International journal of applied exercise physiology

ISSN: 2322-3537
Vol-14 Issue-01 June 2025

Using Dropout Rates to Forecast Academic Failure in Students

[1] Mangalagiri Anitha, [2] N. Sony,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract

A major issue confronting Indonesian colleges is the high rate of student dropout. Universities use a number of academic metrics or criteria when deciding whether or not to accept a student as a dropout. Using data mining or machine learning skills in the classroom is the way to simplify these issues. The most effective way for early academic status prediction is the classification approach methodology that makes use of the Neural Network (NN) technology. The means/average approach, z-score normalization, and information gain will be used to pre-process the supporting data before building the NN model. This will ensure that the optimal parameters are obtained. Optimizing a parameter using Adam optimizer also involves updating the weights repeatedly using the training data. With cross validation as the gold standard, the outcomes of this prediction model are computed. The acquired values are quite precise, reaching a level of 0.937. Grade is the most important factor in determining the likelihood of dropping out, followed by failed courses, student absenteeism, and finally, student age.

Keywords

predictin, neural network, dropout

I. Introduction

A concern that impacts higher education in Indonesia is the high rate of university dropouts. At the present time, Indonesia has what seems to be an extremely high dropout rate. There were 6,951,124 enrolled students and 239,498 who did not complete their degrees at Indonesian universities in 2018, according to the statistics. [1]. A university must take into account a number of academic factors or criteria while making the choice to drop out. As a result of students' decisions to drop out, the institution will also feel the effects. For students, it means less productivity and income throughout their lives, a higher chance of becoming unemployed, and as quandering of economic resources. Financial losses and a diminished academic reputation are two ways in which institutions suffer when students drop out [2]. Various methods have been suggested for gauging how well a pupil has done. When studying it, data mining is a common approach [3]. Educational Data Mining (EDM) [4] is a method for retrieving useful information from a big database of educational records [5], and it has lately seen extensive use in the field of education. In order to make reliable performance predictions for students, EDM is essential [6]. By identifying which children are more likely to fall short of expectations, schools may devote more resources to helping those students succeed [7]. Student performance predictions often make use of predictive modeling. Classification, regression, and categorization are three of the many activities used to construct predictive models. Classification is the most often used [4]. A number of studies have relied on data mining and machine learning techniques in the classroom to make predictions about students' academic behavior [8], [9]. These studies have used Neural Network models trained on students' academic data, which rely on a multilayer perceptron topology, to forecast how well students will do in class. The achieved accuracy is 99.42%, which is quite good. The next group to attempt to include student registration system data using academic data is S.A. Naser et al. [10]. With an accuracy of 84.6%, NN is clearly a top-notch technique. Blended learning performance is also predicted by Zacharis [11] using gradient descent and back-propagation algorithms. When it comes to predicting whether pupils will succeed or fail, the accuracy rate is 98.3%. Using student data that includes demographics, behavior, educational methodology, academic performance, and socioeconomic status, the NN method (as applied by Mayra Alban and David Mauricio [12]) achieves a 96.3% accuracy rate with the multilayer perceptron algorithm and a 96.8% accuracy rate with the radial basis function. At the same time, Ramanathan et al. [13] demonstrate that NN generates accurate predictions by optimizing it using the Lion-Wolf method. Bassi et al. [14] employ cognitive and non-cognitive metrics, together with the student's background information [15], to investigate the predictive quality of NN. The findings demonstrate that this approach can reliably and with little error forecast when a student will finish their course of study. Etebong et al. [16] achieves a 99.99% accuracy rate by combining cognitive and non-cognitive indicators, such as attendance, self-confidence, time management, and IQ scores. Ivan et al. [17] also attempts to use student population segment data and the NN model to make predictions. The model's success rate in categorizing 74.5 percent is the end outcome. Of pupils who do not succeed, yet this outcome is subpar in the eyes of the academics. In their study on the Neural Network approach, Amirah et al. [4] focused on the use of prediction algorithms to find the most relevant features in student data. They found that the major attributes had a high predicted accuracy of impact. Combining elements of both internal and external evaluations, this quality is unique. This demonstrates the significance of the external evaluation, here meaning the value acquired from the final test, in forecasting student achievement. We choose to build our prediction models using Neural Networks (NN) since NN has shown to be very effective in handling data categorization problems in the ever-growing computing domain of big data. Factors such as computing power, algorithm efficiency, and data amount and quality have a significant impact on the performance of NN models [18], [19]. There are a number of studies where the predictions employed are lacking, according to the study mentioned earlier. Several of the aforementioned research detail the student-centered school decision-making process, but none of them have addressed the question of how to design a Neural Network that can withstand several trials. Another limitation is that the research has not systematically summarized all of the elements that contribute to dropout. One study after another focused on dropouts, primarily investigating what causes people to stop attending classes, using the aforementioned dropout model. Predicting whether a student is in dropout status and identifying the elements that impact it may be achieved by integrating the real-world scenario with data on individual attributes and student success, as shown in the study model above. This is due to the fact that these characteristics are associated with students who do not complete their education, and the data is available in real-time from the university's database. Several variables have been discovered as possible predictors of student dropout, and the purpose of this research is to identify which of these factors have the most impact on this phenomenon. Learning how successful artificial neural networks are at generating predictions is another goal; specifically, we want to create a NN that can utilize various student data sources to forecast which students will drop out. Furthermore, educational leaders and all parties involved in this case may benefit from this study's findings. This methodology may be used by university administrators as a proactive tool for monitoring and managing students who are at danger of dropping out. This may also be used as a tool to help execute university policies that aim to raise students' educational level.

