International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

A Comprehensive Study on Student Performance Prediction Through Machine Learning

Dr. Divya Shree
Assistant Professor, Department of Computer Science,
Tika Ram Girls College Sonipat

Abstract—The Machine Learning has been used in educational area necessitated to handle several types of problems such as: to handle the drop out problems/cases, to improve the students' retention cases, knowing in advance at risk students, to predict and analysis the students' performance.Recently, lot of changes have occurred in education sector/system, such as school/university were temporary closed,offline education work moved towards an online education, school/university have reopened, bringing out major changes in the behavior of students which directly or indirectly affects the performance of students. Compatibility of this study to existing study for obtaining best predictive accuracy value model with significant datasets. For predictive analysis the performance of student into three categories such as excellent , average and poor with significant datasets, consequently upon reopening of schools, the aim/objective of this study for considering the selection between 1501 to 9000 range of datasets by determiningthe range on average bases somewhere on the point neither morenor less number of previous researchers and also identifying the exiting the best machine learning algorithms whose accuracy value may be above 90%.From 2019 to 2021 MLP (Multi-layer Perceptron), RF (Random Forest), QDA (Quadratic Discriminant Analysis), LGBM (Gradient Boosting), Support Vector Machine, Linear Regression, BiLSTM (Bidirectional Long Short-Term Memory) algorithms and to provide higher accuracy value that was greater than 90%. After the analysis ofprevious research work there were seven algorithms whoseaccuracy value above than the 90% and also the modest range of datasets (that was greater than 1500 and less than equal to 9000(>1500&<= 9000)) was considered by neither more nor less previous researchers (4 previous researchers) in their studies.

Keywords—Machine Learning, Performance of the Students, Evaluation Matrix, Predictive Analytics, Education System.

I. INTRODUCTION

A. Education System

Schools and University come under the education system. In this system different age of people come to gain an education. The education system stands on three pillars described in Fig. 1.

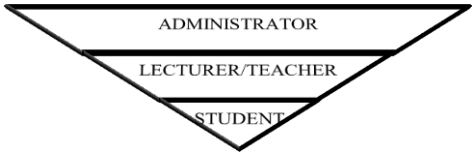

Fig.1. Three Pillars of Education System

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

These three pillars being Teaching style, student's behavior and administration task are interrelated to each other. In its administrator is the person who plan, control and run the academic institution (university administration). Administration helps by developing the child center curriculum, and also assist the teacher/instructor for taking better decision in future and timely providing work andfeedback to and from students vice versa. The teacher/lecturer taught the students and get their responses for feedback in the form of marks or results. With the help of these feedback or responses teacher/lecturer come to know their teaching skill and learning ability of a student. if any type of deficiency is found out in their teaching skill and learning ability of a student then it can be removed by takinga corrective action by administrator at correct time. This process helped for achieving the higher success rate in future[1, 11].

B. Machine learning in Education

Recently, it was difficult to handle the large amount of data manually therefore after some time, machine learning was used/introduced in several area among which educational area is one of them. Machine learning automated the large data and helped in removing the computation complexity. Significant dataset, extracted/selected most Relevant featuresand the best accurate model were the most important factors for providing the sufficient and accurate result into three categories i.e., excellent, average and poor. These factors alsohelped the administrator, parents, students to know about thelagging student so that correct action should be taken at correct time by the administration, parents and students itself and also providing more attention/focused towards lagging students for improving the result or performance in future. Machine learning is also used for several purposes such as toimprove the student's performance, handle the drop out problems, improve student's retention and to analyze the student's performance [10-11].

