TMJM THE MINDANAWAN JOURNAL OF MATHEMATICS
OFFICIAL JOURNAL OF THE MINDANAO STATE UNIVERSITY - ILIGAN INSTITUTE OF TECHNOLOGY

Vol. 6 (2024), no. 2, pp. 155–170 ISSN: 2094-7380 (Print) | 2783-0136 (Online)
http://doi.org/10.62071/tmjm.v6i2.724.

Principal component and multiple regression analysis in modelling of grade and factors affecting the students' performance in Pre-Calculus

Loribee Ann T. Cabangisan[1], Bernadette F. Tubo[2] and Catherine R. Caño[3,*]

[1]Senior High School Department
Notre Dame of Midsayap College, 9410 Midsayap Cotabato, Philippines
loribee.cabangisan@g.msuiit.edu.ph

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

Received: 12th May 2024 Revised: 8th April 2025

Abstract

Predicting students' academic performance is helpful for educational institutions striving to improve students' success and provide support to those at risk of getting a failing grade. This paper presents an empirical study that uses students' academic performance and demographic data to predict the Pre-Calculus grades of Senior High School students in the STEM track during the first quarter of the 2023-2024 academic year, employing both multiple linear regression and principal component regression methods. Multiple regression analysis was used to fit the Pre-Calculus grade using forty-three (43) school-related variables as predictors. A variable selection method based on high loadings of varimax rotated principal components was used to obtain subsets of the predictor variables to be included in the regression model of the Pre-Calculus grade. Result shows that while MLR exhibits slightly higher R^2, lower MSE, and lower MAE compared to MLR-PCA, the differences are negligible. Attending a private school, achieving high grades in core subjects such as Mathematics, Science, English, and Filipino and performing well on assessments such as pre-test, post-tests, and entrance exams play a significant role in the grade of the student in Pre-Calculus. Moreover, the variable regular attendance, fewer past class failures, and shorter commute times seems to contribute improvement of the student's grade in Pre-Calculus.

1 Introduction

In the present information era, education is one of the most essential factors in determining a society's literacy level and the rate at which a country's economy grows. Evaluating students' performance reflects the efficiency of educational institutions responsible for developing successive generations. Focusing on the development of the educational process is one of the utmost necessities that push governments represented by educational institutions to make tremendous

*Corresponding author
2020 Mathematics Subject Classification: 97K80, 62H25, 62J05
Keywords and Phrases: multiple linear regression, principal component analysis, validation, prediction, models

and painstaking efforts to push the educational process towards continuous and escalating development [9]. With this, educational institutions need to develop strategies that enhance their performance system. Those strategies can be planned by analyzing the students' performance since the advanced estimation of the failure rate can help educational institutions make preventive decisions to decrease this rate [4]. Predicting students' performance helps teachers monitor students to support them and integrate the training programs to obtain the best results. One of the benefits it offers is its ability to reduce official warning signs and student expulsions attributed to inefficiencies.

Typically, toward the end of the semester, there is a notable increase in the failure rate among high school students, largely attributed to grade obtained by the student in Pre-Calculus. Hence, it is within the researcher's interest to predict student's performance in Pre-Calculus, aiming to identify those who may require the most guidance from their teachers. A student's academic performance may be affected by several factors like demographic information such as sex, age and address, family backgrounds such as parents' cohabitation status, family size, mother's and father's educational attainment, job, and annual income and social information such as the time hanging out with friends, and many others. One way to objectify this is through the use of learning analytics. Learning analytics is a process of measuring and analyzing learning data collected from the learning environment [12] and one of the leading research topics is predicting students' learning performance.

On the other hand, predictive analytics is the use of statistics and modeling techniques to forecast future outcomes and behavior by analyzing patterns in a given set of current and historical data. In education, the outcome variables can refer to students' performance in the form of marks, numeric values (regression task), decisions, and categorical values (classification task). One of the most prominent modeling technique used to show relationships between response and explanatory variables is regression analysis.

Multiple regression analysis is one of the most widely used methodologies for expressing the dependence of a response variable on several independent (predictor) variables. In spite of its evident success in many applications, however, the regression approach can face serious difficulties when the independent variables are correlated with each other [1]. Multicollinearity, or high correlation between the independent variables in a regression equation, can make it difficult to correctly identify the most important contributors to a modelling process. A key challenge in regression analysis often occurs from the inclusion of variables that contribute little to explaining the response variable.

One method for removing such multicollinearity and redundant independent variables is to use multivariate data analysis (MDA) techniques, and one such method is the Principal component analysis (PCA). Essentially, PCA maximizes the correlation between the original variables to form new variables that are mutually orthogonal, or uncorrelated.

Moreover, in the era of rapid technological advancement, the number of collected data with various types of variables has substantially increased. Consequently, inclusion of numerous variables considerably affect the goodness-of-fit of the prediction model obtained using MLR, as noted by Yang [19].

With consideration of the challenges that may be brought about doing regression analysis, particularly both, Multiple Linear Regression (MLR) and Multiple Linear Regression with Principal Component Analysis (MLR-PCA), this study aims to develop regression models to predict students'performance in Pre-Calculus using data obtained from 225 STEM Grade 11 students enrolled at Notre Dame of Midsayap College during the first quarter of the academic year 2023-2024. Moreover, the study aims to identify the significant explanatory variables that influence students'performance with respect to their Pre-Calculus grade. Additionally, it will conduct a comparative analysis to assess the predictive performance and accuracy of both mod-

els. Baseline comparison are based on key metrics such as adjusted R^2, mean square error (MSE), and mean absolute error (MAE) values.

