An Intelligent Analytic Framework for Predicting Students Academic Performance Using Multiple Linear Regression and Random Forest

Ekemini A. Johnson, Jude A. Inyangetoh, Habeeb A. Rahmon, Tope G. Jimoh, Eduediuyai E. Dan, Mfon O. Esang
Corresponding Author: Ekemini A. Johnson

doi: https://doi.org/10.37745/ejcsit.2013/vol12n35670 Published June 17, 2024

Citation: Johnson E.A., Inyangetoh J.A., Rahmon H.A., Jimoh T.G., Dan E.E., Esang M.O. (2024) An Intelligent Analytic Framework for Predicting Students Academic Performance Using Multiple Linear Regression and Random Forest, *European Journal of Computer Science and Information Technology*, 12 (3),56-70

ABSTRACT: *In the contemporary educational landscape, data-driven decision-making has become pivotal for enhancing student success. This article explores an intelligent analytic framework leveraging Multiple Linear Regression (MLR) and Random Forest (RF) algorithms to predict student performance, providing a comparative analysis of their predictive capabilities. MLR, a statistical technique, models the relationship between students' grades and various factors such as attendance and socio-economic background, offering transparency and interpretability of the impact of each predictor. RF, an ensemble learning method, excels in handling large datasets and capturing non-linear interactions among variables, offering higher accuracy in prediction. The study was conducted using 664 datasets from eight departments of Federal Polytechnic Ukana, following rigorous data preprocessing and normalization. The performance of both models was evaluated based on metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared Score (R^2), and Explained Variance Score (EVS). The results revealed that RF outperformed MLR significantly, with lower error rates and higher predictive accuracy. Scatter plots and bar charts further illustrated the robust performance of RF over MLR. This research underscores the potential of integrating advanced machine learning techniques in educational settings to provide deeper insights into student performance, enabling timely and targeted interventions. The findings advocate for the adoption of RF for more accurate predictions and improved educational outcomes. Future research should explore hybrid models and expand the dataset to validate the applicability of these findings across diverse educational contexts.*

KEYWORDS: academic performance, random forest, multiple linear regression, machine learning and ensemble learning

INTRODUCTION

In the era of data-driven decision-making, the education sector stands at the forefront of leveraging advanced analytical tools to foster student success. As academic institutions strive to improve learning outcomes, predicting student performance has emerged as a critical challenge with far-reaching implications. An intelligent analytic framework utilizing both multiple linear regression (MLR) and random forest algorithms individually offers a promising solution, allowing for a comprehensive comparison of their predictive capabilities.

Multiple linear regression is a statistical technique that models the relationship between a dependent variable and multiple independent variables. In the context of academic performance, MLR can be used to predict students' grades by analyzing various factors such as attendance, study habits, socio-economic background, and previous academic records. By quantifying the impact of each variable, MLR provides educators with a clear understanding of the key determinants of academic success. This clarity enables targeted interventions, ensuring that resources are allocated efficiently to address specific needs.

On the other hand, the random forest algorithm (RF) is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification or mean prediction for regression. This algorithm excels in handling large datasets with high dimensionality and complex interactions among variables. RF is particularly adept at capturing non-linear relationships and interactions that simpler models like MLR might miss.

By using MLR and RF individually within the analytic framework, we can harness the unique strengths of each method. MLR offers a transparent model with easily interpretable coefficients, highlighting the influence of individual predictors. This method's simplicity and interpretability make it an excellent tool for understanding the fundamental relationships between predictors and student performance.

Conversely, the RF algorithm provides a robust approach to prediction, enhancing accuracy by capturing intricate patterns within the data. Its ability to handle non-linear relationships and interactions makes it a powerful tool for predicting student outcomes in complex and diverse educational environments.

Implementing this intelligent analytic framework involves several steps. Initially, data on students' academic history and related factors is collected and preprocessed. The dataset is then split into training and testing subsets to evaluate the models' performance. MLR is applied first to establish a baseline model and identify significant predictors. Subsequently, the RF algorithm is employed to refine predictions and capture more complex patterns. The results from both models are then compared to evaluate their predictive accuracy and effectiveness.

By independently utilizing MLR and RF algorithms, educational stakeholders can gain deeper insights into student performance and make informed decisions that significantly improve student outcomes. This article delves into the methodology, implementation, and potential impacts of such an analytic framework, providing a comprehensive guide for leveraging data science in education.

