



# A comparative study of machine learning and deep learning algorithms for predicting student's academic performance

Megha Bhushan<sup>1</sup> · Satyam Vyas<sup>1</sup> · Shrey Mall<sup>1</sup> · Arun Negi<sup>2</sup>

Received: 29 June 2022 / Revised: 4 July 2023 / Accepted: 7 September 2023 / Published online: 24 September 2023

© The Author(s) under exclusive licence to The Society for Reliability Engineering, Quality and Operations Management (SREQOM), India and The Division of Operation and Maintenance, Lulea University of Technology, Sweden 2023

**Abstract** Education plays a crucial role in personal development and social change. It provides individuals with the necessary skills to succeed in their careers and serve their communities. The Cumulative Grade Point Average (CGPA) is a valuable metric for assessing student's academic performance. It is calculated by averaging the performance of a student in each semester, also known as Semester Grade Point Average (SGPA). Its prediction can provide several benefits to both faculties and students. It enables faculties to promptly identify weak students who are facing difficulties and provide them with necessary intervention and support. Additionally, SGPA empowers students to establish realistic future goals and make well-informed decisions based on it. In this study, various classification and regression algorithms are applied to predict student's SGPA and compared using various evaluation metrics. The results concluded that it is feasible to predict SGPA (of students) using machine learning and deep learning models with low error rate and high accuracy as the predicted results will be equally beneficial for both faculties and students.

**Keywords** Machine learning · Prediction · Deep learning · Classification · Regression · Academics

## 1 Introduction

Education plays an important role in both personal growth as well as social change around the world today. The value

of education cannot be overstated as it equips individuals with the skills necessary for succeeding in their respective careers and concurrently serving their communities' needs. Correspondingly, beyond individual gains lies the underlying national benefit that comes with an educated populace that drives social advancement alongside economic empowerment. In fact, education is critical towards fostering global economic innovation by generating productive citizens who compete favourably within dynamic markets.

A significant parameter used in determining a student's academic performance is the Cumulative Grade Point Average (CGPA) scorecard as it shows a student's overall progress throughout their study years. CGPA has a significant impact on a student's future, as their future prospects and employment opportunities depend on it (Gamboa et al. 2018). CGPA is calculated by averaging the performance of a student in each semester, also known as the Semester Grade Point Average (SGPA). Predicting SGPA has numerous benefits for both faculties and students alike. The prediction of this metric allows faculties to identify students lacking the necessary support they need from time-to-time or those at risk of substantially falling behind others without tailored interventions unique to their learning needs at an early stage. Furthermore, predicting SGPA allows students to set reasonable future goals and make informed decisions based on it.

Machine Learning (ML) and Deep Learning (DL) have shown significant enhancement in various domains such as healthcare (Pathan et al. 2020), education (Shila Pawar 2020), agriculture (Goel et al. 2020), and intelligent transportation systems (Rai et al. 2023). In literature, various methods and techniques for predicting CGPA have been proposed ranging from ML algorithms including Decision Trees (DT) (Alyahyan and Dusteator 2020) and Neural Networks (NN) (Saifuzzaman et al. 2021) (Verma et al. 2022)

✉ Megha Bhushan  
mb.meghabhushan@gmail.com

<sup>1</sup> School of Computing, DIT University, Dehradun, India

<sup>2</sup> Deloitte USI, Gurgaon, India

to statistical techniques such as regression analysis (Akuma and Abakpa 2021a) or Data Mining (DM) approaches like association rule mining (Bhaskaran 2021) etc. Despite varied effectiveness amongst these methods, the onset of COVID-19 has amplified needs for precise predictive modelling.

The data for this study was collected from undergraduate students (B.Tech CSE and IT) studying at DIT University, India through a Microsoft questionnaire designed to capture various influential factors. These factors incorporated several aspects such as academic performance, daily physical activity, social media interaction and study habits like help from outside sources like YouTube and other online platforms, help from friends, SGPA of semesters 1–4 etc. On the collected data, various classification and regression techniques were applied to determine the most effective techniques in predicting SGPA. This work has the potential to aid faculties, students, and policymakers by providing them with more accurate information about student performance.

This paper is structured as follows: Sect. 2 comprises review of related literature. Section 3 provides description of considered ML and DL models. In Sect. 4, the proposed work is discussed. Section 5 presents the findings of this study. Also, a comparison between existing works and the proposed work is depicted in Sect. 6. Finally, Sect. 7 outlines the conclusions and future directions for this research.