C. Measure of Predictive Accuracy of the Student's Performance.

Fig. 2. is described Structure and steps of predictiveanalysis. Large amount of data (school/ university record) notonly in the form of time related data (historical record etc.), but also in the form of web (huge database repository provided by internet), multimedia and hypertext (audio, text video, image) etc. were stored into different-different locations databases (school/university's branches) or flat files. From multiple data sources (database or flat files etc.) only the relevant data were collected, cleaned and integratedinto a single site (single place) in the form of data Warehouse(offline data, online data). The relevant data were retrievedfrom offline and online data (Kaggle, UCI Govt data repository etc.) and transformed it into the form of well- defined structured data (summary data) and then analyzed it. The most relevant features/variables were retrieved/extracted from it by applying the numerous intelligent methods such asSMOTE with FS (feature selection), BiLSTM (BidirectionalLong Short-Term Memory) combined with an attention mechanism and features extraction method, naïve byes, clustering method (k mean (ANN, SVM)) etc. into it [10, 13,18]. Structured data or relevant features were applied to buildand train the machine learning prediction model. In the next step tested the prediction model by adding some query instances into it and generated the prediction result whose accuracy measured by several evaluation metrics such as F- Measure, AUC, Accuracy, Precision, Recall etc. [15]. In the last and final step model had monitored as well as refined [8].

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

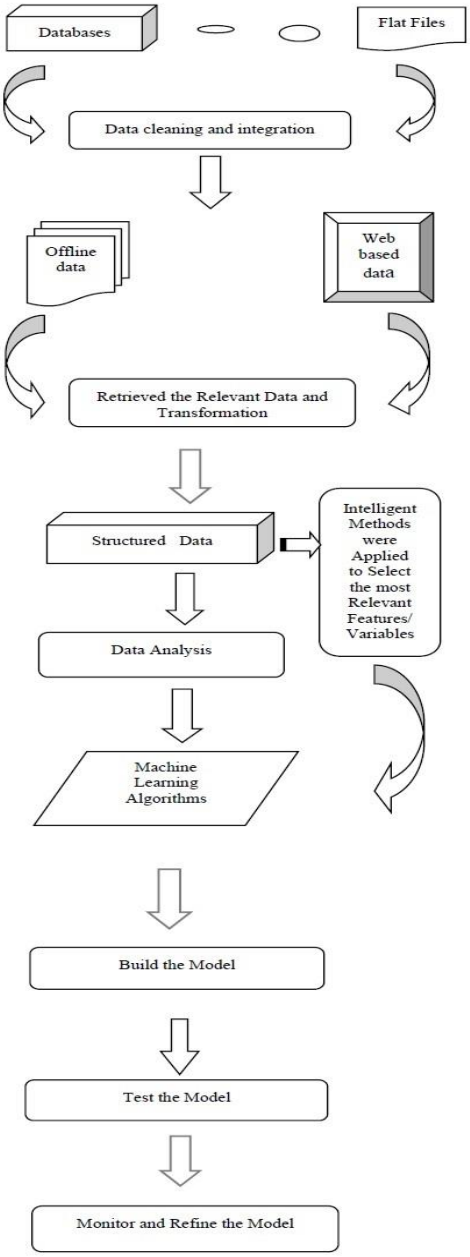

Fig.2. Structure & Steps of Predictive Analytics

This paper reviewing the previous 5 years research work on selecting the best machine learning model and the modest range of dataset was considered by neither more nor less previous researchers in their studies. This paper's author builds a systematic approach of reviewed work which supports the following given objectives.

II. OBJECTIVES OF THIS PAPER ARE:

- To identify the existing study on which it is based,
- To identify the existing best algorithm/machine learning model for predictive analytics the student's performance.

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

III. LITERATURE REVIEW

Reviewed paper reviewed the previous research work which focused on the country that had low literacy rate and helped these country's academics to effectively manage theirstudent performance so that their literacy rate should/could be improved. Not only single aspect/factor instead of the complete program/factors (educator's competence, social-economic data and academic data) needed to predict the student's performance [7]. With the help of recommended system of the Student's success not only analyzed , forecasted but also knew about their reason (behind their success)that helps the education institution and parents to effectively examine the Students performance which depends on various factors such as free time, alcoholic and study time etc.[17].Students based factors(Lack of time , Lack of motivation , Insufficient background knowledge andskills)and MOOC based factors/variables (Isolation and lack of interactivity, Course design, Hidden costs)will resultinto very high degree of drop out cases. To control such dropout cases , some solution being introduced such asClickstream data standardization (MOOC contained severaltraceable events and interactions from various audio & videodevices including user presence time, documents viewed, number of videos watched, frequency of interactions, and links opened - among others.), Student-provided data (restricted to access the private/personal data.), Feature engineering techniques(the techniques explored some more different features such as student's prior experience, test grades etc.), students that were likely to drop were timely identified, Evaluating models and predictors, student interaction through various discussion activities that helped to analyze the student dropout prediction challenges (Availability of publicly accessible dataset, Managed big masses of unstructured data, Student schedule, Lack of enough sample data, Data variance, High data imbalanced etc.) by way of developing an effective and accuracypredictive model [20] as described in given Table 1.

