Journal of Artificial Intelligence and Metaheuristic (JAIM)

♦ ISSN: 2833-5597 Vol: 07, No. 01, 2024 ♦ PP. 20–29

Article DOI: https://doi.org/10.54216/JAIM.070103



FULL LENGTH ARTICLE

Exploring Predictive Models for Students' Performance in Exams: A Comparative Analysis of Regression Algorithms

Faris H. Rizk[1,⋆] **Ahmed Saleh**[2] **Abdulrhman Elgaml**[2] **Ahmed Elsakaan**[2] **Ahmed Mohamed Zaki**[1]

[1] Computer Science and Intelligent Systems Research Center, Blacksburg 24060, Virginia, USA

[2] Department of Communications and Electronics, Delta Higher Institute of Engineering and Technology, Mansoura, 35111, Egypt

Emails: faris.rizk@jcsis.org; Ch1900135@dhiet.edu.eg; Ch1900053@dhiet.edu.eg; Ch1900089@dhiet.edu.eg; azaki@jcsis.org

Received: April 17, 2023 Revised: August 17, 2023 Accepted: January 04, 2024 ⋆ Corresponding author

ABSTRACT

Student-centered analysis of academic performance is also the most important aspect in improving education by being able to determine what measures work best, individualized learning approaches, and intervention programs. In this study, we performed a detailed analysis based on the "Students Performance in Exams" dataset and different regression methods to estimate students' grades. We sought to assess the functioning of numerous metrics and determine an optimal model for this task. Our descriptive analysis identified meaningful trends within this dataset, as it includes central factors like gender, race/ethnic diversity-based status of a student, and parental education level based on which the children are catered to by informing them about important lunches and test preparation courses alongside scores in "Math," "Readings," "Writing," etc. We used a wide range of regression models: XGBoost, CatBoost, GradientBoostingRegressor, etc. Metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Average Marginal Loss were used to assess each model rigorously. Importantly, the XGBoost model gave out an MSE value of 0.028, which was the best among all values obtained from various other models. The superiority of the XGBoost model is supported by the excellent performance that was reported across many metrics. This work can be important for informing educational practitioners and policymakers regarding the best possible accurate and realistic model that would predict the students' outcome results. Educational data analytics incorporating the XGBoost model can be used for the customization of interventions and mapping resource allocation while promoting a results-oriented approach based on data in education. This study is a step towards the accumulation of knowledge on educational data analytics. It can serve as a background for further research aimed at improving predictive models regarding student performance.

Keywords: Predictive Modeling ▪ Education Analytics ▪ Regression Models ▪ Students Performance ▪ Descriptive Analysis ▪ XGBoost

1. INTRODUCTION

In the realm of education, understanding and predicting students' academic performance holds paramount importance for educators, policymakers, and researchers. The ability to foresee students' success not only aids in personalized learning but also facilitates targeted interventions, ultimately fostering a more effective and equitable educational system [1]. With the advent of advanced data analytics, particularly predictive modeling, there exists an unprecedented opportunity to unravel intricate patterns within educational datasets that can inform decision-making processes and improve overall learn-

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

ing outcomes. This research endeavors to contribute to this burgeoning field by employing sophisticated regression algorithms to analyze a dataset titled "Students Performance in Exams." Comprising variables such as gender, race/ethnicity, parental level of education, lunch, test preparation course, and scores in math, reading, and writing, this dataset encapsulates diverse aspects of students' backgrounds and academic experiences [2]. The overarching goal is to discern the most effective regression model for predicting students' performance, thereby enhancing our understanding of the multifaceted factors influencing educational outcomes.

The dataset under consideration reflects the rich diversity inherent in educational settings. The inclusion of demographic variables like gender, race/ethnicity, and parental level of education is pivotal, as it enables a nuanced exploration of how these factors intersect with academic achievement. Moreover, the incorporation of auxiliary information such as lunch and test preparation course affords a comprehensive view of external influences on student performance. As traditional approaches to educational assessment evolve, predictive modeling emerges as a powerful tool to decipher complex relationships within educational data. By leveraging machine learning algorithms, this research seeks to not only describe observed patterns but also to forecast future academic performance. Such predictions can pave the way for proactive interventions, ensuring timely support for students at risk and optimizing educational resources.

Against this backdrop, the primary objective of this study is to identify the most effective regression model for predicting students' scores in exams [3]. To achieve this, an array of regression algorithms, including XGBoost, CatBoost, GradientBoostingRegressor, and others, have been applied to the dataset. The ensuing comparative analysis aims to unveil the strengths and weaknesses of each model, with a particular focus on performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and others [4]. By determining the optimal predictive model, this research aims to contribute valuable insights to educational practitioners, administrators, and policymakers. The findings have the potential to guide the development of targeted strategies for improving students' academic outcomes. Moreover, the exploration of diverse regression algorithms sheds light on the suitability of these models in the context of educational data analytics, advancing our understanding of their applicability and efficacy. The key technical contributions of this work can be summarized as follows:

- **Dataset Preprocessing:** The "Students Performance in Exams" dataset was intricately described, encompassing key demographic variables such as gender, race/ethnicity, and parental level of education. The preprocessing steps involved a meticulous approach to handle missing data, ensuring dataset completeness. For categorical variables like gender and race/ethnicity, encoding techniques such as one-hot encoding or label encoding were applied to prepare them for model training. Additionally, numerical feature normalization was executed using methods like Min-Max scaling or Z-score normalization to maintain consistent scales across variables.
- **Descriptive Analysis Insights:** The descriptive analysis involved a thorough examination of the dataset, featuring advanced statistical measures such as mean, median, standard deviation, and visualizations like histograms, scatter plots, and correlation matrices. The objective was to unravel complex patterns, correlations, and identify potential outliers, thereby providing nuanced insights into the dataset's characteristics. These insights played a pivotal role in guiding subsequent model selection decisions.
- **Regression Model Diversity:** A diverse set of regression models was employed, each chosen for its specific strengths in addressing the dataset's nuances. The models included advanced techniques such as XGBoost, CatBoost, GradientBoostingRegressor, and ensemble methods like ExtraTreesRegressor and RandomForestRegressor. Additionally, traditional models such as LinearRegression, SVR, KNeighborsRegressor, MLPRegressor, and DecisionTreeRegressor were also incorporated. The selection rationale was based on the dataset's non-linearity, potential interactions, and the need for robust predictive performance.
- **Performance Metrics Selection:** A comprehensive set of performance metrics was meticulously selected to gauge the effectiveness of the regression models. Metrics included Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Pearson correlation coefficient (R), R-squared (R^2), Relative Root Mean Squared Error (RRMSE), Nash-Sutcliffe Efficiency (NSE), Willmott Index (WI), and Fitted Time. Each metric was chosen for its ability to provide a multifaceted evaluation of model performance across different dimensions.
- **Results and Model Comparison:** Detailed results of model application were presented, highlighting comparative performance metrics across the diverse set of regression models. The XGBoost model emerged as the standout performer, particularly excelling in minimizing MSE. This validation of the chosen data preparation and modeling approach underscores the efficacy of the XGBoost algorithm in handling the complexities of the dataset. The nuanced technical details contribute to a thorough understanding of the regression models' capabilities and limitations.