## 2 Background

The literature review related to ML and DL algorithms for predicting student academic performance is explained in this section. In (Baashar et al. 2022), several ML algorithms were used to predict the academic performance of postgraduate students. The study used a dataset of 635 master's students from a well-known private university in Malaysia. The best results were obtained using an Artificial Neural Network (ANN) model, which achieved an error rate of only 11% when predicting the students' CGPA. A Gaussian Process Regression (GPR) model also performed well, achieving an accuracy of 71%.

The authors of (Yakubu and Abubakar 2022) aimed to predict students' performance in academics using various early detection indicators (i.e., gender, CGPA, age etc.) for enabling early intervention. It was done using logistic regression, a predictive analysis model appropriate for two dependent variables like CGPA. The dataset was divided into 30% testing data and 70% training data for supervised ML methodology. It achieved training accuracy of 84.7% and testing accuracy of 83.5%. Furthermore, Support Vector Machine (SVM) has also been used for predicting academic success in online learning and achieved an accuracy of 83.5% (Samsudin et al. 2022).

Several studies have been conducted to analyze students' academic performance using ML algorithms. Sharma and Aggarwal (2021) have used Linear Regression (LR) to analyze the effect of parental influence on academic performance. The study found that family size, parents' education, educational support from the family, internet access at home, paid classes and semester-wise marks were significant predictors. The accuracy was measured using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) which were 3.155 and 3.76, respectively. In (Gajwani and Chakraborty 2021), feature selection along with ML algorithm was used to analyze students' performance based on behavioral and academic parameters such as nationality, gender, place of birth, student participation in discussion groups, raising a hand in classes, using external resources, grade along with semester marks. It was found that gradient boosting achieved an accuracy of 75%, followed by the Random Forest (RF) classifier achieving an accuracy of 74.31%.

In (Rai et al. 2021), various ML algorithms were applied to a dataset of university students and found that the RF classifier was the most accurate, with an accuracy of 94%. The study helped faculty to take early action and assist students which were categorized as poor and average. Other studies have taken different approaches to predict students' academic performance. For example, in (Shetu et al. 2021) students' results based on both academic status and environmental attributes were evaluated. It achieved an accuracy of 77% and MAE of 0.1087. Prime data mining methodologies like association, classification, clustering and DT were studied to analyze factors affecting a student's academic performance (Ganorkar et al. 2021).

Also, in (Akuma and Abakpa 2021b) a model using LR Algorithm to predict student performance in academics based on their CGPA and prior academic records was developed. It accurately forecasted the fourth year CGPA of 70 undergraduate computer science students and achieved an accuracy of 87.84% and a correlation of 0.9338. Alangari and Alturki (2020) have explored the use of Educational Data Mining (EDM) to predict students' performance. They tested 15 classification algorithms and found that the Naïve Bayes and Hoeffding tree models could predict students' final GPA with 91% accuracy. These models outperformed others on multiple datasets, achieving an average classifier accuracy of 71%.

In (Yulianto et al. 2020), a comparison between data mining methods such as DT and K-Nearest Neighbors (KNN) was studied. It was found that DT performed worse than KNN, with an accuracy of 54.80%, while the latter with 59.32% accuracy. Furthermore, Lau et al. (2019) have used an ANN technique to model input variables, achieving an accuracy of 84.8%. Also, Islam Rifat et al. (2019) have predicted students' CGPA using a DT algorithm as the base model along with Deep Neural Network (DNN) approach.

**Table 1** Literature review

Article	Dataset description	Techniques	Performance measures
Baashar et al. (2022)	635 students	ANN and GPR	Achieved error rate of 11% and accuracy of 71%
Yakubu and Abubakar (2022)	–	Logistic regression	Achieved training accuracy of 84.7% and testing accuracy 83.5%
Samsudin et al. (2022)	225 student respondents	SVM	Achieved an accuracy of 83.5%
Sharma and Aggarwal (2021)	Around 400 students were randomly selected	LR	MAE and RMSE were 3.155 and 3.76, respectively
Gajwani and Chakraborty (2021)	–	Gradient boosting, RF	Gradient boosting achieved an accuracy of 75%, followed by the RF classifier achieving an accuracy of 74.31%
Rai et al. (2021)	–	RF classifier	RF classifier gives an accuracy of 94%
Shetu et al. (2021)	–	DT	Achieved an accuracy of 77% and an MAE of 0.1087
Akuma and Abakpa (2021a)	70 undergraduate students	LR Algorithm	Achieved an accuracy of 87.84% and a correlation of 0.9338
Alangari and Alturki (2020)	129 entries and 39 attributes were present in the dataset	Naïve Bayes and Hoefding tree models	Achieved an accuracy of 91%
Yulianto et al. (2020)	60 students	DT and KNN	DT gives an accuracy of 54.80%, while KNN gives 59.32% of accuracy
Lau et al. (2019)	Data samples of 1,000 students were randomly mixed	ANN	Achieved accuracy of 84.8%
Islam Rifat et al. (2019)	Dataset contains 398 instances in total	DT algorithm and DNN	MSE for DT-0.0226 and DNN- 0.008

ANN: artificial neural network, GPR: gaussian process regression, SVM: support vector machine, LR: linear regression, RF: random forest, DT: decision tree, NN: neural network, MAE: mean absolute error, KNN: K-nearest neighbor, DNN: deep neural network, MSE: mean square error, “–” indicates that parameter was not stated in the corresponding work

The results obtained for Mean Square Error (MSE) with DT and DNN were 0.0226 and 0.008, respectively. Also, Table 1 summarizes the existing literature review.