II. Methodology

Academic and demographic data from students are two of the many datasets that have been combined in this manner from prior research. An integrated data repository receives the data in a file format when it is extracted from the management system. The approach followed in this investigation is shown in Fig. 1. The next part will go into the flow in more depth.A. Gathering Information Finding and collecting data from a variety of university-based sources is essential for this research since it involves a number of variables that have the potential to influence students' academic achievement. After then, the dataset is updated with all the new information.

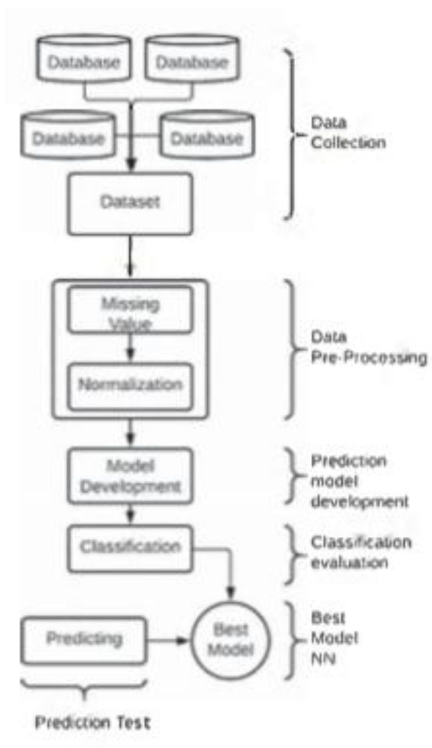

Fig. 1. Methodological Flow for The prediction of Dropout

The sample size for this research was 384 students drawn from the Health research Program at one of the participating universities. Throughout the years 2013–2015, data for this research was gathered from three distinct sources: • Student demographic data, including personal and family details. • Student academic data, including attendance and grades. • A broad survey that was filled out by prospective students. All of the study's variables are shown in Table I. Part B: Preparing Data To prepare raw data for use in the prediction process, data pre-processing involves a number of steps, as shown in Table I: cleaning, integration, normalization, and reduction. Data cleansing entails removing extraneous information and filling in blanks, such as missing numbers, from raw data. Using the data that does include a value—the mathematical mean of all values in a particular set—the Mean technique is used to deal with missing values.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1+x_2+\cdots+x_n}{n} \qquad (1)$$

Were,
i = sequence of data
$\bar{x}$ = the average of observation value
x = observation value
n = the amount of data

TABLE I. Parameter O f Students And Th e ir Data Types

International journal of applied exercise physiology
ISSN: 2322-3537
Vol-14 Issue-01 June 2025

