

# Machine Learning-Based Academic Result Prediction System

Megha Bhushan, School of Computing, DIT University, Dehradun, India\*

 <https://orcid.org/0000-0003-4309-875X>

Utkarsh Verma, School of Computing, DIT University, Dehradun, India

Chetna Garg, School of Computing, DIT University, Dehradun, India

Arun Negi, Deloitte USI, Gurgaon, India

## ABSTRACT

Students' academic performance is a critical issue as it decides his/her career. It is pivotal for the educational institutes to track the performance record because it can help to enhance the standard of their quality education. Thus, the role of the academic result prediction system comes into existence which uses semester grade point average (SGPA) as a metric. The proposed work aims to create a model that can forecast the SGPA of students based on certain traits. It predicts the result in the form of SGPA of computer science students considering their past academic performance, study, and personal habits during their academic semester using different machine learning models, and to compare them based on different accuracy parameters. Some models that are widely used and are found effective in this field are regression algorithms, classification algorithms, and deep learning techniques. The results conclude that deep learning techniques are the most effective in the proposed work because of their high accuracy and performance, depending upon the attributes used in the prediction.

## KEYWORDS

Academic performance, Deep Learning, Educational Data Mining, Machine Learning, Semester Grade Point Average (SGPA)

## INTRODUCTION

Education is a crucial aspect in terms of the economy, due to which researchers are developing several methods to enhance the performance of the students. One way to do so is to track the student's performance. Through research and development in this field, students can be benefitted in numerous ways, like faculty can give special attention to the students whose predicted semester grade point average (SGPA) is low or not up to the mark and this will be very helpful for the student and for the university as well as their entire result percentage will improve, on the other hand students can also track their performance and hence, can improve their study pattern and accordingly perform well.

Educational data mining is an emerging discipline, used to explore the distinctive and increasingly sizable data gathered from various educational institutes, and using data mining techniques to

DOI: 10.4018/IJSI.334715

\*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

comprehend the students and the methods in which they learn. By exploring these huge datasets and using different aforementioned techniques, unique patterns can be identified which will help to study, predict, and improve academic performance of students.

COVID-19 outbreak has brought unique challenges that were not expected earlier, especially in the education system. Due to the pandemic, schools, colleges and universities were closed. Therefore, few universities decided to take the exams in online mode, while some went for promoting the students to next semester by providing them grades based on the assignment marks and the marks obtained by the students in the previous semester only. This cannot be the only factor for the evaluation since there are other personal factors that contribute to a student's result.

Secondly, it was difficult for faculty to identify the students who are at a risk of not performing well due to following variation in students' performance, (i) who perform well in the beginning but degrade their performance by the end., (ii) who perform worse in the beginning but improve their performance by the end, and (iii) who consistently perform either better or worse. Predicting their SGPA can be very helpful to identify the students who need more attention and work hard.

A model is proposed in this work in which the results of students will not be dependent on a single feature, rather the features will include day-to-day personal habits along with academic habits and past academic performance, to predict the future result as fair as possible. Various machine learning (ML) (Kedia & Bhushan, 2022; Kholiya et al., 2021; Singh & Bhushan, 2022; Verma et al., 2019) techniques have been used for predicting the result of the student. It is crucial to assess the data quality (Bhushan & Goel, 2016; Bhushan et al., 2018; Bhushan et al., 2021) used by ML algorithms. Predicting the SGPA will also help the student in planning his/her academic goals and accordingly put in efforts to improve the results. It will also help the faculty as well as the university to maintain a record of responses submitted by the student and hence, to better understand and supervise their students accordingly. The objective of this work is to ensure that each student is being monitored and being given the guidance he or she needs in order to improve their academics.

The remaining paper incorporates related literature review followed by the experiment details along with the results and discussions. Later, the conclusion along with the future scope is discussed.

## RELATED WORK

There are many existing works related to the prediction of students' performance using ML algorithms (Arcinas, 2022; Albreiki et al., 2021; Chakrapani & D, 2022; Gajwani & Chakraborty, 2021; Verma et al., 2022; Yağcı, 2022).

According to Sharma and Aggarwal (2021), a dataset of around 400 students was selected and the analysis was conducted to check the level of parental influence on academic performance. The attributes taken into account for prediction were family size, parents' education, educational support from the family, internet access at home, paid classes and semester wise marks. During the analysis, correlations were found between the attributes, and linear regression (LR) was used as the predictive model. Further, training was done on 90% of the data and accuracy was calculated in terms of mean absolute error (MAE) as well as root mean square error (RMSE) which were 3.155 and 3.76, respectively. Gradient boost and support vector machine (SVM) were also applied but LR proved to be the best with a value of adjusted R square as 0.4771.

The analysis of students' performance based on a subset of behavioral and academic parameters was done using the techniques of feature selection and supervised ML algorithms (logistic regression (LOGR), decision tree (DT), naïve bayes' (NB) classifier and ensemble ML algorithms like bagging and boosting) (Gajwani & Chakraborty, 2021). The attributes such as demographic, behavioral and academic were taken into consideration. These included nationality, gender, place of birth, student's participation in discussion groups, raising hand in classes, using external resources, grade as well as semester marks. The dataset of 500 records was taken from Kaggle which was in turn obtained from learning management system (LMS). Further, 70% records were used in training and 30% for

testing. The correlations between attributes were plotted using box plot and finally, ML algorithms were applied. At the end, results were summarized with gradient boosting achieving the highest accuracy of 75% followed by random forest (RF) classifier achieving the accuracy of 74.31% and LOGR with the accuracy of 73.61%.