### 3 Description of models

ML and DL models can be broadly categorized into two types i.e., regression and classification. Following describes the models used in this work:

#### 3.1 Regression models

Regression is a type of supervised ML algorithm which utilizes labeled data as input. It helps in determining the relation between dependent and independent variables.

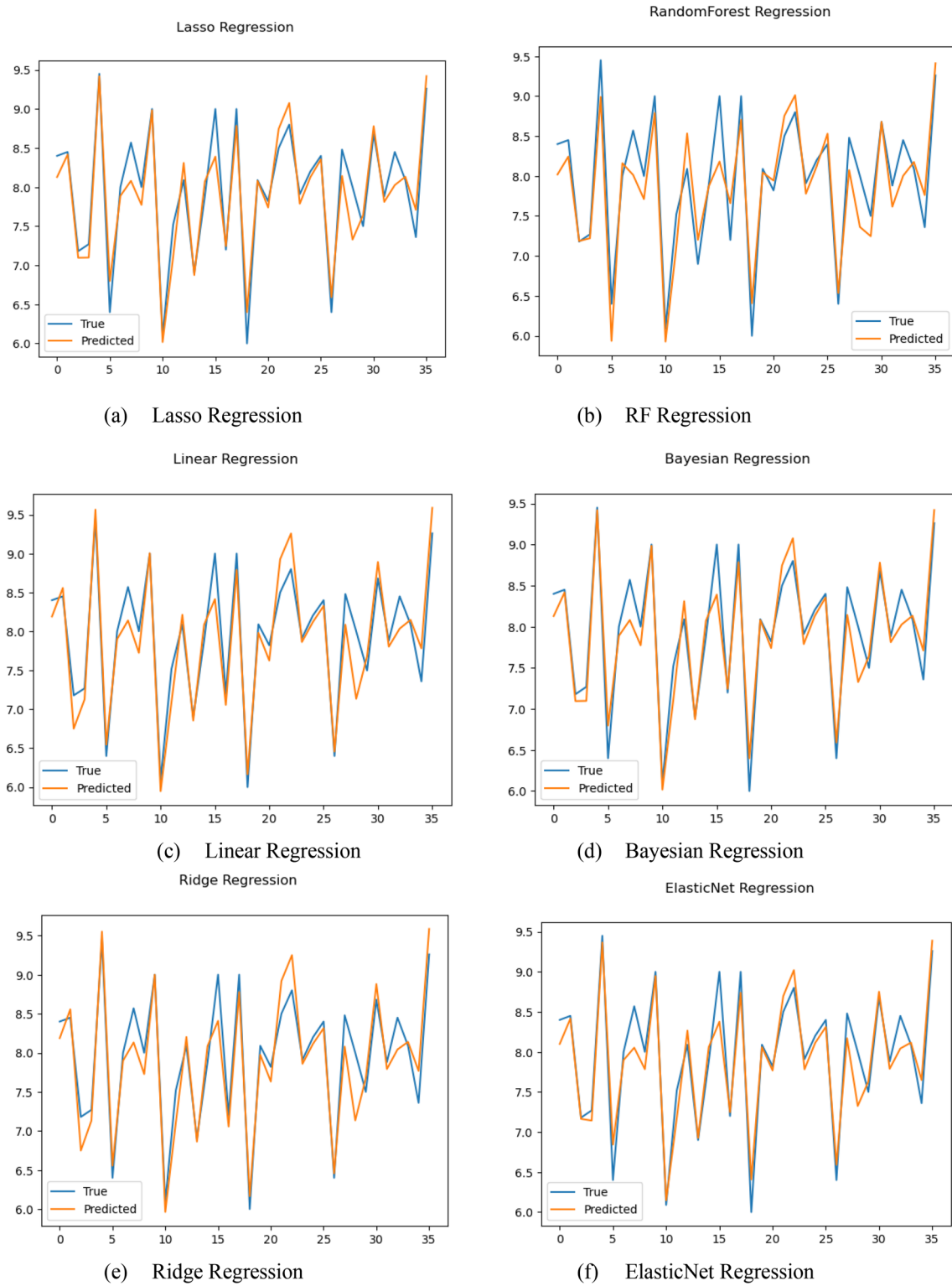
- Linear regression (Su et al. 2012): A statistical method that is used to determine linear relation between dependent and independent variables using a linear equation.
- Ridge regression (McDonald 2009): This type of LR includes a regularization term in the loss function. It

helps to prevent overfitting by reducing the magnitude of the coefficients towards zero.

