



## Research article

# Predicting Student Performance using Linear Regression

Max Kennedy Kassy

Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Riau, Indonesia

email: [max.kennedy@student.pelitaindonesia.ac.id](mailto:max.kennedy@student.pelitaindonesia.ac.id)

## ARTICLE INFO

### Article history:

Received December 14, 2025

Revised June 28, 2025

Accepted July 30, 2025

Available online August 07, 2025

### Kata kunci:

student  
performance  
linear  
regression  
dataset

### Please cite this article in IEEE style as:

M. Kennedy Kassy, "Predicting Student Performance using Linear Regression", Data Science Insights, vol. 3, no. 2, pp. 66–74, Aug. 2025.

## ABSTRACT

This study explores how to measure and predict student performance using various machine learning algorithms to determine the model that produces the best predictions. The collected data is obtained from the Kaggle data science and machine learning community website, obtaining a dataset with 6 attributes, namely: (1) Hours Studied, (2) Previous Scores, (3) Extracurricular Activities, (4) Sleep Hours, (5) Sample Question Papers Practiced, and (6) Performance Index. The data was cleaned and explored using Microsoft Excel, Google Colab and Tableau. Model development using RapidMiner and Google Colab. The algorithms used for the study were: k-NN, SVM, Linear Regression, Generalized Linear Model, Deep Learning. The Root Mean Squared Error (RMSE) results obtained by the algorithm were 2,455 (k-NN), 2,072 (SVM), 2,013 (Linear Regression), 2,030 (Generalized Linear Model), 2,364 (Deep Learning). From the RMSE it can be seen that the algorithm that gets the best results is Linear Regression, after being retested, Linear Regression gets an RMSE of 2.015, and Root Squared (R2) of 0.989, meaning the Linear Regression algorithm has an accuracy of 98.9%.

### Correspondence:

Max Kennedy Kassy  
Institut Bisnis dan Teknologi Pelita Indonesia,  
Pekanbaru, Riau, Indonesia

Data Science Insights is an open access under the with [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## 1. Introduction

Learning or education is the process of gaining new understanding, knowledge, behavior, abilities, values, and attitudes. With good education, one can develop one's character to be the best. In education there are also facilities, namely schools which are places where students study. In the school environment, a system is needed that can detect student performance so that in the learning process the school can find out the size of student performance in their academic pursuits. One element for accreditation assessment is the punctuality of student graduation. The presence of inactive students will certainly affect the punctuality of graduation. Student performance prediction is needed to prevent inactive students [1]. Performance assessments in general have the aim of providing input to students in an effort to improve and increase the quality of student learning, as well as the productivity of an organization.

To predict student performance, it is necessary to build a machine learning model. Machine learning is an application of artificial intelligence (AI) that provides automatic performance systems and learns to improve themselves from experience without being explicitly programmed. Machine learning focuses on developing computer programs that can access data and use it to learn on their own. The application of Machine Learning methods in recent years has grown everywhere in everyday life. The term machine learning is basically the process of computers learning from data. Without data, computers cannot learn anything. Therefore, if we want to learn machine learning, we will definitely continue to interact with data. All machine learning knowledge will definitely involve data. The data may be the same, but the algorithms and approaches are different to get optimal results [2].

Before this research, there had been various other studies with the same or similar topics using the same or different methods, one of which was the research by Esty Purwaningsih and Ela Nurelasari which used the K-Nearest Neighbor method for Classifying Student Graduation Levels, which obtained an accuracy of 96.49% [3]. In addition, there are also those who use the same algorithm to predict student performance, namely Akuma, S. and Abakpa, H who obtained a Mean Absolute Error of 0.1506, a Root Mean Squared Error of 0.2199, and an accuracy of 87.84% [4]. In an article written by Yazan A. Alsariera, Yahia Baashar, Gamal Alkaws, Abdulsalam

Mustafa, Ammar Ahmed Alkahtani, and Nor'ashikin Ali, there is a discussion about the results/accuracy obtained by various other studies on student performance prediction using different algorithms and with the use of different data attributes for the study. The algorithms with the highest and lowest accuracy included are as follows: (1) Artificial Neural Network: 98.3%(Highest), 64.4%(Lowest), (2) Support Vector Machine: 91.3%(Highest), 66%(Lowest), (3) Decision tree: 98.2%(Highest), 66%(Lowest), (4) K-nearest neighbor: 95.8%(Highest), 69%(Lowest), (5) Naive Bayes: 96.9%(Highest), 65.1%(Lowest), and (6) Linear Regression: 76.2%(Highest), 50%(Lowest) [5].

