Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7674227

A REVIEW STUDY OF STUDENT PERFORMANCE USING MACHINE LEARNING AND ITS TECHNOLOGY

Kirti
Computer Science and Application Department, Baba Mastnath University Haryana, India (Keerti.bk11000@gmail.com)

Alok Sharma
Associate Professor, Computer Science Department, Baba Mastnath University Haryana, India (alok.ghritkaushik@gmail.com)

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 determining the range on average bases somewhere on the point neither more nor 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 of previous research work there were seven algorithms whose accuracy 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.

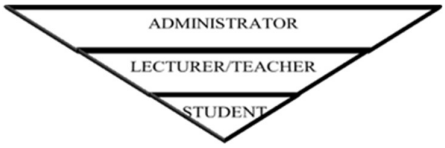

Fig.1. Three Pillars of Education System

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 and feedback 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 taking a 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 features and 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 also helped the administrator, parents, students to know about the lagging 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 to improve 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 predictive analysis. Large amount of data (school/ university record) not only 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 integrated into 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 as SMOTE with FS (feature selection), BiLSTM (Bidirectional Long 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 build and 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].

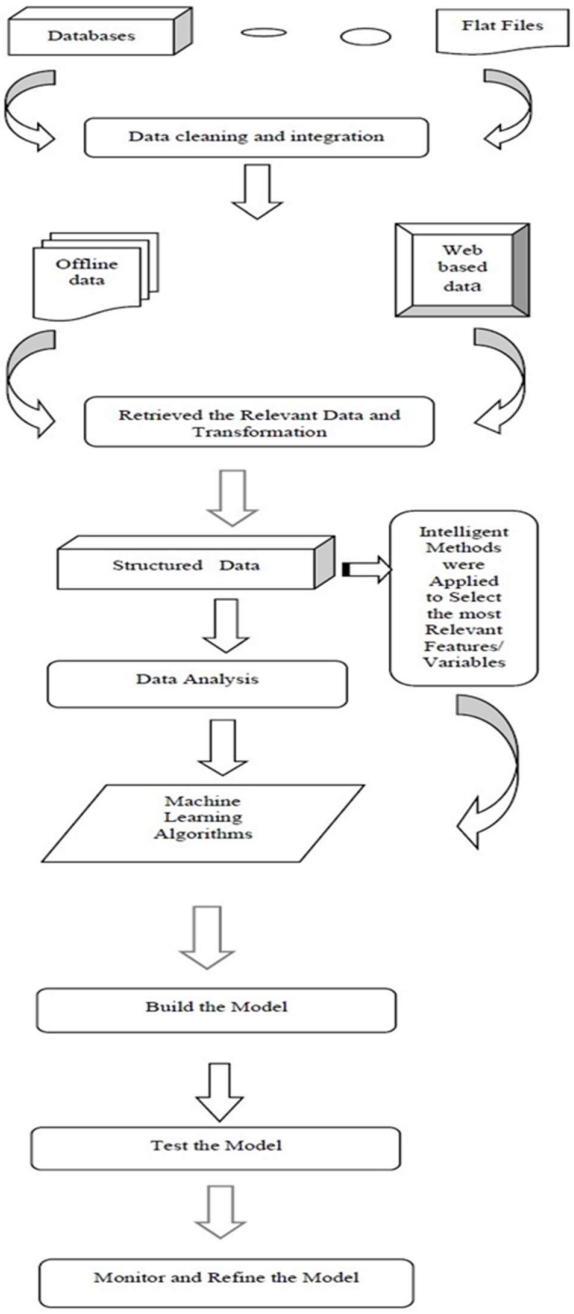

Fig.2. Structure & Steps of Predictive Analytics

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.
- To identify the exiting best model's accuracy or measure/evaluation matrix greater than 90% for improving the predictive model accuracy and generated the high quality of student's performance into three categories such as excellent, average and poor.
- To identify the existing number of instances.
- To identify the existing range of instances that applied by neither more nor less (modest) previous researchers in their studies.

III. LITERATURE REVIEW

A. Literature Review of Previous Reviewed paper.

Review of the paper of previously reviewed research papers.

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 their student 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 and skills)and MOOC based factors/variables (Isolation and lack of interactivity, Course design, Hidden costs)will result into very high degree of drop out cases. To control such drop out cases , some solution being introduced such as Clickstream data standardization (MOOC contained several traceable events and interactions from various audio & video devices 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 accuracy predictive model [20] as described in given Table 1.