- Lasso regression (Ranstam & Cook 2018): It is a type of LR that adds regularization term to the loss function and uses the L1-norm instead of the L2-norm. It can result in some coefficients which are shrunk to zero, effectively performing the variable selection.
- Elastic net regression (Hans 2012): A hybrid method combining Ridge and Lasso Regression by including L1 and L2 regularization terms in the loss function.
- RF regression (Liu et al. 2012): An ensemble learning method that utilizes multiple DTs to make predictions.
- Bayesian ridge regression: It is a type of regression analysis that estimates a probabilistic model of the regression problem.

#### 3.2 Classification models

This supervised learning algorithm divides the dataset into classes based on various parameters. It is trained on a training set and then uses this information to classify the data into different classes.



**Fig. 1** Output of regression models

- RNN (GRU) (Dey & Salemt 2017): GRU is a form of RNN that effectively handles sequential data by utilizing GRUs. These units are a modified version of LSTM

units and they possess fewer parameters, thus, allowing for faster training.

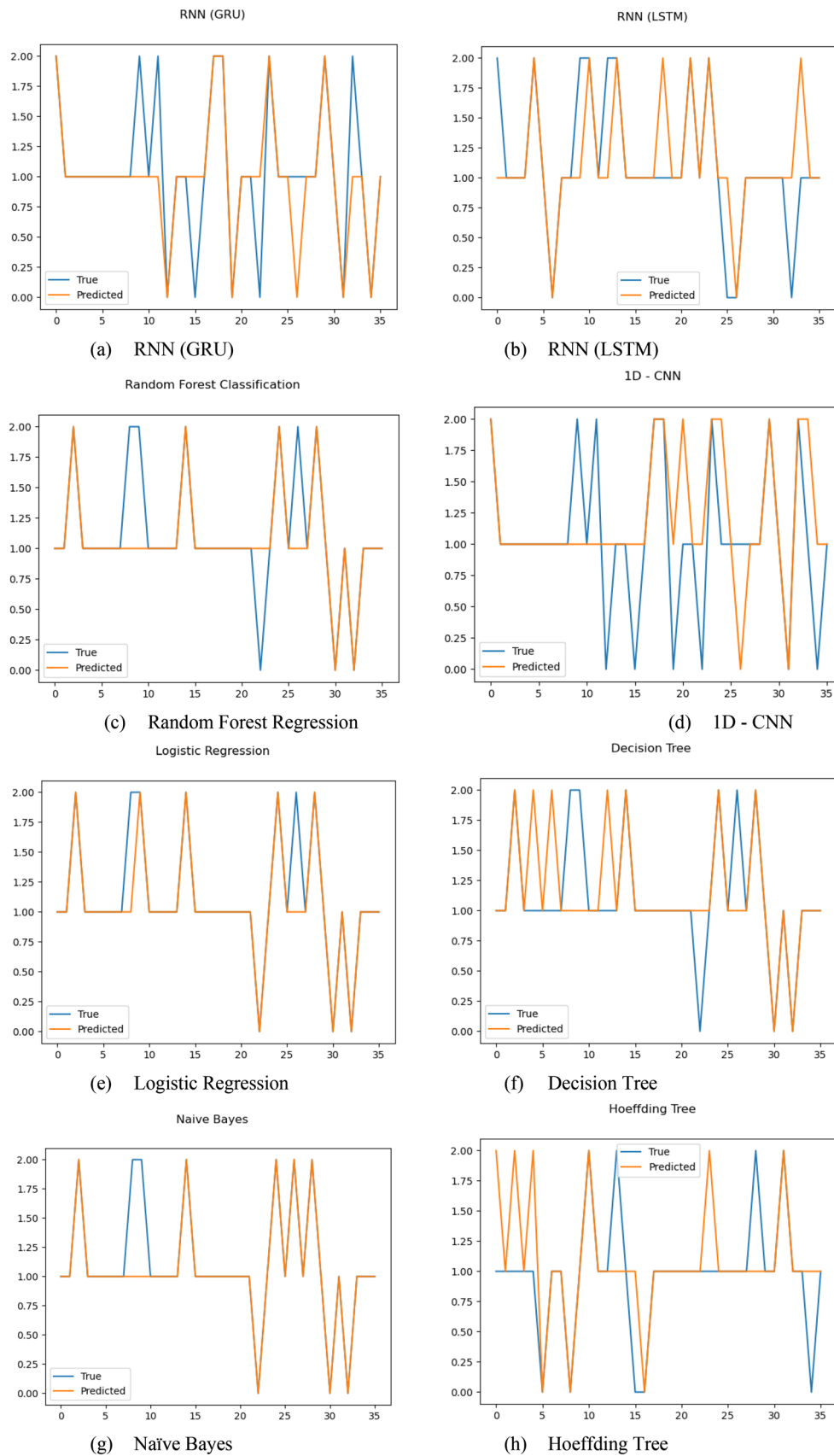


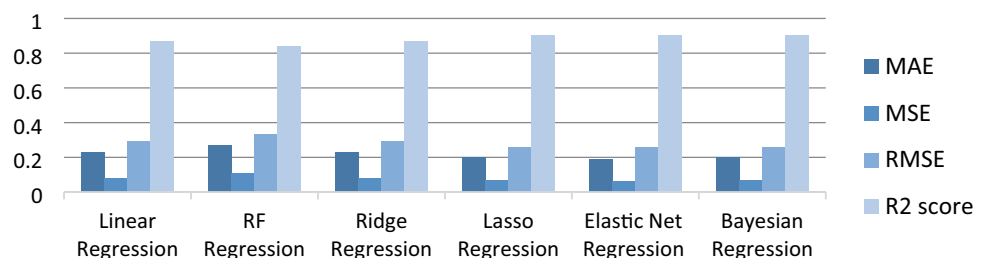
Fig. 2 Output of classification models

- RNN (LSTM) (Staudemeyer & Morris 2019): It is a type of RNN that uses LSTM units to process sequential data. These are designed to retain information for longer periods, making it suitable for tasks such as language modelling and speech recognition
- RF classification (Liu et al. 2012): It is a type of ensemble learning technique that utilizes multiple DTs to make predictions. It can be used for both classifications and regression.
- Convolutional neural networks (CNN) (O’Shea & Nash 2015): CNN constitutes a neural network architecture that exploits convolutional layers in order to extract key features from input data. It is typically employed in image recognition and processing domains
- Logistic regression classification (LaValley 2008): It is a statistical technique used for binary classification. It calculates the likelihood of an instance being in a specific class using a logistic function.
- DT classification (Myles et al. 2004): A tree-like model is generated for prediction. The data is split into subsets based on the values of the input features.