2 Factors Affecting Students' Academic Performance

There are plenty of studies on educational achievements factors. Danilowicz-Gösele et al. (2014), Dooley et al. (2012), and Aina (2011) confirmed the influential role of high school grade point average, establishing its significant dominance over other variables like university program, gender, and neighborhood in explanatory power. Similarly, Dooley et al. (2012) showed that the type of university program, gender, neighborhood, and high school characteristics have weak links with university outcomes. Aina (2011) discussed the potential impact of individual characteristics, academic performance, geographic mobility, and family size on the successful completion of a degree program. Pritchard and Wilson (2003) highlighted an association between GPA and emotional and social elements, such as stress levels and the frequency of alcohol consumption. Al-Qaysi et al. (2020) found that low entry grades and living away from family support significantly impact students' academic performance. Additionally, family expenditure and income were identified as influential factors in academic performance [18] [3].

It is also important to consider the impact of students' legacy data, such as the results of previous assessments, which significantly impact the students' academic performance [18]. Previous grades and class performance were significant factors that could help determine a student's academic success and influence students' academic performance. Also, GPA greatly influences predicting students' academic performance, thus many researchers have used it to analyze students' performance [10]. Several studies in predicting students' academic performance have also revealed that quizzes, tests, mid-term exams, and assignments are significant factors influencing students' academic performance; accordingly, these attributes were used to predict the students' academic performance [16] [8] [11] [17] [14].

3 Multiple Linear Regression Analysis

Regression analysis is a statistical method used to model and quantify the relationship between variables, often identifying cause-and-effect relationships. In this study, the multiple linear regression (MLR) is employed. MLR refers to a statistical technique used to predict the outcome of a dependent variable based on the assumed linear relationship with several independent or predictor variables. A MLR model can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon. \tag{1}$$

where Y is the output variable, likewise called response or dependent variable; $X_j, j = 1, 2, .., k$ are the independent variables, also called regressors or predictors; β_0 is the intercept, representing the estimated value of Y if all independent variables are zero; $\beta_j, j = 1, 2, ..., k$, called the parameters or regression coefficients, represents the estimated change in the response of Y per unit change in X_j when holding other independent variables $X_i (i \neq j)$ constant; and ε gives the random error in Y not explained by the independent variables.

MLR is an extension of a simple linear regression (SLR) and its underlying assumptions and conditions are essentially the same with SLR. It is assumed that ε_i are independent and identically distributed, and have a normal distribution with mean 0 and constant variance σ^2.

Multicollinearity can sometimes exist in regression models. It happens when two or more predictors demonstrate moderate to high correlations with each other. This can cause the coefficient estimates of the model to be unreliable and have high variance. The easiest way to

http://doi.org/10.62071/tmjm.v6i2.724

determine if there is multicollinearity is by examining the variance inflation factor (VIF) value for each predictor variable. A VIF of 1 indicates no correlation, while values between 1 and 5 suggest moderate correlation. However, VIFs exceeding 5 warrant further investigation due to a high risk of multicollinearity, which can negatively impact regression analysis [15]. To deal with multicollinearity, techniques like Principal Component Analysis (PCA) can be employed.

3.1 Hypothesis Testing in Multiple Linear Regression

To test for the significance of the coefficients β_i given in equation (1) to determine if there is a linear relationship between Y and any of the $X_j, j = 1, 2, ..k$, the null and alternative hypotheses are stated, respectively:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1 : \beta_j \neq 0, \text{ for at least one } j.$$

When the null hypothesis H_0 is rejected, it means that at least one of the independent variables significantly contributes to the model. The test statistic F is given by

$$F = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}}$$

where

$$SS_R = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2,$$
$$SS_{Res} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} e_i^2,$$

where SS_R is the sum of squares due to regression and SS_{Res} is the residual sum of squares; e is the residual, $\hat{Y}$ is the fitted value, and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

3.2 Test on Individual Regression Coefficients

The hypotheses for significance testing of any regression coefficient $\beta_j,\ j = 1, 2, ..., k$, are

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0.$$

The test statistic is given by

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}.$$

If H_0 is rejected it implies that the independent variable can be discarded from the model.

4 Principal Component Analysis

Principal component analysis (PCA) is a technique for reducing the dimensions of datasets, increasing interpretability and minimizing information loss. In PCA, a set of correlated variables is transformed into a set of linearly uncorrelated variables, known as principal component (PC). This transformation aims to maximize explained variance while minimizing information loss.

The number of extracted PCs is always less than or equal to the original number of correlated variables.

Suppose that a random vector $\mathbf{X}$ given by $\mathbf{X}' = [X_1, X_2, \cdots, X_p]$ have a variance-covariance matrix of

$$Var(\mathbf{X}) = \mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix}$$

with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$.

Consider the linear combinations,

$$\begin{aligned} Z_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p = \boldsymbol{a_1'X} \\ Z_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p = \boldsymbol{a_2'X} \\ &\vdots \\ Z_k &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p = \boldsymbol{a_p'X} \end{aligned}$$

with $Var(Z_i) = \boldsymbol{a_i'\Sigma a_i}, i = 1, 2, \cdots, p$ and $Cov(Z_i, Z_k) = \boldsymbol{a_i'\Sigma a_p}, i, k = 1, 2, \cdots, p$

The principal components (PC) are those uncorrelated linear combinations $Z_1, Z_2, \cdots, Z_p$ whose variances are as large as possible.

The first PC is the linear combination $\boldsymbol{a_1'X}$ that maximizes $Var(\boldsymbol{a_1'X})$ subject to $\boldsymbol{a_1'a_1} = 1$. The second PC is the linear combination $\boldsymbol{a_2'X}$ that maximizes $Var(\boldsymbol{a_2'X})$ subject to $\boldsymbol{a_2'a_2} = 1$ and $Cov(\boldsymbol{a_1'X}, \boldsymbol{a_2'X}) = 0$. And so on until the ith step where the ith PC is a linear combination $\boldsymbol{a_i'X}$ that maximizes $Var(\boldsymbol{a_i'X})$ subject to $\boldsymbol{a_i'a_i} = 1$ and $Cov(\boldsymbol{a_i'X}, \boldsymbol{a_q'X}) = 0$ for $q < i$.

