Jurnal TAM (*Technology Acceptance Model*)
Jurnal TAM, Volume 15, Number 1, July 2024
E ISSN: 2579-4221; P ISSN: 2339-1103, pp. 8-17
Accredited SINTA 4 Number 225/E/KPT/2022
https://jurnal.ftikomibn.ac.id/index.php/JurnalTam/index

LINEAR REGRESSION MODEL: A STEP ANALYSIS AND ITS APPLICATION FOR EVALUATING THE STUDENT LEARNING PROCESS IN MATH SUBJECT

Erna Kumalasari Nurnawati[1], Ismail Setiawan[2*]
[1]Informatics Department, Institut Sains & Teknologi Akprind, Yogyakarta
[4]System and Information Technology, Universitas Aisyiyah, Surakarta
[1]Jl Kalisahak 28 Balapan, Yogyakarta, Indonesia
[2]Jl. Ki Hajar Dewantara No.10, Jawa, Kec. Jebres, Kota Surakarta, Indonesia
E-mail: ernakumala@akprind.ac.id[1], ismail@aiska-university.ac.id[2*]

Article history:
Received: July 8, 2023
Revised: December 28, 2023
Accepted: July 6, 2024

Corresponding authors
***ismail@aiska-university.ac.id**

Keywords:
Linear Regression;
Model;
Student Learning Process;
Math.

Abstract
The effort to improve the quality of education aims to produce generations or individuals who excel in thinking, actions, and decision-making. The measurement of students' character is based on their school grades. Good or excellent grades are influenced by several factors. Research on student achievement improvement has been conducted across various parts of the world. More specifically, the data used in this research employs linear regression method on public data regarding students' learning achievement in the field of mathematics, which is available in the UCI Machine Learning repository. The utilization of public data aims for the research to be validated by other researchers. Accurately selecting the appropriate factors will enhance the school's success in making decisions to improve student achievements. The results presented in this study indicate that the attribute G3 has the highest correlation with other attributes, followed by G2 and G1.

1. INTRODUCTION

Since its early discovery, mathematics has revealed its uniqueness through its universality as a scientific language that transcends cultural boundaries, providing a consistent and precise foundation in science [1]. Through the development of theorems and proof methods, mathematics facilitates the exploration of analytical thinking and the solution of complex problems, which remain relevant today in the advancement of sophisticated technologies such as artificial intelligence and data analysis. Over time, mathematics has also given rise to mathematical models that provide profound insights into the real world, while its abstractions teach the ability to think abstractly and lead to the beauty of structural patterns and mathematical concepts [2].

Some of the students' issues with mathematics include difficulties in understanding basic concepts, where some students might face challenges in grasping fundamental mathematical concepts such as arithmetic operations, numbers, decimals, and fractions [2]. The lack of motivation contributes to many students feeling demotivated towards mathematics due to the perception that the subject is difficult or irrelevant to everyday life [3]. Additionally, the fear of numbers leads to students experiencing anxiety towards numbers, known as "math anxiety," which can hinder their performance in the subject of mathematics [4]. Furthermore, the severity lies in the lack of practical practice, even though learning mathematics requires active practice and continuous understanding. Insufficient practice and profound understanding can result in difficulties applying concepts in real-life situations [5].

Research related to improving students' understanding of mathematics has been rapidly advancing nowadays. Approaches from the

computational aspect to various variables have been carried out. An approach involving the creation of a smart education platform has been developed based on the processing of extensive data using machine learning to provide relevant content to the learners [6]. The development of learners' soft skills in entrepreneurship has been pursued by establishing an educational system grounded in the results of a machine learning approach [7].

Prior to the onset of online learning, students had observable attitudes toward the process and outcomes of their learning. However, due to the occurrence of COVID-19 and the transition to online learning, new challenges have emerged. One such challenge is the students' lack of understanding and engagement with learning materials, as learning can now take place anywhere and anytime. Research focused on creating decision support systems to assist teachers in understanding students' conditions has also been initiated to detect undesired behaviours as early as possible [8]. On the other hand, a different approach was undertaken by [9], where linear regression was used to eliminate variables that do not influence students' learning development. However, the case study conducted involved pharmacy students. Linear regression was previously employed to identify the causes of low birth weight, where low birth weight contributes to an increased infant mortality rate. Researchers attempted to incorporate maternal education as a factor in the study, but maternal education was not an independent variable [10].

