy poverty, health and education outcomes: Evidence from the developing world," *Energy Economics*, vol. 101, p. 105447, 2021.

[5] C. Cheng, X. Ren, K. Dong, X. Dong, and Z. Wang, "How does technological innovation mitigate CO2 emissions in OECD countries? heterogeneous analysis using panel quantile regression," *Journal of Environmental Management*, vol. 280, p. 111818, 2021.

[6] H.-A. H. Dang and T.-A. Trinh, "Does the COVID-19 lockdown improve global air quality? new cross-national evidence on its unintended consequences,"

Journal of Environmental Economics and Management, vol. 105, p. 102401, 2021.

[7] F. Díez, A. Villa, A. L. López, and I. Iraurgi, "Impact of quality management systems in the performance of educational centers: Educational policies and management processes," *Heliyon*, vol. 6, no. 4, p. e03824, 2020.

[8] H. Doreswamy, K. S., Y. Km., and I. Gad, "Forecasting air pollution particulate matter (PM2.5) using machine learning regression models," *Procedia Computer Science*, vol. 171, pp. 2057–2066, 2020.

[9] D. Filmer, H. Rogers, N. Angrist, and S. Sabarwal, "Learning-adjusted years of schooling (LAYS): Defining a new macro measure of education," *Economics of Education Review*, vol. 77, p. 101971, 2020.

[10] F. Habibi and M. A. Zabardast, "Digitalization, education and economic growth: A comparative analysis of middle east and OECD countries," *Technology in Society*, vol. 63, p. 101370, 2020.

[11] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Systems*, vol. 200, p. 105992, 2020.

[12] M. S. I. Khan, N. Islam, J. Uddin, S. Islam, and M. K. Nasir, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, Part A, pp. 4773–4781, 2022.

[13] K. H. Lee, H. Xu, and B. Wu, "Gender differences in quality of life among community-dwelling older adults in low- and middle-income countries: Results from the study on global ageing and adult health (SAGE)," *BMC Public Health*, vol. 20, no. 1, p. 114, 2020.

[14] H. Liu, M. Alharthi, A. Atil, M. W. Zafar, and I. Khan, "A non-linear analysis of the impacts of natural resources and education on environmental quality: Green energy and its role in the future," *Resources Policy*, vol. 79, p. 102940, 2022.

[15] M. K. Mahalik, H. Mallick, and H. Padhan, "Do educational levels influence the environmental quality? the role of renewable and non-renewable energy demand in selected BRICS countries with a new policy perspective," *Renewable Energy*, vol. 164, pp. 419–432, 2021.

[16] A. A. Maruta, R. Banerjee, and T. Cavoli, "Foreign aid, institutional quality and economic growth: Evidence from the developing world," *Economic Modelling*, vol. 89, pp. 444–463, 2020.

[17] U. Mehmood, "Contribution of renewable energy towards environmental quality: The role of education to achieve sustainable development goals in G11 countries," *Renewable Energy*, vol. 178, pp. 600–607, 2021.

[18] V. Raghupathi and W. Raghupathi, "The influence of education on health: An empirical assessment of OECD countries for the period 1995–2015," *Archives of Public Health*, vol. 78, no. 1, p. 20, 2020.

[19] G. Santos, C. S. Marques, E. Justino, and L. Mendes, "Understanding social responsibility's influence on service quality and student satisfaction in higher education," *Journal of Cleaner Production*, vol. 256, p. 120597, 2020.

[20] M. W. Zafar, M. Shahbaz, A. Sinha, T. Sengupta, and Q. Qin, "How renewable energy consumption contribute to environmental quality? the role of education in OECD countries," *Journal of Cleaner Production*, vol. 268, p. 122149, 2020.

[21] M. W. Zafar, A. Sinha, Z. Ahmed, Q. Qin, and S. A. H. Zaidi, "Effects of biomass energy consumption on environmental quality: The role of education and technology in asia-pacific economic cooperation countries," *Renewable and Sustainable Energy Reviews*, vol. 142, p. 110868, 2021.

[22] C. Zhang, I. Khan, V. Dagar, A. Saeed, and M. W. Zafar, "Environmental impact of information and communication technology: Unveiling the role of education in developing countries," *Technological Forecasting and Social Change*, vol. 178, p. 121570, 2022.