RANDOM FOREST CLASSIFICATION APPROACH FOR THE STUDENT PERFORMANCE ANALYSIS

Harminder Kaur [1], Amandeep Kaur Sohal[2] and [3] Er.Gurjit Kaur Kahlon
Student, Assistant Professor and Assistant Professor
Computer Science and Engineering Department
Guru Nanak Dev Engineering College, Ludhiana-141006, Punjab, India

Abstract: The institutes can improve their services through this service. To evaluate the performance of students in institutes, various techniques are designed. However, for ensuring that good education is provided, the performance of teachers is evaluated. For predicting the student's performance, the existing system applied back propagation approach. This research aims to design the hybrid classifier such that the performance of students can be predicted in highly effective manner. Classification and clustering are integrated in this hybrid model. Clustering is applied using k-means algorithm and final prediction is performed using random forest. The dataset is extracted from UCI repository having attributes such as gender, age, mother education, father education and so on. With respect to certain performance metrics named f-measure, precision, recall and accuracy, the proposed model is implemented in Python.

Keywords: K-mean, Hybrid, Random Forest

I. INTRODUCTION

For every educational institute available today, providing improved and advanced education related to every student is the major objective. For making educational level of improvements, various newly developed and updated modes of education are applied. Through the development of new technologies, the performance of students is evaluated. Data mining approaches help in predicting the performances of students. To measure the performance of students, various attributes are being studied .The data storage has no limit. So, identifying any specific kind of information from the huge databases is very difficult and time consuming. There are databases, hard disks, CD drives etc. available for users to store their important data. Data mining is the technology that is designed to extract the data based on its necessity and relevancy. Knowledge Discovery in Databases (KDD) is the other name for data mining [1].The data that is implicit, previously stored and used most frequently is extracted in a non-trivial manner through this technology. For finding the necessary information, several techniques are designed over the years. For several applications like user modeling, trend analysis, domain modeling and user grouping, data mining is applied [2]. Over the past few years, the Educational Data Mining is growing its trend. New methods are designed and previously designed machine learning, data mining techniques are applied by the researchers so that the collected education data from teaching and learning can be used such that the behavior of students can be understood [3].The performance of students is improved by investigating the educational data. Creating new techniques has become a necessity since the capacity of identifying hidden information related to students and using web-based learning are required as per the advancements in educational institutes and expansion in computing facilities. Focusing on the important patterns and recognizing the important and relevant data from educational information systems is the major objective of this research [4]. The syllabus management, course management, registration and admission system are some of the important applications. The students of various stages of education institutions are handled using these systems and projects [5].For helping the educational fields in managing the smooth working of their institutions, the relevant information is discovered and determined by the researchers. The activities are managed such that the better and improved functioning of institutions can be generated[6].

The student's data is classified using association rules in EDM by studying the data and information about the student [7].The performance data is analyzed and the functioning of educational institutes is managed using this process. The facilities need to be provided appropriately and sufficiently and their education system must not include any issues which can be ensured through this method [8].A common data mining technique through which the educational institutions are classified based on their performance is known as classification. EDM is applicable in varieties of applications. User modeling is one of these applications through which the experience of user, behavior of learning, and the way through which the users can be satisfied can be known by a learner. In education, user behavior modeling is considered to be another application. Similar kind of learning data that are utilized when the user knowledge was predicted help in characterizing the engagement of student. Profiling is another application through which the users are grouped among various categories with the help of prominent properties. Domain modeling that is recently designed by EDM is the model that is used to understand a current topic in a particular level of detail. For studying the information collected related to the student through a survey and performing classification based on the gathered data, Educational Data Mining (EDM) is applied. The performance of student is classified and predicted in their upcoming semesters or examinations through classification.

Identifying the relationship among the personal and social factors of students such that their academic performances can be analyzed is the major objective of this research. The institutions and students work is a synchronized and organized way to facilitate this. The students which are underperforming and which are performing well in academics can be known by analyzing the performance of each student [9].