II. LITERATURE

2.1 Linear Regression Model

A linear regression model describes the relationship between a dependent variable, y, and one or more independent variables, X. The dependent variable is also called the response variable. Independent variables are also called explanatory or predictor variables. Continuous predictor variables are also called covariates, and categorical predictor variables are also called factors. The matrix X of observations on predictor variables is usually called the design matrix [17]. Multiple linear regression model is

$$y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_pX_{ip}+\varepsilon_i, \quad i=1,\cdots,n,$$

were

- n is the number of observations.
- y_i is the ith response.
- β_k is the kth coefficient, where β_0 is the constant term in the model. Sometimes, design matrices might include information about the constant term. However, fitly or stepwise by default includes a constant term in the model, so you must not enter a column of 1s into your design matrix X.
- X_{ij} is the ith observation on the jth predictor variable, j = 1, ..., p.
- ε_i is the ith noise term, that is, random error.

If a model includes only one predictor variable (p = 1), then the model is called a simple linear regression model. In general, a linear regression model can be a model of the form

$$y_i=\beta_0+\sum_{K}{}_{k=1}\beta_kf_k(X_{i1},X_{i2},\cdots,X_{ip})+\varepsilon_i, \quad i=1,\cdots,n,$$

where $f(.)$ is a scalar-valued function of the independent variables, X_{ij}s. The functions, $f(X)$, might be in any form including nonlinear functions or polynomials. The linearity, in the linear regression models, refers to the linearity of the coefficients β_k. That is, the response variable, y, is a linear function of the coefficients, β_k.

2.2 Normalization

Normalization, a vital aspect of Feature Scaling, is a data preprocessing technique employed to standardize the values of features in a dataset, bringing them to a common scale. This process enhances data analysis and modeling accuracy by mitigating the influence of varying scales on machine learning models [17].

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature, respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator, and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

III. RESEARCH METHODS

The research methodology employed in this study is a case study approach, where we will delve deep into the issues faced by a group of students in the subject of mathematics. We will identify the factors influencing their level of understanding and develop suitable solutions based on in-depth analysis of individual cases. We will utilize publicly available data from the UCI Machine Learning repository to gain a comprehensive understanding of the challenges students encounter in comprehending mathematical concepts. This approach allows for other researchers to replicate the study, ensuring fairness and accuracy in measurement. Furthermore, we will analyse this data using both qualitative and quantitative approaches to identify common patterns

and individual differences. From this analysis, we will formulate recommendations and learning strategies tailored to the specific needs of each student, with the aim of enhancing their comprehension of the subject of mathematics.

This research employs a linear regression method to analyse the relationship between two variables deemed to mutually influence each other. In this context, we gather data regarding students' achievements in mathematics (dependent variable) and the factors influencing these achievements, such as study time, practice frequency, and parental support (independent variables). We will apply linear regression analysis to measure the extent to which the independent variables can account for the variation in students' mathematics performance. Through this approach, we will be able to identify the most influential factors on mathematics performance and make predictions about how changes in the independent variables can impact student performance. The outcomes of this research are expected to provide a deeper understanding of the factors affecting students' mathematics performance and offer guidance to educators in enhancing mathematics instruction.

IV. RESULTS

4.1 Data Obtained

The data obtained from the UCI Machine Learning repository (refer to Table 1) is public data created by Paulo Cortez [11]. He aimed to compare the performance of Portuguese students with Irish students. The approach employed involved classification and regression. This study utilized multiple linear regression and was conducted using the Rapid Miner application. Multiple linear regression was used to model the relationship between more than one independent variable and a single dependent variable. In general, the formula for multiple linear regression can be expressed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \in$$

Where:
Y represents the dependent variable to be explained.
β_0 is the constant (intercept).
β_1 〖, β〗_2,…,β_p represents the regression coefficients for each independent variable X_1 〖,X〗_2,…,X_p.
X_1 〖, X〗_2,…,X_p is the independent variable used in the model.
∈ is the error or residual, which is the difference between the actual value (Y) and the value predicted by the model.

The objective of multiple linear regression is to determine the coefficient values β_0,β_1 〖,β〗_2,…,β_p which provides the best-fitting model to explain the variation in the dependent variable (Y) based on the independent variables X_1 〖,X〗_2,…,X_p.