TABLE 1. LITERATURE REVIEW Of PREVIOUS REVIEWEDWORK.

Ref.No.	**Details**			
	Study Basedon	***Year***	***School/University***	***Instances***
[7]	Predicting the student's performance using machine learning methods.	2020	n/a	n/a
[17]	Analysis, forecasting the student's success and also present their reasons.	2021	school	200
[20]	MOOC Dropout Prediction.	2018	Online courses	0ut of 641138 instances 17687 instances had obtained certificate

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

In Education system several changes have occurred fromtime to time. One of the bigger changes was: Used educationsystem with machine learning not only for prediction purposebut also for predictive analytics purpose proved/described byseveral previous researchers in their researches. The prediction system of student's performance with the help of Deep Neural Network applied six algorithm(as Decision Tree (C5.0), Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, Random Forest and Deep neural network in R Programming) with kaggle dataset for trained model and tested .Out of six tested algorithms Deep neural network had outperformed and produced accuracy value of 84% [3].In school the Educator used naïve byes model for selecting the most relevant features and examined the correlation between and predicted performance of students on assessments/results. The Administrator /Investigator used this predicted model for taking right action at right time and achieved higher student's success rate in future [11]. Data mining and machine learning both are used for same purposebut the main difference in between was machine learning automate the work and easily handled the large computation complexity whereas data mining not. Machine learning and data mining used for analysis of the student's performance by using k-mean clustering algorithms with two classification algorithm (ANN, SVM) and collected datasets from ED- Facts (from govt. inventory data). Artificial Neural Network (ANN) achieved higher performance as compared to SupportVector Machine (SVM) in terms of Mean Squared Error (MSE around 5-20%) and Effort Estimation (EE around 15- 27%) [13].Forecasted the Most suitable Educational path for the better career to each and every school students / learners who were convinced for the 12 standard based on their 10 standard performance marks as well as recommended them better academic program for their higher education by used several machine learning methods/approaches. The Light GBM algorithm was the best classification model for arts/humanities and science-based intermediate program whose F-measure values (0.97 and 0.90), ROC-AUC values (0.97 and 0.90), Cohen's Kappa values (0.94 and 0.80) and Log loss values (0.0002 and 0.003). Same as Different coursehave different best algorithm by measure F- Measure, ROC- AUC Value, Cohen Kappa and log loss values. Therefore, all applied machine learning models/method provided averages of evaluation metrics 's different performances, in terms of F-measures value: 97.16%, ROC-AUC value: 97.16%, Cohen's Kappa value: 94.33%, and the Log loss value: 9.88%for all academic programs [4]. Predictive analysis refers to analyze and forecast the student's performance. Student performance's predictive accuracy improved or enhanced byuse of three different datasets and three machine learning algorithms (XG Boost, RF, AdaBoost). Out of three machinelearning model XG Boost algorithm provided the highest andimproved predictive accuracy. For all three datasets Accuracy (Measure Matrix) had increased to (7.35%,4.5%,4.2%) as compared to original Pfa algorithm [10].

IV. ANALYSIS OF PREVIOUS RESEARCH MODELS/ALGORITHMS, TOOLS AND INSTANCES.

A. Machine Learning Best Models/Algorithms and Evaluation Matrix with their Higher Accuracy Value.

In grade k-12 model identified the most relevant featuresfor better/successful performance