Parameter	Data Type
Gender	Categorical (Female = 0 / Male = 1)
Age	Numerical
Address	Categorical (Nearby = 0 / Far = 1)
Live	Numerical (Living with Family=0 / Alone=1)
Mother's Education	Numerical (No school=0 / SD=1 / SMP=2 / SMA=3 / other=4)
Father's Education	Categorical (No school=0 / SD=1 / SMP=2 / SMA=3 / other = 4)
Trip to Campus	Numerical (1 - 3 hours)
Study Time	Numerical (1 - 10 hours)
Educational Support	Categorical (Yes=1 / No=0)
Family's Support	Categorical (Yes=1 / No=0)
Paid Class	Categorical (Yes=1 / No=0)
Futher Study Plan	Categorical (Yes = 1 / No = 0)
Internet Access	Categorical (Yes = 1 / No = 0)
Marital Status	Categorical (Single = 1 / Married = 0)
Failed Course	Numerical
Absence	Numerical
Score 1	Numerical
Score 2	Numerical
Score 3	Numerical
Graduate or drop out	Categorical, (Safe = 0 / Drop Out = 1)

The total is determined by dividing the sum of all numbers by the total number of data points, as shown in equation (1). The Z-Score Normalization is a tool in data normalization that employs the standard deviation to find the distance between a value (from a sample set observation) and the average, as shown in equation (2).

$$z_i = \frac{x_i - \bar{x}}{S}$$

Were,
z_i = z-score value
S = standard deviation

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n-1}}$$

The data that will be produced is based on the forecasts and classifications of safe pupils and those with dropout potentials. Students are considered to be in a safe position if they are not likely to discontinue their education. Conversely, students who are considered to have dropped out of school are unable to continue their education there and are thus required to discontinue their enrollment. Section C: Building Prediction Models An essential part of developing machine learning is determining the optimal NN architecture, which is shown in Fig. 2 of the NN block diagram used to forecast student dropouts. Multiple parameters are examined,

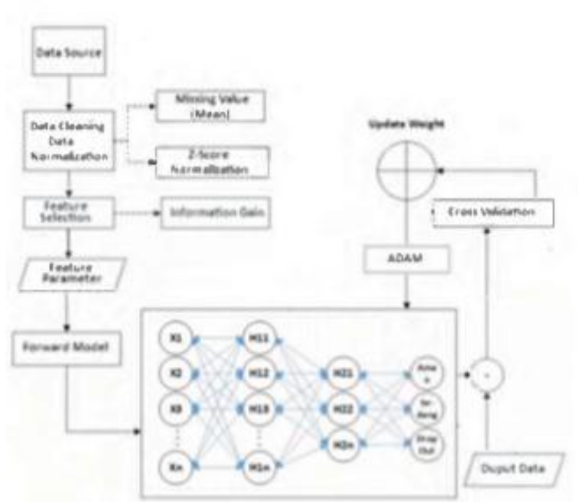

Fig. 2. Block diagram of NN to predict student dropouts.

TABLE II. Number Of Hid d e n Un it s In The Two Hid d e n Layer That Were Tested

International journal of applied exercise physiology
ISSN: 2322-3537
Vol-14 Issue-01 June 2025

No.	Hidden Layer 1	Hidden Layer 2
1.	1	1
2.	5	1
3.	10	5
4.	20	15
5.	40	30
6.	30	40
7.	50	30
8.	50	40
9.	40	50
10.	50	70

it includes: • The total number of hidden levels, which is 2, and • The specific number of concealed units for each layer. The number of units provided in Table II will be used to conduct experiments on each concealed layer. The Adaptive Moment Estimation (ADAM) technique is used to repeatedly update the weights based on training data. Adam adjusts the learning rate for each neural network weight by using the first and second moment gradient estimates. At the outset, a number of parameters must be initialized: mlo = 0, jSi = 0, a = 0,001, m2o = 0, P2 = 0,999, i = 0, and £ = 10-8. After calculating the gradient, Adam Optimizer iteratively adds i to each iteration, where i = i+1

$$g_i = \nabla_\theta f_i(\theta_{i-1})$$

were,
g_i = gradient value
f_i = iteration,

then updates the bias in the first and the second moments

$$m1_i = \beta_1 \cdot m1_{i-1} + (1-\beta_1) \cdot g_i \qquad (5)$$

$$m2_i = \beta_2 \cdot m2_{i-1} + (1-\beta_2) \cdot g_i^2 \qquad (6)$$

and calculates the bias correlation of the first moment and the second moment,

$$\hat{m1}_i = \frac{m1_i}{1-\beta_1^i} \qquad (7)$$

$$\hat{m2}_i = \frac{m2_i}{1-\beta_2^i} \qquad (8)$$

so that the parameters are updated,

$$\theta_i = \theta_{i-1} - \alpha . m1_i : \left(\sqrt{m2_i} + \epsilon\right) \qquad (9)$$