The various methods, through which course management systems (CMS) data can be mined such that new kinds of student behaviors can be provided, are examined using EDM. The learning and sustaining educational processes can be improved for providing help to the staff such that the institutional effectiveness can be improved. Following are some of the common applications of EDM:

a. **Student Retention and Attrition**: The students at risks are detected and institutions are encouraged to be more proactive for responding and identifying those students through EDM. The kinds of students that could drop out of school and would return later can also be predicted through data mining approach. Since data mining tools are applied successfully for supporting student retention endeavors, this research is considered to be necessary.

b. **Personal Learning Environments and Recommender Systems:** EDM can be directly related to through the personal learning scenarios and recommendation systems. For accommodating the system to the learning needs of students, several tools, services and artifacts are designed by the personalized learning scenarios. Since the recommendations coincide with the educational objectives, it is important to adapt recommender systems when applying them in educational contexts. Since the recommender systems are highly domain dependent, assigning the existing recommender system plainly to the educational data is not probable.

c. **EDM and Course Management Systems**: The course management systems and the level with which they can be improved such that student learning outputs and their success can be supported is the major focus of various researches designing EDM. A rearranged data mining toolbox can be built up to be applied within the course administration framework through which data mining information related to their courses can be permitted to non-expert users. An important aspect of student's success within an online educational scenario is the learner commitment. The information mining strategies can be used to investigate the engagement of students with the course substance[10]. Deciding it there are any disengaged learners can be done through this process.

II. LITERATURE SURVEY

Jie Xu, et al. (2016) proposed a new machine learning technique to predict the performance of students in degree programs. The proposed approach had two important features. The multiple base predictors were incorporated to generate a bi-layered structure [11]. This structure was the primary feature of proposed approach. In addition, as per the growing performance conditions of students, a cascade of ensemble forecasters was designed to execute forecasting. For identifying the significance of course, a data-driven approach was designed as a second feature. This approach utilized probabilistic matrix and latent factor models. Various simulations were carried out on a dataset gathered at UCLA over three years. The obtained outcomes depicted that the performance of proposed approach was better in comparison with standard techniques.

S. M. Merchán, et al. (2016) proposed a novel predictive model for predicting academic performance of students. A number of data mining methods were implemented on the data suite of 932 students of Columbia University for evaluating and analyzing their performances [12]. In past, various classification techniques had been implemented for constructing a predictive model in academic atmosphere. On the basis of results obtained in each iteration procedure, the scrutiny of earlier executions was completed as an iterative detection and learning procedure. On the basis of provided input data, the probable results, output description and other aspects, the evaluation of obtained outcomes was carried out. The accuracy of forecasting was an imperative factor which was also used for students' performance evaluation. Taking into account the particular aspects of the population scrutinized and the necessities stated by institute, the assumed applicability was evaluated. In order to prevent any sorts of academic risk and desertion, well-timed decisions were considered significant along with the assistance of students during their learning procedure. This investigative study was developed more through some proposals and opinion of various examiners.

Ishwank Singh, et al. (2016) proposed a simple clustering analysis to develop understanding regarding the behavior of students [13]. For example, if a regular improvement was noticed in the performance of student, a good standard had been set up by the data mining algorithm. During the admission and appointment procedure, this scrutiny was quite supportive. The projects, placements, skillfulness, Xth, XIIth, and B.Tech marks were some of the parameters involved in this analysis. As the execution was simple and the computational competence was high, K-means algorithm was utilized for clustering purpose. In future, more clustering methods can be implemented for enhancing competence levels. In addition, for achieving an improved student performance analysis, the objects occurring within clusters can be classified or ranked.

Ms. Tismy Devasia, et al. (2016) proposed classification within the information of student to forecast the division of students on the basis of earlier existing information. In this study, Naïve algorithm was implemented as various approaches were utilized for knowledge classification inside the area module [14]. In order to predict student's performance on the top of that specific semester, several types of information were gathered from the earlier existing information of students. For encouraging the students of diverse classes regarding their performances, the professors and students could get benefit with the help of this study. The students requiring any particular direction could be enlightened through this approach. Moreover, the failure ratio could be decreased using this study. For the upcoming semester exam, satisfactory measures could be applied using this approach.

Yuni Yamasari, et al. (2106) proposed a novel feature extraction methodology in this study. The student data was collected in a serious way for information retrieval on the basis of class and Bloom's Taxonomy [15]. The proposed approach was applied on this data for carrying out several assessments. It was identified that the precision level was enhanced and execution time was minimized with the help of this approach. In comparison to conventional FCM, the level of accuracy was improved up to 2.3-4.7%. Moreover, in comparison to conventional scheme, the execution time was 2.2-2.7 seconds quicker for the proposed FCM approach. The performance of clustering procedure on the success of student was improved by retrieving features through CBE_FCM and BTBE_FCM approach. Weight was added to every feature to improve the performance of proposed techniques. The association level of student success was regarded as an imperative dynamic in this study.