Before implementing PCA, it is important to ensure that the available data is appropriate for this method. Hence, Kaiser-Meyer-Olkin (KMO) index and Bartlett test are used. The Kaiser-Meyer-Olkin (KMO) index ranges from zero to one. A higher KMO value indicates greater suitability of the data for PCA, while lower values (typically below 0.6), suggest less appropriateness. In addition, PCA is considered suitable if the significance level of the Bartlett test is less than 5%.

5 Methodology

The flowchart in Figure 1 illustrates the overall research methodology, comprising three primary phases: data preprocessing, model building, and model evaluation. Data preprocessing involves data collection, cleaning, coding of categorical variable, and exploratory data analysis. Model building encompasses feature selection and model creation. Finally, model evaluation is conducted using R^2, MSE, and MAE metrics.

5.1 Data

Data were obtained from 225 STEM Grade 11 students enrolled at Notre Dame of Midsayap College. Information were obtained from their current records, such as the advisers' or subject teachers' school forms, and enrollment sheets from the Senior High School (SHS) office. Furthermore, data acquisition for personal information and family background was conducted via online and onsite survey. This research incorporated forty-three (43) independent variables encompassing various aspects of students' profiles, including personal information, family background, educational history, physiological and social factors, as well as residential location. On

http://doi.org/10.62071/tmjm.v6i2.724

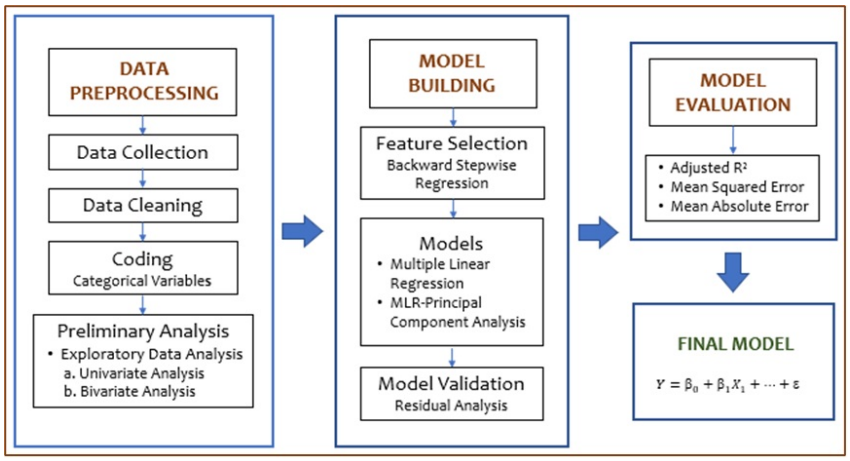

Figure 1: Schematic Diagram of the Research Methodology

the other hand, the dependent variable is the students' academic performance in Pre-Calculus during the first quarter of A.Y. 2023-2024. The complete attributes for each student are summarized and described in Table 1.

5.2 Multiple Linear Regression (MLR) Model

Let Y_i be the GPA of the ith student. The regression model to be fitted is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i, i = 1, 2, ..., 225 \tag{2}$$

where Y_i is the dependent variable and $X_{1i}, ..., X_{ki}$ are the independent variables listed in Table 1. In addition, the coefficient β_0 is the intercept, the parameters $\beta_j, j = 1, 2, ..., k$ are the regression coefficients for the independent variables, and ε_i are independent and identically distributed with mean 0 and a constant variance σ^2.

We typically assess the following assumptions: (1) the linearity in the model parameters, (2) independence of errors, (3) equality of error variances, (4) normality of error terms, and (5) multicollinearity.

The most significant variables were selected using the significant $p-values$ generated by the stepwise regression technique. The model is fitted using the ordinary least squares estimation in R version 4.3.2. All analyses were done using a 5% level of significance.

5.3 MLR with PCA

To perform PCA, only the numerical variables listed in Table 1 are considered. The original data and the average of the original data are chosen. The covariance matrix is computed and the result was used in calculating the eigenvectors and eigenvalues, then the eigenvector with the highest eigenvalue is chosen as the principal component of the data set as it exhibits the most significant relationship between the data set attributes. The eigenvalues are sorted in ascending order to choose the most significant data and discard the least significant one. The variance, covariance, eigenvalues, and eigenvectors were determined. After calculating the principal components (PCs), some of these components are used as predictors in a linear regression model fitted using the typical least squares procedure. It involves having the model construct components from the independent variables that are a linear combination of the independent variables. That is, these selected PCs together with the categorical variables, are