Were,
m1 = The first moment
m2 = The second moment
β_1, β_2 = Exponential decay rates
α = step size
θ = fixed parameters.

$$z_j = f\left(Z_{net_j}\right) = \frac{1}{1+e^{Z_{net_j}}} \qquad (10)$$

$$y_k = f\left(Y_{net_k}\right) = \frac{1}{1+e^{Y_{net_k}}} \qquad (11)$$

were,
e = natural number (2.718281828)
f(*Znetj*) = activation value at the hidden layer j node
f(*Y netj*) = result of activation function.

The resulting f(*Znetj*)value is then sent to all output nodes. The value of f (*Y netj*) is generated from the sum of the weights between hidden and output nodes. Information Gain is used as the best feature ranking method. The formula for calculating entropy is shown in the following equation.

$$\text{Entropy } (S) = \sum_i^c - P_i \log_2 P_i \qquad (12)$$

where, c = number of values in the classification class

P_i = number of samples for class i

$$\text{Gain}(S, A) = \text{ Entropy } (S) - \sum \text{value}(A) \frac{|S_v|}{|S|} \text{ Entropy } (S_v) \qquad (13)$$

were,
A = feature
V = the possible values for feature A
$Values(A)$ = set of possible values for A
$|Sv|$ = the number of samples for the value v
$|S|$ = the number of all data samples
Entropy (Sv) = entropy for the sample having the value v.

The largest Information Gain value ($Gain(S,A)$) is the best selected feature.

ISSN: 2322-3537
Vol-14 Issue-01 June 2025

Classification by Neural Networks (D) Adjusting the weight value of each node in each layer on NN is one of the three processes required for Backpropagation training. The other two phases include input of feedforward data as training and backpropagation as error value. Here is the NN technique that has been proposed: Adjust the computational accuracy and number of iterations after setting the activation function and assigning an intermediate random value (-1,1) to each weight. It is recommended to randomly choose the input sample xi(k) = (x1(k),x2(k),...,x(k)) and the anticipated output (k) = (d1(k),d2(k)),...,dq(k)). • Use this equation to get the hidden layer's input and output neurons:

$$hi_h = \sum_{i=1}^{n} w_{ih} x_i(k) - b_h, h = 1,2,\ldots \quad (14)$$

$$ho_h(k) = f(hi_h(k)), h = 1,2,\ldots, \quad (15)$$

were,
x = input parameter
hi_h = hidden layer connected with input
ho_h = hidden layer connected to output
w = weight
b = bias

To get the do (k) error value for the output neurons, compare the predicted and actual outputs.
• Adjust the connection weight Who (k) using do (k) and hidden output neurons. Fix the connection weight Whi(k) using dh (k) neurons, which are hidden and input neurons.
• Find the overall inaccuracy. • Checking for errors. The method terminates when either the error is low enough or the maximum number of trainings can achieve the requisite precision. In such case, you'll have to choose the next sample and the anticipated outcome before going back to step 3 to begin the iterative process again.

III. Result And Discusscion

In order to develop a model that can predict a student's academic standing, this section details the data mining procedures and experimental findings. In order to achieve data balance, the dataset undergoes data pre-processing. The Information Gain value is used to choose the optimum settings from the available options. In order to construct a prediction model using NN, the fifteen best parameters were chosen from among nineteen available, according to Table III. Gender, age, parental education level, study time, family assistance, paid classes, further study plan, internet access, marital status, statistics on failed courses, attendance, scores 1, 2, and 3, and so on are the factors that will be used. According to Table III, academic factors like grades (with gain information values of 0.715, 0.567, and 0.408), failed courses (which force students to retake the course the following semester), and other similar factors are the most important in determining whether or not students drop out. The pupils' chronically poor attendance rates are another concern. There is a 0.020 information gain value associated with the possibility that students' ages will play a significant role in determining whether or not they will drop out.