TABLE 1. LITERATURE REVIEW Of PREVIOUS REVIEWED WORK.

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 andalso presenttheir reasons.	2021	school	200
[20]	MOOC Dropout Prediction.	2018	Online courses	0ut of 641138 instances17687 instances had obtained certificate

B. Literature Review of Recent/Previous Research Work.

In Education system several changes have occurred from time to time. One of the bigger changes was: Used education system with machine learning not only for prediction purpose but also for predictive analytics purpose proved/described by several 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 purpose but 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 Support Vector 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 course have 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 by use of three different datasets and three machine learning algorithms (XG Boost, RF, AdaBoost). Out of three machine learning model XG Boost algorithm provided the highest and improved 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].

Education system (higher education) had developed from time to time. According to the need of time recommended them (students who enrolled in college) the best educational path. Adaptive recommendation system used five machine learning model (SVM (support vector machine), RF (Random Forest), QDA, LR and KNN (K-Nearest Neighbor). Different best algorithms had Different departments. SVM algorithm with F-measure value 0.73 was selected the best model for the Computer Department, RF algorithm with F- measure value 0.78 was selected the best model for the mechanical Department, QDA algorithm with F-measure value 0.91 value was selected the best model for the Urban Department, LR algorithm with F-measure 0.91 value was selected the best model for the Mining and Petroleum, KNN algorithm with F-measure value 0.91 selected the best model for the Urban Department and the proposed predictions system average performance was 82.57% [6].Prediction of Student Academic Performance had focused on BiLSTM (Bidirectional Long Short-Term Memory) combined with an attention mechanism and feature extraction method which is basically used for effectively predicting the students grade and improving the performance with an accuracy value of (90.16%) [18]. Academic institute spends their lots of money for providing better resources and facilities to candidate/students in future but sometime they suffered huge losses due to chosen improper candidate (drop out reason, at- risk students, student’s retention problem) that is why academic institute want to know/understand in advance the candidate’s (offered admission) decision power. In higher education system predicting the student’s college commitment decisions by use of seven algorithms(Naïve Bayes (NB),Logistic Regression (LG),Decision Trees (DT) , K-Nearest Neighbors (K-NN),Support Vector Machine (SVM),Random Forests (RF),Gradient Boosting (GB)) out of which Logistic regression algorithm had outperformed with an AUC score value of (79.6%) [14].To know the condition of student(who had re-appear) beforehand(at early stage) so that at correct/right time important and corrective/right action may be taken by parents/ teachers and reduce the failure rate at the end of the Course .For this purpose used

131 instances , 22 attribute(used 6 features after feature selection method (gender, number of friends ,caste, family income, mother’s occupation and percentage in XII.)) and six algorithm(Naïve Bayes, Multilayer Perceptron, Logistic regression, Random Forest ,Support Vector Machine and J48

) among which Naïve Byes algorithm / approach used for feature selection purpose .After then the performance of six algorithm compared and predicted the best model (Random Forest classification algorithms and Multi-Layer Perceptron with accuracy of 92.3%). by using Multi-Layer Perceptron (MLP) author got the minimum Relative absolute Error (22.4%) for

identifying whether students got success or not (re-appear) in a course. Multi-layer perceptron provided the best result when used/applied naïve bayes algorithm as feature selection purpose [2].

Decision tree was the most popular method to build best and effective model. In higher education with the help of decision tree algorithm (RepTree, J48, Random Tree) student's success and failure rate predicted and categorized in relation to input features /variables. Decision tree algorithm (J48) provided a high predicted accuracy and directed(destination) an effective road map for listeners(students) and academic stuff(lecturers). RepTree algorithm was the nominated method for the model (The TP rate with the highest value of 0.634, Precision with the highest value of 0.629 , Recall value was of 0.634 and FP rate was value of 0.409) whereas J48 method was the best approach for providing more correlated features to the final class [1].To extract the most relevant attributes in the data four supervised machine learning algorithms were applied out of which Naïve Bayes algorithm reveals that it was the best algorithm to select the most relevant factors .Based on accuracy measure, recall ,precision and ROC curve predicted the student performance either into excellent (in Multivariate Analysis:- A+ and A ; in ITS 472 :- A, A- and B+ and in SAS Programming:- A+, A and A-)or not [8].