LITERATURE REVIEW

The prediction of student academic performance has garnered significant attention in educational research, driven by the need to enhance learning outcomes and identify at-risk students early. This literature review explores the current state of research on predictive modeling in education, with a focus on the application of MLR and RF algorithms.

Early studies in academic performance prediction primarily employed traditional statistical methods, such as linear regression and decision trees, to analyze factors influencing student success. Logistic regression emerged as a powerful tool due to its simplicity and effectiveness in handling binary classification tasks, such as pass/fail outcomes. Researchers have utilized logistic regression to identify critical predictors, including attendance, prior academic records, socio-economic status, and engagement levels, demonstrating its utility in educational settings.

As data availability and computational power have increased, more sophisticated machine learning techniques have been adopted. The RF algorithm, introduced by Breiman (2001), has become particularly popular for its high accuracy and ability to manage large, complex datasets. Studies employing random forest have reported superior performance in predicting academic outcomes compared to traditional methods, highlighting its robustness against overfitting and its capability to capture nonlinear relationships among variables.

Recent literature has seen a comparative analysis of various predictive models to determine the most effective approaches for specific educational contexts. These studies often emphasize the trade-offs between interpretability and predictive power. While MLR offers clear insights into the influence of individual predictors, RFprovides a more nuanced understanding through its ensemble approach, albeit with less interpretability.

Furthermore, contemporary research has explored the integration of these models with other advanced techniques, such as neural networks and ensemble methods, to enhance prediction accuracy. The incorporation of feature selection methods and the use of balanced datasets are also discussed extensively, addressing common challenges like multicollinearity and class imbalance.

Rodríguez-Hernández et al. (2021) used Artificial neural networks in academic performance prediction. The first objective of this study is to test a systematic procedure for implementing artificial neural networks to predict academic performance in higher education. The second objective is to analyze the importance of several well-known predictors of academic performance in higher education. The sample included 162,030 students of both genders from private and public universities in Colombia. The findings suggest that it is possible to systematically implement artificial neural networks to classify students' academic performance as either high (accuracy of 82%) or low (accuracy of 71%). Artificial neural networks outperform other machine-learning algorithms in evaluation metrics such as the recall and the F1 score. Furthermore, it is found that prior academic achievement, socioeconomic conditions, and high school characteristics are important predictors of students' academic performance in higher education. Finally, this study discusses recommendations for implementing artificial neural networks and several considerations for the analysis of academic performance in higher education.

A model for predicting student performance based on supervised machine learning techniques was created by Hashim et al. in 2020. A number of supervised machine learning algorithms, including Decision Tree, Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Sequential Minimal Optimization, and Neural Network, were compared in this study. In order to predict student performance on final exams, the researchers trained a model using datasets from courses in the bachelor study programs at the College of Computer Science and Information Technology, University of Basra, for the academic years 2017–2018 and 2018–2019. The best classifier for precisely predicting students' final grades, according to the results, is the logistic regression classifier (68.7% for passed and 88.8% for failed).

Hashim et al. (2020) predicted students' academic performance using ensemble approaches. We collected educational data from a learning management system (LMS) in order to illustrate the significance of student behavioral aspects in this article. The included dataset was subjected to feature analysis, followed by data preprocessing—a crucial phase in the knowledge-discovery process. The preprocessed dataset is classified using classifiers including Nave Bayes (NB), Decision Tree (ID3), Support Vector Machines (SVM), and K-Nearest Neighbor (KNN) in order to predict student academic achievement. The suggested model's accuracy is increased through the application of ensemble methods. We employed typical ensemble techniques including bagging, boosting, and voting algorithms. By employing group methods, we were able to improve the outcome and show the dependability of the suggested model.

Nabil et al. (2021) predicted students’ academic performance based on courses’ grades using deep neural networks. The main goal of this paper is to explore the efficiency of deep learning in the field of EDM, especially in predicting students’ academic performance, to identify students at risk of failure. A dataset collected from a public 4-year university was used in this study to develop predictive models to predict students’ academic performance of upcoming courses given their grades in the previous courses of the first academic year using a deep neural network (DNN), decision tree, random forest, gradient boosting, logistic regression, support vector classifier, and K-nearest neighbor. In addition, we made a comparison between various resampling methods to solve the imbalanced dataset problem, such as SMOTE, ADASYN, ROS, and SMOTE-ENN. From the experimental results, it is observed that the proposed DNN model can predict students’ performance in a data structure course and can also identify students at risk of failure at an early stage of a semester with an accuracy of 89%, which is higher than models like decision tree, logistic regression, support vector classifier, and K-nearest neighbor.