III. RESEARCH METHODOLOGY

To predict the performance of students, back propagation is known to be an efficient technique which is a neural network based approach. Based on the previous experiences, the back propagation is applied which helps in generating new values for future. Until there is least error in the prediction, the propagating process will run in iterative manner. Classification method is applied in this research to predict the performance of students. With maximum accuracy, the performance of students can be predicted through the classification approach.

Research Gaps

Following are the various research gaps :-

1. This research work is related to student performance analysis. The datasets of student performance analysis is very large in numbers due to which execution time is very high which needs to be reduced.

2. The technique of classification which are proposed in the previous research work give low accuracy which need to improved in the proposed methodology. The research methodology designed for this research includes certain steps which are:

Step 1:- The dataset from UCI repository is extracted and given as input. This dataset is related to the performances of students and teachers over the years for a particular institute. There are 33 attributes and 649 instances available in this dataset. The cleaning of dataset is done and no missing values are available.

Step 2:- To cluster the dataset, the second step applies k-mean clustering method. Here, based on the similarity the data is clustered and to select the number of clusters, k-mean clustering is performed.

Step 3: The output achieved from clustering is given as input to the classification in this step.To perform data classification, the random forest classifier is applied. Here, final prediction is achieved by clustering the input data. To classify the data more accurately, the clustering approach simplifies the data.

Step 4: Evaluation of results with respect to the various performance parameters is done in this final step. Precision, f-measure, recall and accuracy are the parameters used to analyze the performance.

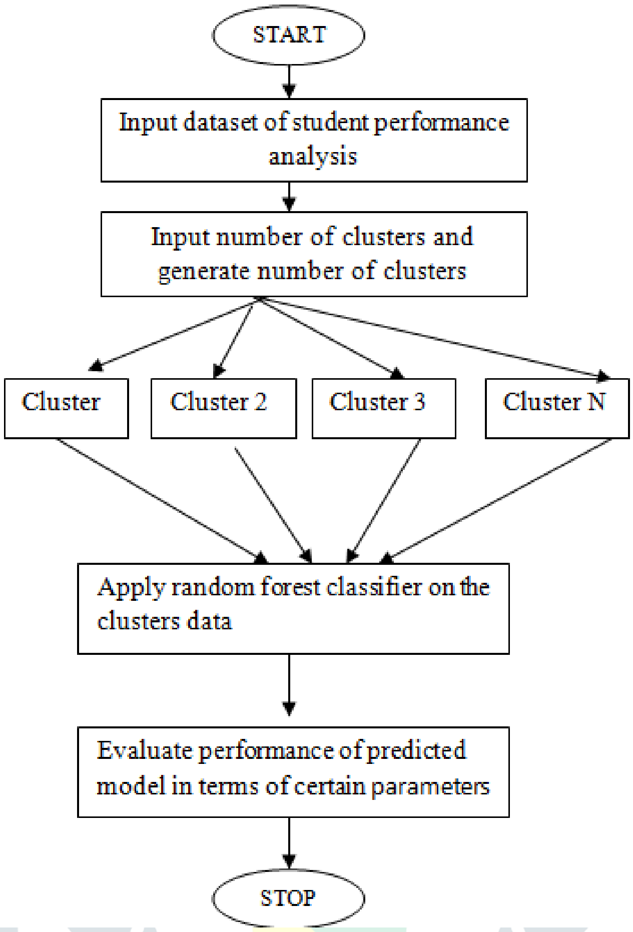
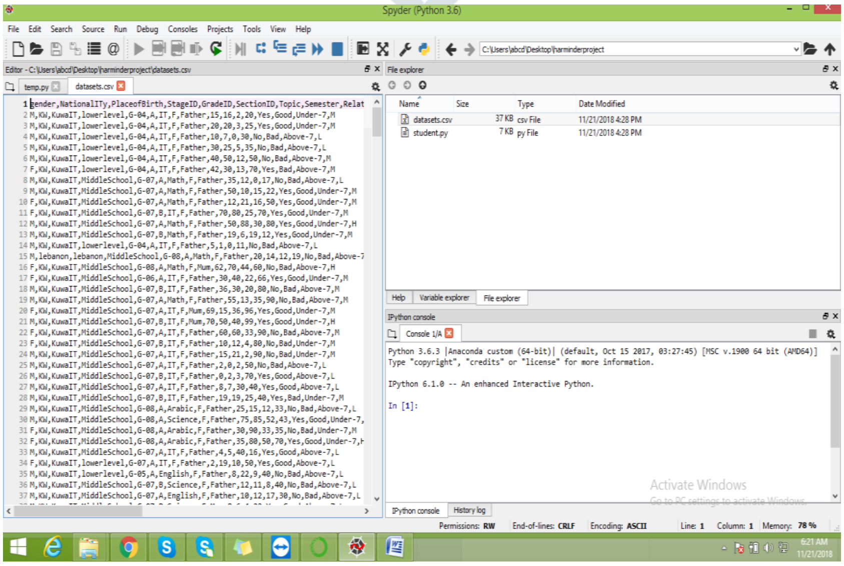