In another study conducted by Stephen Opoku Oppong, it was stated that neural networks are the most frequently used classifiers to predict student performance. Also, 87% of the algorithms used were Supervised Learning, compared to 13% which were Unsupervised Learning [6]. In addition, there is also a study conducted by Hussein Altabrawee, Osama Abdul Jaleel Ali, Samir Qaisar Ajmi regarding student performance prediction using the Artificial Neural Network (ANN), Naive Bayes, Decision Tree, and Logistic Regression machine learning algorithms, where the ANN model obtained the best accuracy results, namely 77.04% compared to other algorithms [7]. Likewise, there is another study by Munise Seçkin Kapucu, İbrahim Özcan, Hülya Özcan, Ahmet Aypay regarding the prediction of elementary and junior high school students' grades in science subjects using the deep learning machine learning method. In this study, researchers found that the average number of books read per year had the most significant impact on students' academic performance in science. Furthermore, deep learning obtained an accuracy value of 90% in predicting student performance in science [8]. Then there is also a study conducted by Bashiru Aliyu Sani, Samaila Baoku I.G, Bashir Jamilu Ahmed and Samaila Musa, with the same topic. The machine learning algorithms used include: Support Vector Machine, Linear Regression, and Stochastic Gradient descent algorithms. These models have been compared using the classification accuracy of Mean Absolute Error, Mean Square Error, and Root Mean Square Error. The data set used to build the model was collected based on a survey given to students and student grade books. The support vector machine model achieved the best performance of 99.1% [9].

The purpose of this study is to analyze the influence of factors that affect student performance. With that, this journal can provide the most suitable algorithm to predict student performance based on the factors mentioned earlier. In this study, the data used is the student performance dataset taken from kaggle which includes data: (1) Hours Studied, (2) Previous Scores, (3) Extracurricular Activities, (4) Sleep Hours, (5) Sample Question Papers Practiced, and (6) Performance Index with Performance Index acting as a label or target model to be built. These data represent all aspects that affect student performance, both large and small, such as time spent studying to time spent sleeping or resting, and extracurricular activities. Although the influence is small, it must still be included in order to achieve the most accurate and satisfying results.

## 2. Research Methods

In this study, the data science process (Figure 1) will be used as a research reference.

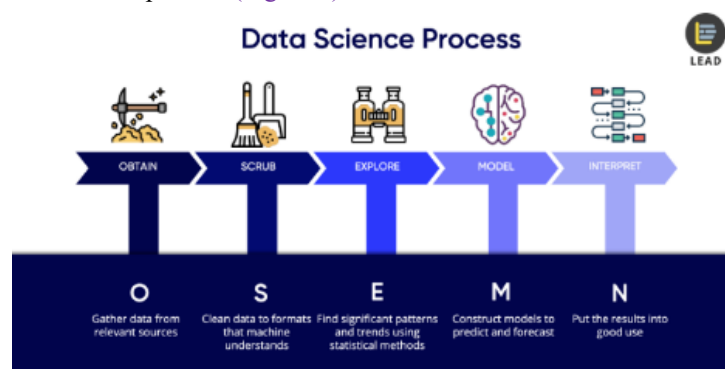


Figure 1. Data Science Process

### 2.1 Obtain Data

The data taken for this study is the Student Performance dataset obtained from the Kaggle website in the form of a .csv file. This dataset is data created to observe factors that affect student academic performance. In this dataset file there are 10,000 data with the following attributes: (1) Hours Studied: shows the amount of time spent studying by students in hours, with a numeric data type. (2) Previous Scores: shows previous grades obtained by students, with a numeric data type. (3) Extracurricular Activities: shows the student's extracurricular status, with a Boolean data type. (4) Sleep Hours: shows the student's sleep time, with a numeric data type. (5) Sample Question Papers Practiced: shows the number of practice papers worked on by students, with a numeric data type. (6) Performance Index: shows student performance, with a numeric data type.

The data taken from this dataset is synthetic or artificial, that is, it is not obtained from the original source, but aims to illustrate the relationship between variables and the Performance Index or student performance. With this research, it is hoped that with a complex system in predicting student

performance, it can maintain student abilities and achievements, ensure that students can graduate on time, and ensure that student abilities are appropriate and match the field taken. In this study, the problem raised is the prediction of student performance. The data used is taken from Kaggle, with a total of 10,000 data. This study will compare the accuracy between several different algorithms using RapidMiner and Google Colab. Google Colab or Google Colaboratory which helps users support all data science and machine learning needs [10].

## 2.2 Scrub Data

In this stage, it is required to look for and resolves abnormalities in the data, such as incorrect data, duplicate data, missing data, and so on. This can be done in Microsoft Excel. This stage is important for the accuracy of the results, because duplicate, incorrect or missing data can affect the results that are not in accordance with the actual ones.

## 2.3 Explore Data

After the cleaning process, using the software Tableau to visualize the data patterns of features. The visualization helps readers understand the data and the relationship between features and target classes more easily. At this step, it is still important to do deeper research and not make decisions based on the visualization alone, because Tableau only provides visualizations according to the data at that time.

## 2.4 Model Data

In this process, a statistical model or machine learning development will be carried out that is able to predict the results. Several algorithms will be using for testing to try the best model to predict student performance. This is done using Google Colab and RapidMiner to get accuracy and deviation values.