Using artificial neural networks, Lau et al. (2019) predicted and categorized the academic performance of their students. Eleven input variables, two levels of hidden neurons, and one output layer make up the neural network model. The backpropagation training rule is implemented using the Levenberg-Marquardt algorithm. The area under the receiver operating characteristics curve, error performance, regression, error histogram, confusion matrix, and error histogram are used to assess the effectiveness of the neural network model. Despite certain drawbacks, the neural network model has an overall strong prediction accuracy of 84.8%.

In order to predict students' academic performance, Albreiki (2021) conducted a mining of student information system records. The primary goal of this research is to determine which characteristics that influence students' performance are most frequently researched and which data mining approaches are most frequently used to find these factors. As a result, a dataset from a nearby university in the United Arab Emirates' student information system was created for this dissertation. The dataset, which had a record count of over 56,000, had 34 attributes relating to student information. According to empirical findings, four categories of student characteristics such as demographics, past performance history, course and teacher information, and some general student information—are in charge of predicting academic success. Furthermore, the findings also showed that artificial neural networks, decision trees, and Naïve Bayes are the most often utilized data mining methods for categorizing and predicting student variables. The best data-mining model for forecasting students' academic success from student information systems was ultimately determined by comparing a set of models.

Tomasevic, et al (2020) aimed of at providing a comprehensive analysis and comparison of state of the art supervised machine learning techniques applied for solving the task of student exam performance prediction, i.e. discovering students at a “high risk” of dropping out from the course, and predicting their future achievements, such as for instance, the final exam scores. For both classification and regression tasks, the overall highest precision was obtained with artificial neural networks by feeding the student engagement data and past performance data, while the usage of demographic data did not show significant influence on the precision of predictions. To exploit the full potential of the student exam performance prediction, it was concluded that adequate data acquisition functionalities and the student interaction with the learning environment is a prerequisite to ensure sufficient amount of data for analysis.

In summary, the literature reflects a dynamic and evolving field, with ongoing efforts to refine predictive models and adapt them to diverse educational environments. This review aims to synthesize these advancements, providing a comprehensive understanding of the current methodologies and identifying gaps for future research. Through this synthesis, we seek to establish a foundation for our comparative study of multiple linear regression and random forest in predicting student academic performance.

Random Forest (RF)

Random Forest is a popular machine learning algorithm and an ensemble learning algorithm used for classification and regression tasks due to its high accuracy, robustness, feature importance, versatility, and scalability (Wainberg et al., 2016). A Random Forest is a tree-based ensemble with each tree depending on a collection of random variables. More formally, for a *p*-dimensional random vector $X = (X_1, . . . ,Xp)^T$representing the real-valued input or predictor variables and a random variable Y representing the real-valued response, we assume an unknown joint distribution $P_{XY}(X,Y)$. The goal is to find a prediction function $f(X)$ for predicting Y. The prediction function is determined by a loss function $L(Y, f(X))$ and defined to minimize the expected value of the loss.

$$E_{XY}(L(Y, f(X))) \qquad \text{Equation 1}$$

where the subscripts denote expectation with respect to the joint distribution of X and Y.
Intuitively, $L(Y, f(X))$ is a measure of how close $f(X)$ is to Y; it penalizes values of $f(X)$ that are a long way from Y. Typical choices of L are *squared error loss* $L(Y, f(X)) = (Y - f(X))^2$ for regression and *zero-one loss* for classification:

$$L\,(Y, f\,(X)) = I\big(Y \neq f(X)\big) = \begin{cases} 0 \; if \; Y = f(X) \\ 1 \;\; otherwise \end{cases} \qquad \text{Equation 2}$$

It turns out that minimizing E_{XY} (L (Y, f (X))) for squared error loss gives the conditional expectation

$$f(x) = \;\; E(Y|X{=}x) \qquad \text{Equation 3}$$

Otherwise known as the *regression function*. In the classification situation, if the set of possible values of *Y* is denoted by Y, minimizing $E_{XY}(L\,(Y, f\,(X)))$ for zero-one loss gives:

$$f\,(x) = argmaxP(Y{=}y|X{=}x) \qquad \text{Equation 4}$$

otherwise known as the *Bayes rule*.