START
Input dataset of student performance analysis
Input number of clusters and generate number of clusters
Cluster
Cluster 2
Cluster 3
Cluster N
Apply random forest classifier on the clusters data
Evaluate performance of predicted model in terms of certain parameters
STOP

Fig 3.1: Proposed Methodology

IV. RESULT AND DISCUSSION

Python can be easily used by Rapid Application Development. The scripting in this tool is used to interlink the already existing components.



```
gender,NationalITy,PlaceofBirth,StageID,GradeID,SectionID,Topic,Semester,Relat
M,KW,KuwaIT,lowerlevel,G-04,A,IT,F,Father,15,16,2,20,Yes,Good,Under-7,M
M,KW,KuwaIT,lowerlevel,G-04,A,IT,F,Father,20,20,3,25,Yes,Good,Under-7,M
M,KW,KuwaIT,lowerlevel,G-04,A,IT,F,Father,10,7,0,30,No,Bad,Above-7,L
M,KW,KuwaIT,lowerlevel,G-04,A,IT,F,Father,30,25,5,35,No,Bad,Above-7,L
M,KW,KuwaIT,lowerlevel,G-04,A,IT,F,Father,40,50,12,50,No,Bad,Above-7,M
F,KW,KuwaIT,lowerlevel,G-04,A,IT,F,Father,42,30,13,70,Yes,Bad,Above-7,M
M,KW,KuwaIT,MiddleSchool,G-07,A,Math,F,Father,35,12,0,17,No,Bad,Above-7,L
M,KW,KuwaIT,MiddleSchool,G-07,A,Math,F,Father,50,10,15,22,Yes,Good,Under-7,M
F,KW,KuwaIT,MiddleSchool,G-07,A,Math,F,Father,12,21,16,50,Yes,Good,Under-7,M
F,KW,KuwaIT,MiddleSchool,G-07,B,IT,F,Father,70,80,25,70,Yes,Good,Under-7,M
M,KW,KuwaIT,MiddleSchool,G-07,A,Math,F,Father,50,88,30,80,Yes,Good,Under-7,H
M,KW,KuwaIT,MiddleSchool,G-07,B,Math,F,Father,19,6,19,12,Yes,Good,Under-7,M
M,KW,KuwaIT,lowerlevel,G-04,A,IT,F,Father,5,1,0,11,No,Bad,Above-7,L
M,lebanon,lebanon,MiddleSchool,G-08,A,Math,F,Father,20,14,12,19,No,Bad,Above-7
F,KW,KuwaIT,MiddleSchool,G-08,A,Math,F,Mum,62,70,44,60,No,Bad,Above-7,H
F,KW,KuwaIT,MiddleSchool,G-06,A,IT,F,Father,30,40,22,66,Yes,Good,Under-7,M
M,KW,KuwaIT,MiddleSchool,G-07,B,IT,F,Father,36,30,20,80,No,Bad,Above-7,M
M,KW,KuwaIT,MiddleSchool,G-07,A,Math,F,Father,55,13,35,90,No,Bad,Above-7,M
F,KW,KuwaIT,MiddleSchool,G-07,A,IT,F,Mum,69,15,36,96,Yes,Good,Under-7,M
M,KW,KuwaIT,MiddleSchool,G-07,B,IT,F,Mum,70,50,40,99,Yes,Good,Under-7,H
F,KW,KuwaIT,MiddleSchool,G-07,A,IT,F,Father,60,60,33,90,No,Bad,Above-7,M
F,KW,KuwaIT,MiddleSchool,G-07,B,IT,F,Father,10,12,4,80,No,Bad,Under-7,M
M,KW,KuwaIT,MiddleSchool,G-07,A,IT,F,Father,15,21,2,90,No,Bad,Under-7,M
M,KW,KuwaIT,MiddleSchool,G-07,A,IT,F,Father,2,0,2,50,No,Bad,Above-7,L
M,KW,KuwaIT,MiddleSchool,G-07,B,IT,F,Father,0,2,3,70,Yes,Good,Above-7,L
M,KW,KuwaIT,MiddleSchool,G-07,A,IT,F,Father,8,7,30,40,Yes,Good,Above-7,L
M,KW,KuwaIT,MiddleSchool,G-07,B,IT,F,Father,19,19,25,40,Yes,Bad,Under-7,M
M,KW,KuwaIT,MiddleSchool,G-08,A,Arabic,F,Father,25,15,12,33,No,Bad,Above-7,L
M,KW,KuwaIT,MiddleSchool,G-08,A,Science,F,Father,75,85,52,43,Yes,Good,Under-7,
F,KW,KuwaIT,MiddleSchool,G-08,A,Arabic,F,Father,30,90,33,35,No,Bad,Under-7,M
F,KW,KuwaIT,MiddleSchool,G-08,A,Arabic,F,Father,35,80,50,70,Yes,Good,Under-7,H
M,KW,KuwaIT,MiddleSchool,G-07,A,IT,F,Father,4,5,40,16,Yes,Good,Above-7,L
F,KW,KuwaIT,lowerlevel,G-07,A,IT,F,Father,2,19,10,50,Yes,Good,Above-7,L
M,KW,KuwaIT,lowerlevel,G-05,A,English,F,Father,8,22,9,40,No,Bad,Above-7,L
M,KW,KuwaIT,MiddleSchool,G-07,B,Science,F,Father,12,11,8,40,No,Bad,Above-7,L
M,KW,KuwaIT,MiddleSchool,G-07,A,English,F,Father,10,12,17,30,No,Bad,Above-7,L
```