There were many students in college who were not successfully completed their courses at exact time (duration of course) and suffered in between the courses. To control such situation, it become necessary to identify at risk students in advance by using dissimilar percentages of Course length. Seven algorithms were used (SVM, Extra tree classifier, KNN, Ada Boost Classifier, Gradient boosting, RF Training predictive
model used DFFNN) out of which RF model was the best prediction model with an accuracy value of 91% [9].To Predict the performance of student in future in degree program on the basis of their past and current performance by applying first ,the most relevant feature based clustering method for choosing right courses with a view to develop sufficient and effective based and ensemble-based prediction of architecture and second, probabilistic matrix factorization and a data-driven approach[12].Predictive analytics reduces the chances of inaccurate result by changed imbalanced dataset into balanced dataset and generated meaningful information to produce high quality of performance. For predictive analysis, constructed six predictive model (Decision Tree (J48), Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF) and Logistic Regression (LR)) with multiclass prediction model to predict the final grades of students on the basis of final examination 's of the previous student performance by applying SMOTE with two FS (feature selection) technique. RF (Random Forest) model to increase
/raise the prediction performance with f-measure value of 99.5% [15].

Data mining and Machine Learning model (decision trees, logistic regression, k-nearest neighbors, Naïve Bayes, support vector machines and random forest) used for Predicting the Student's Retention in higher education not only for final year but also for first, second year of the Course. Random forest model had outperformed and produced accuracy that exceeds 80% and also provides the best result in terms of true positive rates and false positive rate (10 to 15 % in most cases), F-measure ,precision, , κ-statistic and mean squared error[16].In higher education prediction of the Student's Dropout was achieved by use of five machine learning model RF (random forest), DT(decision tree), SVM(support vector machine) , NN (neural

network), LR(Logistic regression) algorithms) and CRISP-DM Methodology with McNemar'test. Among five model RF (random forest) algorithm/model provided the best result with accuracy (value of) 93% [19]. On web-based learning environment it is difficult to know about the mental ability of the learner's that is why the native place of students of real time of two different countries (Indian and Hungarian) easily identified through this advanced technology. In first experiment optimized MLP produced the best prediction accuracy (91.7%) on 10 folds with time 0.2 seconds. Regularization boosted the predicting accuracy and stabilized up to 92.3%. Similarly in experiment 2 SVM (support vector machine) produced 91.1% accuracy along with regularization (alpha=0.0001) on 5 and 337 folds and in experiment three using selected PCA 13 novel features. experiment 3 predicted model enhanced by use of more data pattern as compared to previous experiment data patterns. Therefore, SVM identified that strength raised up to 2.9 and also increased MLP accuracy with 2.3%. Top 11 features identified with info gain and Gain-ratio methods [5] as described in given Table 2.

TABLE 2. LITERATURE REVIEW Of PREVIOUS RESEARCH WORK.

Description									
Ref. No.	***tudy basedon***	***year***	***School/Uni.m (Collage)***	***Instances***	***Methods***	***Best Methods***	***Best method's accuracy/Highest accuracy***	***No. of attri.*** [a]	***Gap /deficiency***
[1]	Questionnaire based study (Provide a road map for both academic stuff and students)	2018	Collage (Higher-Education.)	It collected 161 questionnaires.	Random Tree, J48 and REP Tree.	Rep Tree, J48 algorithms.	Rep Tree TP rate value: 0.634. Precision: 0.629. Recall value: 0.634 and an FP rate value: 0.409. and J48 model was provided more correlate features to the final class.	It contained 5 least effected and 5 most effected attributes on student's success. It Features were: Demographic Data, Social Information, Academic Information, Study Skill, Motivation, Personal Relationship, Health, Time Management, Money Management, Personal Purpose, Career Planning, Resource Needs, Self_ Esteem.	Large dataset and high influenced attributes that effects the accuracy of decision tree.
[2]	Take some	2019	Collage	131 students/	Naïve Bayes,	MLP&RF	MLP&RF	6 important	Limited