In the subsequent sections, we delve into the existing literature on predictive modeling in education, detail the methodology employed, and present the results of our comprehensive analysis. Through this endeavor, we hope to make a meaningful contribution to the ongoing discourse surrounding data-driven approaches to enhance educational practices.

2. LITERATURE REVIEW

For a long time, educational research has been investigating the many things that affect how well students do in school. Test scores are a big part of this. It tells you more about books like the one you are reading and shows them to you. These books are all about how well kids do on tests. It looks at the results from a lot of different studies and research methods to get a full picture of the complicated factors that make studying hard or easy. The main goal is to look at how cognitive and psychological factors, socio-economic factors,

and teaching methods affect test results and how they interact with each other.

It is stressed that having a lot of educational data and technology-enhanced learning tools lets us look at how students learn in new ways, figure out how to make classes better, and make decisions based on that data. Part of its learning analytics method is to use large amounts of data from virtual classes [5]. A deep artificial neural network is taught on carefully chosen clickstream data traits to find kids who might be in danger and learn how to help them right away. Logistic regression and support vector machine models are better at putting things into groups than this model is. It is right 84% to 93% of the time. In line with other studies, this one also looks at how old data and test data can change how accurate a model is. Due to the increased availability of data mining, learning data mining is gaining popularity [6]. The newest directed machine learning methods are used to guess how well students will do on tests. We take a close look at them and compare them. In order for artificial neural networks to do their best at both classification and regression tasks, they need to know how interested the students are and how well they have done in the past. Adding demographic information only helps the guesses a little. People who want to get the most out of guessing how well students will do on tests are told that they need good tools for collecting data and for students to be involved in the learning environment.

Because there is so much data about students, the field of education study has grown very quickly [7]. Educational data mining helps teachers see what their students have done well and what they need to work on. Some new ways of sorting and grouping things are used to make predictions. These tests are done on real-time data from Indian college and university students in Kerala who are learning many different topics. This program uses grouping and classification methods and is more accurate than most classification methods. The connection to the Internet for college students is in real time, and this changes how they learn and also their daily lives. Therefore, [8] investigates the connection between the level of web usage and the level of success that could make machine learning predict how first-year college students will perform. This is bolstered by our attempt to view some important traits by looking at how 4000 students actually conduct themselves on the Internet. The classification techniques include decision trees, neural networks, and support vector machines. However, the study reveals quite unexpected associations between percentage change in the amount of online time and percentage change in performance among some data sets. You can distinguish your children and predict how many grades they will receive if you monitor the amount of time spent on the Internet.

Putting college students into groups based on how well they might do in school is very important [9]. You can guess how well a student will do in school by looking at data from the end of the first year. Our plan is made up of two steps. Things work differently than they did before. Both the average marks and the length of time it takes to finish the degree are used to track growth in school. A group of 2459 students from a European Engineering School at a public research university are used to test the suggested segmentation structure and prediction model. It is possible to guess how well a child will do in school before they even start. Alwarthan et al. systematically reviewed educational data mining studies on higher-education performance prediction, summarizing commonly used techniques and influential features that can inform student support and institutional decision-making [10].

Want to keep kids and help them learn more? Schools need to know how well their students will do in school [11]. There are several ways that data from an Australian university is used to try to guess how well kids will do in school. Much thought goes into the different types of students because the way they are made may affect how eager they are to learn. Models made with tools that use rules and trees are easy for users to understand. In [12], researchers have come up with a new way to teach search skills that will work for both teachers and students. Numerous tasks can be used to learn. Algorithm visualizations are one way to show how hard algorithms work by using ideas from active learning. Teachers can see which students are doing well with the built-in testing tools. These tools also make it easier to grade and give tutors quick, useful comments. One study that was done in a real classroom and looked at the method in detail showed that it helps people learn. To make comparisons, real teachers who are pros are used. These results show that the way judges work automatically and the way people decide are very much the same.

Alturki et al. studied the prediction of master's students' academic performance in Germany using records from more than 700 students. They compared six classifiers, including support vector machines (SVMs) and artificial neural networks (ANNs), and found that academic achievement could be predicted using a small set of readily available features, supporting early-warning applications [13].

Along with other works, it has a part with many studies that check how well students do on tests. Based on these studies, it is clear that school is hard and needs much thought. The study looked at a lot of different areas, from early childhood schooling to higher education. Kind of person, social setting, teaching methods, and drive are all very important. There is new academic talk in these works, and the studies in them change how schools run their policies and programs. Experts, teachers, and politicians all work together to find the best ways to help kids learn and do well in school.