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

of students with accuracy value of 71.0% [11]. In 2017 in degree program to predict thestudent ‘s performance and produced N/A (superior performance to benchmark approaches) [12]. In 2018 Provided a road map for both academic stuff and students andgot RepTree TP rate value: 0.634, Precision (0.629), Recall (refers to a TP rate) with a value of 0.634 and a FP (0.409), J48 model applied provided more correlate features to the final class [1] and also provided MOOC Dropout Prediction.In the next year 2019 accuracy measured by several ways such as 1. Predicting student college commitment decisions with AUC score of 77.79% [14]. 2. Deep neural network applied for prediction of student’s performance 2019 and achieved accuracy of 84% [3]. 3. Students academia performance predicted into excellent or non-excellent with 89.26% accuracy (All classifier overall accuracy was above than 80%) [8]. 4. Adaptive recommendation suitable education path(s) by applied LR and QDA with F measure value of 0.91% [6].5. Some preventive actions taken in advance through which students successfully cope up with the course and applied MLP&RF classification model for achieved accuracy 92.3% (ROC Curve: - MLP-97.5%, RF- 98%.) [2]. 6. Evaluation of student’s performance by applying machine learning algorithm and produced The Mean Square Error: - 5 to 20% and the Effort Estimation wasaround 15-27% [13]. Therefore in 2019 accuracy measured between the range of 77.79% to 98% with Mean Square Error: - 5 to 20% and the Effort Estimation was around 15- 27%.

In 2020 accuracy produced by two ways 1. To recommend the best educational path to each and every student by using various Courses, the science-based intermediate programs & arts/humanities-based intermediateprograms produced best accuracy value by applying Light GBM algorithm (f-measure=100%, Cohen kappa= 100%,ROC Curve= 100%) [4]. 2. and also predicted the student’s native place by applying MLP, SVM model and produced accuracy value of 91.7% ,91.1% respectively. Therefore in 2020 range of accuracy between 91.1% to 100 %. In 2021 accuracy value measured between 80% to 99.5% and also accuracy evaluation metrics increased of (7.35%,4.5%,4.2%)as compared to original PFA algorithm as shown in Table 2.

TABLE 2. ACCURACY VALUES AND EVALUATION MATRIX VALUE FROM 2017 TO 2021.

	Description	
Year	***References No.***	***Accuracy***
N/A	[11]	71.0%
2017	[12]	N/A (superior performance tobenchmark approaches)
2018	[1]	RepTree TP rate value: 0.634, Precision: 0.629, Recall (refers to a TP rate) with a value of0.634 and a FP: 0.409
	[20]	n/a
2019	[14]	AUC score of 77.79%
	[3]	Accuracy = 84%
	[8]	89.26%
	[6]	LR (Linear Regression) andQDA F measure 0.91%

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

	[2]	accuracy: 92.3% (ROC Curve: MLP-97.5%, RF-98%
	[13]	The Mean Square Error = 5- 20% and the Effort Estimation was around 15-27%
2020	[5]	MLP accuracy (91.7%), SVM Accuracy (91.1%)
	[4]	(LGBM) F- measure =100%, cohen kappa= 100%, ROC Curve= 100%
2021	[10]	Accuracy metrics an increase of (7.35%,4.5%,4.2%) as compared to original Pfa algorithm.
	[16]	Accuracy exceeds 80%
	[18]	BiLSTM accuracy =90.16% (With feature selection)
	[9]	RF 91%
	[19]	RF (93%)
	[15]	RF F-measure of 99.5%

B. Machine learning Existing Accuracy Greater than 90%.

Machine learning based on existing model obtained accuracy greater than 90% or that produced higher measure matrix value. LGBM, RF, MPL&SVM, Linear Regression, BiLSTM (With feature selection) were the exiting best model that produced accuracy value greater than 90% [4-5, 9, 15, 19]. QDA F-Measure was the existing measure matrix that produced higher measure matrix [6] as shown in Table 3.

TABLE 3. OBTAINED ACCURACY OR MEASURE MATRIX GREATER THAN 90%.

References No.	**Description** ***Models (Accuracy or Measure/Evaluation Matrix >90%)***
[4, 9, 15, 18, 19]	LGBM, RF, BiLSTM (With feature selection)
[5]	MLP&SVM
[6]	Linear Regression, QDA F-Measure

C. Tools

Machine learning based on existing study used several tools for their studies. These tools helped in handling easily and effectively the large amount of computation complexity in lesser time. Weka 3.8 used as tool in reference number [1, 16], rapid miner 8.3 used as a tool in reference [8], python &its packages (scikit learn) used as tool in existing study reference number [14, 18], Python lib (tensorflow, SKlearn, numpy, seaborn) were the tools of reference number [9, 19], Pycart library tool used by reference number [19] and Python& anaconda IDE used by existing study as their tool
[17] as shown in Table 4.

