Research Article

MACHINE LEARNING BASED STUDENT ACHIEVEMENT PERFORMANCE PREDICTION WEB APPLICATION

Authors: Osman CEYLAN Onur SEVLİ

To cite to this article: Ceylan, O., and Sevli, O., (2024). Machine Learning Based Student Achievement Performance Prediction Web Application, International Journal of Engineering and Innovative Research, 6(2), p 126-134.

DOI: 10.47933/ijeir.1504555

To link to this artice: https://dergipark.org.tr/tr/pub/ijeir/archive

International Journal of Engineering and Innovative Research

http://dergipark.gov.tr/ijeir

MACHINE LEARNING BASED STUDENT ACHIEVEMENT PERFORMANCE PREDICTION WEB APPLICATION

Osman CEYLAN[1*] Onur SEVLİ[2]

[1]Burdur Mehmet Akif Ersoy University, Institute of Science and Technology, Department of Computer Engineering, Burdur, Türkiye.
[2]Burdur Mehmet Akif Ersoy University, Faculty of Engineering and Architecture, Department of Computer Engineering, Burdur, Türkiye.

[*]Corresponding Author: osmanceylan@isparta.edu.tr

(**Received:** 25.06.2024**; Accepted:**10.07.2024)

https://doi.org/10.47933/ijeir.1504555

ABSTRACT: The use of multiple linear regression in our study is of critical importance in terms of determining the factors that have a significant impact on students' course performance success. Machine learning studies that use multiple linear regression models to predict the performance index on student achievement aim to improve educational processes and increase individual student success. Studies in the literature investigate the factors affecting academic success by examining different variables that affect student success performance. The results of these studies showed that a high level of accuracy was achieved and that it could reliably predict student performance. In our research, we built and trained a multiple linear regression model. The data set was divided into training and test sets and the success of the model was evaluated. To calculate the performance of the created model, it was compared with studies in the literature using performance measurement metrics such as Mean Absolute Error (MAE), Mean Square Error (MSE), R-Square (R^2), Root Mean Square Error (RMSE) and Accuracy (ACC). The results showed that the model performed well and could make accurate predictions. Especially the fact that R^2 is 0.99 and ACC value is 0.994 shows that the model is successful in predicting the data accurately. Moreover, in our study, using the Flask web module, a web interface was created that allows predicting another student's performance because of entering new variables.

Keywords: Machine Learning, Multiple Linear Regression, Student Performance Prediction.

1. INTRODUCTION

Machine learning has revolutionized numerous industries by creating analysis and personalized recommendations based on predictive ability based on data set. This achievement also plays an important role in predicting student achievement performance. Researchers have used algorithms such as linear regression to predict students' success in their courses and future semester grades [1]. Linear regression, a supervised learning algorithm, models the relationship between dependent and independent variables to make predictions. Machine learning models are being developed to accurately predict semester grades and final exam results using real-time data sets collected from educational institutions [2]. Nowadays, machine learning models are widely used for training and analysis in various such fields and provide reliable results [3].

Many studies in the literature have shown successful results in predicting student achievement performance using multiple linear regression models. For example, Alharbi et al. (2019) used

linear regression models to predict student performance and achieved an accuracy rate of 92%. Sravani et al. (2020) explained that using educational data mining techniques enabled students to predict student performance with a 90% accuracy rate. Similarly, Asif et al. (2017) concluded that linear regression provided an accuracy rate of 89% in their study where they analyzed the variables affecting student performance. Additionally, Arsad et al. (2014) showed that the accuracy rate of predicting student performance varies between 85% and 95%. These studies conclude that predictions of student performance can easily be carried out successfully with machine learning models.

The aim of our study is to predict student success using the multiple linear regression model and to ensure the accuracy and reliability of the predicted results. In this way, it is aimed to facilitate the measurability of students' course and school success in future semesters and to contribute to their development. In our study, the python module known as Flask was used and a web interface was created. Thanks to this interface, the success performance of new students with different characteristics will be predicted by using the trained linear regression model. By using this web application, real-time forecasting is also aimed by simplifying data entry and presenting instant forecasting results. The aim is to ensure that education processes are guided as a data-based and predictable process in the ever-developing age of technology.

