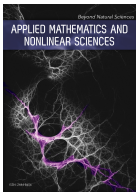
sciendo Applied Mathematics and Nonlinear Sciences 7(2) (2022) 297–306



Applied Mathematics and Nonlinear Sciences

https://www.sciendo.com

Research on predictive control of students' performance in PE classes based on the mathematical model of multiple linear regression equation

Xin Liu[1,3†], Alaa Omar Khadidos[2,3], Mohammed Yousuf Abo Keir[2,3]

[1] Hunan University of Science and Engineering, Institute of Physical Culture, YongZhou 425199, China

[2] Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

[3] Applied Science University, Al Eker, Kingdom of Bahrain

Submission Info

Communicated by Juan Luis García Guirao
Received June 17th 2021
Accepted September 24th 2021
Available online November 22th 2021

Abstract

Aiming to solve the problems in the traditional multiple regression analysis model for predicting college sports performance based on the principles of econometrics, a predictive model that combines genetic algorithm (GA), college sports performance evaluation and regression analysis is proposed. GA is used to conduct dynamic and supervised optimisation evaluation of college sports performance; on this basis, combined with regression analysis and GA's global optimisation capabilities, a complex nonlinear relationship between student sports performance and influencing factors is established; the student's performance is calculated based on the college sports performance. The results show that the method has high prediction accuracy and good stability.

Keywords: regression analysis, college sports performance, college sports, performance, prediction
AMS 2010 codes: 34A34

1 Introduction

At present, the methods for predicting college sports performance mainly include time series models, empirical models based on the principles of econometrics, and neural network models. Among them, the regression analysis model based on the principles of econometrics can comprehensively analyse the influencing factors of college sports performance and provide a basis for quantitative prediction of college sports performance, and it is also the most preferred by the majority of scholars. The author's research found that the predictive model proposed by purely using economic principles may be suitable for one or several schools, but it is unfounded to

†Corresponding author.
Email address: liuxinlunwenfabiao@163.com

sciendo ISSN 2444-8656 doi:10.2478/amns.2021.2.00058 sciendo

apply it to all schools, and the predictive results are often unsatisfactory [1]. To this end, this research proposes a prediction method that combines genetic algorithm (GA), college sports performance evaluation and regression analysis. Based on GA, it dynamically optimises college sports performance and realises supervised evaluation. On this basis, a predictive model of college sports performance is established based on regression analysis. The calculation results show that the model is a high-precision prediction method.

2 Materials and methods

The research on the prediction model of college sports performance can be summarized as time series prediction model, empirical model and neural network model, which is based on the school's nature and economic level and other factors to study the impact of the school's performance in college sports competitions. In subsequent research cases, Bernard and Busse proposed to use the Cobb-Douglas production function to establish a multivariate nonlinear model of the superior number division:

$$Me_t = \beta_0 + \beta_1 \log(POP_t + \beta_2 \log(PGDP_t) + \beta_3 Home_t + \beta_4 P_t + \beta_5 Me_{t-1} \tag{1}$$

Where:

$$Me = \frac{medals_i}{\sum_i medals_i} \tag{2}$$

Me represents the ratio of the number of merits ($medals_i$) obtained by the i-th school in the current college sports to the total merits of the current college sports ($\sum_i medals_i$).

In the formula, t is the time trend; POP is the student population in the current year; PGDP is the GDP per student in the current year; $Home$ is a dummy variable, $Home = 1$ means college physical education examination, $Home = 0$ means non-examination; P is a dummy variable, $P = 1$ is an examination school, $P = 0$ is a non-examination school; β_0 is a constant; β_j ($j = 1.5$) is the coefficient of each explanatory variable. Since then, more people's attention has been focused on this research, which has also made college sports performance prediction a hot research topic. Some scholars pointed out that the shortcomings of the traditional model are explained as follows: the predictive model established solely by the principles of economics may be suitable for one or several schools, but it is unfounded to apply it to all schools. It is proposed that students' college sports performance has an impact on college sports performance. Based on the important influence, based on the Bernard and Busse model, a multivariate nonlinear model based on college sports performance evaluation is established:

$$M_t = \beta_0 + \beta_1 \log(POP_t) + \beta_2 \log(PGDP_t) + \beta_3 Home_t + \beta_4 M_{t-1} + \sum_i^{C-1} \alpha_C D(C) \tag{3}$$

In the formula: C is the student's college sports performance grade, and other parameters are the same as formulas (1) and (2); the college sports performance grade C of each school is obtained by cluster analysis. The research results show that this method has higher prediction accuracy and higher feasibility than traditional regression analysis.