Table 1: Dataset Description

Attribute	Description / Values	Attribute	Description / Values
Age (X_1)	student' age (numerical: from 15 to 19)	mEA (X_{24})	parent's education (0–elementary undergraduate, 1–elementary graduate, 2–HS undergraduate, 3–HS graduate, 4–college level, and 5–college graduate or 6–postgraduate)
PrevGWA (X_2)	students' previous general weighted average (numeric: 70-100)		
PrMath (X_3)	students' previous final grade in Math (numeric: 70-100)	fEA (X_{25})	
PrSci (X_4)	students' previous final grade in Science (numeric: 70-100)	mEmp (X_{26})	parent's employment status (0– unknown or deceased, 1– unemployed, 2–worker or contractual, 3– self-employed, 4–employed/permanent, 5–retired or pensioner)
PrEng (X_5)	students' previous final grade in English (numeric: 70-100)	fEmp (X_{27})	
PrFil (X_6)	students' previous final grade in Filipino (numeric: 70-100)	Grd (X_{28})	guardian (0– relative, 1– father or 2– mother)
EntrExam (X_7)	students' entrance exam percentile (numeric: 0-100)	PStats (X_{29})	parent's cohabitation status (binary: 0– apart or 1– living together
Abs (X_8)	number of absences in Precalculus (numeric: 0-7)	Famsize (X_{30})	family size (binary: 0– less or equal to 3 or 1– greater than 3)
PreTest (X_9)	pre-test raw scores in Precalculus (numeric: 0-25)	FamRel (X_{31})	quality of family relationships (categorical: 0–very bad to 3–very good)
PostTest (X_{10})	post-test raw scores in Precalculus (numeric: 0-25)	HighEd (X_{32})	wants to take higher education (binary: yes or no)
Fail (X_{11})	number of past class failures (numeric: 0 and above)	Health (X_{33})	current health status (categorical:0– very bad to 3– very good)
StudTime (X_{12})	weekly study time in Precalculus (numeric: 0.5-12 hours)	Internet (X_{34})	internet access at home (binary: yes or no)
Travtime (X_{13})	home to school travel time (numeric:8-45 minutes)	Bfast (X_{35})	eat breakfast (binary: yes or no)
Alln (X_{14})	student's weekly allowance (numeric: 200-1500 pesos)	Snack (X_{35})	eat snacks (binary: yes or no)
Sex (X_{15})	student' sex (binary: 0 – Male or 1 – Female	Lunch (X_{36})	eat lunch (binary: yes or no)
PrevSch (X_{16})	student's previous school type (binary: 0– public or 1– private	Romantic (X_{37})	in a romantic relationship (binary: yes or no)
LMod (X_{17})	previous learning modality (0– face to face, 1– modular, 2– online, 3–blended)	GoOut (X_{38})	going out with friends (categorical: 0– never to 3– always)
Section (X_{18})	student's section (categorical: 0 – St. Claire, 1 – St. Gregory, 2 – St. Ignatius, 3 – St. Stephen, 4 – Our Lady of Fatima)	AlcSD (X_{39})	take alcohol during study days (binary: yes or no)
Nursery (X_{19})	attended nursery school (binary: yes or no)	AlcWE (X_{40})	take alcohol during weekends (binary: yes or no)
ALS (X_{20})	attended ALS (binary: yes or no)	FreqSM (X_{41})	frequency of using social media (categorical: 0–never to 3– always)
SchSup (X_{21})	extra educational support or tutor (binary: yes or no)	HighEd (X_{42})	wants to take higher education (binary: yes or no)
mAI (X_{22}) fAI (X_{23})	parent's annual income (categorical: 0– poor (*less than ₱9,100*), 1– low income (*₱9,100-₱18,200*), 2– lower-middle (*₱18,201-₱36,400*), 3– middle (*₱36,401-₱63,700*), 4– upper-middle (*₱63,701-₱109,200*), 5– high income (*₱109,201-₱182,000*), 6– rich (at least *₱182,000 and up*))	Trans (X_{43})	Transportation to school (0– walking, 1– public vehicle, 2– private vehicle)
		Precal GPA	student's first quarter final grade in Precalculus (numeric: from 70-99) **TARGET OUTPUT**

then used as independent variables for doing MLR. The model is then expressed as

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + ... + \beta_k PC_k + \beta_{k+1} X_{15} + \beta_{k+2} X_{16} + ... + \beta_{k+q} X_{43} + \epsilon. \quad (3)$$

http://doi.org/10.62071/tmjm.v6i2.724

where $q \leq 43$. The *pls* package in R was used to perform MLR-PCA. Moreover, PCA gives good result when the following assumptions are satisfied: (1) correlation between features, (2) sensitivity to the scale of the features, (3) no outliers, (4) linear relationship between features, and (5) no missing values.

5.4 Model Evaluation

After creating the models, the accuracy of those models in predicting students' performance was determined. In practice, teachers need to know the performance accuracy in order to reduce the risk of wasting resources through incorrect interventions. Therefore, in this study, a K-fold cross validation with shuffling was used to partition the original dataset into a training dataset and testing dataset. The shuffling mechanism enables overcoming the problem of higher residual errors influenced by a single round of K-fold cross validation. The K-fold cross-validation is described as follows:

1. The entire dataset is randomly split into k-subsets of approximately equal size.
2. A model is trained on $K-1$ of these subsets and tested on the remaining subset. While testing the model, a measure of the model error is obtained.
3. Repeat this process K times until each of the k subsets have used as the test set.
4. Compute the average of the K model errors. This is called the cross-validation error using to estimate the performance of the model.

Evaluating the effectiveness of a multivariate calibration model often involves utilizing the root mean square error of prediction (RMSEP), a commonly employed criterion [16]. Note that the model's predictive ability increases as the RMSEP value decreases.

When performing the K-fold using the *caret* R package, regression-model-accuracy-metrics are computed. Those metrics were used to measure the overall quality of regression models.

Assume n is the number of points in the dataset, $\hat{Y}_i$ is the predicted outcome, Y_i is the observed outcome, and k is the number of independent variables. The adjusted R^2, MSE, and MAE are described as follow.

Adjusted R-squared ($adj\ R^2$) measures the proportion of variation explained by the independent variables that help in explaining the dependent variable. The higher the adjusted R^2, the better the model. It is given by the formula:

$$adj\ R^2 = 1 - \frac{\left(1 - R^2\right)(n-1)}{n-k-1}, \tag{4}$$

Mean Squared Error (MSE) measures the average squared difference between the predicted and the actual target values within a dataset. The lower the MSE, the better the model. It can be calculated through the formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{5}$$

Mean Absolute Error (MAE) measures the average variance between the significant values in the dataset and the projected values in the same dataset. The lower the MAE, the better the model

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|. \tag{6}$$

6 Result and Discussions

6.1 Profile of the Respondents

The respondents of this study were selected through a complete enumeration of the 225 Grade 11 SHS students of the STEM strand. Of these, 139 were female and 86 were male. Their Pre-Calculus GPAs ranges from 78 to 99, with an average GPA of 88.06 and a standard deviation of 4.76.