Name	Size	Type	Date Modified
datasets.csv	37 KB	csv File	11/21/2018 4:28 PM
student.py	7 KB	py File	11/21/2018 4:28 PM

```
Python 3.6.3 |Anaconda custom (64-bit)| (default, Oct 15 2017, 03:27:45) [MSC v.1900 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 6.1.0 -- An enhanced Interactive Python.

In [1]:
```

Permissions: RW End-of-lines: CRLF Encoding: ASCII Line: 1 Column: 1 Memory: 78 %

Fig 4.1: Anaconda default interface

Fig 4.1 shows the default interface of anaconda. Here, a console, editor and interface of anaconda is shown in the default interface.

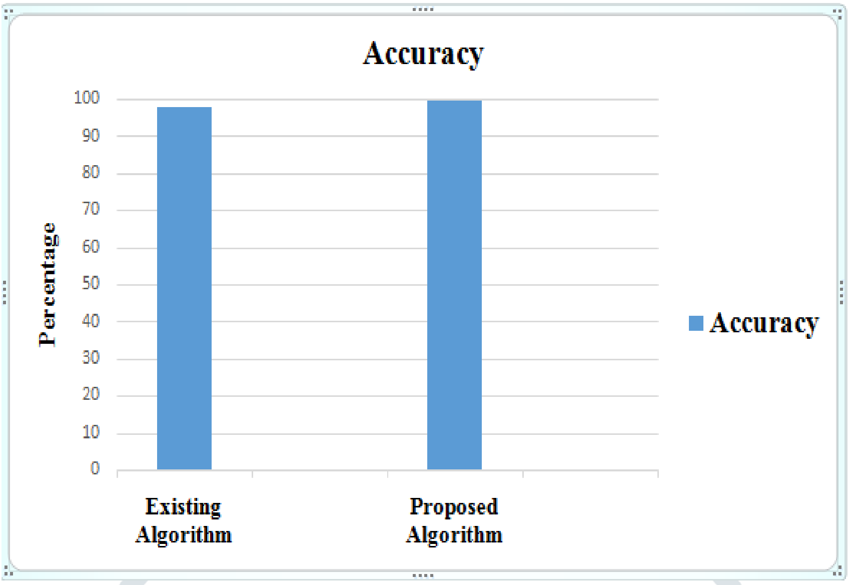


Fig 4.2: Accuracy Comparison

Fig. 4.2 shows the comparative analysis of proposed and existing algorithms in terms of accuracy. Here, in comparison to existing algorithm, higher accuracy is achieved by applying proposed algorithm.