From model (3), it can be found that the evaluation of students' college sports performance is the focus and difficulty of the prediction research. However, the existing evaluations of college sports performance are all unsupervised clustering methods. The disadvantages of this method are based on what data set is used as the cluster analysis, which cluster analysis method to choose, whether the evaluation of outstanding scores and the number of merits is equal, and clustering. It is very difficult to determine the number of classes, and can only be determined based on empirical estimation. These subjective estimation methods will inevitably lead to a decrease in the accuracy of the algorithm. Taking into account the shortcomings of the above cases, the author considers using GA to supervise and evaluate college sports performance. GA transforms the objective function into a

genome group, takes the fitness function as the optimisation goal, and obtains the next-generation optimised gene combination through genetic manipulation, and so on until the optimal convergence goal is met [2].

3 Results of performance evaluation

3.1 GA optimises the general description of college sports performance evaluation

An important reason why GA can be widely used is its global convergence. Due to the diversity of the GA group, it searches in all directions as much as possible. This is a great improvement over the previous gradient method that only searches in one direction. Moreover, GA does not need to have continuity and differentiability restrictions on optimisation problems. In the end, the dynamic optimisation of college sports performance evaluation can be realised. On this basis, prediction is made based on the multivariate nonlinear model of college sports performance, ensuring high prediction accuracy and strong objectivity. The prediction model process based on GA optimised college sports performance evaluation proposed in this research is shown in Figure 1:

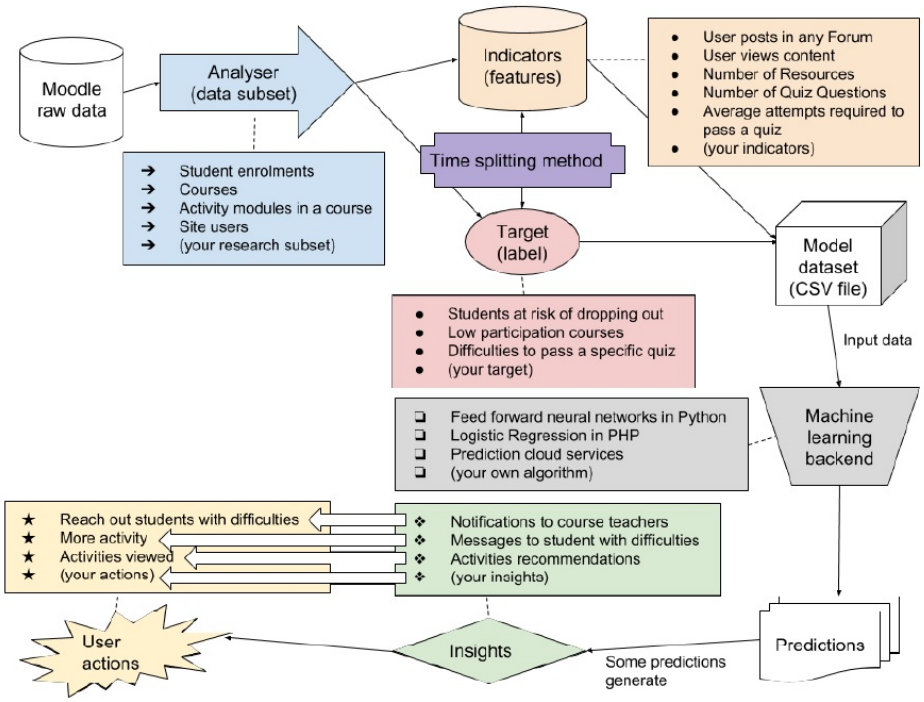


Fig. 1 Forecast model algorithm flow chart.