3. PROPOSED METHODOLOGY

Understanding the intricate process of predictive modeling in the realm of educational data analytics is essential for unraveling the complexities of students' academic performance. As shown in Figure 1, this visual guide serves as a roadmap, presenting the proposed methodology of this paper. The figure outlines the sequential steps taken to analyze the "Students Performance in Exams" dataset and evaluate various regression models. The methodology section of this study further details the research methods approach, describing data description, full detailed analysis, model selection and setup based on regressions, as well as different performance measures. By providing a visual overview of our approach and detailing the research methods, Figure 1 and the methodology section collectively offer readers a clear understanding of the systematic methodology employed in this study. Join us in navigating through the key stages outlined in the figure and

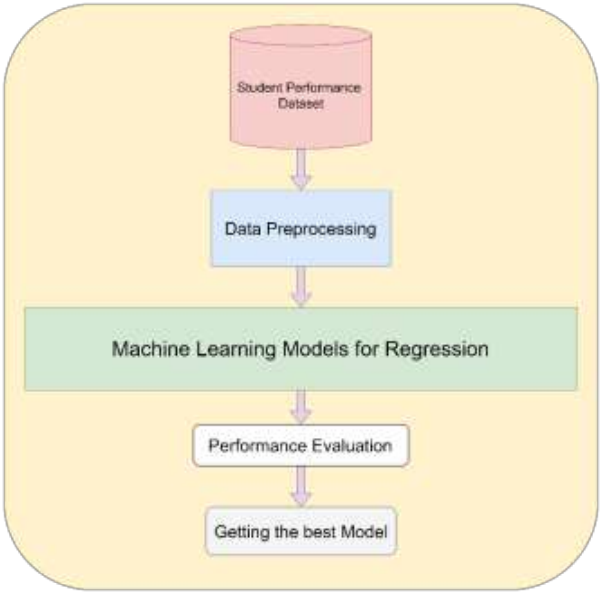


Figure 1. Proposed Methodology of This Paper.

the comprehensive research methods, each contributing to the robust analysis and insightful findings presented in the subsequent sections of this research.

3.1 Dataset Description

The data set in this research is titled "Students Performance In Exams" and has various properties that have been designed to demonstrate different facets of students' academic performance as well as their demographic information. The following fields make up the dataset: In Spirit's opinion, these two cabin girls celebrated their glory days once more during excursions.

- **Gender:** The student gender identity presented in this field might help us to explain why some students perform worse academically than others depending on gender, which is plausible because it always plays a role.
- **Race/Ethnicity:** Combining children into groups of different races and ethnicities so that cultural factors as well as socioeconomic contributors can be studied together with how they perform at school.
- **Parental Level of Education:** This indicates the levels of education possessed by parents and guardians and, consequently, serves as an indicator of their socioeconomic position level, touching on how involved they are in managing learning activities.
- **Lunch:** By clarifying whether a luncheon is free, reduced-price, or regular, researchers and developers in educational studies have used this information as an indication of socioeconomic status.
- **Test Preparation School:** This indicates whether a student has attended prep school and what effect preparatory activities have on academic performance.
- **Math Score:** A figure that indicates good performance in mathematics–a subject area with vast ripple effects on both students' learning and their holistic paths as far as further schooling or employment is concerned.
- **Reading Score:** That requires checking how well the kids get what they read, as that is a lesson in reading comprehension, which facilitates school performance and success across all subjects.
- **Writing Score:** This includes checking how well children can write, and this is vital for a perfect connection with the learning institutions.

After pre-analysis, the dataset went through many careful procedures to ensure good-quality data suited for modeling. As part of these measures, missing values were dealt with, and the categorical variables were encoded using one-hot encoding or label encoding. When it comes to numerical features, they should be normalized in the sense that their errors occur due to scales herein [14].

Race or culture is correlated with academic achievement in mathematics, reading, and writing, as shown in Figure 2. The implication of this is that each box plot drives how scores are distributed across racial or ethnic lines. This helps to compare two or more groups of students so as to find out how well they perform at places in school. The box plots, therefore, represent how scores are distributed among the various racial or ethnic groups. They provide us with valuable information on any possible disparities in the performance of learners at school. This demonstrates the significance of addressing problems with equality in educational settings.

3.2 Descriptive Analysis

To detect the details and underlying trends of information, a detailed descriptive analysis was done. This experimental research involved various statistical approaches and visualizations to provide several varying perspectives on the dataset. With regard to numerical features, several important statistics measures such as the mean value, median value, standard deviation, minimums, and maximum values were discovered–such provided information on central tendency, dispersion and range. Many types of graphs were created, whether it was histograms or box plots to scatter charts and correlation matrices, all for the purpose of deciphering complicated relationships, finding outliers, etc. This holistic descriptive analysis not only creates a picture of the information but also reveals intriguing possible connections, shown above anomalies and areas that require further research.

Figure 3 shows three important academic areas: mathematics, reading and writing. The histograms illustrate the distribution score through these regions. This picture shows the frequency of these scores, which provide a broad overview of the main trends and distinctions among subject areas. Through the distribution of scores within each subject, we can understand how good children do in the classroom from skill level cluster. These lay the foundation for further analysis and interpretation.

Box plots in Figure 4 indicate the association between parents' education and children's achievement in academic reading and math results obtained in writing. The box plot for each of the above represents how these scores are distributed across parents with various level qualifications. This enables us to have a deeper insight into the influence of parental education level on differences in performance. These box plots allow for the implementation of a simplified comprehension in regard