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

TABLE 4. EXISITING STUDY TOOLS.

Sr. No.	Description of Exiting Tools	
	Tools	***References***
1	Weka 3.8	[1, 16]
2	Rapid miner 8.3 software	[8]
	Python & its packages (Scikit learn)	[14, 18]
3	Python library (Tensorflow, SKlearn, Numpy, Seaborn)	[9, 19]
4	Pycart library	[19]
5	Python& anaconda IDE	[17]

D. Instances

The previous research worker mostly considered in their studies less than 1500 instances/dataset and only few researchers considered the range of instances between 1501to 9000 to their studies. In the last a single research paper contained in its studies instances between 31000 to 33000 onthe basis of the research work. For the purpose of this study the data set range which is desirable should not be either lessor more as considered by the previous researchers as described in Table 5.

TABLE 5: RANGE OF INSTANCES WITH REFERENCES.

Instances	References
0-600	[1-3, 5, 10-11, 19]
601-1500	[8, 12-13,15, 18]
1501-3000	[4, 6]
3001-9000	[14, 16]
Upto 33000	[90]

V. CONCLUSION

Machine learning automates the large data and helps the minimizing the computation complexity resultant**.** Machine learning is used in education to handle several problems including dropout cases, retention cases and determining in advance at risk students, predicting and analyzing the student's performance. This analysis is focused on moderaterange of dataset meaning there by neither more nor less dataset as indicated in the previous researchers in their studies. Such study identifies the best algorithms that approves accuracy value above than 90% and also identify the tools that were used by several researcher .The resultant effect is that in2019 MLP (92.3%) ,RF(92.3%) [2] and LR, QDA (91%)[6] , In 2020 MLP (91.7%) , SVM (91.1%) [5] , LGBM (100%) [4] and in 2021 RF(91%)[9] , 99.5% [15] ,93% [19] and BiLSTM (90.16%) [18] had the machine learning algorithms that obtained accuracy value above than the 90%. and Weka 3.8 [1, 16], Rapid miner 8.3 software [8],Python & its packages (scikit learn) [14, 18], Python lib (tensorflow, SKlearn, numpy, seaborn) [9, 19], Pycart library [19], Python& anaconda IDE [17] were the Tools that were applied by the researchers in their works. This study is focused on above objectives which will be used by thispaper/study author in their upcoming research work. These analyses and the use of machine learning model also help theadministrator, teacher and parents to design and develop a student center curriculum and improving the teaching and learning skill as per the student need and also taking correct

action at correct time. It gives more focus on the performance of students so that in future they (students) may be capable for achieving the higher success rate and best performance. Thus, it goes a long way in bringing out the fact that main focus is on above objectives on selecting the best machine learning model and range of instances that will help this paper author in their upcoming research work on predictive analysis of the student's performance with higher accuracy value at the time of school/collages reopening into three categories excellent, average, and poor. All These aims and objects of this study can be achieved only as and when the suggestion enumerated in the foregoing discussion adhere to (follow) and implemented with the same sense and spirit by the institution/administration concerned.

REFERENCES.

[1] Hamoud Khalaf Alaa, Hashim Salah Ali and Awadh Aqeel Wid, "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis," *I.J. Intera. Multi. and Arti. Intell.*, Vol.5, pp. 26-31, 2018. DOI: 10.9781/ijimai.2018.02.004

[2] Aggarwal Deepti, Mittal Sonu and Bali Vikram, "Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques," *International Journal of Recent Technology and Engineering (IJRTE),* Vol. 8, Issue: 2S7, pp.496-503, 2019. DOI: 10.35940/ijrte.B1093.0782S719

[3] Vijayalakshmi V. and Venkatachalapathy K., "Comparison of Predicting Student 's Performance using Machine Learning Algorithms," *I.J. Intell. Sys. and Appl., Vol.* 12, pp. 34-45, 2019. DOI: 10.5815/ijisa.2019.12.04

[4] Dhar Joy and Jodder Kumar Asoke., "An Effective Recommendation System to Forecast the Best Educational Program Using Machine Learning Classification Algorithms," *I. Info.& Eng. Tech. Associ.,* Vol. 25, No. 5, pp. 559-568, 2020. https://doi.org/10.18280/isi.250502