Ensembles construct *f* in terms of a collection of so-called “base learners” $h_1(x)$, . . ., hJ(x) and these base learners are combined to give the “ensemble predictor” *f* (x). In regression, the base learners are averaged

$$f(x) = \frac{1}{J}\sum_{J=1}^{J} h_j\,(x) \qquad \text{Equation 5}$$

$$f(x)\,argmax_{y \in Y} \sum_{J=1}^{J} I\,(y = h_j(x)) \qquad \text{Equation 6}$$

Multiple Linear Regression (MLR)

Multiple linear regression is an extension of simple linear regression that models the relationship between a dependent variable and multiple independent variables. This allows for a more nuanced understanding of how various factors contribute to the outcome.

The equation for multiple linear regression is:

$$Y = \beta 0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \qquad \text{Equation 7}$$

Where Y is the dependent variable, $X_1, X_2, \ldots, X_k$ are the independent variables, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients while ϵ is the error term.

METHODOLOGY

In consultation with stake holders in Federal Polytechnic Ukana, 816 datasets were collected from eight (8) departments of the institution. The data was cleaned and transformed, so that some outliers were identified and resolved, getting rid of 152 data points and leaving 664 data points to be used in this study. The attributes of the data are course 1, course 2, course 3, course 4, course 5, course 6, course 7, course 8, course 9, course10, course 11, course 12, course 13, course 14, course 15, course 16, course 17, course 18, course 19, course 20, course 21, course 22, course 23, course24, course25, total grade point, grade point average (GPA), attendance, extracurricular activities, social life, reading culture, power supply on campus, age, gender, cumulative grade point average. Student's names and registration number were not included and different codes were used for courses offered by students so as to maintain some level of privacy. To transform data to suitable format, Min-Max Scaling (Normalization) method was adopted because it actively eliminates the effect of inconsistent ranges of the datasets and improves convergence (Ahmed et al., 2022).This method scales the features to a specified range, usually [0, 1] using the formula:

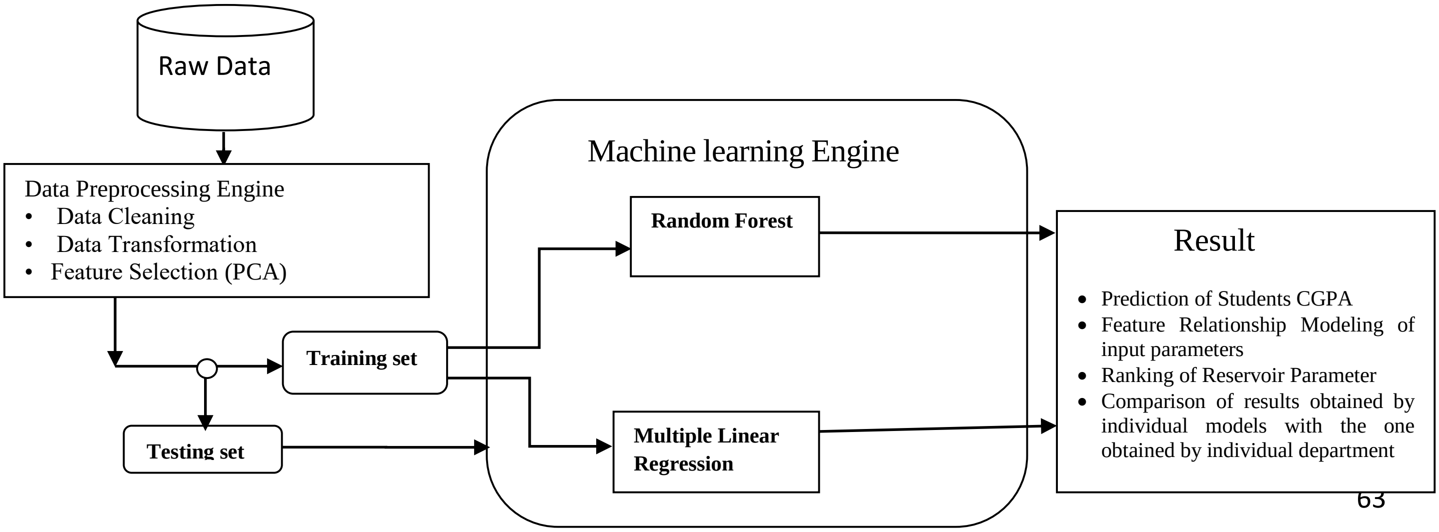
$$X\_normalized \ = \ (X \ - \ X\_min) \ / \ (X\_max \ - \ X\_min) \qquad Equation \ 8$$

Where X is the original feature and X={ $X_1, X_2, \ldots X_n$}, X_min is the minimum value of the feature in the dataset, and X_max is the maximum value of the feature in the dataset.