Table 1. Portuguese Students Data

ID	Data Type	Information
school	Nominal	student's sex (binary: female or male)
sex	Nominal	student's age (numeric: from 15 to 22)
age	Integer	student's school (binary: Gabriel Pereira or Mousinho da Silveira)
address	Nominal	student's home address type (binary: urban or rural)
famsize	Nominal	parent's cohabitation status (binary: living together or apart)
Pstatus	Nominal	mother's education (numeric: from 0 to 4a)
Medu	Integer	mother's job (nominal)
Fedu	Integer	father's education (numeric: from 0 to 4a)
Mjob	Nominal	father's job (nominal)
Fjob	Nominal	student's guardian (nominal: mother, father or other)
reason	Nominal	family size (binary: ≤ 3 or > 3)
guardian	Integer	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
traveltime	Integer	reason to choose this school (nominal: close to home, school reputation, course preference or other)
studytime	Integer	home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour).
failures	Integer	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
schoolsup	Nominal	number of past class failures (numeric: n if 1 ≤ n < 3, else 4)
famsup	Nominal	extra educational school support (binary: yes or no)
paid	Nominal	family educational support (binary: yes or no)
activities	Nominal	extra-curricular activities (binary: yes or no)
nursery	Nominal	extra paid classes (binary: yes or no)
higher	Nominal	Internet access at home (binary: yes or no)
internet	Nominal	attended nursery school (binary: yes or no)
romantic	Nominal	wants to take higher education (binary: yes or no)
famrel	Integer	with a romantic relationship (binary: yes or no)

ID	Data Type	Information
freetime	Integer	free time after school (numeric: fr 1 – very low to 5 – very high)
goout	Integer	going out with friends (numeric: fr 1 – very low to 5 – very high)
Dalc	Integer	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Walc	Integer	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	Integer	current health status (numeric: fro – very bad to 5 – very good)
absences	Integer	number of school absences (numer from 0 to 93)
G1	Integer	first period grade (numeric: from 0 20)

G2	Integer	second period grade (numeric: fro to 20)
G3	Integer	final grade (numeric: from 0 to 20

4.2 Normalization Process

The next step involves normalization by selecting a range transformation method. Normalization with Range Transformation is a data preprocessing technique utilized to scale data values within a specific range. The objective is to render the data more suitable for machine learning algorithms that are sensitive to scale, ensuring that all attributes contribute equally to the analysis.

The normalization process with Range Transformation involves the following steps:

1. Determining the Range: Define the desired new range for the data after normalization. Commonly used ranges are from 0 to 1 or from -1 to 1.
2. Calculating Minimum and Maximum: Calculate the minimum and maximum values of each attribute in the dataset.
3. Transformation: Apply the formula to convert each attribute value into the desired range. The general formula is:

$$new_{value} = \frac{old\_value - \min\ \_value}{\max\ \_value\ - \min\ \_value}\ X(new\_\max - new\_\min) + new\_min$$

Where:

- old-value represents the initial attribute value.
- min_value and max_value are the minimum and maximum values of the attribute.
- new_min and new_max indicate the desired new range.

4. Implementation on the Entire Dataset: Apply the transformation formula to each value within the dataset for all attributes.

The outcome of normalization with Range Transformation is a dataset that possesses a uniform range of values within the specified range. This aids machine learning algorithms in obtaining more balanced information from all attributes and avoids bias towards attributes with larger scales. Table 1 underwent changes in several attributes after the normalization process. These attributes are: age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2, and G3. Further details of the results can be observed in Table 2.

Table 2. Normalization Results

ID	Initial data type	Least	Most	Values	Data type after normalization	Least	Most	Values
school	Polynomial	MS (46)	GP (349)	GP (349), MS (46)	Polynomial	MS (46)	GP (349)	GP (349), MS (46)
sex	Polynomial	Least M (187)	Most F (208)	Values F (208), M (187)	Polynomial	Least M (187)	Most F (208)	Values F (208), M (187)
age	Integer	Min 15	Max 22	Average 16.696	Real	Min 0	Max 1	Average 0.242
address	Polynomial	Least R (88)	Most U (307)	Values U (307), R (88)	Polynomial	Least R (88)	Most U (307)	Values U (307), R (88)
famsize	Polynomial	Least LE3 (114)	Most GT3 (281)	Values GT3 (281), LE3 (114)	Polynomial	Least LE3 (114)	Most GT3 (281)	Values GT3 (281), LE3 (114)
Pstatus	Polynomial	Least A (41)	Most T (354)	Values T (354), A (41)	Polynomial	Least A (41)	Most T (354)	Values T (354), A (41)
Medu	Integer	Min 0	Max 4	Average 2.749	Real	Min 0	Max 1	Average 0.687
Fedu	Integer	Min 0	Max 4	Average 2.522	Real	Min 0	Max 1	Average 0.63
Mjob	Polynomial	Least health (34)	Most other (141)	Values other (141), services (103), ...[3 more]	Polynomial	Least health (34)	Most other (141)	Values other (141), services (103), ...[3 more]
Fjob	Polynomial	Least health (18)	Most other (217)	Values other (217), services (111), ...[3 more]	Polynomial	Least health (18)	Most other (217)	Values other (217), services (111), ...[3 more]
reason	Polynomial	Least other (36)	Most course (145)	Values course (145), home (109)	Polynomial	Least other (36)	Most course (145)	Values course (145), home (109)
ID	**Initial data type**	**Least**	**Most**	**Values**	**Data type after normalization**	**Least**	**Most**	**Values**
guardian	Polynomial	Least	Most	Values	Polynomial	Least	Most	Values