GA uses goodness-of-fit R2 to evaluate the performance and prediction accuracy of college sports performance evaluation and converts this objective function into a fitness function. The algorithm starts by randomly generating a group. Each group of chromosomes in the group represents the student's college sports performance level. Each group of chromosomes is evaluated according to the fitness function, and the corresponding fitness value is obtained. The greater the fitness of the chromosome, the more the representative college sports performance evaluation has been optimised and the prediction effect is better. According to the fitness value, the probability of each chromosome being selected in the selection operation can be calculated. According to the selection probability, a random traversal sampling method is used to select a group of chromosomes to form a new population. According to the crossover probability, the chromosome is selected for GA crossover op-

eration, and finally, according to the mutation probability, the mutation operation is performed on some of the gene positions on the chromosome. This operation makes the college sports performance grade set represented by the chromosome diversity in the entire search process and has a great played an optimisation role, thereby ensuring that the optimal solution can be found. The end condition of the algorithm is to set a maximum number of iterations, epochal, to ensure that the solution obtained by GA after the end condition is reached is the optimal solution [3].

3.2 Chromosome coding scheme

Coding is the prerequisite for GA to solve the problem. This study uses integer coding for college sports performance grades. Before chromosome coding, first of all, the range $[C_{\min}, C_{\max}]$ of the C value of all students' college sports performance grades should be determined. In general, the optimal number of clusters will not exceed $C_{\max} \leq \sqrt{N}$. Therefore, the value range of C can be set to $\left[2, \sqrt{N}\right]$.

Each chromosome represents a set of students' college sports performance grades. The length of the chromosome is the number of students' homes. The genes in the chromosomes represent the college sports performance grades, and the same genes indicate that the college sports performance grades are of the same category. Take an integer k in the value range of C, which means that the students in the set contain k college sports performance levels. The chromosome can be expressed as: $[Z_1, Z_2, Z_3, ..., Z_N], 0 \leq Z_i \leq k-1$.

For example, in this study, $N = 62$ students are selected as the research object, so the best college sports performance level is $2 \leq C \leq 8$. If $k = 6$, then the chromosome code is: (5, 3, 5, 4 ,3, 3, 0, 3, 2,...,1, 3, 4), the length is 62.

3.3 fitness function

According to the code of the chromosome, this code is converted into a dummy variable, to avoid the 'dummy variable trap', Use $k-1$ dummy variables $D_1, D_2, ...D(k-1)$ to represent k categories (as shown in Table 1), perform multiple nonlinear regression analysis according to model (3), and convert the regression model goodness of fit R2 into the objective function shown in (4):

$$obj = 1 - R^2 = \frac{SEE}{SST} = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \hat{y}_i)^2} \tag{4}$$

y_i is the observed value, y_i is the fitted value, and y is the mean value.

Table 1 List of dummy variable settings for college sports performance grades.

Category K		**1**	**2**	**3**	**...**	**k**
Virtual variable	D1	0	1	1	1	1
	D2	1	0	1	1	1
	D3	1	1	0	1	1
	...	...	...	...	...	...
	D (k-1)	1	1	1	0	1

The fitness function is usually used to convert the objective function value to a relative fitness value. To prevent premature convergence, the fitness value can be calculated according to the order of the objective function value in the population. Sort according to the individual objective function value obj from small to large. According to the sequence number of the sort, each level of the individual is given a fitness value. Non-dominated solutions with the same sort are assigned the same fitness value. Equation (5) calculates:

$$FinV(i) = 2 - MAX + 2(MAX - 1)\frac{x_i - 1}{N_{id} - 1} \tag{5}$$

In the formula: MAX represents the selection pressure difference, generally between $[1, 2]$; x_i is the position of individual i in the ordered population; N_{id} is the population number; $FinV(i)$ represents the fitness value of the individual at position i. In this study, the pressure difference is chosen to be E. Since the higher the R2 value, the more accurate the prediction, so the fitness function gives a higher fitness value to the chromosomes with good final prediction results; conversely, the chromosomes with poor prediction accuracy are given a lower fitness value. The essence of using GA to optimise forecasts is to optimise the goodness of fit R2.

3.4 Selection operator

The selection operator is a GA that determines how to select a certain number of good individuals from the parent population based on the set generation gap (GGAP) to inherit into the next generation population. In order to improve global convergence and computational efficiency, the selection method uses random traversal sampling (SUS). SUS is a single-state sampling algorithm with zero deviation and minimum individual expansion. It replaces the single selection pointer used in the roulette method. SUS uses S pointers of equal distance, where S refers to the number of selections required. The population is randomly arranged, S pointers $[ptr, ptr+1, ptr+2, ..., ptr+s-1]$ determine S individuals, and pointer $ptr+i(i=0,1,...,S-1)$ is determined by a random number generated in $[1/S, i+1/S]$. Assuming that $S=6$ individuals are selected from 10 individuals and the random position of the first pointer is 0.04 (Figure 2), then the distance between the pointers is 1/6 = 0.17, so it can be based on the position and cumulative probability of the pointer per the interval can determine the selected individuals are: 1, 2, 3, 4, 7, 8.