[5] Verma Chaman, Stoffova Veronika, Illies Zoltatan, Tanwar Sudeep and Kumar Neeraj, "Machine Learning Based Student's Native Place Identification for Real Time," *IEEE Access*, Vol. 8, 2020. Digital Object Identifier 10.1109/ACCESS.2020.3008830

[6] Ezz Mohamed and Elshenawy Ayman, "Adaptive recommendation system using machine learning algorithms for predicting student's best academic program," *Education and Information Tech*, 2019. https://doi.org/10.1007/s10639-019-10049-7

[7] Enughwure Avwerosuoghene Akpofure and Ogbise Ebitiminipre Mercy, "Application of Machine Learning Methods to Predict Student Performance: A Systematic Literature Review," *I. Research Journal of Eng. and Tech. (IRJET),* Vol. 07, 2020. www.irjet.net

[8] Yaacob Wan Fairos Wan, Nasir Md Azlin Syerina, Yaacob Wan Faizah Wan and Sobri Mohd Norafefah, "Supervised data mining approach for predicting student performance. *Indonesian Journal of Elect. Eng. and Comp. Sci.,* Vol. 16, pp. 1584-1592, No.3, 2019. DOI: 10.11591/ijeecs.v16.i3.

[9] Adan Mumhammad and Ashraf Jawad, "Predicting at risk student at different percentage of course length for early intervention using machine learning model," *IEEE Access*, 2021.

[10] Asselman Amal, khaldi Mohamed and Aammou Souhaib, "Enhancing the prediction of student performance based on the machine learning*," Routledge Taylor and francis group,*

2021.

[11] Harvey L. Julie and Kumar A.P. Sathish, "A Practical Model for Educators to Predict Student Performance in K-12 Education using Machine Learning," *ACEDMIA Accelerating the world's research.*

[12] Xu Jie, Moon Ho Kyeong and Schaar van der Mihaela, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," *IEEE.* DOI 10.1109/JSTSP.2017.2692560

[13] Kumar Mukesh, Singh J. A., "Performance Analysis of Students Using Machine Learning & Data Mining Approach," *International Journal of Engineering and Advanced Technology (IJEAT),* ISSN: 2249 – 8958, Vol. 8, Issue: 3, 2019. Retrieval Number: C5708028319/19©BEIESP

[14] Basu Kanadpriya, Basu Treena, Buckmire Ron and Lal Nishu., "Predictive Models of Student College Commitment Decisions Using Machine Learning," *Vol. 4, MDPI,* 2019. doi:10.3390/data4020065

[15] Bujang Abdul Dianah Siti, Selamat Ali, Ibrahim Roliana, krejcar ondrej, Herrera-Viedma Enrique, Fujita Hamido, and Ghani MD. Azura NOR, "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," *IEEE Access*, Vol. 9, 2021. Digital Object Identifier 10.1109/ACCESS.2021.3093563

[16] Palacios A. Carlos, Reyes-Suárez A. José, Bearzotti A. Lorena, Leiva Víctor and Marchant Carolina, "Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile," *Entropy, MDPI,* 23, 485, 2021. https:// doi.org/10.3390/e23040485

[17] Karthikeyan R., Satheesbabu S. and Gokulakrishnan P., "Machine Learning Based Student Performance Analysis System," *IT in Industry,* 2021, no.1, vol. 9.

[18] Yousafzai Khan Bashir, Khan Afzal Sher, Rahman Taj, Khan Inayat, Ullah Inam, Rehman Ur Ateeq, Baz Mohammed, Hamam Habib and Cheikhrouhou Omar, "Student-Performulator: Student Academic Performance Using Hybrid Deep Neural Network*," MDPI*, 2021, Vol.13, 9775. https://doi.org/ 10.3390/su13179775

[19] Kabathova Janka and Drlik Martin, "Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques," *Appli. Sci., MDPI, 2021.*https://doi.org/10.3390/app11073130

[20] Dalipi Fisnik, Imran Shariq Ali and Kastrati Zenun., "MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges," EDUCON 2018. htttp://10.1109/EDUCON.2018.8363340