- Naïve bayes classification (Webb 2016): A probabilistic algorithm based on Bayes’ theorem that supposes independence between the features. It can be used for both binary and multi-class classification tasks.
- Hoeffding tree algorithm (Kirkby 2007): It is a way to classify stream data using DTs. It employs Hoeffding trees and the Hoeffding bound to select the best splitting attribute from a small data sample.

**Table 2** Results of the Regression Models

Model	MAE	MSE	RMSE	R2 score
Linear regression	0.23	0.08	0.29	0.87
RF regression	0.27	0.11	0.33	0.84
Ridge regression	0.23	0.08	0.29	0.87
Lasso regression	0.20	0.07	0.26	0.90
Elastic Net regression	0.19	0.06	0.26	0.90
Bayesian regression	0.20	0.07	0.26	0.90

**Fig. 3** Comparison of Regression models



## 4 Methodology

The proposed work predicts the SGPA of students using various regression and classification techniques by utilizing data collected through a survey form consisting of questionnaires related to academics and personal routines. The setup used for this experiment included INTEL I7-9750H CPU, NVIDIA GTX 1660ti GPU with 16 GB ram and 1 TB storage with Windows 11, python (3.9.16), VS Code (1.78.2), NumPy (1.24.3), Pandas (1.5.3), TensorFlow (2.10.1), and scikit-learn (1.2.1).

### 4.1 Dataset collection

For this study, the dataset was collected from undergraduate students (B.Tech CSE and IT) studying at DIT University, India through a survey form (Jain 2021). The details of the form are discussed in (Sect. 4.2) and is accessible at <https://forms.office.com/r/pVUWHZ9gWJ>.

### 4.2 Dataset description

The survey form consisted of following attributes:

- Daily social media interaction: “less than 1 h”, “between 1 and 2 h”, “more than 2 h”.
- Physical activity frequency: rated on a scale of 1 to 10.
- Programming language knowledge: rated on a scale of 1 to 10.
- Class Attendance: “85–95%”, “75–85%”, “65–75%”, “55–65%”, “Below 55%”.
- Attention during lectures: rated on a scale of 1–10.
- Daily study duration: “less than 1 h”, “between 1 and 2 h”, “between 2 and 4 h”, “more than 4 h”.
- Study duration during exams: “between 2 and 4 h”, “between 4 and 6 h”, “between 6 and 8 h”, “more than 8 h”.
- Frequency of taking help from external sources (YouTube, Google, Library): rated on a scale of 1–10.
- Frequency of how often student clears doubts of other friends while studying in a group: rated on a scale of 1–10.
- SGPA for semester 1, 2, 3 and 4.