	preventive actions in advance.		/university	instances, 22 attributes & 6 algorithms. - Data preprocessing technique was: naïve byes	Logistic Regression and J48 decision tree, SVM [b], Multi-layer perceptron and RF [c] algorithms. -Feature's Selection method was: Naïve Bayes.		classification accuracy were around: 92.3% (ROC Curve of both methods was around: MLP-97.5% RF-98%.)	features were: 1.Gender (ge) 2. Caste (cst) 3.Percentage in XII (twp) 4. Family income (fmi) 5.Mother's occupation (mo) 6. Number of friends (nf)	dataset with limited pre-processing techniques.
[3]	Deep neural network used to predict the student's performance.	2019	School	Out of 500 instances it used/selected 480 instances.	Decision Tree (C5.0), NB [d], SVM, Deep neural network, RF and KNN [e]	Deep Neural Network.	84% accuracy.	It selected 16 features that grouped into 3 primary classes: 1.Demographic features, 2. Academic background features and 3. Behavioral features.	Less no. of accuracy.
[4]	To Recommend the best educational path to each and every Students.	2020	High school	2404	LGBM, GB, KNN, XGB, CatBoost (CATB), DT (decision tree), LR (Logistic Regression), RF and GNB [h]	Different course has different best algorithm by measure /evaluation matrix.	Science-based intermediate programs & arts/humanities-based intermediate programs were the best Courses with best algorithm i.e. Light GBM algorithm. Evaluation matrix accuracy was: F-measure=100%, cohen kappa= 100%, ROC Curve= 100%	8 features had given below: Physical Science, Mathematics, Bengali, Life Science, Information Technology, English, Geography, and history.	Less prediction accuracy due to limited datasets.
[5]	Predicting the student's native place.	2020	University	Out of 331 instances 168 inliners detected.	Multi-Layer-Perception (MLP) with 3 popular	MLP, SVM both had contained minor	MLP accuracy: 91.7%, SVM accuracy:	It used 37 features: Development availability-	Limited algorithms with a limited

					optimization algorithms: 1.SGD (Stochastic Gradient Descent), 2.LBFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) and 3.Adam (Adaptive Moment Estimation), SVM	difference. (It was almost same)	91.1%	(contained 16 features), Attitude feature - (contained 6 features), Usability - (contained 6), Educational benefit - (contained 9 features)	number of extraction techniques and used fewer samples from two Countries.
[6]	To recommend the best educational path(s)	2019	university	1841 instances had reduced to 1231 instances after applying the discretization process.	SVM, KNN, QDA, RF and LR	LR algorithm for Mining and Petroleum and QDA algorithm for Urban Department.	LR and QDA: F-Measure value of 0.91%	The dataset consists of two main parts: 1. Preparatory year course-related data. 2.Final year grades and scores. This paper contained6 Domains: 1. Urban planning, 2. Mechanical, 3. Mining and petroleum, 4. Electrical, 5. Systems and computers, 6. Production Civil.	Less effective due to Limited dataset.
[8]	Student's performance predicted into excellent or non-excellent.	2019	university	631	Logistic Regression, KNN, Decision Tree and NB Model.	Naïve Bayes	89.26% (Overall accuracy of all classifier above than the 80%)	The most relevant attributes were: student name, gender, student ID, final CGPA, and Course's grades of all Courses were joined by the students.	Limited scope courses used limited and high influenced attributes.

[9]	Using dissimilar percentages of Course's length for Predicting atrisk students	2021	university	32593 instances (3 colum nwith 32593 row) and 47 variables fed to the DFNN	SVM, Extra tree classifier, KNN, Ada boost Classifier, GB[f], RF Training predictive models were used DFNN	RF	91%	Attributes used in it: Demographic, Assessment, Virtual learning environment, Course registration, Course offered.	Insignificant dataset due to lack of activity datasets.
[10]	To raise the predictive accuracy of student's Performance.	2021	Middle and High school.	245 instances (3 different datasets were considered.)	AdaBoost, RF, XGB[g] (3 dataset swere Splitting the sample/instances by use of 10 k fold cross.)	XG Boost	3 dataset's accuracy metrics had increased of (7.35%,4.5%,4.2 %) as compared to original Pfa algorithm.	123 skills in2nd dataset and 119 skills in3rd dataset (each student had contained20 skills and each was answering between 8 to 15 questionsper skill.)	Insignificant and less accurate result due to computation complexity.
[11]	In grade k-12model identified themost relevant features	N/A	school	403	Decision Tree, LR[k] andNaive Bayes techniques.	Naive Bayes	71.0%	27 variables had selected. (Domains of variables were: Student demographics, and assessment data, salariesof educators.)	Lack of classifier techniques (algorithms) for analysis and insufficient model. (Less accurate result by use of LR model.)
[12]	In degre eprogram to predict the student's performance.	2017	University (Degree collage.)	1169	Linear Regression, RF, Logistic Regression, KNN	RF	N/A (Superior performance as compared to benchmark approaches)	First, structure with two layers including multiple base predictors and ensemble predictors Second, probabilistic matrix factorization and a data-driven approach.	Lack of elective Courses with limited recommend ed/ core Courses prediction result to student.
[13]	Evaluation ofthe student's performance by using of machine learn Ing.	2019	school	10000	K-Means-SVM and K-Means-ANN[i]	Performanc eof K-Means-ANN	The Mean Square Error was: 5-20% and the Effort Estimation was:15 to 27%	34 different attributes (School attribute/variable and subject marks attributes)	Insufficient neuron and parameters.