A total of 27 out of the 38 input characteristics were chosen by principal component analysis (PCA) based on their Eigen values and explained variance percentage. Random Forest (RF) and Multiple Linear Regression are the tools utilized in this work. In the training phase, a bootstrap method is used to train each Regressor individually using its own duplicated training data set. Two sets of data: the training and testing sets are created from the data. Twenty percent (20%) of the data are for testing, and the remaining eighty percent (80%) are for the training set.

Architecture of the System

The architecture of the system is depicted in Figure 1.


Raw Data
Data Preprocessing Engine
• Data Cleaning
• Data Transformation
• Feature Selection (PCA)
Training set
Testing set
Machine learning Engine
Random Forest
Multiple Linear Regression
Result
• Prediction of Students CGPA
• Feature Relationship Modeling of input parameters
• Ranking of Reservoir Parameter
• Comparison of results obtained by individual models with the one obtained by individual department

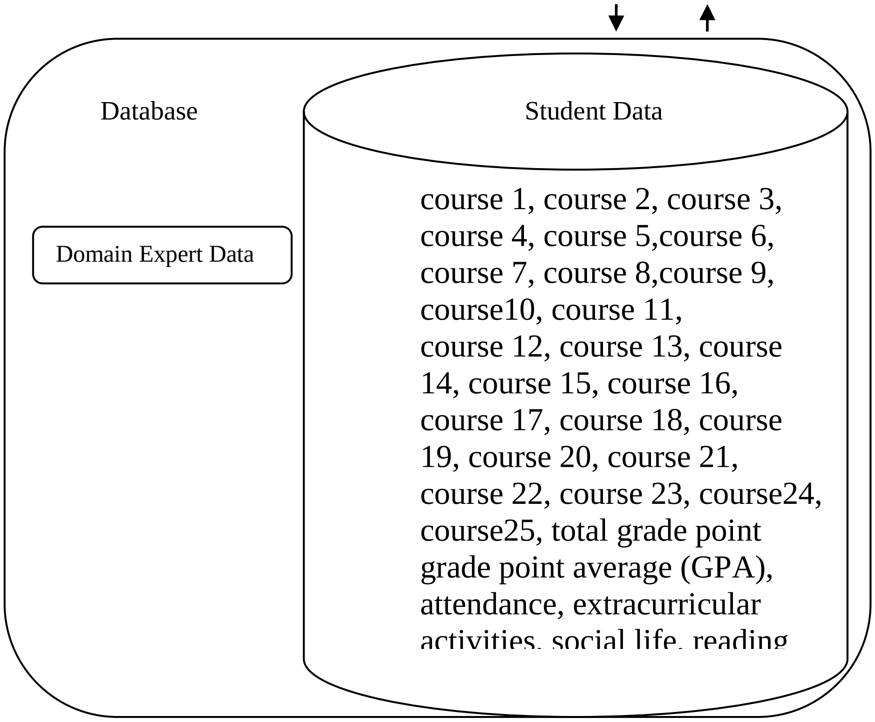

Database
Student Data
Domain Expert Data
course 1, course 2, course 3, course 4, course 5,course 6, course 7, course 8,course 9, course10, course 11, course 12, course 13, course 14, course 15, course 16, course 17, course 18, course 19, course 20, course 21, course 22, course 23, course24, course25, total grade point grade point average (GPA), attendance, extracurricular activities, social life, reading

Figure 1: Architectural design of Student academic performance Prediction using Multiple linear regression and random forest.
Source: The Researcher (2024)

RESULTS

The implementation procedure for the prediction of student academic performance was performed in python programming environment on anaconda software. The datasets collected from Federal Polytechnic Ukana for the purpose of this research was 816. It was stored in Comma-Separated Values (csv) format. Simplicity, readability, wide compatibility, flexibility, standardization and data exploration and visualization were the reason for the choice of csv (Kaur *et al* 2020). The data was cleaned and transformed.

The input features are denoted by x, which includes all columns from index 1 to 27, and the target variable denoted by y is the 28th column. The features that formed the independent variables were course 1, course 2, course 3, course 4, course 5, course 6, course 7, course 8, course 9, course10, course 11, course 12, course 13, course 14, course 15, course 16, course 17, course 18, course 19, course 20, course 21, course 22, course 23, course24, course25, total grade point grade point average (GPA), attendance, extracurricular activities, social life, reading culture, Power supply on campus, age, Gender, while the target variable was the Cumulative Grade Point Average (CGPA) feature.