## 2.5 Interpret

The successfully built model will be applied in various systems or businesses that require it, such as educational institutions, namely schools. The application of this model is carried out responsibly and ethically, and ensures that the model is not biased, discriminatory, or harmful to users.

# 3. Results and Discussion

## 3.1 Data Collection Results

Here is a snippet of the dataset used for this study. This dataset was obtained from the Kaggle website. (<https://www.kaggle.com/>).

	A	B	C	D	E	F
1	Hours Studied	Previous Scores	Extracurricular Activities	Sleep Hours	Sample Question Papers Practiced	Performance Index
2	7	99	Yes	9	1	91
3	4	82	No	4	2	65
4	8	51	Yes	7	2	45
5	5	52	Yes	5	2	36
6	7	75	No	8	5	66
7	3	78	No	9	6	61
8	7	73	Yes	5	6	63
9	8	45	Yes	4	6	42
10	5	77	No	8	2	61
11	4	89	No	4	0	69
12	8	91	No	4	5	84
13	8	79	No	6	2	73
14	3	47	No	9	2	27
15	6	47	No	4	2	33
16	5	79	No	7	8	68

Figure 2. Dataset

From the dataset above (Figure 2) there are 10,000 data with a Performance Index in the range 0-100 as the target of the model to be built.

## 3.2. Data Cleaning Results

In the data cleaning process, first insert the "Student Performance.csv" data set into a Google Colab file, then include the necessary libraries and run this code (Figure 3) to display the top 10 rows of the data set.

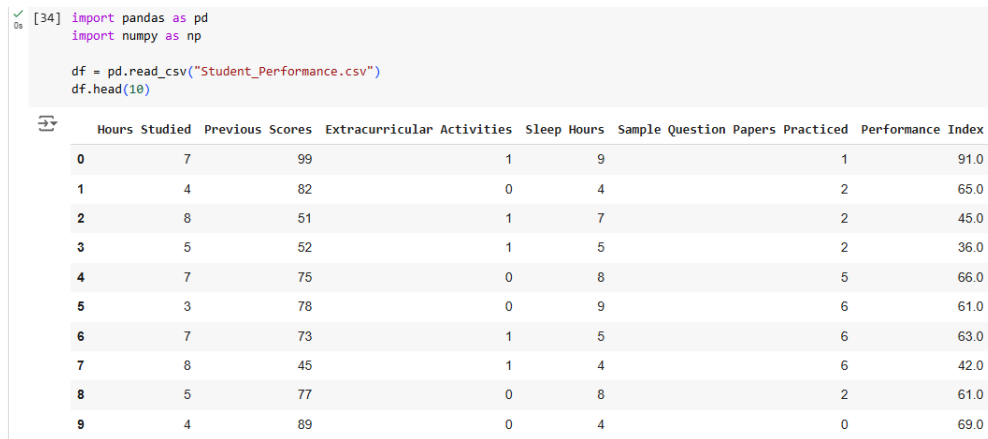


Figure 3. Read Dataset on Google Colab

Then, insert the `df.info()` function into the code line to find out each data type for each variable and `df.shape` to find out each number of rows and columns of the dataset. Then insert the `df.describe()` function into the code line to find out the average value and other values.

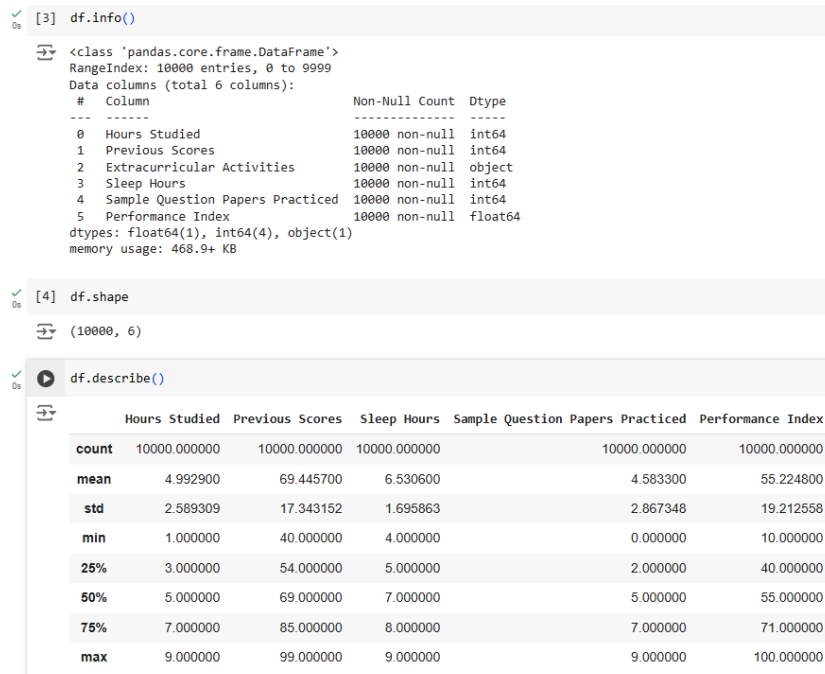


Figure 4. Describe Dataset on Google Colab

From the results obtained from Google Colab (Figure 4), it can be seen that when using the `df.describe()` function, the Extracurricular Activities data does not appear. This is because Extracurricular Activities is data containing the strings "Yes" and "No". In general, this is not a problem, but with the label/target, namely Performance Index being a numeric data, a regression model will be used for model development, while some regression models face problems when data is entered that does not fall into the numeric category.

