TMJM
THE MINDANAWAN
Journal of Mathematics

The Mindanawan Journal of Mathematics
Official Journal of the Department of Mathematics and Statistics
Mindanao State University-Iligan Institute of Technology
ISSN: 2094-7380 (Print) | 2783-0136 (Online)

Vol. 6 (2024), no. 1, pp. 67–78

A Comparative Study of Machine Learning Algorithms for Regression in Predicting the Academic Performance of Students in General Mathematics

Mary Christine G. Ontolan[1], Redeemtor R. Sacayan[2,*], and Bernadette F. Tubo[3]

[1]Notre Dame of Midsayap College, Cotabato City
mg.ontolan@ndmc.edu.ph

[2,3]Department of Mathematics and Statistics
MSU-Iligan Institute of Technology, 9200 Iligan City, Philippines
redeemtor.sacayan@g.msuiit.edu.ph, bernadette.tubo@g.msuiit.edu.ph

Received: 12th May 2024 Revised: 28th August 2024

Abstract

This study explores the application of predictive modeling techniques in assessing the academic performance of Senior High School students in General Mathematics at Notre Dame of Midsayap College, Cotabato City. Employing three distinct machine learning algorithms, namely, multiple linear regression (MLR), random forest regression (RFR), and support vector regression (SVR), the study aims to predict students' General Mathematics grades with some explanatory features like family background, junior and senior high school characteristics. Evaluation of these algorithms' predictive capabilities is conducted utilizing metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and adjusted R^2. Results indicate that multiple linear regression model exhibits superior predictive performance, yielding lower RMSE and MAE values compared to RFR and SVR models, achieving an accuracy prediction of 97.29%.

1 Introduction

In the contemporary educational landscape, there is a discernible surge in scholarly interest directed towards investigating students' academic performance. Educators are actively involved in assessing and tracking students' academic achievements, aiming to acquire deeper insights into their progress and accomplishments. A prevalent methodology involves predicting academic performance through the analysis of diverse explanatory variables. According to Pandey and Taruna (2016) [7], accurate predictions play a vital role in identifying students encountering academic challenges, thereby enabling timely interventions to avert academic setbacks. Moreover, these predictions contribute to refining curricula, enhancing teaching strategies, and implementing targeted educational interventions, thereby ultimately improving the overall educational system (Qasrawi et al., 2021) [9].

Researches conducted by [7, 9], Huang (2011) [4], Ibrahim and Rusli (2007) [5] and Kabakchieva (2013) [6] have extensively examined the multitude of factors impacting aca-

*Corresponding author
2020 Mathematics Subject Classification: MSC62, MSC97
Keywords and Phrases: academic performance, machine learning, multiple linear regression, random forest regression, support vector regression

demic performance. These factors encompass a wide range of variables, including, but not limited to the following: gender, age, residency, prior knowledge, parental employment, financial status, teacher characteristics, absences, and midterm scores. Employing a variety of regression-based machine learning algorithms for predictive modeling, such as linear regression, logistic regression, decision trees, random forest, artificial neural networks, k-nearest neighbor, and support vector machine, these studies have contributed significantly to understanding the complex interplay of factors influencing academic achievement.

One of the most frequently utilized predictive models in statistical analysis is multiple linear regression (MLR). This method is utilized to predict the values of a dependent variable, Y_i, based on a set of k explanatory variables $(X_1, X_2, \ldots, X_k)$. As a supervised machine learning technique, MLR facilitates the assessment of the model's variance and the relative contribution of each independent variable to the overall variance. In multiple linear regression, where k features are considered, the relationship between the dependent variable and the features is expressed as follows:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon_i, \tag{1}$$

where Y_i denotes the dependent variable, β_0 represents the intercept, $\beta_1, \ldots, \beta_k$ denote the model parameters, $X_1, \ldots, X_k$ are the independent variables and ϵ_i is the error term, for $i = 1, 2,...,n$.

Moreover, being a parametric model, MLR requires adherence to five key assumptions. These assumptions encompass: (1) a linear relationship between the independent and dependent variables, (2) absence of multicollinearity in the data, (3) normal distribution of errors or residuals between observed and predicted values, (4) absence of autocorrelation in residuals, and (5) homoscedasticity or constant variances of residuals.