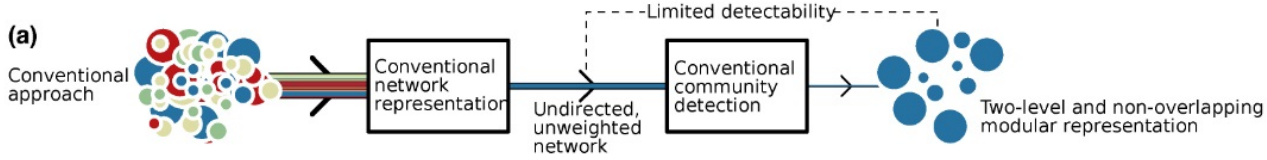


Fig. 2 Schematic diagram of random traversal sampling.

3.5 Mutation operator

Using uniform mutation, its operation refers to replacing the original gene value at each locus in the individual coding string with a random number that is uniformly distributed within a certain range with a certain small probability, that is, depending on the parent individual the mutation probability Pm is operated to prevent premature convergence from producing a locally optimal solution instead of the overall optimal solution [4].

The specific operation processes of uniform mutation are: 1. Specify each locus in the individual code string as a mutation point in turn; 2. For each mutation point, take a random number from the value range of the corresponding gene with the mutation probability Pm Replace the original value.

3.6 Crossover operator

Single-point crossover means that only one crossover point is randomly set in the individual code string, and then part of the chromosomes of two paired individuals are exchanged at this point. Here, a crossover position is randomly set for individuals in the group, and the operation is performed according to the crossover probability Pc. The two paired chromosomes exchange part of their genes at the crossover position by a single point crossover, and a new generation of groups is generated through exchange. Figure 3 is a schematic diagram of a single point crossover operation.

The specific implementation process of single-point crossover: 1. Randomly pair individuals in pairs. If the group size is M, there are [M/2] pairs of paired individual groups; 2. For each pair of paired individuals, randomly Set the position after a certain locus as the crossover point. If the length of the chromosome is N, there

are N-1 possible crossover point positions; 3. For each pair of individuals, the crossover probability Pc is Part of the chromosomes of two individuals are exchanged at the intersection point, resulting in two new individuals [5].

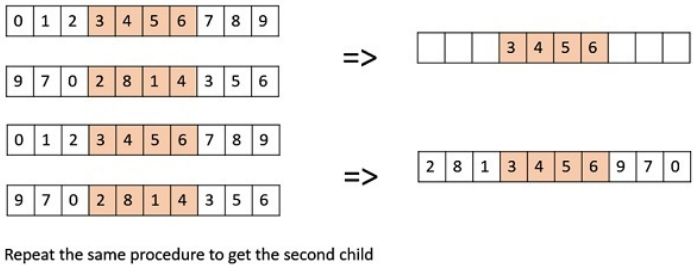


Fig. 3 Schematic diagram of single-point crossover operation.

4 Discussion

To evaluate the prediction accuracy and the pros and cons of the model, this study introduces the following errors:

A. Root mean square error:

$$PMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2} \tag{6}$$

B. Mean absolute percentage error:

$$MAPE = \frac{1}{N}\sum_{i}^{N}\frac{|y_i - \hat{y}_i|}{|y_i|} \tag{7}$$

C. Mean absolute error:

$$MAE = \frac{1}{N}\sum_{i}^{N}|y_t - \hat{y}_t| \tag{8}$$

D. Pearson correlation coefficient:

$$PEA = \frac{\sum_{i=1}^{N}(y_i - \overline{y_i})(\hat{y}_i - \overline{\hat{y}_i})}{\sqrt{\sum_{i=1}^{N}(y_i - \overline{y_i})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}_i})^2}} \tag{9}$$

In formulas (6)–(9): $y_i, \overline{y_i}$ is the actual value and the predicted value respectively.