		other (32)	mother (273)	mother (273), father (90), ...[1 more]		other (32)	mother (273)	mother (273), father (90), ...[1 more]
traveltime	Integer	Min 1	Max 4	Average 1.448	Real	Min 0	Max 1	Average 0.559
studytime	Integer	Min 1	Max 4	Average 2.035	Real	Min 0	Max 1	Average 0.345
failures	Integer	Min 0	Max 3	Average 0.334	Real	Min 0	Max 1	Average 0.111
schoolsup	Polynomial	Least yes (51)	Most no (344)	Values no (344), yes (51)	Polynomial	Least yes (51)	Most no (344)	Values no (344), yes (51)
famsup	Polynomial	Least no (153)	Most yes (242)	Values yes (242), no (153)	Polynomial	Least no (153)	Most yes (242)	Values yes (242), no (153)
paid	Polynomial	Least yes (181)	Most no (214)	Values no (214), yes (181)	Polynomial	Least yes (181)	Most no (214)	Values no (214), yes (181)
activities	Polynomial	Least no (194)	Most yes (201)	Values yes (201), no (194)	Polynomial	Least no (194)	Most yes (201)	Values yes (201), no (194)
nursery	Polynomial	Least no (81)	Most yes (314)	Values yes (314), no (81)	Polynomial	Least no (81)	Most yes (314)	Values yes (314), no (81)
higher	Polynomial	Least no (20)	Most yes (375)	Values yes (375), no (20)	Polynomial	Least no (20)	Most yes (375)	Values yes (375), no (20)
internet	Polynomial	Least no (66)	Most yes (329)	Values yes (329), no (66)	Polynomial	Least no (66)	Most yes (329)	Values yes (329), no (66)
romantic	Polynomial	Least yes (132)	Most no (263)	Values no (263), yes (132)	Polynomial	Least yes (132)	Most no (263)	Values no (263), yes (132)
famrel	Integer	Min 1	Max 5	Average 3.944	Real	Min 0	Max 1	Average 0.736
freetime	Integer	Min 1	Max 5	Average 3.235	Real	Min 0	Max 1	Average 0.559
goout	Integer	Min 1	Max 5	Average 3.109	Real	Min 0	Max 1	Average 0.527
Dalc	Integer	Min 1	Max 5	Average 1.481	Real	Min 0	Max 1	Average 0.12
Walc	Integer	Min 1	Max 5	Average 2.291	Real	Min 0	Max 1	Average 0.323
health	Integer	Min 1	Max 5	Average 3.554	Real	Min 0	Max 1	Average 0.639
absences	Integer	Min 0	Max 75	Average 5.709	Real	Min 0	Max 1	Average 0.076
G1	Integer	Min 3	Max 19	Average 10.909	Real	Min 0	Max 1	Average 0.494
G2	Integer	Min 0	Max 19	Average 10.714	Real	Min 0	Max 1	Average 0.564
G3	Integer	Min 0	Max 20	Average 10.415	Real	Min 0	Max 1	Average 0.521

4.2 Correlation Matrix

The next stage involves determining the Correlation matrix, which is a table indicating the correlation coefficients between two or more variables. Correlation is a statistical measure that assesses the extent to which two variables are related to each other. Correlation can be positive (meaning when one variable goes up, the other variable tends to go up as well), negative (meaning when one variable goes up, the other variable tends to go down), or zero (indicating no clear correlation relationship).

The correlation matrix is typically displayed in the form of a square table, with variables placed in rows and columns. The main diagonal of this matrix contains the correlation coefficients between each variable and itself, which always has a value of 1 (self-correlation). The rest of the matrix contains the correlation coefficients between pairs of different variables. The correlation matrix is highly valuable in statistical analysis as it can provide insights into the relationships among various variables.

For instance, in data analysis, the correlation matrix can assist in identifying variables with strong relationships, enabling those variables to be further explored or analysed together. It can also aid in selecting variables to be included in regression analysis models or other statistical models. The correlation matrix values range between -1 and 1, with -1 indicating a perfect negative correlation, 1 indicating a perfect positive correlation, and 0 indicating no correlation, as shown at figure 1.

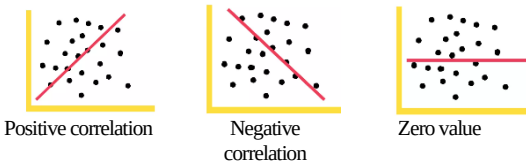

Figure 1. Correlation Type

The closer the value is to 0, the weaker the correlation relationship. Creating a correlation matrix involves plotting a positive curve line. This is done to identify which attributes impact each other and their values will increase as data expands.