**Table 3** Results of the classification models (in %)

Model	Accuracy	F1 score	Recall
RNN (GRU)	83	82	83
RNN (LSTM)	80	79	80
RF classification	88.8	87	88
CNN	69.4	66	69
Logistic regression classification	94.4	94	94
Decision tree classification	80.5	80	80
Naïve bayes classification	94.4	94	94
Hoeffding tree	77.7	78	77

### 4.3 Implementation

Initially, the data was imported to Python, and a label encoder (Bisong 2019) was used to convert the non-numeric values in the dataset to numeric classes. Label encoder transforms the non-numeric values to numeric classes ranging from 0 to n-1 classes. Afterwards, to fill in the missing values mean imputation technique was utilized (Jadhav et al. 2019). Furthermore, a split ratio of 0.2 was used for all the experiments (Roshan et al. 2022). The following techniques were used in this study:

#### 4.3.1 Regression

In this study, a relationship was established between students' personal daily routines, previous SGPA scores, and Semester 4 SGPA using regression models for prediction purposes. The comparison between true and predicted values for various regression models is shown in Fig. 1.

#### 4.3.2 Classification

In this study, the SGPA of students was divided into 3 classes i.e., 0, 1 and 2 for SGPA in range “0–6.5”, “6.5–8.5” and “8.5 above”, respectively, using these the model was trained to predict class category for 4th semester SGPA of students. The comparison between true and predicted class values for various classification models is shown in in Fig. 2.

## 5 Result

The results obtained have been discussed in this section. This study used different classification and regression models to predict students' 4th semester SGPA using the collected data. These models were evaluated and compared based on their MAE, MSE, RMSE along with accuracy values.

The regression models were evaluated using MAE, MSE, RMSE and R2 score as evaluation metrics (Hodson 2022). The results obtained after applying different regression techniques are summarized in Table 2, and a graphical comparison is shown in Fig. 3.

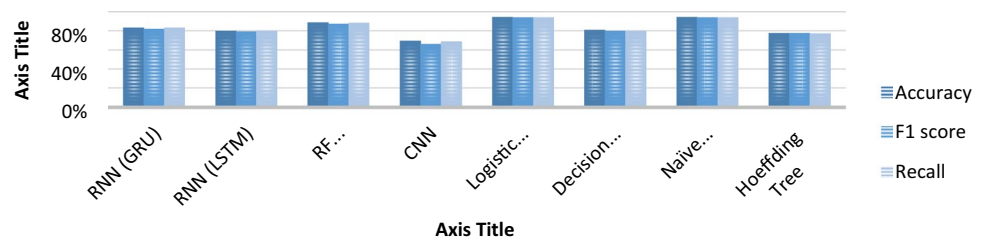
Figure 3 concludes that Elastic Net regression has achieved the lowest MAE, MSE, RMSE and highest R2 score among all, proving to be the best among the regression techniques at predicting the SGPA of students.

The performance of the classification models was evaluated based on accuracy, F1 score and Recall. The results obtained after applying various classification techniques are summarized in Table 3, and a graphical comparison is shown in Fig. 4.

Figure 4 concludes that Naïve Bayes and Logistic regression classification have achieved the highest accuracy, F1-score and recall values among all applied classification models.

## 6 Comparative analysis

This section compares existing studies, namely (Gajwani and Chakraborty 2021) and (Shetu et al. 2021), with the proposed work as shown in Table 4. The comparison was based on dataset, techniques applied, and various performance metrics. The proposed study has utilized 10 features in the considered dataset whereas (Gajwani and Chakraborty 2021) and (Shetu et al. 2021) have used 7 and 9, respectively. Also, Gajwani and Chakraborty (2021) and Shetu et al. (2021) have applied 5 and 1 ML models, respectively, in their work while the presented work has applied 14 ML and DL models. Furthermore, the proposed study has achieved better accuracy i.e., 94.4% than other studies. The highest values of F1 Score and Recall were achieved by Logistic Regression Classification and Naïve Bayes Classification of 94 and 94