[14]	Predicting ofstudent's college commitment decisions.	2019	University(collage)	7976	Naive Bayes (NB), Logistic Regression (LR), Decision Trees (DT) ,KNN, SVM, RF, Gradient Boosting(GB)	LR (Logistic regression) classifier	AUC score value of 77.79%	It used 15 features: Direct Legacy, Financial aid intent, scholarship, Ethnic Background,	Lack of ensemble technique that require increase amount of storage and computation time.
								Extracurricular Interests, First Generation, GPA, Top Academic interest, numerical, Class Size, interview, Permanent State /Region,Level of Financial need reader Academic rating.	
[15]	To predict the Student's Grades by used o f Multiclass Model	2021	University (Collage)	1282 instances In this study used: 641 students/instances (CSA &ICS).	Decision tree(j48) , Naïve Bayes (NB), SVM, KNN, RF andLR [1].	RF	F-Measure:99.5%	Wrapper-Subset Evaluation based: 6 FS 1 Variables had used, Classifier Subset Evaluation based: 5 FS 2 variables (with best firstsearch) hadused Filter-based: (Info-Gain Attribute Eval) 4 FS 3(with rankersearch method) variables had extracted and used.	Lack of predictive techniques to optimize the result for prediction student's grades & Lack of sampling techniques and different evaluation metrics for analyze multiclass imbalance datasets.
[16]	Student's Retention cases.	2021	Higher education.	6656	Decision Trees, SVM, KNN, naiveBayes, Logistic Regression and RF.	Random Forest technique.	Model's accuracy abovethan the 80%that produced the better output than the first three models (global, first-level, and second-level)	Out of 165 features the first 20 features which had contained the highest IG scores in each case were selecting.	Third level model has the less accurate model due to excessive difference between the number of instances used to train the model.

[18]	Student's Academic Performance with the use of BiLSTM method.	2021	School	1044	Classifier Model (SVM,LR, NB, KN, RF) and DL Model (RNN, CNN, LSTM, BiLSTM [j] and the Exiting Baseline -Methods were: ML(SVM) (Imran e tal.2019), ML(SVM) (Sultan et al.2019)	BiLSTM method combined with attention mechanism and produced Better/superior performanceas comparedto the existing state-of-the-art.	The BiLSTM method achieved the accuracy of 90.16% (With features selection) and 88.46% (Without features selection.)	Out of the 12 features suchas: G1_f,G2_f, Absent,Father_job, Fails, Mjob, F_education, Schools up, Medu, Study hour, HealthTop 10 features were Selecting.)	Limited Dataset as well as limited features.
[19]	Student's Dropout prediction	2021	University (collage)	261	Method used in was CRISP-DM Methodology with Mc Nemar'test and classification results were obtained by using RF, DT,SVM, NN (neural network), LR (Logistic regression) algorithms	RF	93%	It used following features: Access, exams, tests, assignments, project (Features aggregation was based onthe Course topic and weeks during which semester wereexamined.)	Insufficient to accurately identified at risk students.
[20]	MOOC Dropout Prediction (Review and Research Challenges)	2018	Online courses	It used 641138 instances out of which17687 instances hadobtained certificate.	N/A	N/A	N/A	N/A	Insufficient to accurate identify the MOOC drop out students.
[21]	Forecasting college commitment decisions of students.	2021	Higher education	In this study used: 280 students	RF Training predictive models were used DFNN	Deep Neural Network.	88% accuracy.	Direct Legacy, Financial aid intent, scholarship, Ethnic	Inadequate for effectively identifying at-risk students.
[220]	Assessment of student achievement using learning algorithms.	2021	Online courses	1028	BiLSTM method.	N/A	N/A	N/A	Lack of predictive techniques to optimize the result for prediction student's grades

a. Attributes, b support vector machine, c Random Forest, d Naïve Bayes, e K-nearest neighbor, f Gradient Boosting, g XG Boost, h Gaussian Naïve Bayes, i Artificial Neural Network, j Bidirectional Long Short-
Term Memory, k Linear Regression, l Logistic Regression, m University.