A principal component Analysis (PCA) was conducted on the features and 27 out of the 36 input features were selected based on their Eigen values and Explained Variance Percentage. The decision of using 27 input features was arrived at using domain expert knowledge and literature source. According to Araújo and Santos (2018), features with eigen values of 0.5 and above are stable ; hence the decision of using 27 features.

The Random Forest and Multiple linear regression algorithms were trained and tested with the transformed data set using the ratio of 80:20 respectively.
The random forest model gave the results captured on table 1 while results of Multiple linear regression is captured on Table 2. The percentage errors of Models for prediction are shown in Table 3. The Scatter Plots of Field Results against Computed Results for Random Forest and Multiple Linear Regression is depicted in Figure 2 and 3 respectively.

Table 1: Performance metrics of Random Forest

Performance metrics of Random Forest
Mean_Squared_Error: 0.006
Mean_Absolute_Error: 0.043
R_Squared_Score: 0.905
Explained Variance Score: 0.911
Median Absolute Error: 0.019

Table 2: Performance metrics of Multiple Linear Regression

Performance metrics of Stacking Model
Mean_Squared_Error: 0.043
Mean_Absolute_Error: 0.160
R_squared_Score: 0.400
Explained Variance Score: 0.443
Median Absolute Error: 0.124

Table 3: Percentage Errors of Prediction by Random Forest and Multiple Regression

	RF	MLR
PE	**0.597**	**4.300**

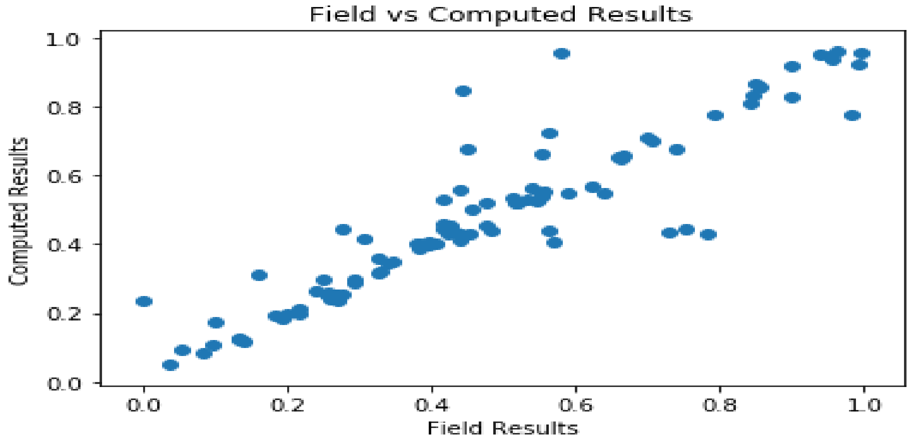


Figure 2: Scatter Plot of Field Results against the Computed Results in RF.
Source: The Researcher (2024)

In Figure 2, the relationship between variables is high, positive and linear. There are five (5) outliers with the farthest point apart being 0.3 units. The points form a tight cluster around the diagonal line (indicating a strong positive correlation between field and computed results). The model shows a relatively tight and evenly distributed cluster around the diagonal line.

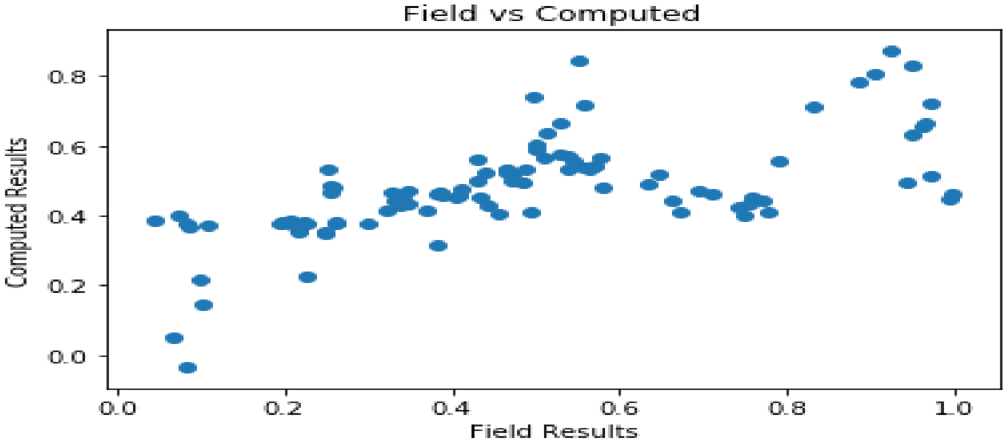


Figure 3: Scatter Plot of Field Results against Computed Results in Multiple Linear Regression Model
Source: The Researcher (2024)

Figure 3: shows that the relationship between variables is low, positive and linear. There are five (5) outliers with the furthest points apart being 1.5 units. The points form a cluster around the diagonal line. The model shows an evenly distributed cluster around the diagonal line.