Various ML algorithms have been applied on the dataset of the university students to predict their performance (Rai et al., 2021). The students were split into 3 different classes based on their performance as good, average and poor. The dataset was collected from the UCI student performance dataset which consist of 831 samples with 22 student attributes comprising study hours, study field, success rate, gender, internal assessment marks, end-semester marks, year back status, family income, family education and medium of education. The dataset was pre-processed by handling missing values and label encoding, and then splitting into train and test sets. Supervised ML algorithms such as SVM classifier and RF classifier were used. RF proved to be the best with an accuracy of 94% and SVM with 79%. This analysis helped faculty to take early actions and assist the students belonging to the poor as well as the average category to improve their results.

According to Shetu et al. (2021), the academic result of students studying in a private institution was evaluated based on both environmental attributes and academic status. Three attribute selection methods i.e., info gain ranking filter, gain ratio feature evaluator and correlation ranking filter were used to identify the attribute with the most significant effect on the students' result. DT algorithm was also used. As a result, about 77% instances were correctly classified, 22.57% were incorrectly classified and MAE came out to be 0.1087.

As per Ganorkar et al. (2020), the dataset was collected from the student database of RKDF Institute of Science and Technology. Prime data mining methodologies which were applied were association, classification, clustering and DT. The attributes considered in the dataset were student id, gender, date of birth, college grade point average (GPA), grade etc. The objectives of this review paper are to find the factors affecting a student's academic performance, the student's weakness along with their strength so that the students who are low-grade performers could be assisted to gain better achievement in their academics.

A comparison was done between data mining methods such as DT C4.5 and k-nearest neighbors (KNN) (Yulianto et al., 2020). The dataset was collected from various sources such as by making a questionnaire and gathering information from the literature work. After simplifying the received data, following attributes were considered: student id, name, religion, gender, school origin, domicile, distance of student's residence from campus, number of siblings, parents' job, number of vehicles owned by student families, scholarship received, study time and graduation status. Two classification techniques such as DT C4.5 and KNN were applied, and then later, were compared to find the highest accuracy using hypertext pre-processor (PHP) programming language. It was found that KNN performed better than DT C4.5 with an accuracy of 59.32% while the latter with 54.80% accuracy.

Artificial neural network (ANN) technique was used to model 11 input variables with 2 layers of hidden neurons and one output layer (Lau et al., 2019). The dataset used consisted of 1000 students from a private Chinese Institution. National university entrance exam result and socio-economic background of the students along with previous attained cumulative grade point average (CGPA) were the attributes considered. Further, Levenberg-Marquardt algorithm is employed as the backpropagation training rule. The accuracy obtained was 84.8%.

The results of students studying in a private university in Bangladesh were evaluated by Rifat et al. (2019). The students' CGPA was predicted by taking DT algorithm as base model and using deep neural network (DNN) approach with major attributes such as student name, student id, gender, SGPA and course grade. The results obtained were mean square error (MSE) (DT-0.0226, DNN- 0.008).

Techniques such as regression, DT, gradient booster and DNN were used for the prediction by Kumar and Garg (2018). The dataset was provided by DIT University, Dehradun which includes the result data of engineering students (B.Tech. IT, 2017-18 Batch). Attributes considered were schooling

marks, continuous assessment and final evaluation. The best accuracy was bagged by the gradient boost model with 98.26% followed by neural network with accuracy of 97.05%.

Existing research works based on predicting academic results were reviewed by Dhamija et al. (2017). Class behavior, exam result, class attendance and extra-curricular activities were major attributes used in this work. Waikato environment for knowledge analysis (WEKA) tool along with many other data mining algorithms were discussed. In WEKA, the past result is given as input and then a predictive model is generated using an appropriate algorithm. By this, it can be classified which students need extra attention or have low predicted grades.

The data is collected from BMSIT & M for the 2014-18 batch (Pushpa et al., 2017). The data consists of three features, internal score, external score, and total score for each subject. The final feature shows the status of the result, whether the student passed or failed the semester. The ML based techniques such as SVM, NB algorithm, RF, gradient boosting were used in this work. The accuracies of the models were identified as SVM- 87.5%, NB – 87.5%, RF - 89% and gradient boosting-82%.

A review of 25 research works was conducted by Kumar et al. (2017), to understand techniques for prediction used in education. It identified and comprehended different attributes of the students which are used for predicting their performance. The major techniques addressed were classification, regression, KNN, DT, NB and clustering. Attributes included were personal, family, institutional and social. KNN and NB possessed the best results in most of the works, accuracy ranged from 90-98%.

The dataset was chosen from the University of Bangladesh BSMRSTU for result prediction (Sikder et al., 2016). Various techniques such as data mining, neural network, MATLAB tool, Levenberg Marquart backpropagation algorithm were used. In total, 14 attributes were considered including test marks, attendance, lab performance, previous result, social media interaction and study time. In this work, a student's yearly performance is predicted in the form of CGPA using a neural network and then compared with real CGPA. The best accuracy attained was 88% and the average was 74%. Neural networks proved out to be the best method when compared with other ML models by having least error.