Attribute selection, also known as feature selection, is a process in data analysis and machine learning where we choose the most relevant or significant subset of attributes (variables) available in a dataset. The primary goal of attribute selection is to reduce data dimensionality, avoid issues related to "curse of dimensionality" or attributes that don't provide significant information, and enhance model performance and interpretability is present at figure 2.

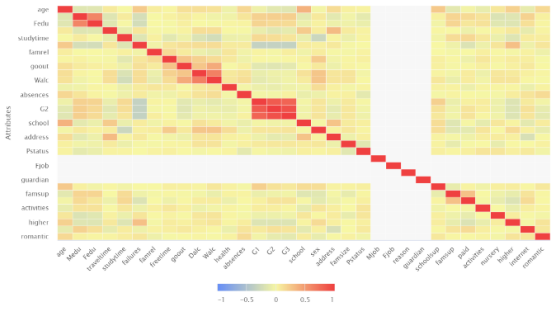

Figure 2. Matrix Correlation

It can be observed that G1, G2, and G3 exhibit the highest values compared to the other attributes. When comparing these three attributes, their values can be seen in Table 3. The correlation value of G2 with G3 is the highest, scoring 0.904, followed by G2 with G1 scoring 0.85, and finally G1 with G3 scoring 0.80. From this, it is evident that the attribute G2 has the most influence among the attributes. However, since we are interested in the G3 value or the end-of-term grade, G2 becomes an independent attribute.

Table 3. Matrix Correlation

	G1	G2	G3
G1	1	0.852118	0.801468
G2	0.852118	1	0.904868
G3	0.801468	0.904868	1

To formulate a multiple linear regression equation, we need to examine each attribute that holds the most weight concerning the G2 attribute. Feature selection techniques can offer a solution, specifically the "Select by Weight" feature. This technique is frequently used in data processing and machine learning to choose the most important or significantly weighted attributes for decision-making. The technique employed here is based on a specific weight selection. This approach involves assigning specific weights to each attribute based on domain knowledge, analysis, or certain assumptions. For instance, if you know in a particular analysis that some attributes are more critical than others, you can assign higher weights to those attributes. In this case, attribute selection for use in multiple linear regression is considered.

To determine the appropriate attributes, the research employs the select by weight method, selecting the attribute with the highest value among others. This technique is commonly used in data processing and machine learning to select the most important or significantly weighted attributes for decision-making [12].

Based on the weight testing results, some attributes have values that are not suitable for linear regression processing. As a result, the two highest-weighted attributes were selected, and their types allow for linear regression processing. The weight test results can be observed in Figure 4. Initially, this study was limited to using 10 attributes. However, based on the weight search results, the top 3 and 4 attribute values have demonstrated polynomial data types. Consequently, it was decided to test these two highest-weighted attributes to assess the root mean squared error (RMSE), squared correlation, and squared error values.

Root Mean Squared Error (RMSE) is a commonly used evaluation metric in regression analysis or forecasting to measure how accurately a prediction model estimates actual values [13]. RMSE quantifies the extent of the difference between predicted values and actual values, providing an average error value in the same units as the target variable.

How to calculate RMSE is as follows:

1. Calculating the Difference in Squares: For each data point, calculate the difference in the squares between the predicted 〖y_hat〗 _i values and true value y_i:

$$squared\ Error_i = (y_{hat_i} - y_i)^2$$

$$RMSE = sqrt\,(\sum(\mathrm{y_{hat}} - \mathrm{y})^2/n$$

2. Calculating Average of Squared Difference: Compute the average of all the squared errors calculated in the previous step:

$$Mean\ squared\ Error = \frac{1}{n}\sum_{i=1}^{n}(y_{hat_i} - y_i)^2$$

3. Calculating RMSE: The root of the Mean Squared Error gives the RMSE value, which is the average of the square roots of the difference between the predicted values and the actual values:

$$RMSE = \sqrt{Mean\ Squared\ Error}$$

The lower the RMSE value, the more accurate the prediction model is in estimating the true value. RMSE also has easier interpretation in the context of data, because the units are the same as the units of the target variable. Figure 3. represent the weight of each attribute.

Squared Correlation, or Squared Correlation Coefficient, refers to the squared correlation coefficient

between two variables. The correlation coefficient is a statistical measure that quantifies the extent to which two variables are linearly related to each other [14]. Squaring the correlation coefficient provides information about the proportion of variation in one variable that can be explained by the variation in another variable.