Another widely employed and efficient algorithm for classification and regression tasks based on model aggregation principles is the random forest (RF) introduced by Breiman (2001) [2]. It entails constructing multiple decision trees and combining their outputs to enhance model generalization. Renowned for their versatility, random forests excel in effectively modeling intricate nonlinear relationships while demonstrating resilience against overfitting and robustness in the presence of data noise. Furthermore, they offer unbiased error rate estimation and facilitate the determination of variable importance ([2]; Chagas et al., 2016 [3]; Zhang et al., 2019 [12]). Despite their widespread adoption, recent theoretical and methodological advancements in random forests have prompted ongoing exploration, as highlighted by Biau and Scornet (2015) [1].

Support Vector Regression (SVR), conversely, represents one of the powerful machine learning algorithms employed for regression tasks. It serves as an extension of the Support Vector Machine (SVM) algorithm, primarily utilized for classification tasks. Fundamentally, SVR aims to identify the optimal hyperplane that best fits the training data while maximizing the margin between the data points and the hyperplane. This is accomplished by mapping the input variables to a high-dimensional feature space using a kernel function and identifying the hyperplane that maximizes the margin between the hyperplane and the closest data points, while simultaneously minimizing the prediction error (Sethi, 2023 [11]). Commonly utilized kernel functions in SVR include linear, polynomial, radial basis function (RBF), and sigmoid kernels. One of the primary advantages of SVR lies in its capacity to model nonlinear relationships between the input features and the target variable, rendering it particularly useful for tasks characterized by nonlinearity between the input and output variables. Additionally, SVR demonstrates resilience to outliers in the training data and can accommodate datasets with a large number of features. It also offers flexibility in model complexity through the regularization parameter, aiding in the prevention of overfitting. However, SVR may necessitate meticulous tuning of hyper-parameters, such as the choice of kernel function and regularization

parameter, to achieve optimal performance. Moreover, it may encounter computational complexity, particularly with large datasets, as training time escalates with dataset size.

In this study, these three regression-based algorithms will be employed to predict academic performance among senior high school students at Notre Dame of Midsayap College in General Mathematics of the K to 12 Basic Education Curriculum of the Philippines. Their predictive ability will be compared in terms of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), and their adequacy will be evaluated in terms of adjusted R^2.

2 Methodology

Figure 2.1 presents the systematic methodology employed for predicting academic performance, encompassing several stages from data collection, data pre-processing, model building to model evaluation.