According to Halde et al. (2016), the dataset was taken from the students of Thadomal Shahani Engineering College. Techniques used were ANN, LR, NB, DT and SVM. The experiment was performed on the real time data collected from the final year students. The matriculate and pre-university examination scores, five semester scores along with data on the motivation level, information processing ability and other learning and study skills were taken as input to the model to predict the CGPA. Further, accuracy without psychological factors was 0.9310 and accuracy with psychological factors was 0.9999.

A dataset of 300 students at Sacred Hearts Girls High School in Kerala and the results were evaluated using neural network (multi-layer perceptron (MLP) training using K fold cross validation) with the help of WEKA data mining tool and association rule mining (Sebastian & Puthiyidam, 2015). Attributes were related to academic details (interest of study, unit-test marks, attendance, and assignment) and personal attributes (residence, parent education, family status). The obtained results include K fold cross validation with 91% accuracy and association rule mining with 62% accuracy.

Real data history was extracted from UiTM's student's information management system (SIMS) in (Arsad & Buniyamin, 2014), where the data included student identity number, gender, CGPA obtained in previous semesters, grade point (GP) of all subjects attempted at every semester as well as GP of English courses. The techniques used in this work were ANN and LR. In ANN, three models with different inputs and same output were developed. The correlation coefficients of these three models were 0.9254, 0.9225 and 0.5221, respectively. In LR, for matriculation students R squared score in model A and B are both 0.720 and 0.752 respectively while model C gave R squared score of 0.35 which proved to be poor correlation. As for the diploma students, the R squared score was 0.728 and 0.755 while R squared score for model C was 0.318.

The dataset from Sofia University was used in (Kabakchieva, 2013). Various classifiers such as Bayesian network, NB, DT and KNN were used for the result prediction. Different attributes were

considered such as age, gender, ethnicity, education, work status and disability. It presents the initial results from a data mining research project implemented at a Bulgarian University, aimed at revealing the high potential of data mining applications for university management. The accuracy with DT was observed as 66% and with KNN was 60%.

A dataset was collected from 3 phases, personal data, pre-university data and university data (Kabakchieva, 2012). The attributes were age, gender, previous exam score, entrance exam score, current semester marks and number of backs. The algorithms applied were DT, neural network and KNN algorithm using WEKA tool. After applying the algorithms, the highest accuracy achieved was 73.59% by neural networks model followed by DT model with an accuracy of 72.74% and KNN with an accuracy of 70.49%.

Table 1 summarizes the existing related work in the area of student result prediction. It contains the attributes of the dataset used in the corresponding work and the results obtained after applying the relevant techniques are shown in column third.

## PROPOSED WORK

The proposed work predicts the SGPA/result of computer science students using regression, classification and deep learning (DL) (Nalavade et al., 2020) techniques, by using academic and personal factors. It aims to predict the SGPA of the student based on his/her attributes.

### Experimental Setup

The setup required to conduct experiments include hardware, RAM of 8 GB, hard disk of 512 GB and processor of 3<sup>rd</sup> generation Intel Core i5. Further, software requirements include Spyder IDE version 4.0.1 which is a free integrated development environment included in Anaconda Navigator and python 3.7.4.version is used as a programming language.

### Dataset

The data of 2,115 computer science engineering students studying at DIT University, Dehradun, was collected through Microsoft form which is shown in Figure 1. It includes both academic as well as personal attributes required to predict the result of the students. The list of attributes considered is presented in Table 2.

### Steps of Implementation

**Importing the Libraries and Dataset.** NumPy, Pandas, Matplotlib, Scikit-learn and TensorFlow are the libraries used in the proposed work.

**Data Pre-Processing.** It comprises of following steps:

1. Label encoding.
2. Handling missing values.
3. Feature scaling.
4. Splitting the dataset into a testing and training set.

**The Learning Process.** It discusses various ML techniques that were used to predict the SGPA/result of students.

**Regression.** It is a supervised learning algorithm and works with labeled datasets (Sharma & Aggarwal, 2021). It identifies the relation between dependent and independent variables. The technique is used for predictive modelling in ML and it predicts the continuous outcomes. In this work, regression is applied to predict the SGPA 4 of the student using various personal factors like daily social media interaction and attention during lectures which were collected through survey (see