TABLE III. PARAMETER RANKING BASED ON INFORMATION GAIN VALUE

Parameter	Info. Gain	Gain Ratio	Gini
Score 3	0,715	0,359	0,345
Score 2	0,567	0,284	0,287
Score 1	0,408	0,205	0,211
Failed Courses	0,080	0,079	0,051
Attendance	0,033	0,017	0,020
Age	0,020	0,010	0,012
Further Study Plan	0,015	0,052	0,010
Mother'sEducation	0,010	0,005	0,006
Father'sEducation	0,009	0,004	0,005
Educational Support	0,008	0,013	0,005
Marital Status	0,006	0,007	0,004
Paid Classes	0,006	0,006	0,003
Study Time	0,005	0,003	0,003
Gender	0,005	0,005	0,003
Internet Access	0,003	0,005	0,002
Trip To Campus	0,003	0,002	0,002
Family's Support	0,003	0,003	0,002
Live	0,002	0,005	0,001
Address	0,002	0,002	0,001

Using the Classification Accuracy (CA) approach, we attempted to ascertain the number of neurons in

International journal of applied exercise physiology
ISSN: 2322-3537
Vol-14 Issue-01 June 2025

hidden layer 1 and hidden layer 2 in accordance with Table II. In order to determine how accurate the training data is, it is tested using two hidden layers. The purpose of this is to test whether the system's accuracy improves after training with more data after adding a hidden layer. The column with 50 hidden layer 1 neurons and 70 second hidden layer neurons has the greatest CA value of 0.938, according to Table IV. To aid in the creation of the NN model, we used the amount of neurons in the hidden layers. This demonstrates that the accuracy of the prediction is significantly affected by the quantity of hidden layer neurons. For the purpose of early prediction of student dropouts, we constructed the NN model scheme. We used Cross Validation as a Loss Function with K-Fold = 5 to validate the model's predictions using data from many students. We utilize 80% of the 384 students' data for training and the rest for testing. Based on both real data and predictions made by our model, the data testing results display the dropout levels. A Precision of 0.937 and a Recall of 0.938 are derived from these findings. Figures 3 and 4 show that out of 250 data points, 94.3% were false negatives and 92.4% were real positives. In addition, there are 15 data points representing real negative outcomes (5.7% of the total) and 9 data points representing false positive results (7.6% of the total).

Conclusion

We have seen that looking at a number of academic indicators or variables makes it a challenging endeavor to forecast whether or not a student would fail in college. In order to best anticipate student dropouts as early as feasible, it is helpful to have understanding of data mining or machine learning approaches in education. We have used 384 student records from a college to perform our study. To anticipate the pupils' academic standing, we have implemented a categorization strategy based on the Neural Network (NN) technology. We have also shown that various methods, such as normalizing data and overcoming missing values, as well as choosing the optimal features using the hidden layer neuron count, may significantly enhance prediction accuracy. The study relies heavily on two crucial activities: information collection and pre-processing data. The trustworthiness and accuracy of the data utilized have a direct bearing on the final product. In summary, the following are the key takeaways from the classification findings and the Neural Network (NN) method: • We have shown the efficacy of the NN algorithm in predicting students' academic outcomes, specifically in distinguishing between 'Safe' and 'Dropout' pupils. • We have shown the method's usefulness in choosing parameters from a set of options. Without sacrificing the n N classification performance, we were able to decrease the number of characteristics to be employed from the original set of 19 parameters to the optimal 15. • We demonstrated that achieving higher accuracy in predictions is possible with the right amount of neurons in the hidden layers. When looking at the acquired categorization model, there are a few key numbers that pertain to student failure, such as score 1, score 2, and score 3, which represent the grade point averages from the last three semesters. Factors such as age, class attendance, number of failed courses, family support, parents' educational backgrounds, and plans for future study are all linked to student failure. Faculty, staff, and parents may use this model as a red flag when it comes to pupils who may struggle academically.

TABLE IV. Number Of Hid d e n Un it s In The Two Hid d e n Layer That Were Tested

No.	Hidden Layer 1	Hidden Layer 2	CA
1.	1	1	0,674
2.	5	1	0,674
3.	10	5	0,896
4.	20	15	0,914
5.	40	30	0,927
6.	30	40	0,932
7.	50	30	0,919
8.	50	40	0,924
9.	40	50	0,919
10.	50	70	0,938

		Predicted		
		0.0	1.0	Σ
Actual	0.0	110	15	125
	1.0	9	250	259
	Σ	119	265	384

Fig. 3. Confusion Matrix Instance Model NN

		Predicted		
		0.0	1.0	Σ
Actual	0.0	92.4 %	5.7 %	125
	1.0	7.6 %	94.3 %	259
	Σ	119	265	384