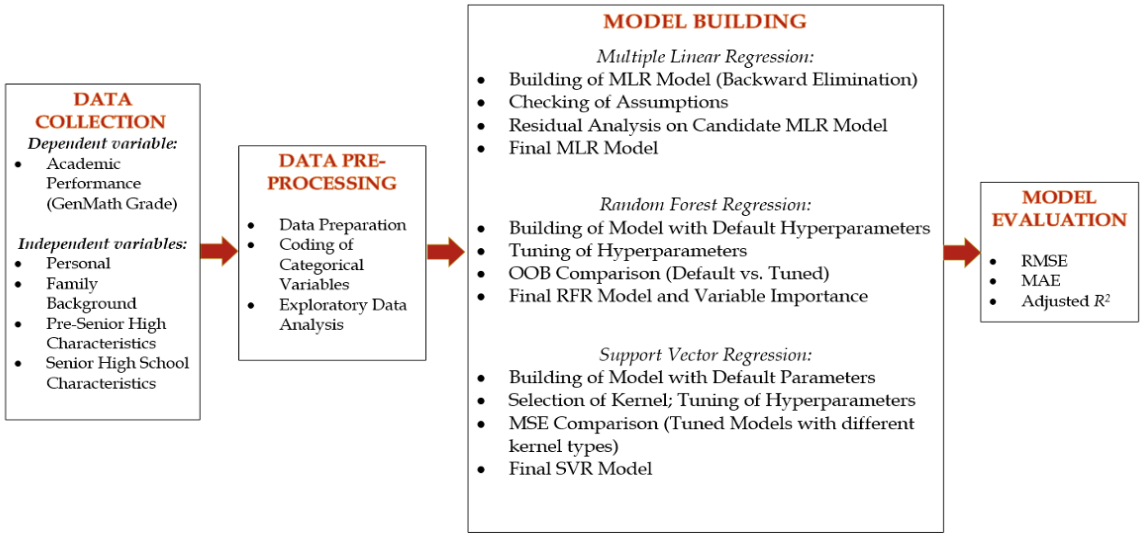

Figure 2.1. Systematic Methodology for Predicting Academic Performance

2.1 Data Collection and Pre-Processing

The research data were collected from Notre Dame of Midsayap College- Senior High School Department during the Academic Year 2022-2023. The information were collected from the adviser's record, General Mathematics (GenMath) teacher's records, enrollment and school forms, and guidance office, following protocols ensuring consent and strict confidentiality. A total of $n = 575$ observations were considered with primary focus on the actual final grade of the students in GenMath as the dependent variable and 15 independent variables categorized into four (4) general characteristics (personal, family background, pre-senior high school, and senior high school) as shown in Table 2. For the subsequent model development stage, the observations were divided into two sets: a training set and a test set. The training set, consisting of two-thirds (2/3) of the total observations consisting 383 observations was utilized to construct all three models. Conversely, the remaining one-third (1/3) of the observations consisting 192 observations served as the test set, employed to assess the performance of the constructed models in terms of RMSE, MAE, and adjusted R^2.

Subsequently, data pre-processing is conducted to ensure the quality and integrity of the dataset. During the pre-processing stage, various tasks such as data preparation, encoding of categorical variables, and exploratory analysis were executed for model building.

Table 2.1. Independent Variables of the Study

General Features	Independent Variable and Description	Type of Variable	Line Code and Domain
Personal	Gender (*gen*)	Categorical	0−Male
			1−Female
	Age (*age*)	Continuous	15 − 21 years
Family Background	Father has source of income (*fsi*)	Categorical	0−None
			1−Yes
	Mother has source of income (*msi*)	Categorical	0−None
			1−Yes
Pre-Senior High Characteristics	Type of Previous School (*st*)	Categorical	0−Public
			1−Private
	Grade 10 grade in Math (*Gr10M*)	Continuous	75 − 100
	Grade 10 GPA (*Gr10G*)	Continuous	75 − 100
	Academic Awardee (*award*)	Categorical	0−No
			1−Yes
Senior High School Characteristics	Entrance Exam Percentage (*EEP*)	Continuous	0 − 100
	Strand(*str*)	Categorical	0−TVL
			1−HUMSS
			2−ABM
			3−STEM
	Teacher in General Mathematics (*teacher*)	Categorical	0−Teacher0
			1−Teacher1
			1−Teacher2
	1^{st} Quarter Exam Score (*Q1S*)	Continuous	0 − 100%
	2^{nd} Quarter Exam Score (*Q2S*)	Continuous	0 − 100%
	1^{st} Quarter Grade in General Mathematics (*Q1G*)	Continuous	75 − 100
	Number of Absences (*Ab*)	Discrete	0 − 7

2.2 Model Building

Following data pre-processing, the model building phase involves employing the machine learning algorithms for regression to predict student's academic performance.

2.2.1 Multiple Linear Regression (MLR)

Let Y_i represent the grade in GenMath of the i^{th} student, where $i = 1, 2, 3, \ldots, n$, with $n = 383$. By equation (1), the linear regression model is given by:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 \mathbf{1}_{gen}(x) + \beta_{10} X_{10} \mathbf{1}_{msi}(x) + \beta_{11} X_{11} \mathbf{1}_{fsi}(x) + \beta_{12} X_{12} \mathbf{1}_{award}(x) + \beta_{13} X_{13} \mathbf{1}_{st}(x) + \beta_{14} X_{14} \mathbf{1}_{teacher}(x) + \beta_{15} X_{15} \mathbf{1}_{str}(x) + \epsilon_i, \quad i = 1, 2, 3, \ldots, n$$

where Y_i = Grade in GenMath; X_1 = Age; $X_2 = Gr10G$ (Grade 10 GPA); $X_3 = Gr10M$ (Grade 10 Math grade); $X_4 = EEP$ (entrance exam percentage); $X_5 = Q1S$ (1^{st} quarter exam score); $X_6 = Q2S$ (2^{nd} quarter exam score); $X_7 = Q1G$ (1^{st} quarter grade in GenMath); $X_8 = Ab$ (number of absences);

$$X_9\mathbf{1}_{gen}(x) = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female;} \end{cases}$$

$$X_{10}\mathbf{1}_{msi}(x) = \begin{cases} 1 & \text{if the mother has a source of income} \\ 0 & \text{otherwise;} \end{cases}$$