Table 1. Summary of the existing related work

Citation	Attributes	Results
Sharma and Aggarwal (2021)	Family size, parent's education, educational support from the family, paid classes, internet access at home and semester wise marks	<ul style="list-style-type: none"> <li>● The accuracy was calculated in terms of MAE and RMSE which were 3.155 and 3.76, respectively.</li> <li>● Gradient boost and SVM were also applied but LR proved to be the best with the value of adjusted R square 0.4771.</li> </ul>
GajwaniandChakraborty (2021)	Nationality, gender, place of birth, student's participation in discussion groups, raising hand in classes, using external resources, grade along with semester marks.	<ul style="list-style-type: none"> <li>● Gradient boosting achieved the highest accuracy of 75%</li> <li>● RF classifier achieving the accuracy of 74.31%</li> <li>● LOGR with the accuracy of 73.61%.</li> </ul>
Citation	Attributes	Results
Rai et al. (2021)	Study hours, study field, success rate, gender, internal assessment marks, end semester marks, year back status, family income, family education, medium of education.	<ul style="list-style-type: none"> <li>● RF proved to be best with an accuracy of 94%</li> <li>● SVM with 79% accuracy.</li> </ul>
Shetu et al. (2021)	CGPA, gender, attendance, class test, drug addiction, social media, attention, depression, extra-curricular activities	<ul style="list-style-type: none"> <li>● Highest accuracy was achieved with DT algorithm</li> <li>● Correctly classified instances were 77.43%</li> <li>● Incorrectly classified instances were 22.56%</li> <li>● MAE was 0.1087</li> </ul>
Ganorkar et al. (2020)	Student identification, date of birth, gender, place of birth, speciality, enrolment year, year of graduation, city, contact, school type, matriculation year, grade, college GPA	<ul style="list-style-type: none"> <li>● A review paper of data mining techniques discussed various research works done in this specific area.</li> </ul>
Yulianto et al. (2020)	Student id, name, religion, gender, school origin, domicile, distance of student residence to campus, number of siblings, parents' job, number of vehicles owned by student families, scholarship received, study time and graduation status	<ul style="list-style-type: none"> <li>● KNN performed with an accuracy of 59.32%</li> <li>● DT C4.5 gave an accuracy of 54.80%.</li> </ul>
Lau et al. (2019)	National University entrance exam result, socio-economic background of the students, previously obtained CGPA taking in consideration their credit hours.	<ul style="list-style-type: none"> <li>● 84.8% accuracy was achieved with neural networks.</li> <li>● Error and precision were 15.02% and 86.3%, respectively.</li> <li>● Area under curve was observed 0.86 which proved to be a good result</li> </ul>
Rifat et al. (2019)	Student name, gender, SGPA, course grade	<ul style="list-style-type: none"> <li>● Accuracy was compared on the bases of baseline model- DT.</li> <li>● MSE of DT was 0.0226 and DNN was 0.008.</li> </ul>
Kumar and Garg (2018)	Schooling marks, continuous assessment, final evaluation	<ul style="list-style-type: none"> <li>● Highest accuracy achieved by gradient boost model is 98.26%</li> <li>● Accuracy of neural network is 97.05%</li> </ul>
Citation	Attributes	Results
Dhamija et al. (2017)	Class behaviour, exam result, class attendance, extra-curricular	<ul style="list-style-type: none"> <li>● In WEKA, the past result is given as the input and using an appropriate algorithm, a predictive model is generated.</li> <li>● It can classify which students need more attention or have low predicting grades.</li> </ul>
Pushpa et al. (2017)	Internal score, external score, total score	<ul style="list-style-type: none"> <li>● Accuracy of SVM is 87.5%</li> <li>● Accuracy of NB is 87.5%</li> <li>● Accuracy of RF is 89%</li> <li>● Accuracy of gradient boosting is 82%</li> </ul>
Kumar et al. (2017)	Personal attributes, family attributes, institutional attributes, social attributes	<ul style="list-style-type: none"> <li>● KNN and NB proved to be the best.</li> </ul>
Sikder et al. (2016)	Test marks, attendance, lab performance, previous result, social media interaction, study time	<ul style="list-style-type: none"> <li>● Best accuracy of 88% was achieved by neural networks.</li> <li>● Average accuracy achieved was 74%</li> <li>● Neural networks proved to be the best method when compared with other ML models by having least error.</li> </ul>
Halde et al. (2016)	Pre-university percentage, matric percentage, semester CGPA, motivation, concentration, management, study aids, study main ideas	<ul style="list-style-type: none"> <li>● Accuracy without psychological factors was 0.9310</li> <li>● Accuracy with psychological factors was 0.9999</li> </ul>
Sebastian and Puthiyidam (2015)	Interest of study, unit-test marks, attendance, assignment, residence, parent education, family status	<ul style="list-style-type: none"> <li>● 91% accuracy was achieved with K-fold cross validation.</li> <li>● 62% accuracy was achieved with association rule mining.</li> </ul>
Arsad and Buniyamin (2014)	Previous CGPA	<ul style="list-style-type: none"> <li>● ANN and LR gave the same results taking MSE in consideration.</li> </ul>
Kabakchieva (2013)	Age, gender, ethnicity, education, work status, disability	<ul style="list-style-type: none"> <li>● 66% accuracy with DT.</li> <li>● 60% accuracy with KNN.</li> </ul>
Kabakchieva (2012)	Age, gender, previous exam score, entrance exam score, current semester marks, number of backs	<ul style="list-style-type: none"> <li>● Highest accuracy of 73.59% was achieved by the neural networks model.</li> <li>● DT gave an accuracy of 72.74%</li> <li>● KNN gave an accuracy of 70.49%.</li> </ul>

Note: MAE-mean absolute error, RMSE-root mean squared error, SVM-support vector machine, LR- linear regression, RF- random forest, LOGR- logistic regression, CGPA- cumulative grade point average, DT- decision tree, GPA- grade point average, KNN- k-nearest neighbors, MSE- mean square error, SGPA- semester grade point average, NB- naïve bayes', WEKA- waikato environment for knowledge analysis.