To solve this problem, it is required to change Extracurricular Activities into numeric data. This can be done using Microsoft Excel software, where the "Find and Replace" feature is used (Figure 5) to find and change all cells that matches the search query. To suit the needs of this study, "Yes" will be replaced by 1, and "No" will be replaced by 0, to represent the boolean values of True and False. (Figure 6).

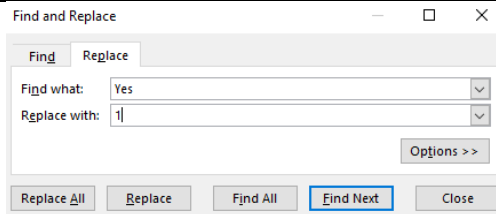


Figure 5. Find and Replace in Microsoft Excel

df.describe()

	Hours Studied	Previous Scores	Extracurricular Activities	Sleep Hours	Sample Question Papers Practiced	Performance Index
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	4.992900	69.445700	0.494800	6.530600	4.583300	55.224800
std	2.589309	17.343152	0.499998	1.695863	2.867348	19.212558
min	1.000000	40.000000	0.000000	4.000000	0.000000	10.000000
25%	3.000000	54.000000	0.000000	5.000000	2.000000	40.000000
50%	5.000000	69.000000	0.000000	7.000000	5.000000	55.000000
75%	7.000000	85.000000	1.000000	8.000000	7.000000	71.000000
max	9.000000	99.000000	1.000000	9.000000	9.000000	100.000000

Figure 6. After changing Extracurricular Activities data

After changing the Extracurricular Activities value, then, standardize the values of all variables. This serves to ensure that all values in the dataframe variables are on the same scale, and also reduces the impact of outliers, thereby improving the results of the experiment and making them more reliable.

This process will be done using Google Colab, using StandardScaler from the sklearn.preprocessing library. Before doing Standard Scaling, first separate the Performance Index attribute from the others to ensure that the Performance Index is not changed. Then, after that use the StandardScaler() function as follows (Figure 7).

```
[9] from sklearn.preprocessing import StandardScaler

independent = ['Hours Studied', 'Previous Scores', 'Sleep Hours', 'Sample Question Papers Practiced', 'Extracurricular Activities']
x = df[independent]
y = df['Performance Index']

scaler = StandardScaler()
scaler.fit(x)
scaledData = scaler.transform(x)
scaledData = pd.DataFrame(scaledData, columns=x.columns)
scaledData.head()
```

	Hours Studied	Previous Scores	Sleep Hours	Sample Question Papers Practiced	Extracurricular Activities
0	0.775188	1.704176	1.456205	-1.249754	1.010455
1	-0.383481	0.723913	-1.492294	-0.900982	-0.989654
2	1.161410	-1.063626	0.276805	-0.900982	1.010455
3	0.002742	-1.005963	-0.902594	-0.900982	1.010455
4	0.775188	0.320275	0.866505	0.145333	-0.989654