$$X_{11}\mathbf{1}_{fsi}(x) = \begin{cases} 1 & \text{if the father has a source of income} \\ 0 & \text{otherwise;} \end{cases}$$

$$X_{12}\mathbf{1}_{award}(x) = \begin{cases} 1 & \text{if the student is an academic awardee} \\ 0 & \text{otherwise;} \end{cases}$$

$$X_{13}\mathbf{1}_{st}(x) = \begin{cases} 1 & \text{if the student is from a private school} \\ 0 & \text{if the student is from a public school;} \end{cases}$$

$$X_{14}\mathbf{1}_{teacher}(x) = \begin{cases} 0 & \text{teacher 0} \\ 1 & \text{teacher 1} \\ 2 & \text{teacher 2;} \end{cases}$$

$$X_{15}\mathbf{1}_{str}(x) = \begin{cases} 0 & \text{TVL strand} \\ 1 & \text{HUMSS strand} \\ 2 & \text{ABM strand} \\ 3 & \text{STEM strand;} \end{cases};$$

and the error term, ϵ_i, being independent and identically distributed random variables, $\epsilon_i \sim N(0, \sigma^2)$. The linear regression model was estimated using the ordinary least squares (OLS) procedure processed in R software, version 4.3.0 (R CoreTeam, 2023). In all analyses, we apply a 5% significance level to show significance of tested associations.

2.2.2 Random Forest Regression (RFR)

For constructing the Random Forest Regression (RFR) model, the `Ranger` package within R was utilized. This package includes the use of ranger() function that automatically produces an RFR model with hyperparameters including mtry, num.trees, and min.node.size, which further can be tuned. Employing the expand.grid() function enabled the selection of optimal values for each hyperparameter, enhancing the model's performance. Additionally, a repeated k-fold cross- validation resampling technique with 10 folds and repeated 5 times was implemented, ensuring robustness and reliability of the model. The splitrule utilized was based on variance which ensures that the splits made at each node result in child nodes with lower variance in the target variable. Subsequently, the tuned RFR model derived with the lowest out-of-bag prediction error (OOB RMSE) was the final RFR model and was employed for prediction tasks, leveraging its enhanced performance to make accurate predictions.

2.2.3 Support Vector Regression (SVR)

The Support Vector Regression (SVR) model was developed using the `e1017` package in R. This package streamlined the process of selecting appropriate kernels and fine-tuning hyper - parameters. To search for the optimal values of the hyperparameter, the range argument in the tune() function was utilized to input a range of values for each corresponding hyperparameter. Furthermore, the performance error (MSE) of the tuned SVR models with different kernels were

compared and the tuned model with the lowest error (MSE) was the final SVR employed for prediction tasks.

2.3 Model Evaluation

Once the predictive models are constructed, the subsequent step involves a comprehensive evaluation to gauge their efficacy. This crucial phase is integral to assessing the accuracy and reliability of the predictive models. The performance of the predictions is quantified using the Root Mean Squared Error (RMSE), defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

and the Mean Absolute Error (MAE), given by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$

where y_i represents the observed value, $\hat{y}_i$ denotes the predicted value of the dependent variable for the i-th observation obtained from the predictive model, with $i = 1, 2, \ldots, n$ and n is the total number of observations in the test set.

Additionally, the adequacy of the models is evaluated using the adjusted R^2, which quantifies the proportion of variability in the dependent variable that is explained by the predictors, while accounting for the number of predictors in the model. The adjusted R^2 is calculated as:

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - p - 1}\right),$$

where R^2 is the coefficient of determination, n is the total number of observations and p is the number of predictors in the model.

3 Results and Discussion

3.1 Multiple Linear Regression

Table 3.1 presents a summary of the regression estimates for the model parameters in predicting students' General Mathematics grade (*response variable*) based on the fifteen (15) independent variables listed in Table 2.1.