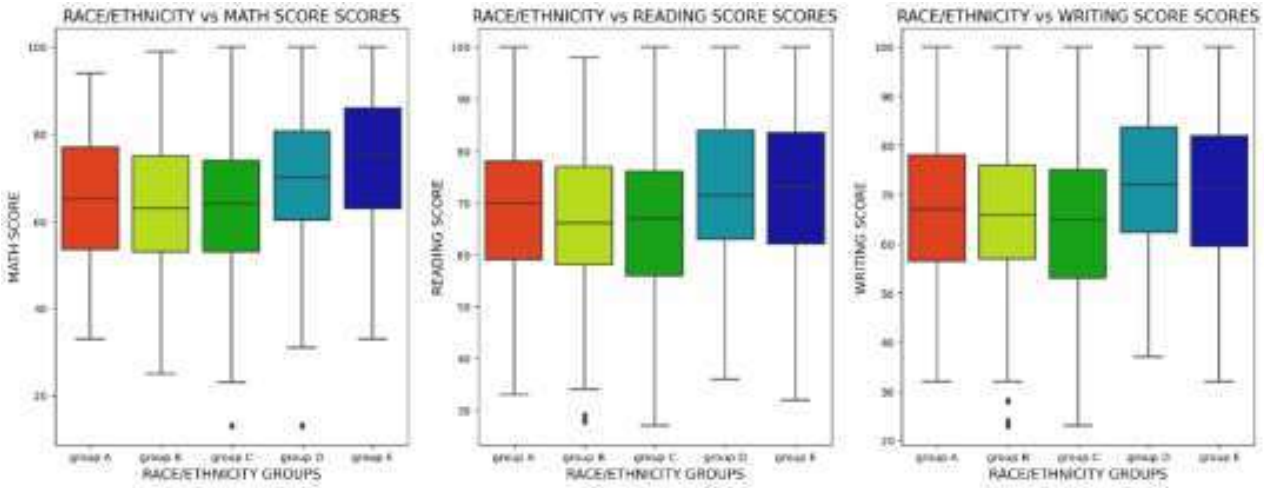


Figure 2. A box plot for scores on race/ethnicity versus reading, writing and math skills.

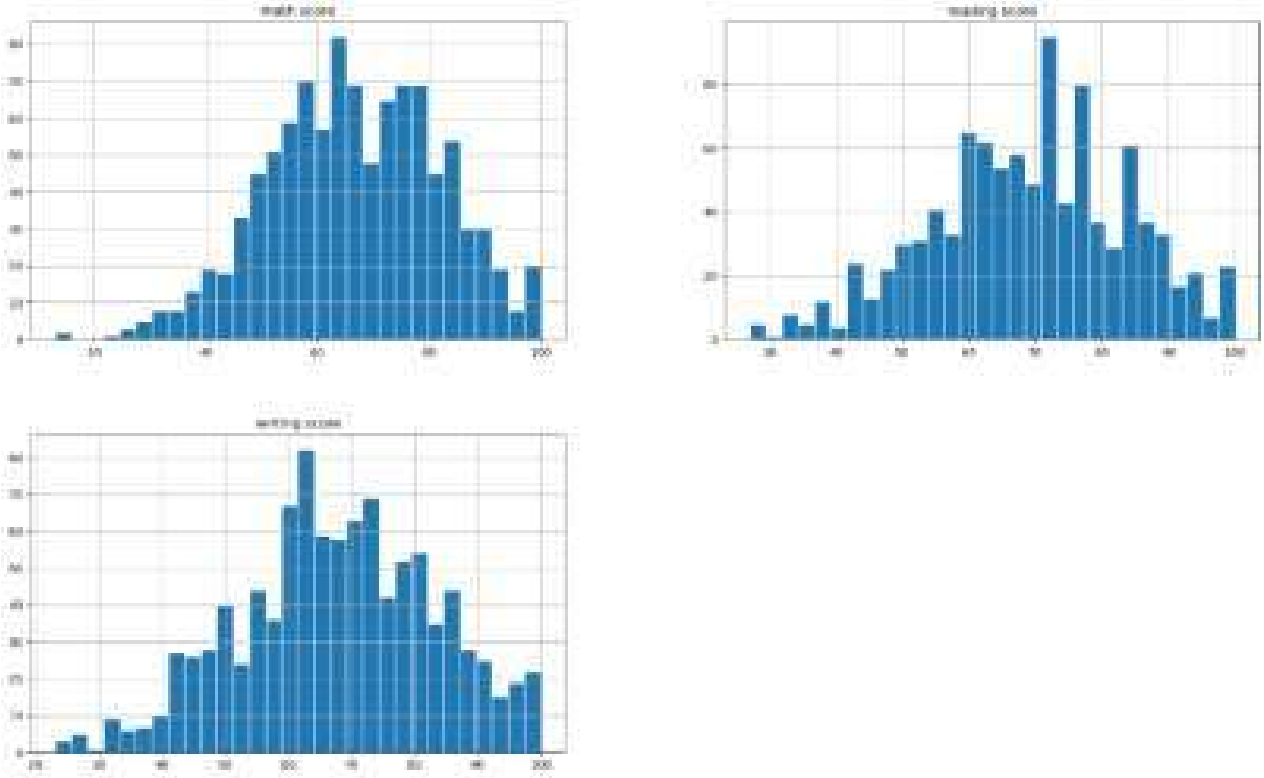


Figure 3. Histograms features the scores of reading, writing and math.

to how family background affects performance in academics; they reveal the distribution patterns from various parental education levels. This contributes to the fairness of talks about educational attainment.

To view the connection matrix of math, reading, and writing results, one should refer to Figure 5. It shows a heatmap. The type of heatmap used to create this illustration is a Correlation Map, which is made up of different colors at various positions in order to show how some quantitative variables' internal relationships here were observed. It is a shortcut method for the visualization of these relationships. If there is any purpose in finding possible trends and connections between different subjects, this does not harm this means. This elicits further studies about the backgrounds that influence student's performance in school.

3.3 Models for Regression

A number of regression models from a broad spectrum were carefully selected to facilitate the prediction of test performance [15]. A number of considerations, for instance, mathematical elasticity, calculative quickness, and previous experience in applying regression models to analyze educational data, made it possible. All the regression models used in this research were very distinct. They presented non-parametric methods like simple linear regression techniques and the more complicated ensemble learning algorithms. In particular, the following regression methods were set up and used with great care: Exercise is not a guarantor of losing or gaining weight; it simply reduces rather than increases the number of calories one takes in.