Figure 7. Standard Scaling in Google Colab

### 3.3. Data Exploring Results

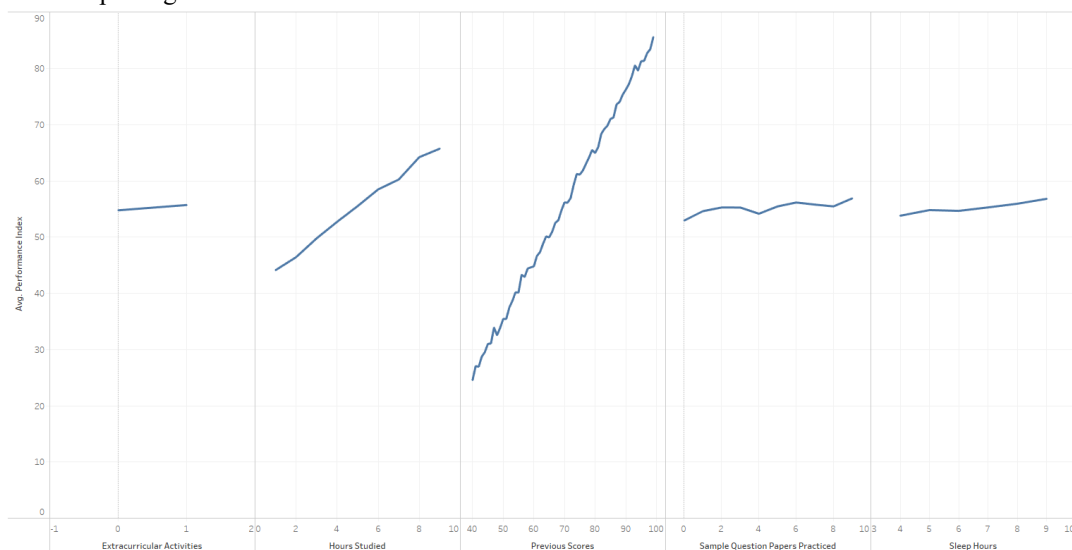


Figure 8. Student Performance Data Visualization

From the results shown (Figure 8), 5 different types of graphs can be seen, with each graph representing each factor used to measure the student's performance index, in order from left to right, namely: (1) Extracurricular Activities, (2) Hours Studied, (3) Previous Scores, (4) Sample Question

Papers Practiced, and the last (5) Sleep Hours. From the graph above, it can be seen based on its gradient, that the factor with the smallest influence is Extracurricular Activities, and conversely the one with the largest influence is Previous Scores.

### 3.4. Data Model Results

#### 1) K-Nearest Neighbor (k-NN) Algorithm

The basic concept of K-NN is to find the closest distance between the data to be evaluated with its k nearest neighbors. The value of the distance between the test data and the training data is sorted from the lowest value. The sorting process is carried out to select a minimum distance of K pieces [11]. Using the K-NN algorithm in Rapid Miner (Figure 9), the data is divided by a ratio of 7 (Training): 3 (Testing), and the number of k will be determined automatically. The root mean squared error (RMSE) obtained is 2,455 (Figure 10).

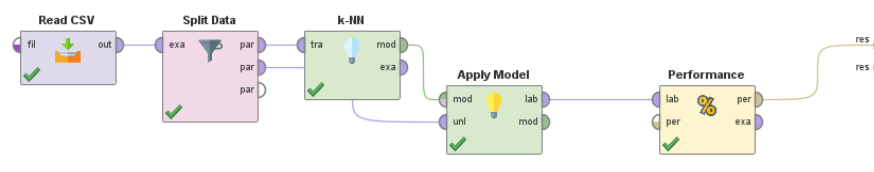


Figure 9. k-NN Algorithm Design in RapidMiner

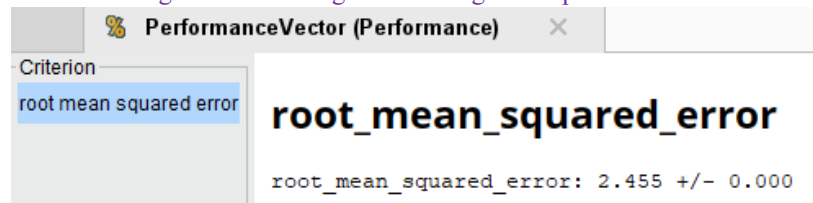


Figure 10. RMSE Deviation of the k-NN Algorithm

#### 2) Support Vector Machine (SVM) Algorithm

Support Vector Machine is a learning system whose classification uses a hypothesis space in the form of linear functions in a high-dimensional feature space, trained with a learning algorithm based on optimization theory by implementing learning derived from statistical learning theory [12]. As before, the data will be separated in a ratio of 7:3 for training and testing the model (Figure 11). The results obtained from this model are an RMSE deviation of 2,072 (Figure 12).