2. LINEAR REGRESSION ANALYSIS

Predicting an unknown output value using the values of input variables is the basic task of the prediction process. Methods used in the field of linear regression, which is a prediction process, can be divided into two types. The method that aims to establish a connection between the dependent variable and a single independent variable is called simple linear regression. The other method, which aims to reveal the relationship between the dependent variable and two or more independent variables, is called multiple linear regression [4]. The purpose of multiple linear regression is to create a linear equation that accurately represents the relationship between two or more independent variables and a dependent variable in a set of data [1]. The supervised learning algorithm method, known as linear regression, includes both univariate and multivariate algorithms [5].

It is often used in the method of univariate linear regression to create a model that represents the connection between the dependent variable (y) and the single independent variable (x). This technique assumes a linear relationship between y and x, which can be represented by the following equation:

Equation (1) shows the relationship between the dependent variable (y) and the independent variable (x). The intercept term (β_0) represents the starting point of the relationship, while the slope coefficient (β_1) represents the size of the change in y for each unit change in x. This equation allows predictions to be made about the output variable using only a single predictor variable [5].

$$y = \beta_0 + \beta_1 x \quad (1)$$

It is often used in the multivariate linear regression method to create a model that represents the connection between the dependent variable (y) and at least two independent variables (x_1, x_1,..., x_p). The mathematical expression of this method is given by equation (2).

$$y=\beta_0+\beta_0 x_1+\beta_2 x_2+\cdots+\beta_n x_p+\epsilon, \quad (2)$$

where the coefficients, the coefficients β_0, β_1, β_2, βn represent the input variables x_1, x_2, x_3, x_n, respectively. The intercept of the line is denoted by ϵ [5].

3. MATERIAL AND METHODS

This study was conducted on a data set of 10,000 samples to make predictions about student achievement. Our study, in which multiple linear regression models were used, aimed to improve educational processes and individual student achievements. The prediction results for new independent variables can be seen with the web interface created using the prediction results obtained because of the study.

3.1. Data Set

The student performance data set will be used to make predictions about student success. This dataset of 10,000 student records was created specifically to investigate various factors affecting academic performance [6]. Each record in the dataset contains information about a performance index as well as various predictors. These determinants include the number of hours devoted to study (Study Hours), previous test scores (Previous Scores), participation in extracurricular activities (Extracurricular Activities), average daily sleep hours (Sleep Hours), and the number of sample questionnaires studied (Sample Questionnaires Applied). The performance index serves as a representation of the student's overall academic performance and ranges from 10 to 100; higher values indicate superior performance.

The dataset used in our study was taken from the Kaggle website [6]. The dataset consists of 10.000 student records, each containing information about several predictors, including a performance index and achievement. The dataset, consisting of a total of 6 features and 10.000 records, is detailed in Table 1 below.

Table 1. Dataset features.

Features	Description/values
Hours Studied	The total number of hours each student spent studying.
Previous Scores	These are the scores students received in previous tests.
Extracurricular Activities	Whether the student participates in extracurricular activities (Yes or no).
Sleep Hours	Student's average number of hours of sleep per day.
Sample Question Papers Practiced	Number of sample question papers studied by the student.
Performance Index	A measure of each student's overall performance. The performance index represents a student's academic performance and is rounded to the nearest whole number between 10-100.

3.2. Data Preprocessing

Repetitive data in the data set was removed by data preprocessing. After this process, the number of samples in the data set decreased to 9.873. As a result of examining each feature in the data set, it was checked whether the features contained the number of unique data, the number of missing data and non-numeric data. This process is important to understand the structure of the data set. It was seen that the values in the Extracurricular Activities column in the data set were categorical data, and the No and Yes values of the Extracurricular Activities features were changed to 0 and 1, respectively. The current state of the dataset is shown in Figure 1.

#	Column	Non-Null Count	Dtype
0	Hours_Studied	9873 non-null	int64
1	Previous_Scores	9873 non-null	int64
2	Extracurricular_Activities	9873 non-null	object
3	Sleep_Hours	9873 non-null	int64
4	Sample	9873 non-null	int64
5	Performance_Index	9873 non-null	float64

Figure 1. Current state of the dataset.