**Fig. 4** Accuracy comparison of classification models

**Table 4** Comparative analysis

Parameters	Gajwani and Chakraborty (2021)	Shetu et al. (2021)	Proposed work
No of features in dataset	7	9	10
ML models applied	DT Naïve bayes Logistic regression RF Gradient boosting	DT	LR Logistic regression classification Elastic net regression Naïve bayes classification Ridge regression DT classification Lasso regression RF regression and classification Bayesian regression Hoeffding tree
DL techniques applied	No	No	RNN (LSTM) RNN (GRU) CNN
Accuracy (in %)	RF–74.31 DT–73.61 Naïve bayes –72.92 Boosting gradient–75	DT–77	RNN (GRU)–83 RNN (LSTM)–80 RF classification–88.8 CNN–69.4 Logistic regression classification–94.4% DT classification–80.5 Naïve bayes classification–94.4 Hoeffding Tree–77.7
F1 score and recall (in %)	–	–	RNN (LSTM)—79 and 80 RNN (GRU)—82 and 83 RF classification—87 and 88 CNN- 66 and 69 Logistic regression classification—94 and 94 DT classification—80 and 80 Naïve bayes Classification—94 and 94 Hoeffding tree—78 and 77

LR: linear regression, DT: decision tree, RNN: recurrent neural network, GRU: gated recurrent unit, LSTM: long short term memory, RF: random forest, CNN: convolutional neural network, “–” indicates that parameter was not stated in the corresponding work

respectively, however, these metrics were not computed by Gajwani and Chakraborty (2021) and Shetu et al. (2021).

## 7 Conclusion and future scope

The proposed work predicts student’s SGPA using various regression and classification techniques. The data used for this study was collected from undergraduate students (B.Tech CSE and IT) studying at DIT University through a Microsoft survey form consisting of questionnaires based on academic parameters and daily personal routine. Using the collected data, regression and classification techniques

were applied and the results were evaluated based on MAE, MSE, RMSE and R2 score for the regression models and accuracy, F1 score and recall for the classification models. Among regression models, Elastic Net Regression yielded the lowest MAE, MSE, RMSE and highest R2 score values and proved to be the best regression technique for predicting student SGPA, while Naïve Bayes and Logistic Regression yielded the highest accuracies among all other classification models. This work is helpful for both faculties as well as students as it can predict SGPA (of students) using ML and DL models with low error rate and high accuracy.



In future scope, this research could be expanded by collecting more data and using more algorithms to improve the accuracy of predictions. Other factors that may affect a student's performance in academics can be considered and included in the analysis.

**Acknowledgements** The authors would like to thank School of Computing, DIT University for helping them to collect the dataset.