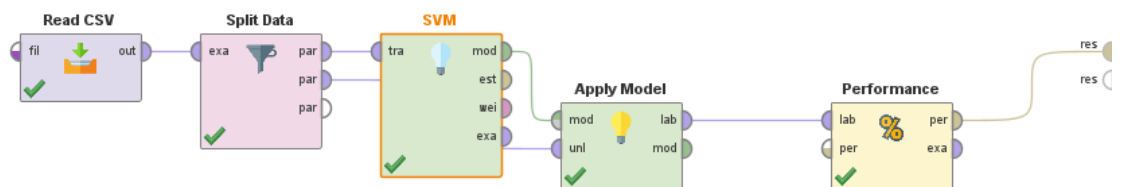


Figure 11. Support Vector Machine Algorithm Design in RapidMiner

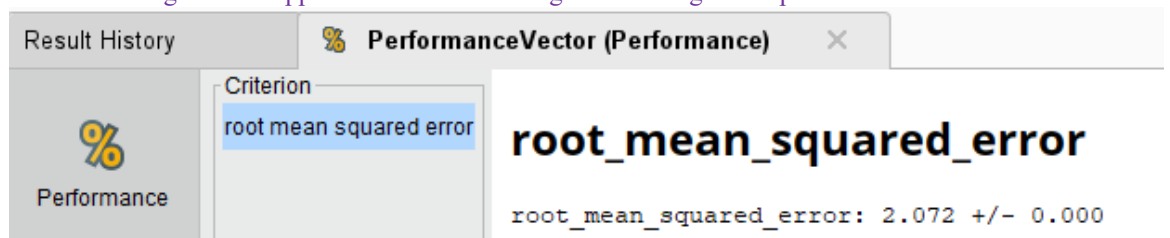


Figure 12. RMSE Deviation of the Support Vector Machine Algorithm

#### 3) Linear Regression Algorithm

Linear regression (Figure 13) is one of the widely applied statistical techniques to investigate the linear relationship between a single dependent variable and one or more independent variables [13]. Regression measures how much a variable can affect another variable so that the value of a variable can be predicted based on another variable [14]. With this algorithm, RMSE result of 2.013 is obtained (Figure 14).

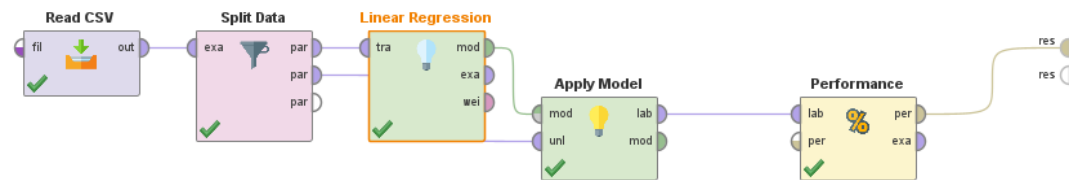
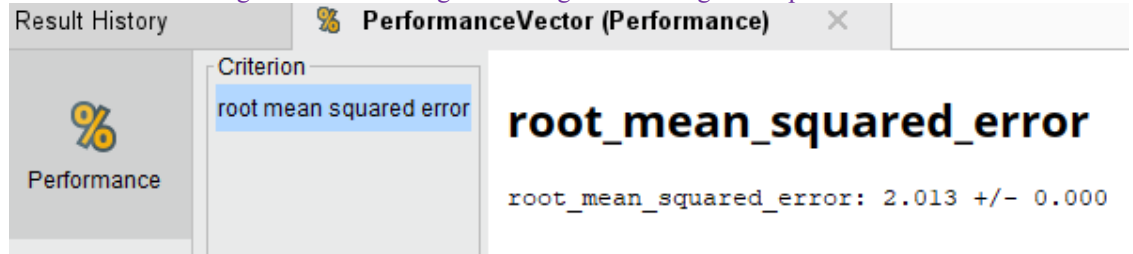


Figure 13. Linear Regression Algorithm Design in RapidMiner



Gambar 14. RMSE Deviation of the Linear Regression Algorithm

#### 4) Generalized Linear Model Algorithm

Generalized Linear Model (Figure 15) is an extension of the linear regression model, and is useful for modeling response variables that do not follow a normal distribution. More specifically, the Generalized Linear Model is useful for modeling data that is a proportion of a total, for modeling counts, and for modeling skewed continuous responses [15]. The Generalized Linear Model algorithm produces an RMSE of 2,030 (Figure 16).

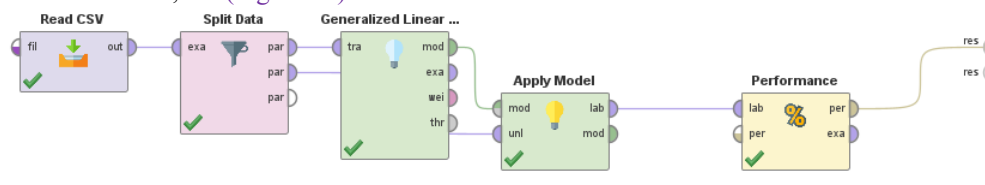


Figure 15. Generalized Linear Model Algorithm Design in RapidMiner

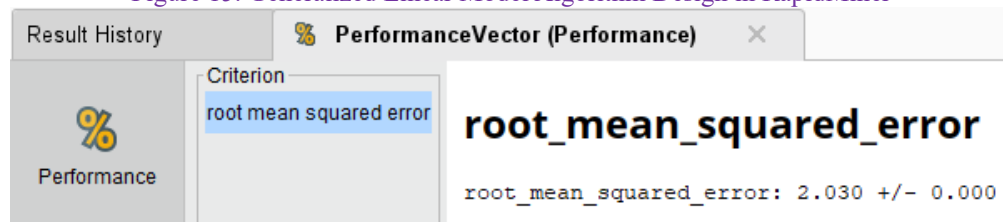


Figure 16. RMSE Deviation of the Generalized Linear Model Algorithm

#### 5) Deep Learning

Deep Learning (Figure 17) is connected with the use of Neural Networks to improve things like speech recognition, computer vision, and natural language processing. In information technology, a Neural Network is a system of programs and data structures that approximates the operation of the human brain. Neural Networks typically involve a large number of processors operating in parallel, each with its own small knowledge base and access to data in its local memory [16]. This Deep Learning model obtained an RMSE of 2,364 (Figure 18).