This study uses the actual data of college physical education from 2014 to 2018 as sample data, selects 62 schools (regions) as the research object, and uses the 2018 college physical education performance to test the effect of the prediction model. The software to realize the algorithm is MATLAB software, and the control parameters of GA are set as: initial population number $M = 50$; chromosome length $N = 62$; crossover rate $Pc = 0.7$; mutation rate $Pm = 0.01$; generation gap is $GGAP = 0.9$.

4.1 Determination of the number of college sports performance levels

To compare the influence of the number of college sports performance levels on the multiple regression model, the GA optimised multiple regression nonlinear model is used to calculate all the best fit goodness R2 within the range of the number of college sports performance levels C. The calculation results are shown in Figure 4. The data of Jiangxi province, Henan province, Heilongjiang province and Jiangsu province are the 4 sub-maps in Figure 4 respectively.

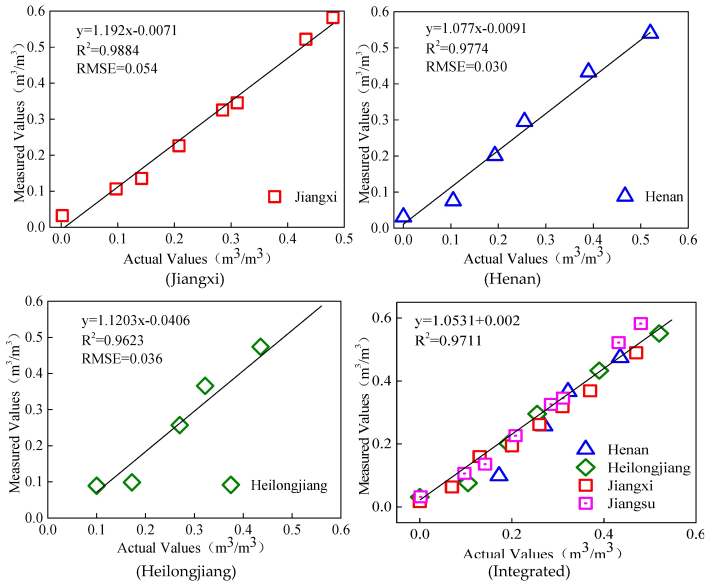


Fig. 4 Schematic diagrams of the relationship between the number of college sports performance levels and the goodness of fit R2.

It can be seen from Figure 4 that for the prediction of merit number, when the number of college sports performance grades is $C = 7$, the goodness of fit R2 is the largest, that is, the best college sports performance grade for the student (region) to obtain the merit number should be divided into 7 categories; for the prediction of excellent performance, when the college sports performance level $C = 4$, the goodness of fit R2 is the largest, that is, the best college sports performance level for students (regions) to obtain excellent results should be divided into 4 categories [6].

4.2 Forecast results

According to the above analysis, the number of college sports performance grades of the student (region) merit number prediction model is set to 7; the number of college sports performance grades of the excellent performance prediction model is set to 4, and the sample data is subjected to regression analysis (Table 2).

According to the results in Table 2, the excellent results in 2018 can be predicted (Table 3). Finally, respectively calculate the prediction results of the literature and the prediction ability evaluation indicators of the prediction results proposed in this study (Table 4). It can be seen from Table 4 that the prediction model proposed in this study has obvious advantages in predicting excellent performance; in the prediction of excellent performance, except for the slightly smaller MAE index, other indicators are better than the former.

From Table 4, it can be found that for the FCM-regression model, because the university sports performance evaluation based on unsupervised fuzzy C-means clustering is difficult to objectively describe, it has limited ability to effectively optimise the combination of student (regional) university sports performance and its predictive ability Naturally, there is no guarantee, making the prediction accuracy relatively low [7].

The GA-regression model proposed by this research can realise the supervised calculation of the student

(regional) college sports performance grade through GA, and can dynamically mine the best college sports performance evaluation [8] so that the prediction model based on college sports performance can be optimised. At the same time, the subjectivity of the prediction model is reduced, and the accuracy and stability of the superior and prediction are higher [9].