6.2 MLR Model

The stepwise regression test were ran using the *tidyverse*, *caret*, *leaps* and *MASS* packages in R to determine the most significant variables to be included in the MLR model. The 5-fold cross-validation result reveals that the best model among the predetermined has six variables with the lowest RMSE of 3.176 and MAE of 2.567 as displayed in Table 2.

Table 2: Cross-validation result

Number of Predictors	RMSE	RSquared	MAE
1	3.732679	0.3989773	2.995473
2	3.388385	0.4978250	2.746738
3	3.365244	0.5088224	2.717543
4	3.243238	0.5446184	2.645031
5	3.222374	0.5555085	2.598396
6	3.176353	0.5663271	2.568619
7	3.225739	0.5523457	2.609510
8	3.277973	0.5397162	2.661987
9	3.250775	0.5463422	2.650597
10	3.224743	0.5485697	2.636595
11	3.203502	0.5573902	2.573942
12	3.239670	0.5475802	2.655184

Among the 43 predictor variables, the students' pre-test score (PreTest, X_9) and post-test score (PostTest, X_{10}), previous general weighted average (PrevGWA, X_2), the type of school they previously attended (PrevSch, X_{16}), entrance exam score (EntrExam, X_7), and the number of absences in Pre-Calculus (Abs, X_8) were found to be significant as depicted in Table 3.

Table 3: Summary of the MLR Model

Variables	Coefficients	Std. Error	t value	p value	VIF
(Intercept)	31.40519	6.03738	5.202	<0.0001	
PreTest (X_9)	0.26489	0.07439	3.561	0.00045	1.3509
PostTest (X_{10})	0.41225	0.06962	5.921	<0.0001	1.5753
PrevGWA (X_2)	0.48252	0.07100	6.796	<0.0001	1.4567
PrevSch (X_{16})	1.80625	0.42678	4.232	<0.0001	1.4953
EntrExam (X_7)	0.04279	0.01219	3.510	0.00055	1.0476
Abs (X_8)	-1.32876	0.57556	-2.309	0.02190	1.0296

The proposed MLR model is given in equation (7),

$$\hat{Y} = 31.405 + 0.265X_9 + 0.412X_{10} + 0.483X_2 + 1.806X_{16} + 0.043X_7 - 1.329X_8 \quad (7)$$

http://doi.org/10.62071/tmjm.v6i2.724

where

$$X_{16} = \begin{cases} 0 & \text{if a student was previously enrolled in a public school} \\ 1 & \text{if a student was previously enrolled in a private school} \end{cases}$$

Before adopting the final model, the proposed model underwent a validation process. Residuals are normally distributed as assessed by Shapiro–Wilk's test ($p-value = 0.2989$). Breusch-Pagan test result (BP value$= 4.053, p - value = 0.6695$) suggests that the variances are homogeneous among the samples. The Durbin-Watson (D-W) test, used to determine the magnitude of autocorrelation, has a value of 2.07 with $p - value = 0.58$ suggesting that the independence of error assumption is not violated. The MLR also assumes that the independent variables are not highly correlated. The VIF values of each predictor shown in Table 3 are less than 5, implying no multicollinearity among variables.

Since all the regression assumptions are satisfied, then the MLR model is valid and reliable enough to make predictions and inferences. The MLR model (equation 7) indicates key factors contributing to NDMC students' success in Precalculus performance during the first quarter. The analysis reveals that a strong academic background from Grade 10, high entrance exam scores, and consistent attendance are crucial for better performance. Good pre-test and post-test scores also play an important role, showing the value of a strong foundation in the subject. Additionally, students from private schools tend to excel, possibly due to differences in educational quality and resources.

6.3 MLR-PCA Model

MLR-PCA uses the principal components as the predictor variables for regression analysis instead of the original features. It is important to note that PCA is sensitive to the scale of data and it is only applicable to numerical variables. Thus, in this study, PCA is only processed to 14 numerical explanatory variables. The idea is to transform the predictor variables via PCA, the factors in the principal components will then be merged to the original categorical features and then fit a Least Squares model using the transformed and original categorical features. Hence, the number of transformed variables is smaller than the number of predictors.

To check the data suitability for principal component analysis, Kaiser-Meyer-Olkin (KMO) test was conducted and the results are shown in Table 4. The KMO overall measure of sampling adequacy ($KMO = 0.82 > 0.6$) suggests that the data is appropriate for the for PCA. Moreover, the Bartlett's test of sphericity indicates that the data exhibits significant correlations, hence, we can run a PCA.

Table 4: KMO Statistics for Sampling Adequate and Bartlett's test for Homogeneity.

KMO and Barlett's test		
Kaiser-Meyer-Olkin Factor Adequacy		0.79
Barlett's Test of Sphericity	Chi-Square	2279.24
	df	105
	P-value	<0.001

To determine the minimum number of principal components that account for most of the variation in the data, the scree plot in Figure 2 was used. The first four principal components have eigenvalues greater than 1. These four components explain 67.5% of the variation in the data as seen in Table 5.