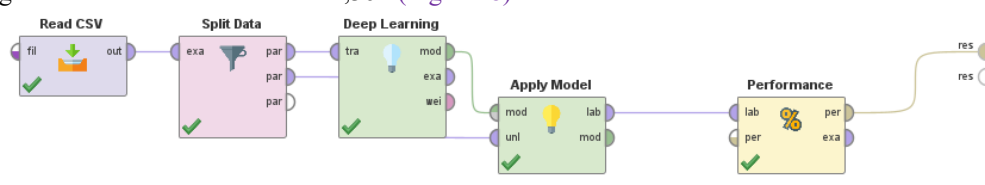


Figure 17. Deep Learning Algorithm Design in RapidMiner

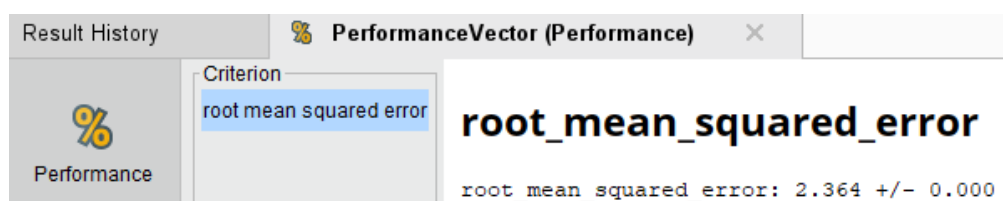


Figure 18. RMSE Deviation of the Deep Learning Algorithm

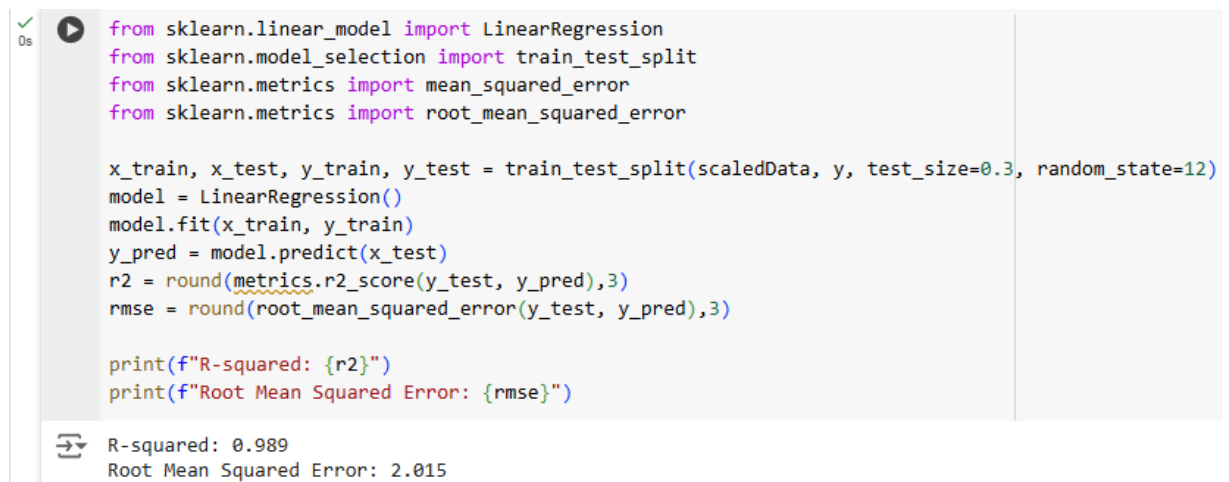


6) Table 1 shows the Summary of every Algorithm's Root Mean Squared Error Of the five algorithms tested using RapidMiner, the one with the smallest RMSE or deviation is chosen to be used as the main algorithm. The main algorithm will then be retested in Google Colab.

Table 1. Algorithm RMSE

Method	RMSE
Linear Regression	2.013 +/- 0.000
Deep Learning	2.364 +/- 0.000
k-NN	2.455 +/- 0.000
Support Vector Machine	2.072 +/- 0.000
Generalized Linear Model	2.030 +/- 0.000

#### 7) Linear Regression Algorithm Results using Google Colab



```

from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn.metrics import root_mean_squared_error

x_train, x_test, y_train, y_test = train_test_split(scaledData, y, test_size=0.3, random_state=12)
model = LinearRegression()
model.fit(x_train, y_train)
y_pred = model.predict(x_test)
r2 = round(metrics.r2_score(y_test, y_pred),3)
rmse = round(root_mean_squared_error(y_test, y_pred),3)

print(f"R-squared: {r2}")
print(f"Root Mean Squared Error: {rmse}")

```

R-squared: 0.989  
 Root Mean Squared Error: 2.015

Gambar 19. Linear Regression Testing on Google Colab

Based on the test (Gambar 18), the data is divided into a ratio of 7:3 as in RapidMiner, and the Root Mean Squared Error obtained is 2,015 and the Root Squared (R-Squared or R2) obtained is 98.9. This can be interpreted as the accuracy of this model is 98.9%.