Table 2 Summary of the regression results of the share of outstanding and excellent results in college sports from 2014 to 2018

	Regression results of performance share		**Regression results of outstanding performance share**	
	Coefficient	**t**	**Coefficient**	**t**
log (POP)	0.700580484	4.046	0.415217178	3.612307
log (PGDP)	0.00492357	3.8	0.507549691	1.092639
Home	0.720660842	1.656	0.950142728	3.393209
Mt-1	0.935333311	0.9	0.733616507	4.407553
β_0	0.819858339	0.645	0.949132464	2.307292
D1	0.233461872	1.489	0.861656496	0.10588
D2	0.522969528	4.718	0.159542912	1.948906
D3	0.887297088	3.593	0.108918231	3.609763
D4	0.673300591	4.348	0.916780756	1.823459
D5	0.681120134	1.667	0.871739745	1.593532
D6	0.767696616	2.757	0.6677149	1.354243
Statistical test	$R^2 = 0.9544$	$F = 491.43223$	$R^2 = 0.9254$	$F = 444.43223$

Table 3 List of the classification results of the merits of each school (region) and college sports performance.

Classmate code	**Performance prediction**		
	Actual	**Prediction**	**Strength ranking**
A	98.55	98	1
B	98	98	1
C	98	98	1
D	98	98	1
E	97.25	97	2
F	97.11	97	2
G	97.1	97	2
H	96.12	96	3
I	92.12	92	4
J	92	92	4
K	90	90	5
L	90	90	5
M	89.11	89	6
N	89.12	89	6
O	87.12	87	7
P	86	86	8

Table 4 Summary of the results of the two models' prediction statistical indicators.

	Model	RMSE	MAPE	MAE	PEA
Grades	FCM	7.123	0.5446	4.789	0.954
	GA	6.785	0.5214	3.256	0.957
Excellent results	FCM	3.456	0.4851	2.0657	0.925
	GA	3.278	0.4712	2.2145	0.952

GA, genetic algorithm.

5 Conclusion

GA can realise effective supervision and calculation of students' (regional) college sports performance grades, and can dynamically mine the best college sports performance evaluations so that the prediction model (3) based on college sports performance can be optimised. At the same time, the objectivity of the prediction model is improved, and the accuracy and stability are high in the prediction of the number of excellent (excellent grades). Using GA optimised multiple regression nonlinear model, it is possible to calculate the number of college sports performance grades of college sports students (regions). In the student (region) merit number prediction, the number of college sports performance grades is 7; in the student (region) excellent performance prediction, the college sports performance grade number is 4.

References

[1] Zhang, S. J. Li, Y. X. Ma, S. B., Yan, H. J. & Mao, H. E. Optimization on key parameters for the metal deformation of rapid shear extrusion bonding. Journal of Plasticity Engineering, 24(3) (2017), 69-77.
[2] Shaofei Wu, Jun Liu, Lizhi Liu. Modeling method of internet public information data mining based on probabilistic topic model, The Journal of Supercomputing, 75(2019), 5882–5897.
[3] Besbes, O. & Zeevi, A. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. Management Science, 61(4) (2015), 723-739.
[4] Ageeva, H., & Kharin, Y. Ml estimation of multiple regression parameters under classification of the dependent variable. Lithuanian Mathematical Journal, 55(1) (2015), 1-13.
[5] Liska, G. R. Silveira, E. C. D. Reis, P. R., M. Â. Cirillo, & Gonzalez, G. G. H. Selecting a binomial regression model on the predation rate of euseius concordis (chant, 1959). Coffee ence, 10(1) (2015), 113-121.
[6] Shaofei Wu, Qian Zhang, Wenting Chen, Jun Liu, Lizhi Liiu, Research on trend prediction of internet user intention understanding and public intelligence mining based on fractional differential method, Chaos, Solitons and Fractals, 128(2019), 331-338.
[7] Zhou, D. & Wu, J. On the stochastic restricted two-parameter ridge type estimator in a linear regression model. Far East Journal of Mathematical Sciences, 102(2) (2017), 421-424.
[8] Ramazan Sari, Some Properties Curvture of Lorentzian Kenmotsu Manifolds. Applied Mathematics and Nonlinear Sciences, 2020. 5(2):pp. 283-292.
[9] G. Gopi Krishna, S. Sreenadh, A. N. S. Srinivas, Entropy Generation in Couette Flow Through a Deformable Porous Channel, Applied Mathematics and Nonlinear Sciences. 2019. 4(2):pp. 575-590.