Table 5 lists the variables associated with each principal component (PC) and their corresponding eigenvectors. PC1 accounts for the most variation in the data, explaining 35.42% of

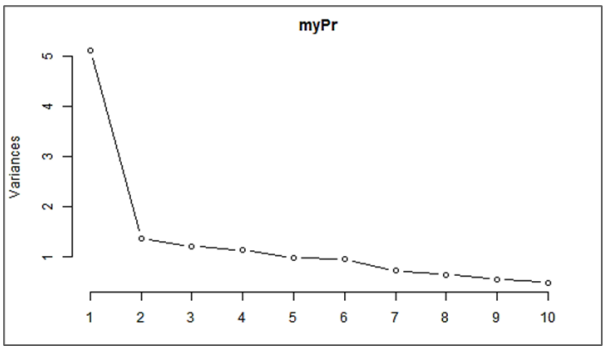

Figure 2: Validation Plot.

the variance. PC2 explains 14.2%, while PC3, PC4, and PC5 account for 9.4%, 8.48%, and 7.87% of the variance, respectively. Together, these five components explain 75.38% of the total variation in the original data. The table also shows that variables have both very high and very low loadings within each component, all contributing to the overall variance.

Table 5: Eigenvalue and Eigenvectors

Eigenvectors	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
PreTest	-0.2110	0.0699	0.3389	-0.3316	-0.3156
PostTest	-0.2887	0.0526	0.1561	-0.2148	-0.2721
PrevGWA	-0.4296	-0.0230	0.1577	0.0787	0.0869
PrMath	-0.3739	0.0543	0.2112	0.0393	0.0258
PrSci	-0.3846	-0.0313	0.0002	-0.0025	0.1377
PrEng	-0.3975	-0.0711	-0.1510	0.0823	0.0426
PrFil	-0.3877	0.0026	-0.2577	0.0627	0.0662
EntrExam	-0.2630	0.0928	0.3467	-0.3128	-0.0565
Abs	0.0653	-0.1397	-0.4072	-0.5190	0.0396
Fail	0.1011	-0.0540	-0.5307	-0.4122	0.0498
Age	-0.0238	-0.1109	-0.2562	0.5232	-0.4484
StudTime	0.0296	0.6841	-0.1453	0.0146	-0.0255
Alln	0.0314	0.6856	-0.1270	0.0267	-0.0022
TravTime	-0.0467	0.0352	0.1771	0.1032	0.7642
Eigenvalues	2.2269	1.4098	1.14743	1.08991	1.0496
Var Proportion	0.3542	0.1420	0.09404	0.08485	0.0787
Cum Proportion	0.3542	0.4962	0.59023	0.67508	0.7538

Next, these PCs were merged to the remaining 29 categorical data for MLR-PCA model. The general model was developed using the *pcr* function software that uses *scale* argument to make sure everything is measured the same way. It also uses validation argument that enables cross-validation. This helped in determining the number of components to use for prediction. The 5-fold cross-validation result reveals that the best model has three variables with the lowest RMSE of 3.3438 and MAE of 2.6670 as shown in Table 6.

Backward stepwise regression was applied to further identify the variables that contribute statistical significance to the model. The significant variables were PrevSch, PC1, and PC3. Given that PC1 and PC3 aggregate multiple underlying variables, they were further extracted, and their coefficients were combined into a linear combination, as detailed in Table 7. The coefficients from PC1 and PC3 were integrated into the final coefficients for each predictor in

http://doi.org/10.62071/tmjm.v6i2.724

Table 6: Cross Validation Results.

Number of Predictors	RMSE	RSquared	MAE
1	3.4452	0.4857	2.7317
2	3.4191	0.4938	2.6999
3	3.3438	0.5200	2.6670
4	3.3876	0.5088	2.7050
5	3.3957	0.5091	2.6929
6	3.3630	0.5161	2.6732
7	3.4074	0.5014	2.7217

the model.

Table 7: Summary of the MLR-PCA Model

Variables	**Coefficients**	**Std. Error**	**t value**	**P-value**	**VIF**
(Intercept)	86.8591	0.36042	240.993	<.0001	-
PrevSch (X_{16})	1.88601	0.45553	4.14	<.0001	1.042915
PC1	-1.56792	0.09866	-15.892	<.0001	1.042814
PC3	0.72597	0.18752	3.871	<.0001	1.000101
Extracted Variables		**Coefficients**			
Components	**PC1**	**PC3**	**Total**		
Age (X_1)	0.03731	-0.18592	-0.14861		
PrevGWA (X_2)	0.67357	0.11448	0.78805		
PrMath (X_3)	0.58624	0.15332	0.73956		
PrSci (X_4)	0.60302	0.00014	0.60316		
PrEng (X_5)	0.62324	-0.10962	0.51362		
PrFil (X_6)	0.60788	-0.18708	0.4208		
EntrExam (X_7)	0.41220	0.25169	0.66389		
Abs (X_8)	-0.10222	-0.29561	-0.39783		
PreTest (X_9)	0.33083	0.24603	0.57686		
PostTest (X_{10})	0.45266	0.11332	0.56598		
Fail (X_{11})	-0.15835	-0.38527	-0.54362		
StudTime (X_12)	-0.04641	-0.10541	-0.15182		
TravTime (X_{13})	0.07322	-0.12849	-0.20171		
Alln (X_{14})	-0.04923	-0.09219	-0.14142		

The MLR-PCA model is reflected in Equation (8).

$$\begin{aligned}\hat{Y} &= 86.859 + 1.886X_{16} - 0.149X_1 + 0.788X_2 + 0.740X_3 \\ &+ 0.603X_4 + 0.514X_5 + 0.421X_6 + 0.664X_7 - 0.398X_8 \\ &+ 0.577X_9 + 0.566X_{10} - 0.544X_{11} - 0.152X_{12} \\ &- 0.202X_{13} - 0.141X_{14}.\end{aligned} \tag{8}$$

where

$$X_{16} = \begin{cases} 0 & \text{if a student was previously enrolled in a public school} \\ 1 & \text{if a student was previously enrolled in a private school} \end{cases}$$

The proposed model in Equation 8 underwent a validation process. Residuals are normally distributed as assessed by Shapiro–Wilk's test ($p - value = 0.2266$). Breusch-Pagan (BP) test

result ($BP\ value = 1.711, p - value = 0.6345$) implies that the variances are homogeneous among the samples. The Durbin-Watson (D-W) test was used to determine the magnitude of autocorrelation. The D-W value of 2.09 ($p - value = 0.644$) and autocorrelation of -0.0348 indicates that the residuals are uncorrelated; that is, the independence of error assumption is not violated. Moreover, the VIF values of each predictor is shown in Table 7, and all values are lesser than 5. Thus, no multicollinearity assumption is satisfied.