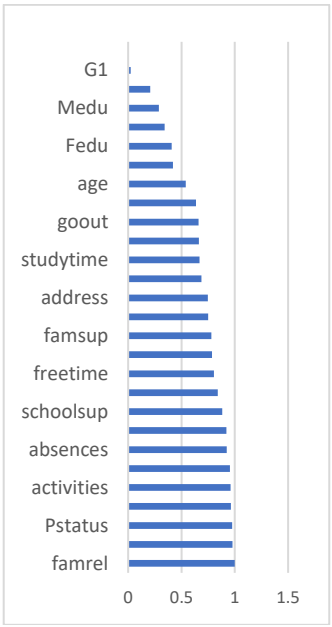

Figure 3. The weight value of each attribute

How to calculate Squared Correlation is as follows:

1. Calculate Correlation Coefficient: First of all, calculate the correlation coefficient between two variables. You can use the Pearson Correlation formula or other suitable methods depending on the type of relationship and the type of data.
2. Square the Correlation Coefficient: After getting the correlation coefficient, square the value:

   Squared Correlation=Pearson Correlation^2

   Whereas Pearson Correlation represents the coefficient value of correlation between two variables.

The result of Squared Correlation is a value between 0 and 1. A value of 0 indicates no linear relationship between the variables, while a value of 1 signifies a perfect linear relationship between the variables (one variable can be fully explained by the other variable). Squared Correlation is often used in regression analysis to indicate how well the variability in the dependent variable is explained by the independent variable. However, it's important to note that Squared Correlation does not provide information about the direction of the relationship or causal impact.

The final test involves determining the squared error value. Squared Error (SE) measures the squared difference between predicted and actual values in regression analysis or forecasting. The computation of Squared Error is as follows:

1. Calculate Difference: For each data point, calculate the difference between the predicted values y_{hat_i} and true value y_i

   $$Error_i = y_{hat_i} - y_i$$

2. Square Difference: Square each difference calculated in the previous step:

   $$Squared\ Error_i = (Error_i)^2$$

3. Compute Average of Squared Difference: Compute the average of all the squared errors calculated in the previous step:

   $$Mean\ Squarred\ Error = \frac{1}{n}\sum_{i=1}^{n} Squared\ Error_i$$

   Here, n is the number of data points in the dataset.

SquarEd Error provides an indication of the extent of difference between predicted and actual values. Nilai squared error yang lebih tinggi menunjukkan bahwa prediksi memiliki kesalahan yang lebih besar dalam A higher squared error value indicates that the prediction has a larger error in estimating the actual value [15]. Mean Squared Error (MSE) represents the average of all squared errors within the dataset [16]. It offers insight into the average prediction error in the model. A lower MSE value reflects better accuracy of the model in predicting actual values. SE and MSE are common evaluation metrics used in regression analysis or forecasting to measure the predictive performance of a model.

The data testing process using RapidMiner was initiated by reading the dataset with the read CSV operator. Data preprocessing involved selecting all available attributes (33) to undergo linear regression processing. The primary objective of this research was to assess the impact of attributes on the final end-of-term grade (G3). The initial test employed simple linear regression, selecting the single highest-weighted attribute with a positive value for further testing. Subsequently, a multiple linear regression test was conducted with attribute selection using weighting. Interestingly, the attribute initially chosen with the highest positive weight did not appear in the weighted selection.

Simple linear regression yielded an RMSE value of 2.96, while multiple linear regression resulted in a value of 5.15. For squared error testing, the simple linear model still produced smaller values compared to multiple linear regression, with respective scores of 8.50 and 26.57. The squared correlation values for simple linear regression indicate differences in RMSE and SE testing, consistently lower than in multiple linear regression. In this test, its value was unexpectedly higher at 0.49 compared to the value of 0.04 in multiple linear regression as shown at figure 4.