The correlation matrix presented in Figure 2 shows the relationships between independent variables affecting student performance. Correlation coefficients take values between -1 and 1 and indicate the direction and strength of the relationship between these values. Examining the correlation matrix, Previous_Scores is the independent variable with the strongest relationship with Performance_Index (0.92), indicating that the student's previous test performance contributes greatly to the overall performance index. Hours_Studied, on the other hand, has a moderate positive relationship (0.38) and indicates that study time positively affects performance. The other independent variables have a weak relationship with Performance_Index and do not show a significant impact. This correlation analysis is important to identify independent variables that affect student performance and to take them into account in the model development process.

Figure 2. It is a correlation matrix that shows the relationships between independent variables affecting student performance.

The model was evaluated by regression analysis using the Ordinary Least Squares (OLS) method on the created linear regression model. P>|t| using the formula, if there are independent variables with alphas greater than their critical values (5%), they must be determined and removed from the model. This process must be calculated each time and continued until there

is no independent variable greater than 5%. However, as seen in Figure 3, the P value in the OLS regression results table is calculated significantly lower than the alpha critical value in the model we developed. This showed that none of the independent variables in the model should be removed. Thus, the independent variables included in the initially established model were included in our model.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.989
Model:	OLS	Adj. R-squared:	0.989
Method:	Least Squares	F-statistic:	1.724e+05
Date:	Tue, 07 May 2024	Prob (F-statistic):	0.00
Time:	19:47:56	Log-Likelihood:	-21065.
No. Observations:	9873	AIC:	4.214e+04
Df Residuals:	9867	BIC:	4.219e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-34.0689	0.128	-265.875	0.000	-34.320	-33.818
x1	2.8527	0.008	358.940	0.000	2.837	2.868
x2	1.0183	0.001	857.427	0.000	1.016	1.021
x3	0.6167	0.041	14.981	0.000	0.536	0.697
x4	0.4803	0.012	39.623	0.000	0.457	0.504
x5	0.1939	0.007	27.017	0.000	0.180	0.208

Omnibus:	3.123	Durbin-Watson:	2.003
Prob(Omnibus):	0.210	Jarque-Bera (JB):	3.224
Skew:	0.014	Prob(JB):	0.200
Kurtosis:	3.084	Cond. No.	451.

Figure 3. OLS regression results.

3.3. Model Performance Measurement

Evaluating the performance of the model created for a machine learning study and measuring its prediction success is an important criterion for the reliability of the results obtained. There are many machine learning studies conducted today, and for the validity of these studies to be accepted, the success of the model must be above a certain threshold value. While this threshold value should be quite high in the field of health, it is expected to be slightly lower in other areas. The purpose of performance evaluation is to measure the performance of the model on real data and share the results. As a result, it is necessary to evaluate the performance of the machine learning model and reveal similar success for real-life problems and new examples [4]. The description and formulas of some basic metrics used to evaluate the performance of a model developed using machine learning are listed below.

Mean Absolute Error (MAE): The average absolute error calculation given by Equation (3) is found by averaging the absolute differences between the actual values and the predicted values.

$$MAE = \frac{1}{2}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

Mean Squared Error (MSE): The average absolute error calculation given by Equation (4) is calculated by finding the average of the squares of the differences between the actual values and the predicted values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (4)$$

Root Mean Squared Error (RMSE): The square root calculation of the average absolute error given by Equation (5) is calculated by taking the square root of the MSE. In this way, the results are reduced to a certain extent to get an idea about the typical size of the prediction errors.

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{n}} \quad (5)$$

R-squared (R^2 Score): Calculating the coefficient of determination given by Equation (6) is a measure of the extent to which the independent variables can predict the variability in the dependent variable for a sample not included in the data set. It has a value range between 0 and 1.

$$R^2\ Score = 1 - \frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2} \quad (6)$$

Where $\hat{y}_i$ and y_i are the predicted and actual values. $\overline{y}_i$ refers to the average value of y_i. The proximity between the predicted and actual values shows that the error values are directly proportional to the decrease. These values and the above formulas are important for measuring the prediction performance of the created models.

Accuracy (ACC): ACC is a widely used performance measurement metric to evaluate the performance of classification models. It measures the proportion of correct predictions out of the total number of predictions made by the model.