Upon analysis, all the regression assumptions were met. Thus, the MLR-PCA model is reliable enough to make predictions and inferences.

The MLR-PCA model in Equation 8 highlights several factors that enhance a student's first quarter performance in Precalculus at NDMC. Attending a private school during junior high contributes positively to learning outcomes. High previous general weighted averages and good grades in core subjects such as Math, Science, English, and Filipino demonstrate consistent academic achievement and mastery of essential skills. Moreover, achieving high scores on pre-tests, post-tests, and entrance exams indicate preparedness and academic capability. Beyond academic factors, regular attendance, fewer past class failures, and shorter commutes to school also improve students' performance. Additionally, efficient study habits, reflected in less study time and a younger age, contribute to better focus and effective learning. Notably, an average weekly allowance appears to positively affect performance, potentially reflecting students' time management and budgeting skills.

6.4 Model Evaluation

The final models, along with their optimal features for both MLR and MLR-PCA, are given, respectively,

$$\hat{Y}_{\text{MLR}} = 31.405 + 0.483X_2 + 0.043X_7 - 1.329X_8 + 0.265X_9 + 0.412X_{10} + +1.806X_{16} \quad \text{and}$$

$$\begin{aligned}\hat{Y}_{\text{MLR-PCA}} = 86.859 &- 0.149X_1 + 0.788X_2 + 0.740X_3 + 0.603X_4 + 0.514X_5 + 0.421X_6 + \\ &0.664X_7 - 0.398X_8 + 0.577X_9 + 0.566X_{10} - 0.544X_{11} - \\ &0.152X_{12} - 0.202X_{13} - 0.141X_{14} + 1.886X_{16}.\end{aligned}$$

The values for the metrics $adj\ R^2$, MSE, and MAE values are shown in Table 8.

The MLR model has five (5) positive coefficients and one (1) negative coefficient. This infers that the increase in students' pre-test and post-test scores, previous GWA and entrance exam scores would also cause an increase in their Pre-Calculus grades. Moreover, a student who graduated from private school is more likely to have higher grade in Pre-Calculus. In contrast, a higher number of absences negatively affects a student's performance in Pre-Calculus.

The MLR-PCA model identifies nine (9) factors that positively influence and six (6) factors that negatively impact the first-quarter Precalculus performance among STEM students at NDMC during the 2023-2024 academic year. A higher previous general weighted average, previous Math, Science, English and Filipino averages, higher entrance exam, lower number of absences in class, higher scores in pretest and posttest, a lower number of past failures, and a shorter travel time to school would cause an increase in their Precalculus grade. Additionally, efficient study habits, often characterized by less study time can positively influence focus and learning outcomes. However, a higher average weekly allowance may have a negative impact on performance, potentially indicating challenges in time management and budgeting skills. Moreover, Pre-Calculus grade would be more likely to increase if a student graduated from a private school.

Based on Table 8, compared to MLR-PCA, MLR exhibits slightly higher adjusted R^2, lower MSE, and lower MAE. However, these differences are minimal, suggesting that both MLR and MLR-PCA models perform comparably in this case study.

http://doi.org/10.62071/tmjm.v6i2.724

Table 8: Final Fitted Models

Fitted Models	Best Features	Adj. R^2	MSE	MAE
MLR	PreTest (X_9), PostTest (X_{10}), PrevGWA (X_2), PrevSch (X_{16}), EntrExam (X_7), Abs (X_8)	0.6827	7.7751	2.2096
MLR-PCA	PrevSch (X_{16}), Age (X_1), PrevGWA (X_2), PrMath (X_3), PrSci (X_4), PrEng (X_5), PrFil (X_6), EntrExam (X_7), Abs (X_8), PreTest (X_9), PostTest (X_{10}), Fail (X_{11}), StudTime (X_{12}), TravTime (X_{13}), Alln (X_{14})	0.5895	9.3095	2.5041

7 Concluding Remarks

This study seeks to identify the primary explanatory variables that shows significant influence on students' performance in Pre-Calculus. The findings of this research are context-specific, focusing on providing insights for teachers at Notre Dame of Midsayap College. The primary goal is to help teachers identify significant variables for improvement and success of their students in Pre-Calculus. The predictive models were obtained using Multiple Linear Regression (MLR) and Multiple Linear Regression with Principal Component Analysis (MLR-PCA) .

The result derived using MLR shows that the following variables are statistically significant in the first-quarter Precalculus performance of the students: (1) students' previous general weighted average, (2) entrance exam score, (3) number of absences, (4) pretest and posttest scores, and (5) the type of school they previously attended.

Moreover, using MLR-PCA, the variables that showed significant relationship with the Pre-Calculus performance of the students are as follows: (1) age, (2) previous general weighted average, (3) previous Math, Science, English, and Filipino grades, (4) entrance exam percentile, (5) number of absences, (6) study time in Precalculus, (7) pretest and posttest scores, (8) the type of school they previously attended, (9) number of failures, (10) weekly allowance and (11) the students' travel time to school.