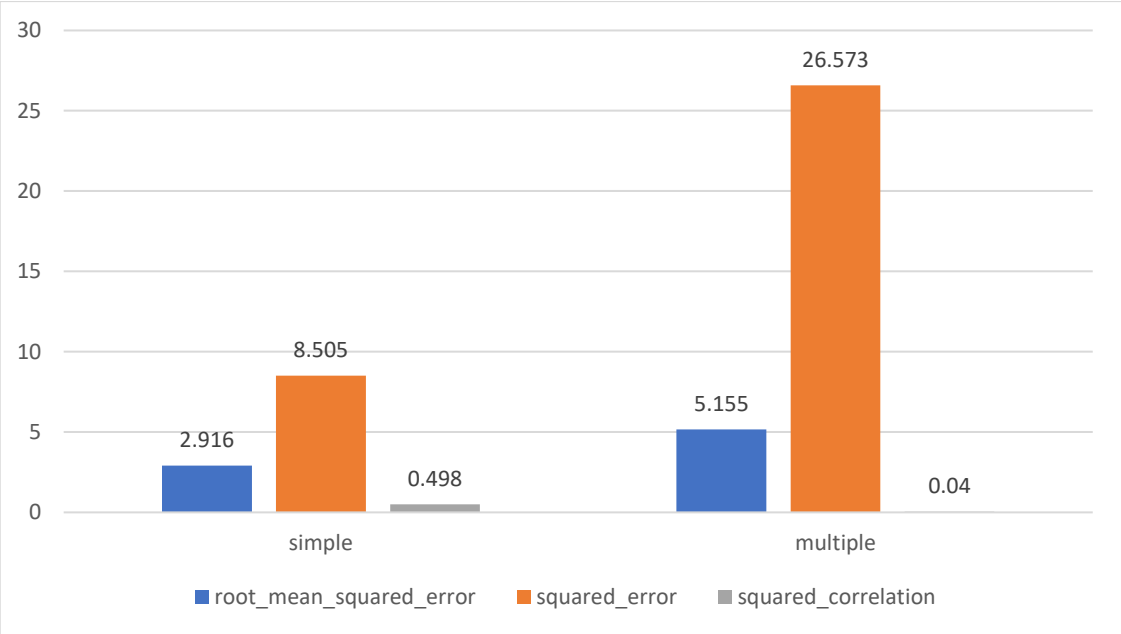

Figure 4. Test Result

The measurement results using simple linear regression can actually still be subjected to further testing. This is due to the fact that in simple linear regression, there are several attributes with positive correlation values other than G2 and G1. The correlation values of these other attributes with G3 are displayed in Table 4.

Table 4. Attributes Correlation Values

Atribut	Dependent	Correlation Value	RMSE	SE	SEC
G3	G2	0.904868	2.916	8.505	0.498
G3	G1	0.801468	2.311	5.341	0.772
G3	Medu	0.217147	1.88	3.54	0.854
G3	Fedu	0.152457	2.503	6.266	0.771
G3	studytime	0.103456	1.413	1.996	0.95
G3	famrel	0.101996	1.663	2.764	0.902
G3	absences	0.098483	2.295	5.265	0.655
G3	freetime	0.09782	0.929	0.863	0.947

The testing outcomes for each of these attributes can be observed in Figure 5.

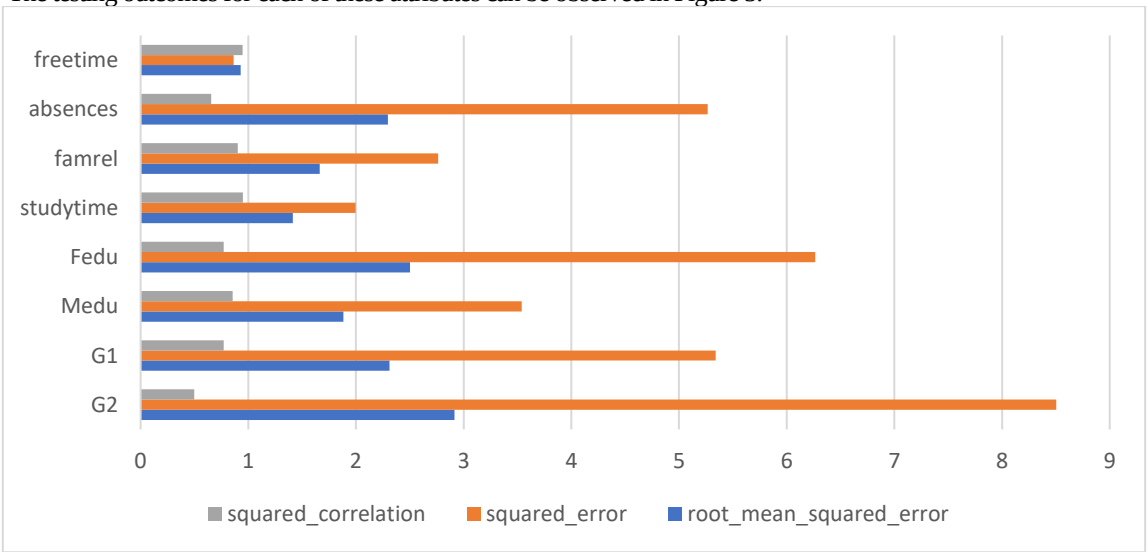

Figure 5. Testing Outcomes.

V. CONCLUSION

Based on the results of the discussion, it can be concluded that the quality of student study results in mathematics subjects is influenced by many things based on the student's background. The highest correlation was obtained by the final exam score (G3) of 0.904, the exam score in class 2 (G2) was 0.85 and the exam score in class 1 (G1) was 0.80. This shows that students tend to be more serious in higher classes. Apart from this, students' mathematics scores are also determined by supporting factors such as activities outside of school and absence from class. The test results on the model obtained RMSE results of 2.96, SE of 8.50 and SC of 0.49.