1. **XGBoost:** XGBoost is an ensemble learning method, specifically a gradient boosting algorithm [16]. The model

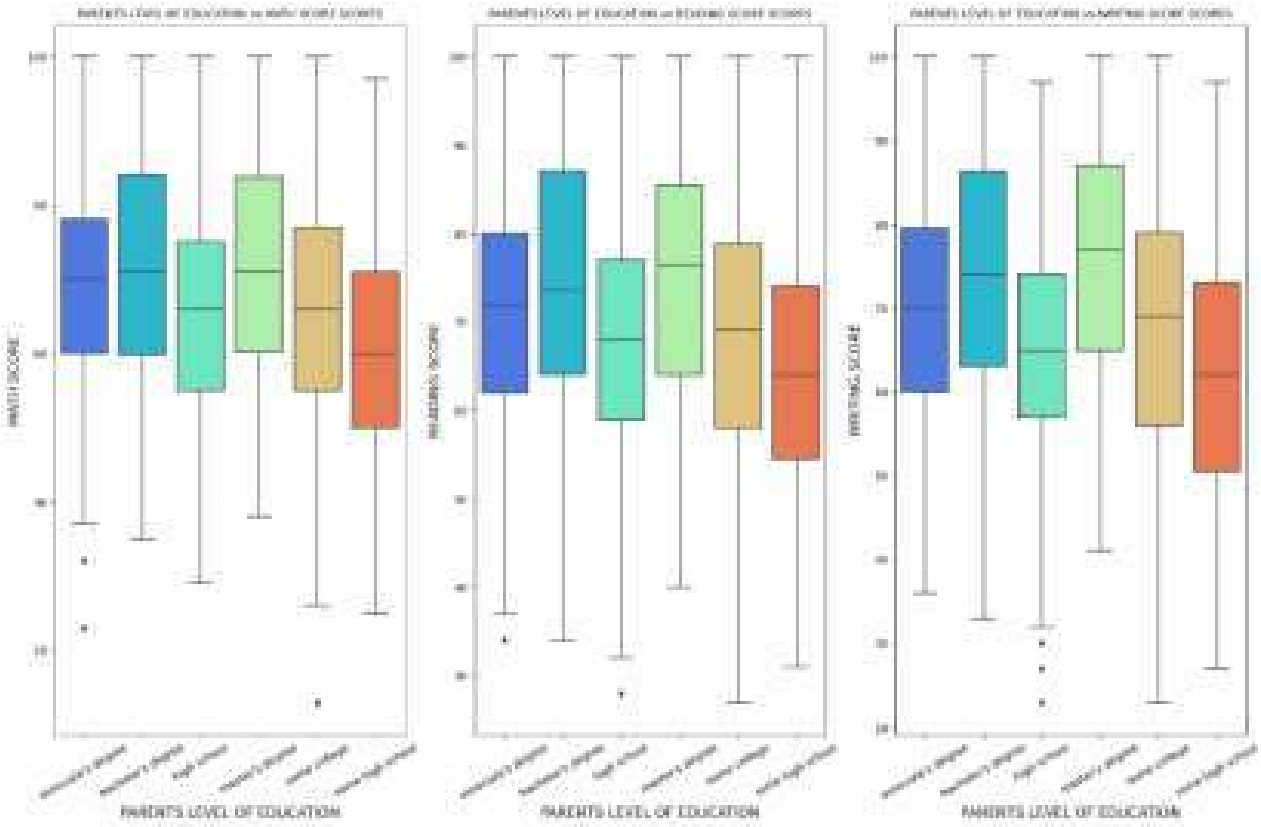


Figure 4. Child's math, reading and writing scores compared to the level of education his/her parents have.

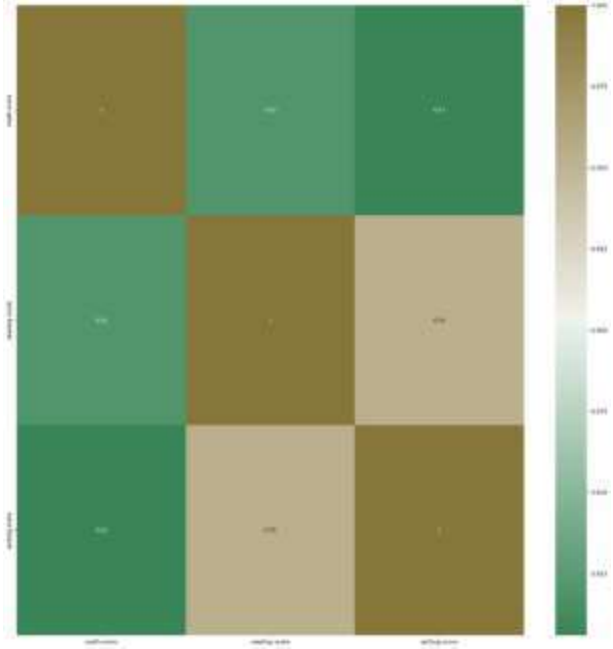


Figure 5. A heatmap of scores in reading, writing, and math.

prediction is based on the gradient-boosting from multiple weak learners (usually decision trees). The final prediction is given by:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \tag{1}$$

where f_k is the k-th weak learner.

2. **CatBoost:** CatBoost is similar to XGBoost but is designed to handle categorical features efficiently [17]. The prediction is given by:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \tag{2}$$

where f_k is the k-th weak learner.

3. **Gradient Boosting Regressor:** Gradient Boosting Regressor builds an additive model in a forward stage-wise manner [18]. The prediction is given by:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \tag{3}$$

where f_k is the k-th weak learner.

4. **Pipeline:** The pipeline is a concept in scikit-learn for streamlining a lot of the routine processes, including combining different models [19]. The mathematical equation is not directly applicable here as it's more of a software engineering concept.

5. **K-Neighbors Regressor:** K-Nearest Neighbors Regression predicts the output based on the average of the output values of its k-nearest neighbors [20]. The prediction is given by:

$$\hat{y}_i = \frac{1}{k} \sum_{j=1}^{k} y_{N_j}, \tag{4}$$

where N_j is the j-th nearest neighbor.

6. **Extra Trees Regressor:** Extra Trees Regressor is similar to Random Forests but builds multiple decision trees with random splits [21]. The prediction is an average of the predictions from all the trees.

7. **Linear Regression:** Linear Regression predicts the output as a linear combination of input features [22]. The prediction is given by:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_n x_{in}, \tag{5}$$

where $b_0, b_1, \ldots, b_n$ are the coefficients.

8. **Support Vector Regression (SVR):** SVR maps input features into a higher-dimensional space and finds a hyperplane that best represents the data [23]. The prediction is given by:

$$\hat{y}_i = \sum_{j=1}^{n} (\alpha_j - \alpha'_j) K(x_i, x_j) + b, \tag{6}$$

where K is the kernel function, α and α' are the Lagrange multipliers, and b is a bias term.