The grouped bar chart of the MSE, MAE,R^2,EVS and MedAE of RF, Multiple linear Regression models is depicted is Figure 4. The plot shows the visualization of the performance of the two (2) models. The information from the bar chart allows for a comparison of the performance of each model using a particular performance metric. The heat map as a visualization tool is depicted in Figure 5. It tabulates the performance metrics of the two models, creating a room for easy comparison of the performance of the models.

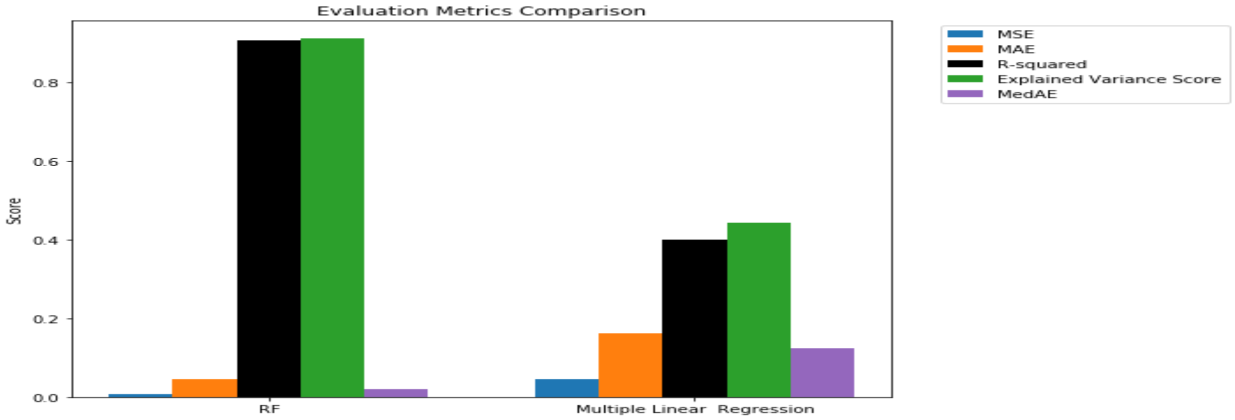


Figure 4: The Bar Chart of the MSE, MAE, R^2, EVS and MedAE of RF,MLR models. Source: The Researcher (2024)