Precision-Recall Analysis:- The precision-recall are the another parameters which define the accurate prediction of the target set. The value of precision-recall is shown in figure below

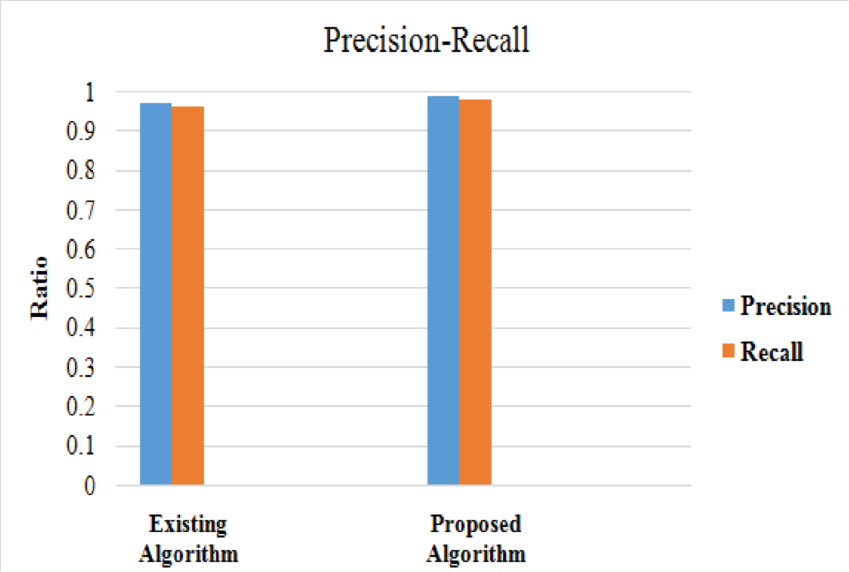


Fig 4.3 Precision-Recall Analysis

As shown in fig.4.3, the precision-recall value of existing algorithm and proposed algorithm is compared for the performance analysis. It is analyzed that proposed algorithm has more precision-recall value as compared to existing algorithm

F Measure: - The F measure is the parameter which defines the average value of the precision and recall. The F measure of the proposed algorithm is shown in figure given below

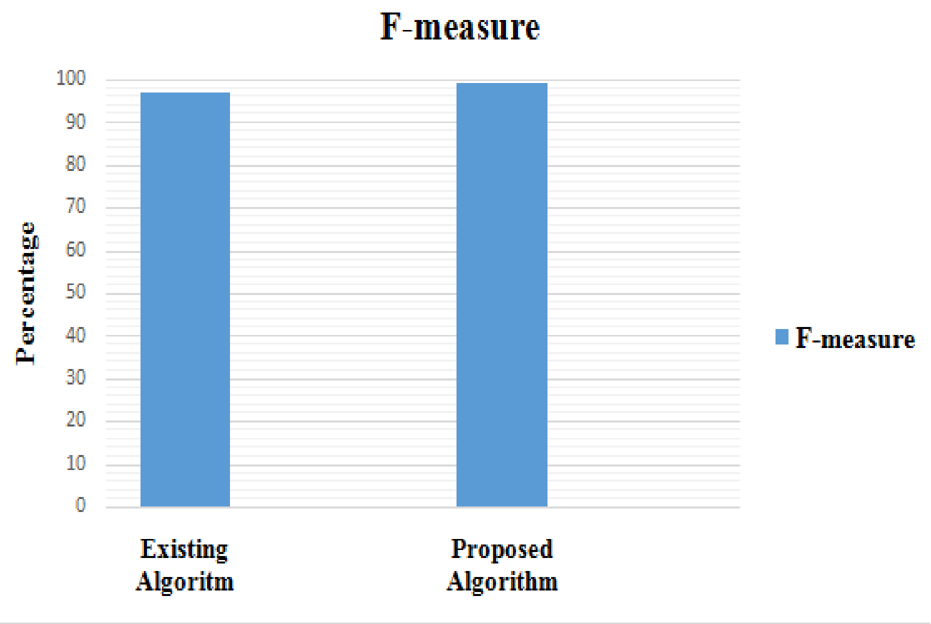


Fig 4.4 F1 Measure

As shown in figure 4.4, the existing and proposed algorithms are compared in terms of F measure. The existing algorithm has low F measure score and compared to proposed algorithm

Table 1: Performance Analysis

Parameter	**Existing Algorithm**	**Proposed Algorithm**
Accuracy	97.86 %	99.5 %
Precision	0.97	0.99
Recall	0.96	0.98
F Measure	97 %	99 %

Table 1 shows the comparative analysis of existing and proposed algorithms with respect to various performance metrics in a tabular form. Here, it is concluded that the overall performance of proposed method is better than the existing algorithm.