Table 2. Attributes of the dataset

Attributes	Values
Daily social media interaction	< 1 hour 1-2 hours > 2 hours
Physical activity frequency	Rating from 1-10
Programming language knowledge	Rating from 1-10
Class Attendance	85% - 95% 75% - 85% 65%-75% 55%-65% Below 55%
Attention during lectures	Rating from 1-10
Daily study duration	< 1 hour 1-2 hours 2-4 hours > 4 hours
Study duration during exams	2-4 hours 4-6 hours 6-8 hours > 8 hours
Frequency of taking help from external sources (YouTube, Google, Library)	Rating from 1-10
Frequency of how often student clears doubts of other friends while studying in a group	Rating from 1-10
SGPA 1	Float value ranging from 0-10
SGPA 2	Float value ranging from 0-10
SGPA 3	Float value ranging from 0-10
SGPA 4	Float value ranging from 0-10

Table 2) as well as taking the continuous academic progress of the student through his/her previous SGPAs. Following are the applied models:

1. Multiple linear regression (MLR) is a statistical technique that uses more than one independent variable to predict the dependent output variable (Sharma & Aggarwal, 2021). Its aim is to predict and model the linear relationship between the input and output variables.
2. Random forest regression is a supervised ML technique and here ensemble learning is used (Obata et al., 2021). Ensemble learning is a method which aggregates the predictions from several ML models to make even more accurate predictions than a single model.

**Classification.** It is a supervised learning algorithm, where the dataset is divided into classes based on different parameters (Gajwani & Chakraborty, 2021). In this process, the dataset is trained on a training set and accordingly, it divides the data into different classes. In this work, the previous SGPA of the students was divided in 3 classes, according to which the model was trained and predicted the SGPA 4.

The three classes of SGPA are shown in Table 3.

Figure 1. Survey form

**Student's Information**  
 Fill this form to the best of your knowledge.  
 Please use the following scale for rating:  
 Excellent/Regular: 8-10    Good/Mostly: 6-7  
 Average/Quite often: 3-5    Poor/Not really: 1-2

3. Your daily social media interaction \*

Less than 1 hour  
 1-2 hours  
 More than 2 hours

4. Rate your programming skills (Fluency with languages: C/C++/Python/Java etc. ) (Only for CSE, IT, BCA and MCA students)

1 2 3 4 5 6 7 8 9 10

5. Rate your physical activity (Dedicated workout session/sports) \*

1 2 3 4 5 6 7 8 9 10

6. Your class attendance \*

85% - 95%  
 75% - 85%  
 65% - 75%  
 55% - 65%  
 Below 55%

7. Rate your attention during lectures \*

1 2 3 4 5 6 7 8 9 10

8. How often do you take references from Google, YouTube, library or any other platform to clear your concepts? \*

1 2 3 4 5 6 7 8 9 10

Table 3. SGPA divided into classes

SGPA Range	Class
0-6.5	0
6.5-8.5	1
8.5 above	2

The models applied are as follows:

1. Random forest classification is a supervised ML technique, used for classification (Gajwani & Chakraborty, 2021). This algorithm makes decision trees on data and then predicts, and finally selects the best solution by means of voting/majority.
2. Logistic regression classification is a supervised ML technique used for classification and works on predictive modelling; therefore, it is called logistic regression, but is used for classifying the samples, hence, it is a classification algorithm (Gajwani & Chakraborty, 2021). It gives the probabilistic values ranging from 0 to 1.

3. Decision tree classification is a supervised ML method and a tree-structured classifier (Shetu et al., 2021). It consists of 2 nodes, decision node and leaf node. Decisions are made by decision nodes and the outputs of those decisions are reflected in leaf nodes.
4. Naïve bayes classification which is based on Bayes' theorem is a supervised ML method and used for classification (Pushpa et al., 2017). It is a probabilistic classifier, i.e., it predicts based on the probability of an object.

**Artificial Neural Network.** It's a DL technique. Neural networks learn through a process called back-propagation (Lau et al., 2019). It uses a set of training data that matches known input to desired outputs.

**ANN with principal component analysis (PCA).** It is one of the most used unsupervised dimensionality reduction techniques. PCA reduces the dimension of N dimensional data set by projecting it into M dimensional subspace where  $M < N$  (Lee & Jemain, 2021). In this work, PCA is applied to reduce the dimensionality of data to only 2 columns. After applying PCA, ANN was used to build and train the model in order to fetch the predictions.

## Results and Discussions

After applying all the aforementioned models, the results were calculated and were compared as shown in this section.

The regression models were evaluated on the basis of MAE, MSE and RMSE. RMSE is the standard deviation of the residuals which are also called as prediction errors. Residuals are basically the measure of distance between the regression line and the data points. A lower RMSE indicates a better model but RMSE of zero will denote over-fitting. The RMSE comparison of different models has been shown in Figure 2.

The results obtained after applying different regression techniques are summarized in Table 4.

In Figure 2, RF regression has low RMSE as compared to MLR proving to be the best among regression. The classification models and ANN were evaluated on the basis of accuracy which was calculated on the number of correctly classified instances. The accuracy comparison of different models has been represented in Figure 3. The results obtained after applying different classification techniques and ANN are summarized in Table 5.

Figure 3 represents that ANN along with PCA has the highest accuracy among all classification models, proving to be the best among classification.

## Comparative Analysis

A comparative analysis of the proposed work with the existing works has been shown in Table 6. It is observed that the proposed work considered more attributes which affect the student's result, than the existing works. Moreover, both regression and classification algorithms along with ANN were also applied in the proposed work. The proposed work resulted in better accuracy when compared with the existing works.