Table 3.1 Summary of Regression Coefficients of the General MLR Model

coefficient	Estimated	Std. Error	t-values	$Pr(>\|t\|)$
Intercept	36.443264	1.5704004	22.305	<0.0001
$X_1 = Age$	0.0142147	0.0556617	0.255	0.7986
$X_2 = Gr10G$	0.0064260	0.0189722	0.339	0.7350
$X_3 = Gr10M$	0.0073532	0.0144480	0.509	0.6111
$X_4 = EEP$	0.0023008	0.0025288	0.910	0.3635
$X_5 = Q1S$	−0.0005642	0.0032654	−0.173	0.8629
$X_6 = Q2S$	0.0784145	0.0027619	28.391	<0.0001
$X_7 = Q1G$	0.5506897	0.0107730	51.118	<0.0001
$X_8 = Ab$	−0.6870333	0.0442047	−15.542	<0.0001
$X_9 = gen1$	−0.0417961	0.0814128	−0.513	0.6080
$X_{10} = msi1$	0.2034823	0.0779261	2.611	<0.0010
$X_{11} = fsi1$	0.2193287	0.1179605	1.859	0.0638
$X_{12} = award1$	0.0962651	0.1220015	0.789	0.4306
$X_{13} = st1$	−0.1168197	0.0777821	−1.502	0.1340
$X_{14}\mathbf{1}_{teacher}(x = 1)$	−0.8267482	0.1234413	−6.698	<0.0001
$X_{14}\mathbf{1}_{teacher}(x = 2)$	−1.1279946	0.2110915	−5.344	<0.0001
$X_{15}\mathbf{1}_{str}(x = 1)$	−0.0400467	0.2123870	−0.189	0.8505
$X_{15}\mathbf{1}_{str}(x = 2)$	−0.1118369	0.1969729	−0.568	0.5705
$X_{15}\mathbf{1}_{str}(x = 3)$	−0.1370003	0.1546933	−0.886	0.3764

Findings reveal a noteworthy association between six variables, specifically $X_6 = Q2S$, $X_7 = Q1G$, $X_8 = Ab$, $X_{10}\mathbf{1}_{msi}(x = 1)$, $X_{14}\mathbf{1}_{teacher}(x = 1)$, and $X_{14}\mathbf{1}_{teacher}(x = 2)$, and Y_i (grade in GenMath), as evidenced by the p-values less than 0.05.

Moreover, the study intends to employ backward elimination to systematically refine the selection process, ensuring that only significant independent variables are retained for inclusion in the final model.

Table 3.2 Summary of Regression Coefficients of the Final MLR Model

coefficient	Estimated	Std. Error	t-values	$Pr(>\|t\|)$
Intercept	36.443264	0.692283	52.542	<0.0001
$X_6 = Q2S$	0.078959	0.002428	32.516	<0.0001
$X_7 = Q1G$	0.550649	0.008813	62.478	<0.0001
$X_8 = Ab$	−0.684771	0.042783	−16.005	<0.0001
$X_{10} = msi1$	0.187267	0.074565	2.5110	0.0124
$X_{12} = award1$	0.187361	0.082109	2.2820	0.0231
$X_{14}\mathbf{1}_{teacher}(x = 1)$	−0.815910	0.091358	−8.9310	<0.0001
$X_{14}\mathbf{1}_{teacher}(x = 2)$	−1.109468	0.120284	−9.2240	<0.0001

Table 3.2 summarizes the derived regression coefficients for the MLR final model. It can be deduced from the results that for students whose mother have source of income and was an academic awardee during Grade 10, their GenMath grade will increase by 0.187267 and 0.187361, respectively. Also, for students whose GenMath teacher is either teacher1 or teacher2, their GenMath grade will decrease by 0.815910 or 1.109468. Furthermore, for students with higher 2nd quarter examination scores and 1^{st} quarter grades, and fewer to no absences, their GenMath grade will increase by 0.078959, 0.550649, and 0.684771, respectively. It can also be noted that the intercept of the refitted model, 36.443264, was also significant.

Figure 3.1 illustrates the results of the residual analysis conducted to verify whether the model adheres to the underlying assumptions of the regression analysis. The findings indicate

that the errors are approximately normally distributed, albeit with the presence of some outliers. Furthermore, the analysis confirms that the errors exhibit independence and maintain constant variances across observations.