V. CONCLUSION

The prediction analysis is approach which can predict future possibilities based on the current information. The prediction analysis can be applied with the approach classification. The classification techniques can classify data into certain target sets. In this existing system, the technique of back propagation is applied for the student performance prediction. In this research work, hybrid classification approach will be designed based on the decision tree and random forest classifier. The decision tree classifier will works like the meta classifier and random forest will works like base classifier. The proposed algorithm will be implemented in Python and results will be analyzed in terms of accuracy, precision, recall and f measure. The overall results which are improved with the proposed algorithm is about the 3 to 4 percent

References

[1] U.Fayyad,"The KDD process for extracting useful knowledge from volumes of data", Commun. ACM," vol. 39, no. 11, pp. 27–34, 1996.

[2] T. Mishra, D. Kumar, and S. Gupta, “Mining students’ data for prediction performance,” Int. Conf. Adv. Comput. Commun. Technol.
ACCT, pp. 255–262, 2014.

[3] Y. Meier, J. Xu, O. Atan, and M. Van Der Schaar, “Personalized grade prediction: A data mining approach,” Proc. - IEEE Int. Conf. Data
Mining, ICDM, vol. 2016–Janua, pp. 907–912, 2016.

[4] Y.-h. Wang and H.-C. Liao, “Data mining for adaptive learning in a tesl-based e-learning system,” Comput. Educ., vol. 123, no. October
2017, pp. 97–108, 2018.

[5] S. D. GHEWARE, A. S. KEJKAR, and S. M. TONDARE, “Data Mining :Task ,Tools, Techniques and Applications,” Ijarcce, vol. 3, no.
10, pp. 8095–8098, 2014.

[6] S. L. Prabha, “Educational Data Mining Applications,” nternational J. Oper. Res. Appl., vol. 1, no. 1, pp. 23–29, 2014.

[7] R. S. J. d Baker and K. Yacef, “The State of Educational Data Mining in 2009: A Review and Future Visions,” J. Educ. Data Min., vol. 1,
no. 1, pp. 3–17, 2009.

[8] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.,
vol. 40, no. 6, pp. 601–618, 2010.

[9] M. P. G. Martins, V. L. Migueis, and D. S. B. Fonseca, “Educational data mining: A literature review [Data Mining Educacional: Uma
Revisão da Literatura],” Iber. Conf. Inf. Syst. Technol. Cist., vol. 2018–June, pp. 1–6, 2018.

[10] T. Devasia, T. P. Vinushree, and V. Hegde, “Prediction of students performance using Educational Data Mining,” Proc. 2016 Int. Conf.
Data Min. Adv. Comput. SAPIENCE 2016, pp. 91–95, 2016.

[11] Jie Xu, Kyeong Ho Moon, and Mihaela van der Schaar, “A Machine Learning Approach for Tracking and Predicting Student Performance
in Degree Programs”, 2016, IEEE Journal of Selected Topics in Signal Processing, Volume: 11 , Issue: 5

[12] S. M. Merchán, and J. A. Duarte, “Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance”,
IEEE Latin America Transactions, Vol. 14, No. 6, June 2016

[13] Ishwank Singh, A Sai Sabitha, Abhay Bansal, “Student Performance Analysis Using Clustering Algorithm”, 2016, 6th International
Conference - Cloud System and Big Data Engineering (Confluence)

[14] Ms.Tismy Devasia, Ms.Vinushree T P, Mr.Vinayak Hegde, “Prediction of Students Performance using Educational Data Mining”, 2016,
International Conference on Data Mining and Advanced Computing (SAPIENCE)

[15] Yuni Yamasari, Supeno M. S. Nugroho, I N. Sukajaya, Mauridhi H. Purnomo, “Features Extraction to Improve Performance of Clustering
Process on Student Achievement”, 2016, International Computer Science and Engineering Conference (ICSEC)