#### 4. Conclusion

Learning or education is the process of acquiring new understanding, knowledge, behavior, abilities, values, and attitudes. In a school environment, a system is needed that can detect student performance so that in the learning process the school can find out the size of student performance in their academic pursuits. The Student Performance dataset contains 10,000 data with the following attributes: (1) Hours Studied (2) Previous Scores (3) Extracurricular Activities (4) Sleep Hours (5) Sample Question Papers Practiced (6) Performance Index. The analysis begins with data cleaning, and it was found that the dataset had no abnormal values. The next step is to train the model using RapidMiner with the following algorithms: k-NN, SVM, Linear Regression, Generalized Linear Model, Deep Learning. The RMSE results obtained by each algorithm are 2,455 (k-NN), 2,072 (SVM), 2,013 (Linear Regression), 2,030 (Generalized Linear Model), 2,364 (Deep Learning).

#### References

- [1] S. Wiyono and T. Abidin, "Implementation Of K-Nearest Neighbour (KNN) Algorithm To Predict Student's Performance," *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 9, no. 2, 2018, doi: 10.24176/simet.v9i2.2424.
- [2] Admin AWS, "Apa itu Machine Learning?," Amazon Web Services.
- [3] E. Purwaningsih and E. Nurelasari, "Penerapan K-Nearest Neighbor Untuk Klasifikasi Tingkat Kelulusan Pada Siswa," *Syntax: Jurnal Informatika*, vol. 10, no. 01, 2021, doi: 10.35706/syji.v10i01.5173.
- [4] S. Akuma and H. Abakpa, "Predicting Undergraduate Level Students' Performance Using Regression," *Nigerian Annals Of Pure And Applied Sciences*, vol. 4, no. 1, 2021, doi: 10.46912/napas.224.



- 
- [5] Y. A. Alsariera, Y. Baashar, G. Alkaws, A. Mustafa, A. A. Alkahtani, and N. Ali, "Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance," 2022. doi: 10.1155/2022/4151487.
  - [6] S. O. Oppong, "Predicting Students' Performance Using Machine Learning Algorithms: A Review," *Asian Journal of Research in Computer Science*, vol. 16, no. 3, 2023, doi: 10.9734/ajrcos/2023/v16i3351.
  - [7] H. Altabrawee, O. A. J. Ali, and S. Q. Ajmi, "Predicting Students' Performance Using Machine Learning Techniques," *Journal Of University Of Babylon For Pure And Applied Sciences*, vol. 27, no. 1, 2019, doi: 10.29196/jubpas.v27i1.2108.
  - [8] M. Seckin Kapucu, I. Ozcan, H. Ozcan, and A. Aypay, "Predicting Secondary School Students' Academic Performance in Science Course by Machine Learning," *International Journal of Technology in Education and Science*, vol. 8, no. 1, 2024, doi: 10.46328/ijtes.518.
  - [9] B. Aliyu Sani, S. Baoku I.G, B. Jamilu Ahmed, and S. Musa, "Comparative Between Three Machine Learning Algorithms to Predict and Improve Students' Academic Performance," *International Journal of Science for Global Sustainability*, vol. 8, no. 4, 2023, doi: 10.57233/ijsgs.v8i4.365.
  - [10] Kompas.com, "Mengenal Google Colab, Fungsi dan Manfaatnya," 2023.
  - [11] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Computer Engineering, Science and System Journal*, vol. 4, no. 1, 2019, doi: 10.24114/cess.v4i1.11458.
  - [12] A. miftahul I. Habiba, A. Prasetiadi, and C. Ramdani, "Analisis Kesehatan Terumbu Karang Berdasarkan Karakteristik Sungai, Laut, Dan Populasi Area Pemukiman Menggunakan Machine Learning," *IJIS - Indonesian Journal On Information System*, vol. 5, no. 2, 2020, doi: 10.36549/ijis.v5i2.119.
  - [13] S. Huang, "Linear regression analysis," in *International Encyclopedia of Education: Fourth Edition*, Elsevier, 2022, pp. 548–557. doi: 10.1016/B978-0-12-818630-5.10067-3.
  - [14] H. Hasanah, A. Farida, and P. P. Yoga, "Implementation of Simple Linear Regression for Predicting of Students' Academic Performance in Mathematics," *Jurnal Pendidikan Matematika (Kudus)*, vol. 5, no. 1, 2022, doi: 10.21043/jpmk.v5i1.14430.
  - [15] P. K. Dunn, "Generalized linear models," in *International Encyclopedia of Education: Fourth Edition*, 2022. doi: 10.1016/B978-0-12-818630-5.10077-6.
  - [16] A. Banafa, "What is Deep Learning?," in *Quantum Computing and Other Transformative Technologies*, 2023. doi: 10.1201/9781003339175-12.
-