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 features for better/successful performance of students with accuracy value of 71.0% [11]. In 2017 in degree program to predict the student ‘s performance and produced N/A (superior performance to benchmark approaches) [12]. In 2018 Provided a road map for both academic stuff and students and got 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 was around 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 intermediate programs 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 3.

TABLE 3. ACCURACY VALUES AND EVALUATION MATRIX VALUE FROM 2017 TO 2021.

Year	**Description**	
	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%
	[2]	accuracy: 92.3% (ROC Curve:MLP-97.5%, RF-98%
	[13]	The Mean Square Error = 5- 20% and the Effort Estimationwas around 15-27%
2020	[5]	MLP accuracy (91.7%), SVMAccuracy (91.1%)
2021	[4]	(LGBM) F- measure =100%, cohen kappa= 100%, ROC Curve= 100%
	[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 4.

TABLE 4. 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 5.

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]

TABLE 5. EXISITING STUDY TOOLS.

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 on the basis of the research work. For the purpose of this study the data set range which is desirable should not be either less or more as considered by the previous researchers as described in Table 6.

TABLE 6: 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. EDUCATION SYSTEM PROBLEMS.

A. Machine Learning Handled the Following Education System Problems of Previous Research Work.

The foregoing study dealt with the previous research work that mainly focused on student performance in relation to machine learning being used to handle the several types of education problems mainly relating to the cases about student's performance /success rate [2-3],[7-8, 12, 15].It also relates to the improvement in predicting accuracy of student performance [10].This study is also based on analysis and forecasting the student's performance [10, 13].It also includes the studies on drop out[19-20] along with recommendation for the best educational path[4][6] suggesting the ways and means on student retention problem
[16] of risk student their success and their failure[1]. It also deals with selection of most relevant features [11] and identifying the native place of the students at real time as shown in Table 7.

TABLE 7. PREVIOUS RESEARCH WORK BASED ON THE FOLLOWING CASES.

Sr. No.	Description	
	Cases	*References*
1	Drop out.	[19-20]
2	Retention.	[16]
3	At risk.	[9]
4	Analysis and forecasting.	[10, 13, 17]
5	Success/fail.	[1]
6	Recommended and adaptive recommended the best educational path.	[4, 6]
7	Identified the native place at real time.	[5]
8	Prediction.	[2-3, 7-8, 10, 12, 15]
9	Selected the Most relevant features.	[11]

VI. LIMITATION

A. Limitation of Previous Research Work.

The studies deal with the research work suggesting some solution by researcher by applying preventing action in advance where students may successfully cope with the problem facing in the Course [2]. Also recommended the best educational path [4][6] for student's academia performance, excellent or non-excellent [8]. It also deals with dissimilar percentages of course length for predicting at risk student [9], improved Student academic performance using BiLSTM method [18], provide a road map for both academic stuff and students [1], Evaluation of student's performance by using machine learning [13]and MOOC Dropout prediction [20]. But due to limited dataset [2, 4, 6, 8-9, 12, 18] and insufficient dataset [1, 13, 20] these methods would be least sufficient. Some previous researchers concentrated on minimizing the problems faced in this field by applying deep neural network for prediction of student performance [3] recommending the best educational path [4] to improve the predictive accuracy of student's performance [10]. In grade k-12 model identifying the most relevant feature for successful

performance of the students [11], student's drop out prediction [19] but due to incomplete/less predictive accuracy [3-4, 10-11, 19] and computation complexity [10] the researchers were not fully successful. The researcher also tried to manage this problem as shown in Table 8.

TABLE 8. LIMITATION OF PREVIOUS RESEARCH WORK.

Sr. No.	Deficiency in the Existing Research	References
1	Limited dataset.	[2, 4, 6, 8-9, 12, 18]
2	Less/limited accuracy.	[3-4, 10-11, 19]
3	Computation complexity.	[10]
4	Insufficient dataset.	[1, 13, 20]
5	Limited algorithms.	[5, 11]
6	Limited pre-processing or extraction techniques.	[2, 5]
7	Lack of ensemble techniques.	[14]
8	Less /limited predictive sampling techniques.	[15]

VII. 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 moderate range 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 this paper/study author in their upcoming research work. These analyses and the use of machine learning model also help the administrator, 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.

ACKNOWLEDGEMENTS

I am thankful to my university computer science teaching staff who provides me their great guidance and supports.

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