**Table 4. Results of regression models**

Name of the Model	MAE	MSE	RMSE
MLR	0.39	0.30	0.55
RF Regression	0.16	0.046	0.21

Figure 2. Comparison of RMSE

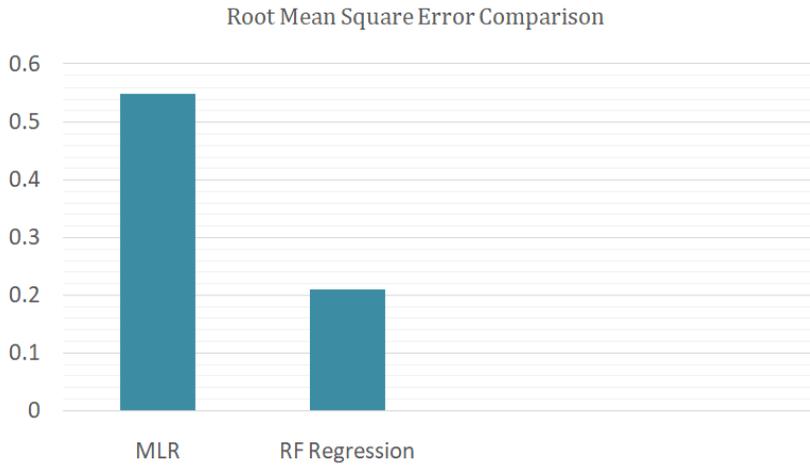


Table 5. Results of classification models and ANN

Name of the Model	Accuracy
Random forest classification (RFC)	81.8%
Logistic regression classification (LRC)	81.8%
Decision tree classification (DTC)	72.7%
Naïve bayes' classification (NBC)	72.72%
Artificial neural network (ANN)	81.8%
ANN with PCA	86.3%

Figure 3. Accuracy comparison

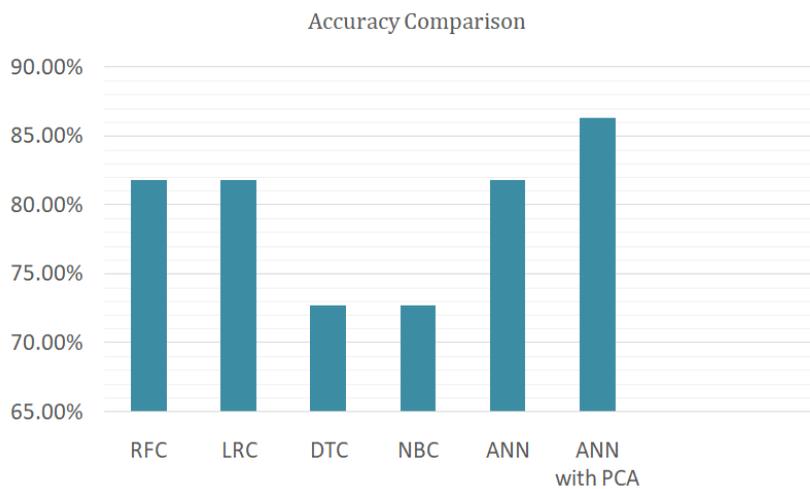


Table 6. Comparative analysis

Parameters	Gajwani and Chakraborty (2021)	Shetu et al. (2021)	Proposed Work
Attributes	<ul style="list-style-type: none"> <li>• Nationality</li> <li>• Gender</li> <li>• Place of birth</li> <li>• Participation of students in discussion groups</li> <li>• Raising hand in classes</li> <li>• Using external resources</li> <li>• Grade and semester marks.</li> </ul>	<ul style="list-style-type: none"> <li>• CGPA</li> <li>• Gender</li> <li>• Attendance</li> <li>• Class test</li> <li>• Drug addiction</li> <li>• Social media</li> <li>• Attention</li> <li>• Depression</li> <li>• Extra-curricular activities</li> </ul>	<ul style="list-style-type: none"> <li>• Student's participation in doubt discussion groups</li> <li>• Using external sources (YouTube, Library)</li> <li>• Attention and attendance in class</li> <li>• Social Media interaction</li> <li>• Physical Activity frequency</li> <li>• Study duration</li> <li>• Programming language knowledge</li> <li>• Semester marks</li> </ul>
ML Models applied	<ul style="list-style-type: none"> <li>• DT classifier</li> <li>• NB classifier</li> <li>• LOGR classifier</li> <li>• RF classifier</li> <li>• Gradient Boosting</li> </ul>	<ul style="list-style-type: none"> <li>• DT classifier</li> </ul>	<ul style="list-style-type: none"> <li>• DT classifier</li> <li>• NB classifier</li> <li>• LOGR classifier</li> <li>• RF Regression and Classification</li> <li>• MLR</li> </ul>
DL techniques applied	No	No	Yes
Accuracy achieved	<ul style="list-style-type: none"> <li>• RF – 74.31%</li> <li>• DT – 73.61</li> <li>• NB – 72.92%</li> </ul>	<ul style="list-style-type: none"> <li>• DT (correctly classified instances) – 77%</li> </ul>	<ul style="list-style-type: none"> <li>• RF – 81.8%</li> <li>• DT – 72.7%</li> <li>• NB – 72.72%</li> <li>• ANN – 81.8%</li> <li>• ANN with PCA – 86.3%</li> </ul>

## CONCLUSION AND FUTURE SCOPE

This work defined the factors responsible for a student's academic performance and compared various ML techniques that predict a student's result. The future challenges include finding more attributes that affect the result of the students and also to hyper-tune the parameters using advanced hyper-parameter tuning techniques like Hyperopt in order to get the better results. Data gathering can be improved with focus on getting a larger number of data points. Exploratory data analysis can be done using libraries like seaborn to visually understand the relationship between different parameters. Various other ML techniques like boosting algorithms can also be used. Moreover, in the future an application can be built to be used by stakeholders (students, faculty as well as, training and placement officers), that can help them to evaluate a student's performance and hence work on improvement.