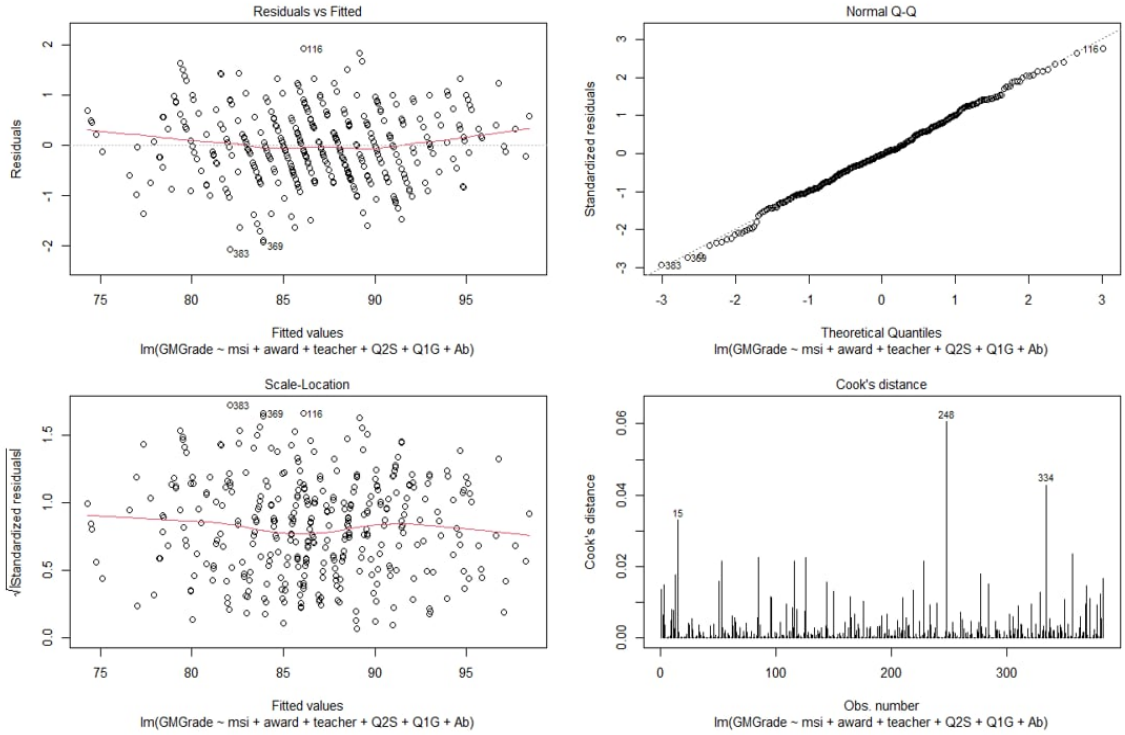

Figure 3.1. Residual Analysis

Therefore, the derived final regression model is given by

$$\hat{Y} = 36.443264 + 0.078959X_6 + 0.550649X_7 - 0.684771X_8 + 0.187267X_{10}\mathbf{1}_{msi}(x=1) + 0.187361X_{12}\mathbf{1}_{award}(x=1) - 0.815910X_{14}\mathbf{1}_{teacher}(x=1) - 1.109468X_{14}\mathbf{1}_{teacher}(x=2),$$

where $X_6 = Q2S$ (2^{nd} quarter exam score), $X_7 = Q1G$ (1^{st} quarter grade in GenMath), $X_8 = Ab$ (number of absences).

3.2 Random Forest Regression

Table 3.3 shows the optimal parameter values in the RFR model with their corresponding out-of-bag (OOB) RMSE to evaluate the accuracy of the model. It can be noted that from the result, `mtry` = 9, `num.trees` = 340, and `node_size` = 4 has the lowest OOB RMSE. Consequently, this resulted in a tuned RFR model shown in Table 3.4. Finally, the optimized RFR model was then utilized for prediction.

Moreover, the important variables in the model building are shown in Figure 3.2. Based on this result, out of the 15 identified features, the most important features in predicting the academic performance of the students in General Mathematics using the RFR tuned model are *Q1G, Q1S, Q2S, Ab, Gr10G, Gr10M, EEP, str*, and *teacher*. These 9 important variables reflected the `mtry`=9 parameter of the final RFR model, while the least important features are *age, st, gen, msi, fsi* and *award*.