9. **MLP Regressor (Multi-Layer Perceptron):** The prediction in a Multi-Layer Perceptron is given by [24]:

$$\hat{y}_i = \phi(x_i) = f^{(L)}\left(f^{(L-1)}\left(\cdots f^{(1)}(x_i;\theta^{(1)})\cdots;\theta^{(L-1)}\right);\theta^{(L)}\right), \tag{7}$$

where $f^{(l)}$ is the activation function in layer l and $\theta^{(l)}$ are the weights.

10. **Random Forest Regressor:** Random Forest Regressor combines the predictions from multiple decision trees [25]. The prediction is an average of the predictions from all the trees.

11. **Decision Tree Regressor:** Decision Tree Regressor predicts the output based on a tree structure [26]. The prediction is given by traversing the tree from the root to a leaf node.

Every regression model was meticulously organized with a customized set of hyperparameters, and the accuracy of predictions increased through retesting. This entire process assisted in overcoming underfitting by moderating prediction statistics so as to produce more accurate estimates during testing and cross-validation instead or vice versa depending upon differentials related to errors made during such processes discussed hereafter below briefly.

3.4 Performance Metrics

In the realm of model evaluation for our study, the choice and application of evaluation metrics play a paramount role in scrutinizing the predictive performance of the selected models. A diversified set of metrics, as shown in Table 1, is employed to encompass various facets of model accuracy, precision, and generalization [27].

The following metrics are important parts of our full review framework: So, in this case, it is utterly irresponsible not to acknowledge the fact that as much as one can nullify a possible preterm birth.

1. **Mean Squared Error (MSE):** This is how the mean square difference between what really happened and that which was expected can be determined. A smaller MSE indicates a good model.

Table 1. Criteria for evaluating regression result.

Metric	Formula
RMSE	$\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{V}_n - V_n)^2}$
RRMSE	$\frac{RMSE}{\sum_{n=1}^{N}\hat{V}_n} \times 100$
MAE	$\frac{1}{N}\sum_{n=1}^{N}\lvert\hat{V}_n - V_n\rvert$
MBE	$\frac{1}{N}\sum_{n=1}^{N}(\hat{V}_n - V_n)$
NSE	$1 - \frac{\sum_{n=1}^{N}(V_n - \hat{V}_n)^2}{\sum_{n=1}^{N}(V_n - \bar{V}_n)^2}$
WI	$1 - \frac{\sum_{n=1}^{N}\lvert\hat{V}_n - V_n\rvert}{\sum_{n=1}^{N}\lvert V_n - \bar{V}_n\rvert + \lvert\hat{V}_n - \bar{V}_n\rvert}$
R^2	$1 - \frac{\sum_{n=1}^{N}(V_n - \hat{V}_n)^2}{\sum_{n=1}^{N}\left(\sum_{n=1}^{N} V_n - V_n\right)^2}$
r	$\frac{\sum_{n=1}^{N}(\hat{V}_n - \bar{\hat{V}}_n)(V_n - \bar{V}_n)}{\sqrt{\left(\sum_{n=1}^{N}(\hat{V}_n - \bar{\hat{V}}_n)^2\right)\left(\sum_{n=1}^{N}(V_n - \bar{V}_n)^2\right)}}$

2. **Root Mean Squared Error (RMSE):** This is also called the square root of MSE, and it indicates how large the errors are on average. A lower RMSE is more desirable, by analogy with MSE.

3. **Mean Absolute Error (MAE):** This function determines the average absolute bouncing $(y - x)$ between what actually occurred and what was anticipated. It helps you realize the scope of such an average misconduct.

4. **Mean Bias Error (MBE):** This is the average of discrepancies between what was presumed and observed. For positive values, the estimate is too high, and for negative ones–it is underestimated.

5. R **(Correlation Coefficient):** This number tells you how powerful and in what relation a linear relationship between two factors is. There is a very strong positive relationship indicated by a figure close to 1.

6. R^2**:** Also known as the coefficient of determination, reveals that many variations observed in a dependent variable can be determined by using some variables, such as an independent one. A higher R^2 value shows a better model fitting.

7. **RRMSE:** Means Relative Root Mean Squared Error, which represents the relation between both ranges from smallest value to biggest one. It gives normalized measures of the precision with which an estimate was made.

8. **NSE (Nash-Sutcliffe Efficiency):** This metric tests the accuracy of model forecasts in relation to annual average metrics actually observed. Closer values to 1 are indicative of the model's better functioning.

9. **WI (Willmott Index):** This verification measure determines to what extent actual and estimated figures correspond. Having a WI value of 1 indicates a perfectly precise agreement. Zero WI means a need for more agreement.

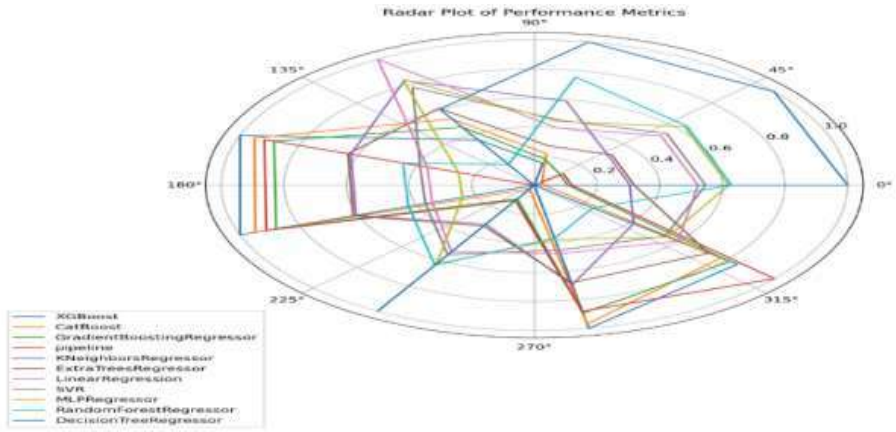


Figure 6. Radar plot of the performance metrics.

10. **Fitted Time:** This is the time it would be trained to fit or work on that data. The smaller numbers are considered faster.

4. RESULTS

In this section, we provide a detailed breakdown of students' performance prediction results based upon the application of an array of regression models. The various performance metrics, namely Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and others are used to see how well each model captures the complex patterns hidden in this data set.