REFERENCES

[1] Li H, Zhang M, Hou S, Huang B, Xu C, Li Z, et al. Examining the Dynamic Links among Perceived Teacher Support, Mathematics Learning Engagement, and Dimensions of Mathematics Anxiety in Elementary School Students: A Four-wave Longitudinal Study. Contemp Educ Psychol [Internet]. 2023;102211. Available from: https://www.sciencedirect.com/science/article/pii/S0361476X23000656

[2] Jung H, Wickstrom MH. Teachers creating mathematical models to fairly distribute school funding. J Math Behav [Internet]. 2023;70:101041. Available from: https://www.sciencedirect.com/science/article/pii/S0732312323000111

[3] Iyamuremye E, Ndayambaje I, Muwonge CM. Relationships of mathematics achievement with self-determined motivation and mathematics anxiety among senior two students in Northern Rwanda. Heliyon [Internet]. 2023;9(4):e15411. Available from: https://www.sciencedirect.com/science/article/pii/S240584402302618X

[4] Cuder A, Živković M, Doz E, Pellizzoni S, Passolunghi MC. The relationship between math anxiety and math performance: The moderating role of visuospatial working memory. J Exp Child Psychol [Internet]. 2023;233:105688. Available from: https://www.sciencedirect.com/science/article/pii/S0022096523000644

[5] Cerón-García MC, López-Rosales L, Gallardo-Rodríguez JJ, Navarro-López E, Sánchez-Mirón A, García-Camacho F. Jigsaw cooperative learning of multistage counter-current liquid-liquid extraction using Mathcad®. Educ Chem Eng [Internet]. 2022;38:1–13. Available from: https://www.sciencedirect.com/science/article/pii/S1749772821000488

[6] Zheng L, Wang C, Chen X, Song Y, Meng Z, Zhang R. Evolutionary machine learning builds smart education big data platform: Data-driven higher education. Appl Soft Comput [Internet]. 2023;136:110114. Available from: https://www.sciencedirect.com/science/article/pii/S1568494623001321

[7] Malik A, Onyema EM, Dalal S, Lilhore UK, Anand D, Sharma A, et al. Forecasting students' adaptability in online entrepreneurship education using modified ensemble machine learning model. Array [Internet]. 2023;19:100303. Available from: https://www.sciencedirect.com/science/article/pii/S2590005623000280

[8] Gupta S, Kumar P, Tekchandani R. A machine learning-based decision support system for temporal human cognitive state estimation during online education using wearable physiological monitoring devices. Decis Anal J [Internet]. 2023;8:100280. Available from: https://www.sciencedirect.com/science/article/pii/S2772662223001200

[9] Olsen AA, McLaughlin JE, Harpe SE. Using multiple linear regression in pharmacy education scholarship. Curr Pharm Teach Learn [Internet]. 2020;12(10):1258–68. Available from: https://www.sciencedirect.com/science/article/pii/S1877129720302136

[10] Smolen HJ, Yuan M, Hawthorne ME, Wang Q, Kelton K. PIH10 - Development, Validation, And Analysis Of A Linear Regression Model Predicting Child's Birthweight From Mother's Race, Education Level, Smoking Status, And Gestation Age. Value Heal [Internet]. 2015;18(3):A105. Available from: https://www.sciencedirect.com/science/article/pii/S1098301515006737

[11] Cortez P, Silva AMG. Using data mining to predict secondary school student performance. 2008;

[12] Oshan TM, Li Z, Kang W, Wolf LJ, Fotheringham AS. mgwr: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. ISPRS Int J Geo-Information. 2019;8(6):269.

[13] Karunasingha DSK. Root mean square error or mean absolute error? Use their ratio as well. Inf Sci (Ny). 2022;585:609–29.

[14] Jumayev S, Khudayberganov S, Achilov O, Allamuratova M. Assessment criteria for optimization of parameters affecting to local wagon-flows at railway sites. In: E3S Web of Conferences. EDP Sciences; 2021. p. 5022.

[15] Ćalasan M, Aleem SHEA, Zobaa AF. On the root mean square error (RMSE) calculation for parameter estimation of photovoltaic models: A novel exact analytical solution based on Lambert W function. Energy Convers Manag. 2020;210:112716.

[16] Nainggolan R, Perangin-angin R, Simarmata E, Tarigan AF. Improved the performance of the

K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method. In: Journal of Physics: Conference Series. IOP Publishing; 2019. p. 12015.

[17] Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*. IRWIN, The McGraw-Hill Companies, Inc., 1996.