4. RESULTS AND DISCUSSIONS

4.1. Experimental Study and Findings

In our study, a multiple linear regression model was created and trained. The data set was divided into training and test sets and the model was evaluated on these data. The performance of the model was measured with various metrics such as MAE, MSE, R^2, RMSE, ACC and the values of the calculated metrics are given in Table 2. The results obtained show that the performance of the model is high, and the model can make accurate predictions.

Table 2. Calculated metrics for performance.

Metrics	Value
MAE	0.019
MSE	0.01
R^2	0.99
RMSE	0.02
ACC	0,994

In Table 2, the accuracy value showing the overall success of the model is 99.4%. The test data set error values and accuracy value performance results obtained in the output of our model are given in Table 2. These results are an indication that the data preprocessing processes, and our model were successful. It is seen that the error values are quite low, and R^2 and accuracy are

high. Moreover, the high accuracy of the model also shows that the generalization error should below.

The equation of the resulting linear model is;

It consists of the multivariate correlation between the independent variables of previous scores (x_1), hours worked (x_2), sleeping hours (x_3), applied sample question papers (x_4) and extracurricular activities (x_5) and the dependent variable of the student performance index (y). This resulting equation is very close to the modeled linear model. The resulting equation is shown in equation (7).

$$y= 1.018\ x_1 + 2.854\ x_2 + 0.472\ x_3 + 0.193\ x_4 + 0.59\ x_5 - 33.70 \quad (7)$$

Considering the coefficients of the independent variables, it can be concluded that the most weight is on the hours studied and there is an increase of 2.854 coefficients per hour worked, compared to a coefficient of 0.472 per hour of sleep. For this reason, a student who thinks that he will sleep first and then study, needs to study to be successful. It can be seen in the linear equation created by the model that there is an almost 1:1 ratio between the previous scores and the student success index. According to this equation, it is also seen that the number of sample question applications contributes little to student performance success.

4.2. Developed Web Interface Application

Different visuals of the web interface developed with the Flask module for Student Performance Index estimation are shown in Figure 4 and Figure 5 below. The designed web interface was run on the IP and Port number on localhost "127.0.0.1:50000". In this web interface, it is written which parameters can be entered in the relevant field. According to the written numerical values, our model makes a prediction with an accuracy value of 99.4%.

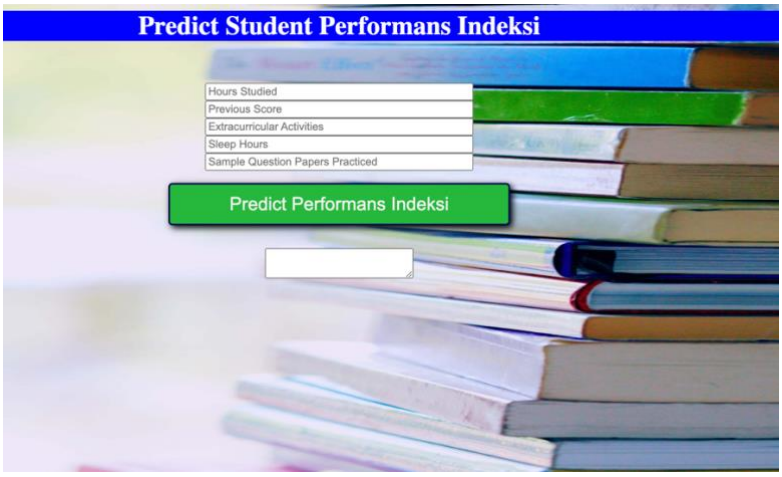

Figure 4. Web interface page for student performance Index estimation

Using the linear equation given in Equation (3), the values corresponding to the variables of hours worked, previous scores, extracurricular activities, sleep hours, and applied sample question papers parameters were entered in the web interface as 12, 92, 0, 1, 1, respectively. Then, the student performance index was estimated to be approximately 95 and this result is shown in Figure 5.