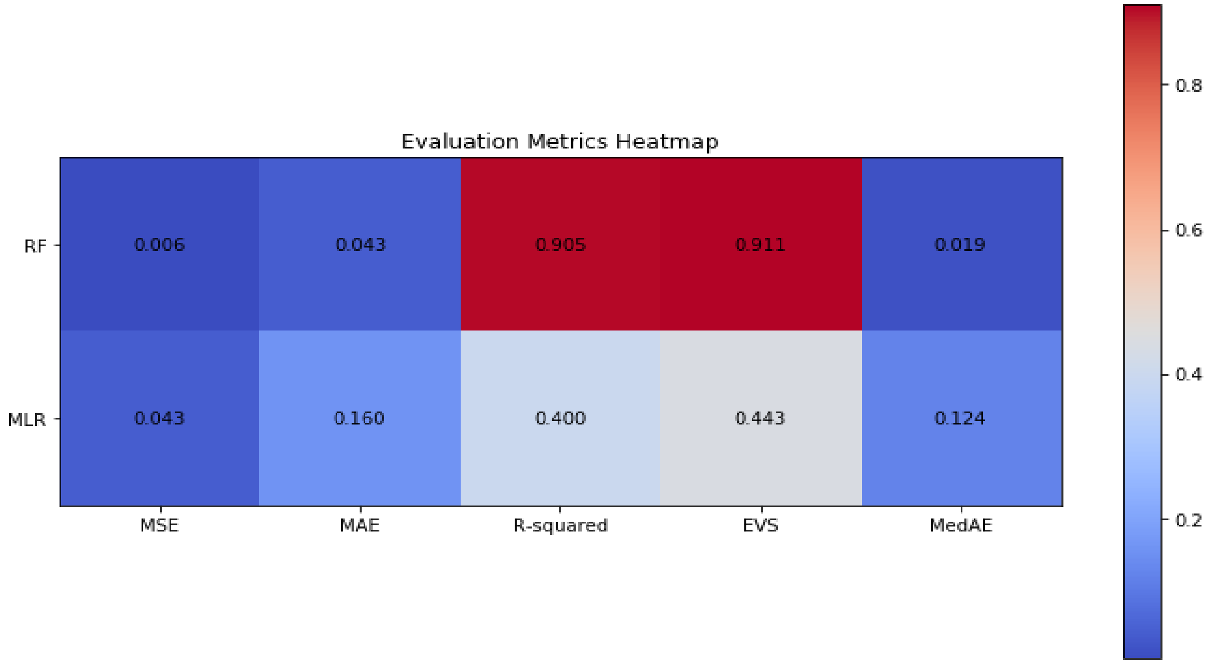


Figure 5: Heat Map Visualization for the two Models
Source: The Researcher's (2024)

CONCLUSION

This study set out to evaluate the predictive power of two distinct predictive models: Multiple Linear Regression (MLR) and Random Forest (RF) in predicting student academic performance at Federal Polytechnic Ukana. Through comprehensive data collection, preprocessing, and model implementation, we aimed to compare these models' effectiveness in a structured, intelligent analytic framework.

The results demonstrate a clear distinction in performance between the two models. Random Forest significantly outperformed Multiple Linear Regression across several key metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared Score (R^2), Explained Variance Score (EVS), and Median Absolute Error (MedAE). The RF model's ability to handle large datasets with high dimensionality and capture non-linear relationships proved advantageous, yielding a more robust and accurate prediction of students' cumulative grade point averages (CGPA).

In contrast, while MLR provided a straightforward and interpretable model, it fell short in predictive accuracy. The linear nature of MLR limited its ability to capture the complex interactions and non-linear patterns within the educational data, leading to higher error rates and lower explanatory power.

The scatter plots and performance metrics visualized in bar charts and heat maps further elucidate these differences, underscoring the superior performance of Random Forest. The percentage errors of prediction by RF and MLR, depicted as 0.597% and 4.300% respectively, highlight the substantial improvement achieved through the ensemble learning approach of Random Forest.

This study's findings hold significant implications for educational institutions. By leveraging advanced machine learning techniques like Random Forest, educators and administrators can gain deeper insights into student performance, identify at-risk students more accurately, and implement timely, targeted interventions. The high predictive accuracy of the RF model can support data-driven decision-making, ultimately enhancing educational outcomes and student success.

In conclusion, this comparative analysis underscores the potential of intelligent analytic frameworks in education. The clear superiority of Random Forest over Multiple Linear Regression in this context paves the way for more sophisticated, data-driven approaches to understanding and improving student academic performance.

Future studies

Future research could explore the integration of these models with other advanced techniques, such as neural networks or hybrid models, to further enhance prediction accuracy. Additionally, expanding the dataset to include a more diverse range of institutions and student populations could validate and extend the applicability of these findings.

Reference

Ahmed, H. A., Ali, P. J. M., Faeq, A. K. and Abdullah, S. M. (2022). An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method. *Aro-The Scientific Journal of Koya University*, 10(2): 29-37.

Araújo, J. M. and Santos, T. L. M. (2018). Control of a class of second-order linear vibrating Systemswith time-delay: *Smith predictor approach. Mechanical Systems and Signal Processing*, 108: 173-187.

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. Education Sciences, 11(9), 552.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020). Student performance prediction model based on supervised machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 928, No. 3, p. 032019). IOP Publishing.

Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. SN Applied Sciences, 1(9), 982.

Kaur, A., Ayyagari, S., Mishra, M. and Thukral, R. (2020). A Literature Review on Device-to-Device Data Exchange Formats for IoT Applications.*Journal of Intelligent Systems and Computing*, 1(1): 1-10.

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. IEEE Access, 9, 140731-140746.

Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. Computers and Education: Artificial Intelligence, 2, 100018.

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & education, 143, 103676.

Wainberg, M., Alipanahi, B., and Frey, B. J. (2016). Are random forests truly the best classifiers?. *The Journal of Machine Learning Research*, *17*(1), 3837-3841.