4.1 Comparative Performance Metrics

Table 2 summarizes the results in full detail that provides a comparative analysis of each regression-based model. Let's delve into the intricacies of these metrics: No one had competed with Connranny in such a way.

Figure 6 showcases a radar plot comparing performance metrics of various regression models predicting exam scores. The visual presentation includes metrics like Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), providing a comprehensive view. The plot effectively illustrates strengths and weaknesses across models, simplifying the understanding of overall performance in one coherent picture.

4.2 Discussion of Findings

A full review of the performance indicators shows interesting patterns and trends within all data. In this regard, the most prominent advantage of XGBoost is its ability to predict complex interactions between several factors that influence student exam scores consistently and with minimal MSE. What is surprising, however, the closely aligned performance of CatBoost and GradientBoostingRegressor requires more research on interdependence between their algorithms in connection with adjustment over educational datasets. The pipeline model demonstrates the computational efficiency that does not compromise on accuracy, which is an attractive alternative when real-time predictions are most needed. Though KNeighborsRegressor and ExtraTreesRegressor provide acceptable accuracy, their increased Fitted Time may indicate the test's scalability problems when dealing with big educational datasets. The linear models, LinearRegression and SVR, provide useful insights but demonstrate the shortcomings of such approaches in making sense out of non-linearities implicit to educational data.

Figure 7 is a box plot illustrates prediction error distribution across analyzed regression models, offering insights into variability, central tendency, and comparability. Critical for assessing consistency, replicability, and predictive validity.

Figure 8 explores how regression models predict students' scores, revealing consistency and performance alignment through a Correlation Heatmap. The varying colors provide insights into correlation patterns, aiding in assessing predictive model reliability. Join us in decoding visual relationships for a deeper understanding of predictive modeling.

This detailed analysis serves as a benchmark in terms of performance for each model and reveals the key strengths and

Table 2. Regression result.

Models	MSE	RMSE	MAE	MBE	R	R^2	RRMSE	NSE	WI	Fitted Time
XGBoost	0.0280	0.1634	0.1409	-0.0135	0.4831	0.2334	25.0909	0.2286	0.5529	0.0317
CatBoost	0.0302	0.1737	0.1414	-0.0119	0.4798	0.2302	31.1343	0.2265	0.5513	11.0233
GradientBoostingRegressor	0.0304	0.1743	0.1411	-0.0126	0.4754	0.2260	31.2447	0.2210	0.5522	8.3450
Pipeline	0.0304	0.1744	0.1398	-0.0170	0.4775	0.2280	31.2623	0.2201	0.5566	0.1464
KNeighborsRegressor	0.0309	0.1759	0.1441	-0.0090	0.4577	0.2095	31.5278	0.2068	0.5428	21.7646
ExtraTreesRegressor	0.0310	0.1760	0.1418	-0.0111	0.4584	0.2101	31.5470	0.2058	0.5501	7.3560
LinearRegression	0.0315	0.1776	0.1427	-0.0074	0.4402	0.1938	31.8357	0.1912	0.5471	0.0160
SVR	0.0316	0.1778	0.1431	-0.0095	0.4419	0.1952	31.8699	0.1895	0.5459	0.7035
MLPRegressor	0.0318	0.1784	0.1431	-0.0089	0.4331	0.1876	31.9729	0.1842	0.5461	6.5425
RandomForestRegressor	0.0318	0.1784	0.1454	-0.0154	0.4458	0.1987	31.9861	0.1836	0.5388	0.5905
DecisionTreeRegressor	0.0329	0.1813	0.1472	-0.0112	0.4157	0.1728	32.5013	0.1571	0.5331	0.2056

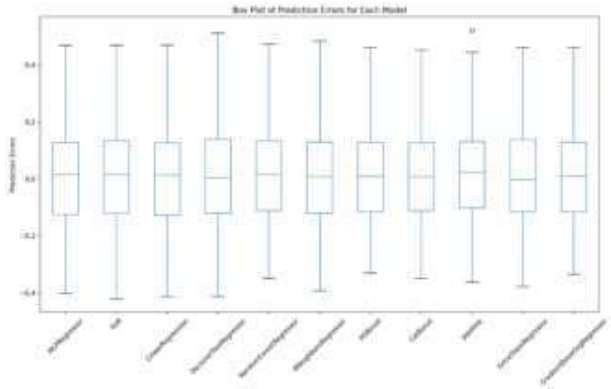


Figure 7. Box plot of prediction errors for each model.

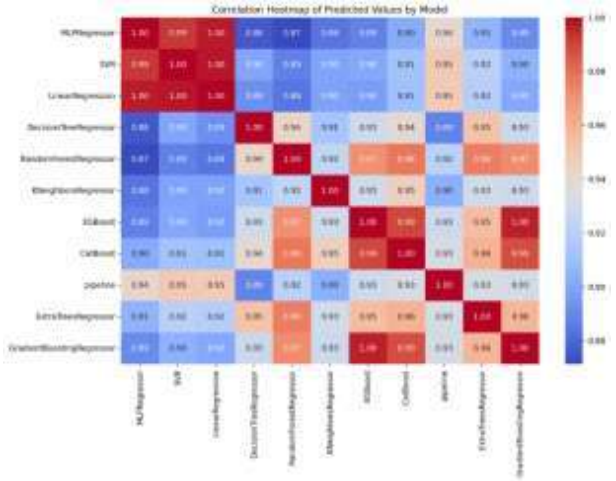


Figure 8. Correlation heatmap.

weaknesses. The findings clearly support XGBoost as the best model in predicting students' performance results, creating an example of adoption of more sophisticated regression models to educational analytics. This awareness provides educators, policymakers and researchers with the right knowledge to be able to make informed decisions as they implement predictive modeling in this fast-changing terrain of education.