Table 3.3 Search for Optimal Hyperparameters in RFR Model

	mtry < *int* >	**nodesize** < *int* >	**num.tress** < *dbl* >	**OOB.RMSE**< *dbl* >
1	9	4	340	0.9252133
2	9	3	340	0.9257400
3	11	4	340	0.9292397
4	9	4	170	0.9302685
5	10	4	170	0.9308921
6	9	5	340	0.9313859
7	11	5	340	0.9319652
8	9	3	170	0.9319670
9	11	6	340	0.9330323
10	12	5	340	0.9332151

Table 3.4 Ranger Result for the Tuned RFR Model

```
Ranger result
Call:  ranger(formula=GMGrade ~.  , data=train, num.tress=340, mtry=9,
min.node.size=4, importance = "impurity")

Type:                                Regression
Number of Trees:                     340
Sample size:                         383
Number of independent variables:     15
Mtry:                                9
Target node size:                    4
Variable importance mode:            impurity
Splitrule:                           variance
OOB prediction error (MSE):          0.8912145
R squared (OOB):                     0.9607587
```

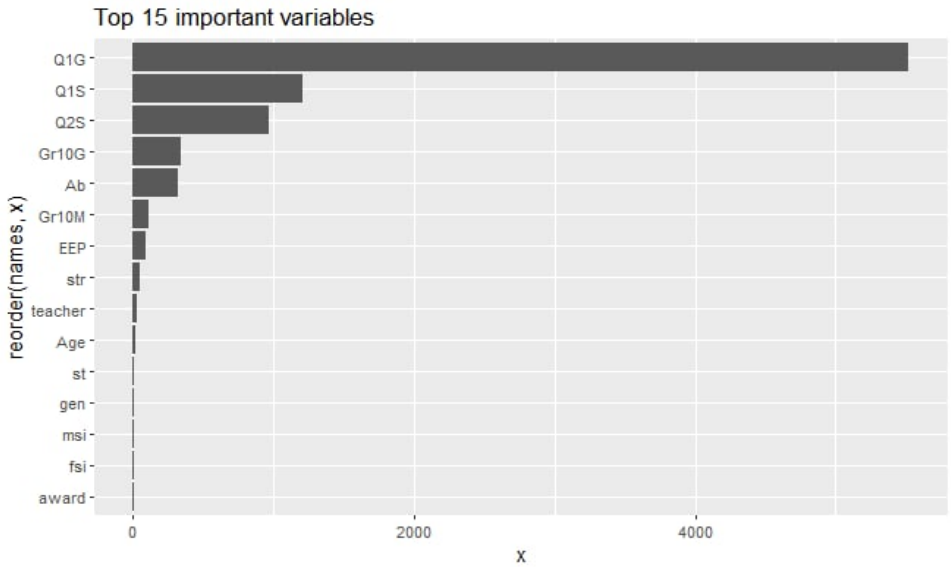

Figure 3.2. Variable Importance in RFR Model

3.3 Support Vector Regression

As depicted in Table 3.5, the SVR Model with a Linear kernel, utilizing the optimal hyperparameter cost (C) = 0.5, yielded MSE of 0.53399238. Conversely, the SVR model employing a non-linear Radial kernel, with cost (C)=1.0 and sigma = 0.01, 0.15, 0.2, and 1, produced MSE

of 0.9298082. Considering that the SVR model with a Linear kernel exhibited the lowest MSE value, it is therefore selected as the final SVR model.

Table 3.5. Result of the Search for Optimal Kernel and Hyperparameters

Tuned SVR Model	Hyperparameters	MSE
With a linear kernel	$C = 0.1$	0.5530744
	$C = 0.5$	0.5399238
	$C = 1.0$	0.5402018
	$C = 10$	0.54007479
With a radial kernel	$C = 0.1$, Sigma = 0.01, 0.15, 0.2, 1	3.0380280
	$C = 0.5$, Sigma = 0.01, 0.15, 0.2, 1	1.1311485
	$C = 1.0$, Sigma = 0.01, 0.15, 0.2, 1	0.9298082
	$C = 10$, Sigma = 0.01, 0.15, 0.2, 1	0.9620150

3.4 Model Comparison

Presented in Table 3.6 is the comparison of the performance of the three (3) regression-based models that are used in this study: multiple linear regression (MLR), random forest regression (RFR), and support vector regression (SVR).

The Multiple Linear Regression (MLR) model, identified six (6) significant features, namely: the mother's source of income (*msi*), student being an academic awardee (*award*), teacher in General Mathematics (*teacher*), 2^{nd} quarter exam score (*Q2S*), 1^{st} quarter grade in general mathematics (*Q1G*), and the student's number of absences (*Ab*) has demonstrated superior predictively capability. This model achieved lower RMSE and MAE values, as well as higher adjusted R^2 outperforming the other two regression-based models.