Fig. 4. Confusion Matrix Proportion of Predicted Model NN

Those involved with higher education and this case stand to benefit from the findings of this study. This approach may be used by university administrators as a proactive tool for monitoring students who are prone to dropping out. The university's strategies aimed at raising students' educational attainment may find an ally in this. Finally, we want to proceed with more trials utilizing data from a variety of educational levels to see whether the NN method yields the same performance outcomes. We want to expand upon this study in future research by using it as a warning system to alert us to new student data events at the input. Faculty, staff, and parents may now respond more quickly to student attrition thanks to this innovation.

REFERENCES

[1] Kemenristekdikti, "Ministry of Research, Technology and Higher Education, Statistik Pendidikan Tinggi Indonesia 2018, Jakarta Higher Education Database: Pusdatin Iptek Dikti, Secretariat General." 2018.

[2] M. R. Larsen, H. B. Sommersel, and M. S. Larsen, Evidence on dropout phenomena at universities. Danish Clearinghouse for educational research Copenhagen, 2013.

[3] P. Thakar, "Performance analysis and prediction in educational data mining: A research travelogue," arXivPrepr. arXiv1509.05176, 2015.

[4] A. M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques," Procedia Comput. Sci., vol. 72, pp. 414-422, 2015, doi: 10.1016/j.procs.2015.12.157.

[5] M. F. Sikder, M. J. Uddin, and S. Halder, "Predicting students yearly performance using neural network: A case study of BSMRSTU," in 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016, pp. 524-529.

[6] D. Zacharias and A. Athanasios, "Monitoring of the Results through a Survey Concerning the Socio-Economic Characteristics of the Elderly Using Geographic Information Systems (GIS): A Case Study in Greece," Int. J. Innov. Econ. Dev., vol. 6, no. 3, pp. 36-45, 2020, doi: 10.18775/ijied.1849-75517020.2015.64.2004.

[7] V. K. Pal and V. K. K. Bhatt, "Performance Prediction for Post Graduate Students using Artificial Neural Network," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 7S2, 2019.

[8] J. L. Rastrollo-Guerrero, J. A. Gomez-Pulido, and A. Duran- Dominguez, "Analyzing and predicting students' performance by means of machine learning: a review," Appl. Sci., vol. 10, no. 3, p. 1042, 2020, doi: 10.3390/app10031042.

[9] S. R. Qwaider, S. S. Abu-Naser, and I. S. Zaqout, "Artificial Neural Network Prediction of the Academic Warning of Students in the Faculty of

ISSN: 2322-3537
Vol-14 Issue-01 June 2025

Engineering and Information Technology in Al-Azhar University-Gaza," 2020.

[10] S. S. Abu-Naser, I. S. Zaqout, M. Abu Ghosh, R. R. Atallah, and E. Alajrami, "Predicting student performance using artificial neural network: In the faculty of engineering and information technology," 2015, doi: 10.14257/ijhit.2015.8.2.20.

[11] N. Z. Zacharis, "Predicting student academic performance in blended learning using Artificial Neural Networks," Int. J. Artif. Intell. Appl., vol. 7, no. 5, pp. 17-29, 2016, doi: 10.5121/ijaia.2016.7502.

[12] M. Alban and D. Mauricio, "Neural networks to predict dropout at the universities," Int. J. Mach. Learn. Comput., vol. 9, no. 2, pp. 149-153, 2019, doi: 10.18178/ijmlc.2019.9.2.779.

[13] L. Ramanathan, "Angelina Geetha, Khalid, M., Swarnalatha, P.: Student performance prediction model based on lion-wolf neural network," Int. J. Intell Eng. Syst., vol. 10, no. 1, pp. 114-123, 2017, doi: 10.22266/ijies2017.0228.13.

[14] J. S. Bassi, E. G. Dada, A. A. Hamidu, and M. D. Elijah, "Students Graduation on Time Prediction Model Using Artificial Neural Network," IOSR J. Comput. Eng., vol. 21, no. 3, pp. 28-35, 2019.

[15] P. R. Shetgaonkar, "Predicting the impact of different Variables on Students Academic Performance using Artificial Intelligence," Int. J. Comput. Sci. Inf. Technol., vol. 6, no. 2, pp. 1367-1370, 2015.