## COMPLIANCE WITH ETHICAL STANDARDS

### Acknowledgment

The authors acknowledge School of Computing, DIT University for assisting them in the collection of dataset.

### Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

### Competing Interests

All authors of this article declare that there are no competing interests.

### Conflicts of Interest

The authors declare that they have no conflicts of interest.

**Research involving Human Participants and/or Animals:** Not applicable

**Informed consent:** Not applicable

## REFERENCES

- Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, 11(9), 552. doi:10.3390/educsci11090552
- Arcinas, M. M. (2022). Design of machine learning based model to predict students academic performance. *ECS Transactions*, 107(1), 3207–3214. doi:10.1149/10701.3207ecst
- Arsad, P. M., & Buniyamin, N. (2014). Neural network and linear regression methods for prediction of students' academic achievement. In *Proceedings of the 2014 IEEE Global Engineering Education Conference (EDUCON)* (pp. 916-921). IEEE. doi:10.1109/EDUCON.2014.6826206
- Bhushan, M., & Goel, S. (2016). Improving software product line using an ontological approach. *Sadhana*, 41(12), 1381–1391. doi:10.1007/s12046-016-0571-y
- Bhushan, M., Goel, S., & Kumar, A. (2018). Improving quality of software product line by analysing inconsistencies in feature models using an ontological rule-based approach. *Expert Systems: International Journal of Knowledge Engineering and Neural Networks*, 35(3), e12256. doi:10.1111/exsy.12256
- Bhushan, M., Kumar, A., Samant, P., Bansal, S., Tiwari, S., & Negi, A. (2021). Identifying quality attributes of FODA and DSSA methods in domain analysis using a case study. In *Proceedings of the 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 562-567). IEEE. doi:10.1109/SMART52563.2021.9676289
- Chakrapani, P., & Chitradevi, D. (2022). Academic performance prediction using machine learning: A comprehensive & systematic review. In *Proceedings of the 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)* (pp. 335-340). IEEE. doi:10.1109/ICESIC53714.2022.9783512
- Dhamija, P., Nandal, R., & Sehrawat, H. (2017). A review paper on prediction analysis: predicting student result on the basis of past result. *International Journal of Engineering and Technology (IJET)*, 9(2), 1204-1207.
- Gajwani, J., & Chakraborty, P. (2021). Students' performance prediction using feature selection and supervised machine learning algorithms. In *Proceedings of the International Conference on Innovative Computing and Communications* (pp. 347-354). Springer, Singapore. doi:10.1007/978-981-15-5113-0\_25
- Ganorkar, S. S., Tiwari, N., & Namdeo, V. (2020). Analysis and prediction of student data using data science: A review. *Smart Trends in Computing and Communications: Proceedings of SmartCom 2020*, 443-448.
- Halde, R. R., Deshpande, A., & Mahajan, A. (2016). Psychology assisted prediction of academic performance using machine learning. In *Proceedings of the 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 431-435). IEEE. doi:10.1109/RTEICT.2016.7807857
- Kabakchieva, D. (2012). Student performance prediction by using data mining classification algorithms. *International Journal of Computer Science and Management Research*, 1(4), 686–690.
- Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, 13(1), 61–72. doi:10.2478/cait-2013-0006
- Kedia, S., & Bhushan, M. (2022). Prediction of mortality from heart failure using machine learning. In *Proceedings of the 2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET)* (pp. 1-6). IEEE. doi:10.1109/ICEFEET51821.2022.9848348
- Kholiya, P. S., Kapoor, A., Rana, M., & Bhushan, M. (2021). Intelligent process automation: The future of digital transformation. In *Proceedings of the 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 185-190). IEEE. doi:10.1109/SMART52563.2021.9676222
- Kumar, M., Singh, A. J., & Handa, D. (2017). Literature survey on student's performance prediction in education using data mining techniques. *International Journal of Education and Management Engineering*, 7(6), 40–49. doi:10.5815/ijeme.2017.06.05
- Kumar, V., & Garg, M. L. (2018). Comparison of machine learning models in student result prediction. In *Proceedings of the International Conference on Advanced Computing Networking and Informatics* (pp. 439-452). Springer, Singapore.

- Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences*, 1(9), 1–10. doi:10.1007/s42452-019-0884-7
- Lee, L. C., & Jemain, A. A. (2021). On overview of PCA application strategy in processing high dimensionality forensic data. *Microchemical Journal*, 169, 106608. doi:10.1016/j.microc.2021.106608
- Nalavade, A., Bai, A., & Bhushan, M. (2020). Deep learning techniques and models for improving machine reading comprehension system. *International Journal of Advanced Science and Technology*, 29(04), 9692–9710.
- Obata, S., Cieszewski, C. J., Lowe, R. C. III, & Bettinger, P. (2021). Random forest regression model for estimation of the growing stock volumes in Georgia, USA, Using Dense Landsat Time Series and FIA Dataset. *Remote Sensing (Basel)*, 13(2), 218. doi:10.3390/rs13020218
- Pushpa, S. K., Manjunath, T. N., Mrunal, T. V., Singh, A., & Suhas, C. (2017). Class result prediction using machine learning. In *Proceedings of the 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)* (pp. 1208-1212). IEEE. doi:10.1109/SmartTechCon.2017.8358559
- Rai, S., Shastry, K. A., Pratap, S., Kishore, S., Mishra, P., & Sanjay, H. A. (2021). *Machine learning approach for student academic performance prediction. Evolution in Computational Intelligence*. Springer.
- Rifat, M. R. I., Al Imran, A., & Badrudduza, A. S. M. (2019, May). Edunet: A deep neural network approach for predicting CGPA of undergraduate students. In *Proceedings of the 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)* (pp. 1-6). IEEE.
- Sebastian, S., & Puthiyidam, J. J. (2015). Evaluating students performance by artificial neural network using weka. *International Journal of Computer Applications*, 119(23), 36–39. doi:10.5120/21380-4370
- Sharma, D., & Aggarwal, D. (2021). A predictive approach to academic performance analysis of students based on parental influence. In *Proceedings of the International Conference on Innovative Computing and Communications* (pp. 75-84). Springer, Singapore. doi:10.1007/978-981-15-5113-0\_6
- Shetu, S. F., Saifuzzaman, M., Moon, N. N., Sultana, S., & Yousuf, R. (2021). Student's performance prediction using data mining technique depending on overall academic status and environmental attributes. In *Proceedings of the International Conference on Innovative Computing and Communications* (pp. 757-769). Springer, Singapore. doi:10.1007/978-981-15-5148-2\_66
- Sikder, M. F., Uddin, M. J., & Halder, S. (2016). Predicting students yearly performance using neural network: A case study of BSMRSTU. In *Proceedings of the 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 524-529). IEEE. doi:10.1109/ICIEV.2016.7760058
- Singh, S. N., & Bhushan, M. (2022). Smart ECG monitoring and analysis system using machine learning. In *Proceedings of the 2022 IEEE VLSI Device Circuit and System (VLSI DCS)* (pp. 304-309). IEEE. doi:10.1109/VLSIDCS53788.2022.9811433
- Verma, K., Bhardwaj, S., Arya, R., Islam, U. L., Bhushan, M., Kumar, A., & Samant, P. (2019). Latest tools for data mining and machine learning. *International Journal of Innovative Technology and Exploring Engineering*, 8(9S), 24–28.
- Verma, U., Garg, C., Bhushan, M., Samant, P., Kumar, A., & Negi, A. (2022). Prediction of students' academic performance using machine learning techniques. In *Proceedings of the 2022 International Mobile and Embedded Technology Conference (MECON)* (pp. 151-156). IEEE. doi:10.1109/MECON53876.2022.9751956
- Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1), 1–19. doi:10.1186/s40561-022-00192-z
- Yulianto, L. D., Triayudi, A., & Sholihati, I. D. (2020). Implementation educational data mining for analysis of student performance prediction with comparison of k-nearest neighbor data mining method and decision tree C4.5. *Jurnal Mantik*, 4(1), 441–451.

*Megha Bhushan is an Associate Professor in the School of Computing, DIT University, Dehradun, India. She has received her ME and Ph.D. degrees from Thapar University, Punjab, India. She was awarded with a fellowship by UGC, Government of India, in 2014. In 2017, she was a recipient of Grace Hopper Celebration India (GHCI) fellowship. She has published 5 national patents and 1 international patent has been granted. She has published many research articles in international journals and conferences of repute. Further, she is the editor of many edited books with different publishers such as CRC Press, Taylor & Francis Group, Wiley-Scrivener and Bentham Science. Her research interests include Artificial Intelligence, Knowledge representation, Expert systems, and Software quality. She is also the reviewer and editorial board member of many international journals.*

*Utkarsh Verma has graduated in B. Tech (CSE) from School of Computing, DIT University, Dehradun, India in 2022. He is currently working as a Data Scientist at Celebal technologies, India. His research areas include Artificial Intelligence, Machine Learning and Educational Data Mining.*

*Chetna Garg has graduated in B. Tech (CSE) from School of Computing, DIT University, Dehradun, India in 2022. She is currently working as a Data Engineer at Acko General Insurance, India. Her research areas include Artificial Intelligence, Machine Learning and Educational Data Mining.*

*Arun Negi is currently a Manager at Deloitte USI, Gurgaon, India. He has completed a course in Business Management from IIM, Ahmedabad and has obtained B. Tech degree from Jawaharlal Nehru University, New Delhi, India. He has 13+ years of diverse experience in cyber risk services. He has worked on various network security technologies and platforms for Fortune 500 clients. He has been granted one national and international patent each. He has published many research articles in international journals and conferences of repute. His research areas include Artificial Intelligence, Machine Learning, Software Product Line, Cloud Computing, and Cyber Security.*