5. CONCLUSION

In the work we presented here, by means of one dataset, "Students Performance in Exams," and several regression algorithms, this paper aimed to address students' performance modeling for various examination series so as to get the most optimized model among all those available. Our results accentuate the importance of adopting sophisticated machine-learning technologies in education analysis. As shown in our results, XGBoost performed consistently better than other regression algorithms from all of the performance measures we evaluated it on. This, however, indicates that the XGBoost classifier performs optimally in capturing the relationships within the data set and forecasting students' scores. Such attractiveness and performance of XGBoost become an asset for educational practitioners who want to guide themselves in the right direction with its aid relying on predictive analytics tools targeting student success interventions.

Our analysis gives great insight into dataset character and the factors affecting students' performance. The patterns and correlations found through descriptive analysis were noteworthy, highlighting the relationship between demographic factors, test preparation, and academic performance. Through analyzing a wide range of different regression models, we were able to gain an appreciation of their merits and flaws when used in predicting the learning process. Our results have significance that transcends the mere academic environment. Predictive modeling techniques educational institutions can utilize in identifying at-risk students, personalizing learning experiences to help them and intervene when they are tackling the problem. Policymakers can utilize these findings to inform evidence-based interventions and distribute resources optimally within the education system. Turning to the future, other parameters, such as socio-economic factors, school environment and behavioral elements, could be studied in order to obtain higher levels of accuracy for models. Additionally, discussing the generalizability of our findings in other educational environments and populations would shed light on predicting model validation possibilities in education. In this way, such a study serves as an example proving the predictive modeling's capability to forecast and explain students' achievements. Deploying powerful machine learning algorithms that rely on educational datasets opened a window to explore unexplored insights and possibilities, which lead others in diverse teaching environments towards success.

REFERENCES

[1] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Computers in Human Behavior*, vol. 73, pp. 247–256, 2017.

[2] Q. Liu, Z. Huang, Y. Yin, E. Chen, H. Xiong, Y. Su, and G. Hu, "EKT: Exercise-aware knowledge tracing for student performance prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 1, pp. 100–115, 2021.

[3] O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, A. Dakkak, and Y. El Allioui, "A multiple linear regression-based approach to predict student performance," in *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)*, edited by M. Ezziyyani, Springer International Publishing, pp. 9–23, 2020.

[4] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481–5487, 2022.

[5] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Computers in Human Behavior*, vol. 104, p. 106189, 2020.

[6] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Computers & Education*, vol. 143, p. 103676, 2020.

[7] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining

approach," *Journal of Medical Systems*, vol. 43, no. 6, p. 162, 2019.

[8] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," *Computers in Human Behavior*, vol. 98, pp. 166–173, 2019.

[9] V. L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," *Decision Support Systems*, vol. 115, pp. 36–51, 2018.

[10] S. A. Alwarthan, N. Aslam, and I. U. Khan, "Predicting student academic performance at higher education using data mining: A systematic review," *Applied Computational Intelligence and Soft Computing*, vol. 2022, no. 1, p. 8924028, 2022.

[11] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, D. J. Murray, and Q. Long, "Predicting academic performance by considering student heterogeneity," *Knowledge-Based Systems*, vol. 161, pp. 134–146, 2018.

[12] F. Grivokostopoulou, I. Perikos, and I. Hatzilygeroudis, "An educational system for learning search algorithms and automatically assessing student performance," *International Journal of Artificial Intelligence in Education*, vol. 27, no. 1, pp. 207–240, 2017.

[13] S. Alturki, L. Cohausz, and H. Stuckenschmidt, "Predicting master's students' academic performance: An empirical study in Germany," *Smart Learning Environments*, vol. 9, p. 38, 2022.

[14] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics–challenges in topic discovery, data collection, and data preparation," *International Journal of Information Management*, vol. 39, pp. 156–168, 2018.

[15] F. Raposo, R. Borja, and C. Ibelli-Bianco, "Predictive regression models for biochemical methane potential tests of biomass samples: Pitfalls and challenges of laboratory measurements," *Renewable and Sustainable Energy Reviews*, vol. 127, p. 109890, 2020.

[16] Y. Qiu, J. Zhou, M. Khandelwal, H. Yang, P. Yang, and C. Li, "Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration," *Engineering with Computers*, vol. 38, no. 5, pp. 4145–4162, 2022.

[17] G. Huang, L. Wu, X. Ma, W. Zhang, J. Fan, X. Yu, W. Zeng, and H. Zhou, "Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions," *Journal of Hydrology*, vol. 574, pp. 1029–1041, 2019.

[18] A. Keprate and R. M. C. Ratnayake, "Using gradient boosting regressor to predict stress intensity factor of a crack propagating in small bore piping," in *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 1331–1336, 2017.

[19] M. Xie and Z. Tian, "A review on pipeline integrity management utilizing in-line inspection data," *Engineering Failure Analysis*, vol. 92, pp. 222–239, 2018.

[20] Y. Song, J. Liang, J. Lu, and X. Zhao, "An efficient instance selection algorithm for *k* nearest neighbor regression," *Neurocomputing*, vol. 251, pp. 26–34, 2017.

[21] M. Reza and M. A. Haque, "Photometric redshift estimation using ExtraTreesRegressor: Galaxies and quasars from low to very high redshifts," *Astrophysics and Space Science*, vol. 365, no. 3, p. 50, 2020.

[22] A. F. Schmidt and C. Finan, "Linear regression and the normality assumption," *Journal of Clinical Epidemiology*, vol. 98, pp. 146–151, 2018.

[23] F. Zhang and L. J. O'Donnell, "Chapter 7–support vector regression," in *Machine Learning*, edited by A. Mechelli and S. Vieira, Academic Press, pp. 123–140, 2020.

[24] B. T. Pham, M. D. Nguyen, K.-T. T. Bui, I. Prakash, K. Chapi, and D. T. Bui, "A novel artificial intelligence approach based on multi-layer perceptron neural network and biogeography-based optimization for predicting coefficient of consolidation of soil," *CATENA*, vol. 173, pp. 302–311, 2019.

[25] R. Couronné, P. Probst, and A.-L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," *BMC Bioinformatics*, vol. 19, no. 1, p. 270, 2018.

[26] E. Pekel, "Estimation of soil moisture using decision tree regression," *Theoretical and Applied Climatology*, vol. 139, no. 3, pp. 1111–1119, 2020.

[27] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," in *Proceedings of the 8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2022)*, pp. 1–25, 2022.