Subsequent residual analysis assessed the models' performance using R^2, MSE, and MAE metrics. While MLR exhibits marginally higher adjusted R^2, lower MSE, and lower MAE than MLR-PCA, the differences are negligible. This indicates that both models perform similarly well. However, since the Kaiser-Meyer-Olkin (KMO) test and Bartlett's indicated the suitability of the data for PCA suggesting the presence of correlations among certain explanatory variables, hence MLR-PCA may be a more effective predictive model. By addressing multicollinearity,

MLR-PCA can better capture the relationships between numerical variables and accurately predict first-quarter Precalculus grades in Notre Dame of Midsayap College STEM students for the 2023-2024 academic year.

This study hopes to inspire further research into the development of related models aimed at predicting students' performance in relevant courses, particularly in the fields of science and mathematics.

Acknowledgements

The authors gratefully acknowledge the support provided by Notre Dame of Midsayap College. We also extend our sincere thanks to the anonymous referees for their insightful comments and suggestions, which significantly enhanced the clarity and rigor of this manuscript.

References

[1] S. A. Abdul-Wahab, C. S. Bakheit, and S. M. Al-Alawi, *Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting it concentration*, Environmental Modelling and Softwares **20** (2005), no. 10, 1263–1271, doi.org/10.1016/j.envsoft.2004.09.001.

[2] C. Aina, *The Determinants of Success and Failure of Italian University Students. Evidence from administrative data*. RePEc: Research Papers in Economics **119** (2010), no. 2, 85–108.

[3] N. R. Aljohani, A. Daud, R.A. Abbasi, J.S. Alowibdi, M. Basheri and M.A. Aslam, *An integrated framework for course adapted student learning analytics dashboard*, Computers in Human Behavior, **92** (2019), 679-690 , doi:10.1016/j.chb.2018.03.035.

[4] O. El Aissaoui, El Alami El Madani, Y., L. Oughdir, A. Dakkak, Y. El Allioui, *A Multiple Linear Regression-Based Approach to Predict Student Performance*, In: Ezziyyani, M. (eds) Advanced Intelligent Systems for Sustainable Development (AI2SD'2019). AI2SD 2019. Advances in Intelligent Systems and Computing **1102** (2020). Springer, Cham., doi:10.1007/978-3-030-36653-7_2.

[5] N. Al-Qaysi, N. Mohamad-Nordin, N., and Al-Emran, *A Systematic Review of Social Media Acceptance From the Perspective of Educational and Information Systems Theories and Models*, J. Educ. Comput. Res. **57** (2020), no. 8, 2085–2109, doi:10.1177/0735633118817879.

[6] K. Danilowicz-Gösele, *Determinants of students' success at university*, (2014), https://www.econstor.eu/handle/10419/102397.

[7] M.D. Dooley, A. A. Payne, and A. Robb, *Persistence and academic success in university*, Canadian Public Policy-analyse De Politiques **38** (2012), no. 3, 315–339, doi:10.1353/cpp.2012.0028.

[8] G. Elakia and N. J. Aarthi, *Application of data mining in educational database for predicting behavioural patterns of the students*. Elakia et al,/(IJCSIT) International Journal of Computer Science and Information Technologies **5** (2014), no. 3, 4649–4652.

[9] S. Li and T. Liu, *Performance prediction for higher education students using deep learning*, Complexity, (2021), 1–10.

[10] N. M. Suhaimi, S. Abdul-Rahman, S. Mutalib, N. A. Hamid, and A. Hamid, *Review on predicting students' graduation time using machine learning algorithms*, International journal of modern education and computer science **11** (2019), no. 7, 1–13, doi:10.5815/ijmecs.2019.07.01.

[11] T. Papamitsiou, V. Terzis, and A. A. Economides, *Temporal learning analytics for computer-based testing*, In Proceedings of the fourth international conference on learning analytics and knowledge, (2014), pp. 31–35, doi:10.1145/2567574.2567609.

[12] A. Peña-Ayala (Ed.). *Learning Analytics: Fundaments, Applications, and Trends*, Studies in Systems, Decision and Control, Springer International Publishing, 2017, doi:10.1007/978-3-319-52977-6.

[13] M. E. Pritchard and G. S. S. Wilson, *Using Emotional and Social Factors to Predict Student Success*, Journal of College Student Development **44** (2003), no. 1, 18–28, doi:10.1353/csd.2003.0008.

[14] A. M. Shahiri, W. Husain, and N. A. Rashid, *A Review on Predicting Student's Performance Using Data Mining Techniques*, Procedia Computer Science **72** (2015), 414–422, doi:10.1016/j.procs.2015.12.157.

[15] N. Shrestha, *Detecting Multicollinearity in Regression Analysis*, American Journal of Applied Mathematics and Statistics **8** (2020), no. 2, 39–42, .

[16] C. Tucker, B. Pursel, and A. Divinsky, *Mining Student-Generated Textual Data In MOOCS and Quantifying Their Effects on Student Performance and Learning Outcomes*, 2014 ASEE Annual Conference & Exposition Proceedings, doi:10.18260/1-2--22840.

[17] V. Vijayalakshmi, K. Venkatachalapathy, and V. Ohmprakash, *A Comparison of Classification Techniques on Prediction of Student Performance in Educational Data Mining*, no. September, 2017. Res Militaris, vol.12, n°6, Winter 2022 294.

[18] 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*, Comput. Human Behav. **104** (2020), 106189, doi:10.1016/j.chb.2019.106189.

[19] S. J. H. Yang, O. H. T. Lu, A. Y. Q. Huang, J. C. H. Huang, H. Ogata, and A. J. Q. Lin, *Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis*, Journal of Information Processing **26** (2018), 170–176, doi:10.2197/ipsjjip.26.170.