Following MLR, the SVR model with Linear kernel and a cost parameter of $C = 0.5$ exhibits the second-highest level of predictive performance. The RFR model demonstrates comparatively lower performance, a characteristic attributed to its inherent strength in capturing complex, nonlinear relationships within multi-dimensional datasets.

Table 3.6. Model Comparison using MLR, RFR and SVR

Model	Features/ Parameters	Measures	Performance
MLR	$X_6 = Q2S, X_7 = Q1G, X_8 = Ab$ $X_{10}\mathbf{1}_{msi}(x = 1)$, $X_{12}\mathbf{1}_{award}(x = 1)$, $X_{14}\mathbf{1}_{teacher}(x = 1)$, $X_{14}\mathbf{1}_{teacher}(x = 2)$	RMSE	0.6677996
		MAE	0.5303493
		adjusted R^2	97.2944%
RFR	**mtry**=9, **num.trees**=340 and **nodesize**=4	RMSE	0.9701349
		MAE	0.7118426
		adjusted R^2	94.2901%
SVR	Linear Kernel with cost $C = 0.5$	RMSE	0.6716097
		MAE	0.5318759
		adjusted R^2	97.26348%

4 Conclusions and Recommendations

In this empirical study, three regression-based machine learning algorithms are employed to predict the academic performance of the grade 11 students of Notre Dame of Midsayap College in the General Mathematics subject. The methodologies employed are multiple linear regression (MLR), random forest regression (RFR) and support vector regression (SVR). The performance of the three (3) algorithms were evaluated using the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and adjusted R^2 metrics.

Findings reveal that the multiple linear regression model demonstrates superior predictive performance, yielding a lower RMSE and MAE values compared to RFR and SVR models with higher accuracy prediction of 97.29%. From the comprehensive analysis, factors such as student's maternal source of income, student's academic achievements, teacher classification in General Mathematics, second quarter exam score and first quarter grade in General Mathematics and number of absences emerged as significant features in predicting students' academic performance in General Mathematics at Notre Dame of Midsayap College, Cotabato City.

Based on the results of this study, it is recommended to explore other machine learning algorithms for regression like Decision Trees, Gradient Boosting Machines (GBM), K-Nearest Neighbors (KNN). Additionally, delving into the assessment of several feature importance measures could enhance the predictive capabilities for students' academic performance. Such endeavors hold the potential to deepen understanding and refine the predictive models employed in educational contexts.

Acknowledgements

This research work is supported by Notre Dame of Midsayap College.

References

[1] Biau, G. and Scornet, E. (2015). *A Random Forest Guided Tour. arXiv.org.*

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

[3] Chagas, C. da, de Carvalho Junior, W., Bhering, S. B., and Calderano Filho, B. (2016), *Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions.* CATENA, 139, 232–240.

[4] Huang, Shaobo (2023). *Predictive modeling and analysis of student academic performance in an engineering dynamics course".* All Graduate Theses and Dissertations, Spring 1920 to Summer 2023.

[5] Ibrahim, Zaidah and Rusli, Daliela (2007). *Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression.* 21st Annual SAS Malaysia Forum.

[6] Kabakchieva, D. (2013). *Predicting student performance by using data mining methods for classification.* Cybernetics and Information Technologies, 13(1), 61–72.

[7] Pandey, M. and Taruna, S. (2016). *Towards the integration of multiple classifier pertaining to the Student's performance prediction.* Perspectives in Science, 8, 364-366.

[8] Probst, P., Wright, M.N. and Boulesteix, A. L. (2019). *Hyperparameters and tuning strategies for random forest.* Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 9(3), e1301

[9] Qasrawi, R., Vicuna Polo, S., Al-Halawa, D. A., Hallaq, S., and Abdeen, Z. (2021), *Predicting school children academic performance using machine learning techniques*, Advances in Science, Technology and Engineering Systems Journal, Vol. 6(5), pp. 8–15.

[10] Sadrnia, L. (2023). *The Future of Marketing: How Predictive Modeling Optimizes Campaign Strategies.* iBusiness, 15, 249-262.

[11] Sethi, A. (2023). *Support Vector Regression Tutorial for Machine Learning. Analytics Vidhya. https : //www.analyticsvidhya.com/blog/*2020/03/$support - vector - regression - tutorial - for - machine - learning/$

[12] Zhang, H., Zimmerman, J., Nettleton, D., and Nordman, D. J. (2020), *Random Forest Prediction Intervals.* The American Statistician, 74(4), 392–406.