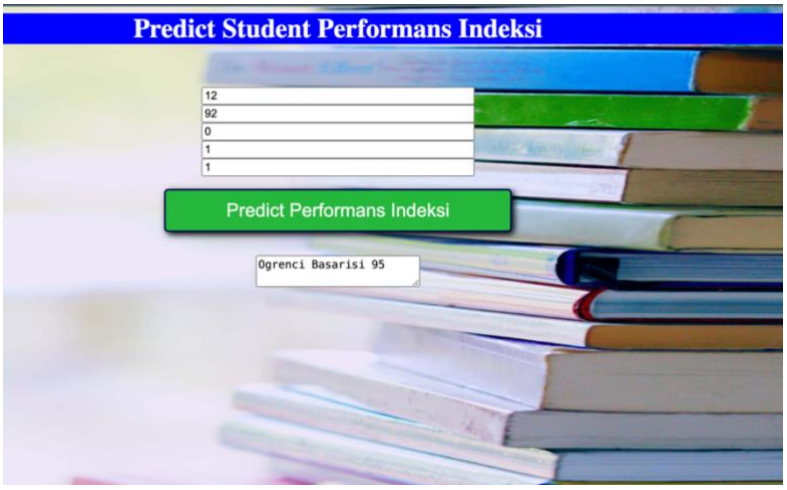

Figure 5. Estimation result of the student performance index according to the entered values

5. CONCLUSIONS

Linear regression analysis is a statistical method that investigates and illustrates the association between a dependent variable and one or more independent variables. Multiple linear regression, another statistical technique, is utilized to study the connection between multiple independent variables and a sole dependent variable. The main aim of linear regression is to identify the most precise model that predicts the dependent variable using the values of the independent variables.We conducted a study where we developed and trained a multiple linear regression model. The dataset was split into training and test sets, and the model's performance was assessed using several metrics such as MAE, MSE, R^2, RMSE, and accuracy. The aim of this study is to select the most important ones among all the previously described features to create a multiple linear regression model that allows predicting the student's performance index. The most accurate predictions are made using different data sets with the linear model created with these features. The outcomes revealed that the model performed exceptionally well, demonstrating its ability to make precise predictions. In particular, R^2 is 0.99 and the ACC value is 0.994, indicating that the model has a high accuracy in predicting the data. These accuracy values appear to be higher than the values previously reported in the literature. The prediction accuracy value obtained shows the impact of different independent factors on student success using the multiple linear regression model. In addition, the Flask web application developed to estimate the student performance index using new data sets allowed student performance to be predicted with high accuracy based on the entry of new variables. It is thought that the results obtained from our study will be guiding for future studies on predicting student success and exam grades by using more features.

REFERENCES

[1] El Aissaoui, O., El Alami El Madani, Y., Oughdir, L., Dakkak, A., & El Allioui, Y. (2019, July). A multiple linear regression-based approach to predict student performance. In International conference on advanced intelligent systems for sustainable development (pp. 9-23). Cham: Springer International Publishing.

[2] Dhilipan, J., Vijayalakshmi, N., Suriya, S., & Christopher, A. (2021, February). Prediction of students performance using machine learning. In IOP conference series: Materials science and engineering (Vol. 1055, No. 1, p. 012122). IOP Publishing.

[3] Chauhan, N., Shah, K., Karn, D., & Dalal, J. (2019, April). Prediction of student's performance using machine learning. In 2nd International Conference on Advances in Science & Technology (ICAST).

[4] S. Kour, R. Kumar and M. Gupta, "Analysis of student performance using Machine learning Algorithms," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 1395-1403, doi: 10.1109/ICIRCA51532.2021.9544935.

[5] Abirami, T., & Vadivel, R. (2023). Student semester marks prediction using linear regression algorithms in machine learning. World Journal of Advanced Research and Reviews, 18(1), 469-475.

[6] Dataset: https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression/data

[7] Asif, R., Hina, S., & Haque, S. I. (2017). Predicting student academic performance using data mining methods. Int. J. Comput. Sci. Netw. Secur, 17(5), 187-191.

[8] B. Sravani and M. M. Bala, "Prediction of Student Performance Using Linear Regression," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9154067.

[9] Arsad, P. M., & Buniyamin, N. (2014, April). Neural Network and Linear Regression methods for prediction of students' academic achievement. In 2014 IEEE Global Engineering Education Conference (EDUCON) (pp. 916-921). IEEE.