**Funding** The authors have no relevant financial or non-financial interests to disclose.

#### Declarations

**Conflict of Interest** The authors declare that they have no conflicts of interest.

## References

- Akuma S, Abakpa H (2021) Predicting undergraduate level students' performance using regression. *Niger Ann Pure Appl Sci* 4:109–117. <https://doi.org/10.46912/napas.224>
- Akuma S, Abakpa H (2021) Predicting undergraduate level students' performance using regression. *Niger Ann Pure Appl Sci* 4:129–139. <https://doi.org/10.46912/napas.224>
- Alangari N, Alturki R (2020) Predicting students final GPA using 15 classification algorithms. *Roman J Inf Sci Technol* 23:238–249
- Alyahyan E, Dustearo D (2020) Decision trees for very early prediction of student's achievement. In: 2020 2nd International Conference on Computer and Information Sciences, ICCIS 2020. <https://doi.org/10.1109/ICCIS49240.2020.9257646>
- Baashar Y, Hamed Y, Alkawsy G, Fernando Capretz L, Alhussian H, Alwadain A, Al-amri R (2022) Evaluation of postgraduate academic performance using artificial intelligence models. *Alex Eng J* 61:9867–9878. <https://doi.org/10.1016/j.aej.2022.03.021>
- Bhaskaran, S.S., 2021. Investigation of Student Performance with Contextual Factors Using Association Rules in Higher Educational Institutions (HEIs). [https://doi.org/10.1007/978-981-16-0171-2\\_41](https://doi.org/10.1007/978-981-16-0171-2_41)
- Bisong E (2019) Introduction to Scikit-learn. Build Mach Learn Deep Learn Models Google Cloud Platform. [https://doi.org/10.1007/978-1-4842-4470-8\\_18](https://doi.org/10.1007/978-1-4842-4470-8_18)
- Dey R, Salemt FM (2017) Gate-variants of gated recurrent unit (GRU) neural networks. *Midwest Symp Circuits Syst*. <https://doi.org/10.1109/MWSCAS.2017.8053243>
- Gajwani J, Chakraborty P (2021) Students' performance prediction using feature selection and supervised machine learning algorithms. *Adv Intell Syst Comput* 1165:347–354. [https://doi.org/10.1007/978-981-15-5113-0\\_25/COVER](https://doi.org/10.1007/978-981-15-5113-0_25/COVER)
- Gamboa RA, Namasivayam S, Singh R (2018) Correlation study between CGPA and PO attainments: a case study for taylor's university school of engineering. *Redesig Learn Greater Soc Impact*. [https://doi.org/10.1007/978-981-10-4223-2\\_1](https://doi.org/10.1007/978-981-10-4223-2_1)
- Ganorkar SS, Tiwari N, Namdeo V (2021) Analysis and prediction of student data using data science: a review. *Smart Innov Syst Technol* 182:443–448. [https://doi.org/10.1007/978-981-15-5224-3\\_44/COVER](https://doi.org/10.1007/978-981-15-5224-3_44/COVER)
- Goel R, Jain A, Verma K, Bhushan M, Kumar A, Negi A (2020) Mushrooming trends and technologies to aid visually impaired people. In: International Conference on Emerging Trends in Information Technology and Engineering, Ic-ETITE 2020. Doi <https://doi.org/10.1109/IC-ETITE47903.2020.437>
- Hans C (2011) Elastic net regression modeling with the orthant normal prior. *J Am Stat Assoc* 106(496):1383–1393
- Hodson TO (2022) Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geosci Model Dev* 15:5481–5487. <https://doi.org/10.5194/GMD-15-5481-2022>
- Islam Rifat MR, Al Imran A, Badrudduza ASM (2019) EduNet: A Deep Neural Network Approach for Predicting CGPA of Undergraduate Students. 1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019. <https://doi.org/10.1109/ICASERT.2019.8934616>
- Jadhav A, Pramod D, Ramanathan K (2019) Comparison of performance of data imputation methods for numeric dataset. *Appl Artif Intell*. 33:913–933. <https://doi.org/10.1080/08839514.2019.1637138>
- Jain N (2021) Survey versus interviews: comparing data collection tools for exploratory research. *The Qualitative Report* 26:541–554. <https://doi.org/10.46743/2160-3715/2021.4492>
- Kirkby RB (2007). Improving hoeffding trees (Doctoral dissertation, The University of Waikato).
- Lau ET, Sun L, Yang Q (2019) Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Appl Sci* 1:1–10. <https://doi.org/10.1007/S42452-019-0884-7/FIGURES/6>
- LaValley MP (2008) Logistic Regression. *Circulation* 117:2395–2399. <https://doi.org/10.1161/CIRCULATIONAHA.106.682658>
- Liu Y, Wang Y, Zhang J, (2012) New machine learning algorithm: Random forest. *Lecture Notes in Computer Science* (including subseries *Lecture Notes in Artificial Intelligence* and *Lecture Notes in Bioinformatics*) 7473 LNCS, 246–252. [https://doi.org/10.1007/978-3-642-34062-8\\_32/COVER](https://doi.org/10.1007/978-3-642-34062-8_32/COVER)
- McDonald GC (2009) Ridge regression. *Wiley Interdiscip Rev Comput Stat* 1:93–100. <https://doi.org/10.1002/WICS.14>
- Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD (2004) An introduction to decision tree modeling. *J Chemom* 18:275–285. <https://doi.org/10.1002/CEM.873>
- O'Shea K, Nash R (2015) An Introduction to Convolutional Neural Networks. *Int J Res Appl Sci Eng Technol* 10:943–947. <https://doi.org/10.22214/ijraset.2022.47789>
- Rai S, Shastri KA, Pratap S, Kishore S, Mishra P, Sanjay HA (2021) Machine learning approach for student academic performance prediction. *Adv Intell Syst Comput* 1176:611–618. [https://doi.org/10.1007/978-981-15-5788-0\\_58/COVER](https://doi.org/10.1007/978-981-15-5788-0_58/COVER)
- Rai P, Kothari EE, Bhushan M, Negi A (2023) Vehicle theft identification using machine learning and OCR. *Springer, Singapore*, pp 753–763. [https://doi.org/10.1007/978-981-99-0085-5\\_61](https://doi.org/10.1007/978-981-99-0085-5_61)
- Ranstam J, Cook JA (2018) LASSO regression. *Br J Surg* 105:1348–1348. <https://doi.org/10.1002/BJS.10895>
- Roshan V, Stewart JHM, Joseph R, Stewart HM (2022) Optimal ratio for data splitting. *ASA Data Sci J* 15:531–538. <https://doi.org/10.1002/SAM.11583>
- Saifuzzaman M, Parvin M, Jahan I, Moon NN, Narin Nur F, Shetu SF (2021) Machine Learning Approach to Predict SGPA and CGPA. In: ICAICST 2021–2021 International Conference on Artificial Intelligence and Computer Science Technology 211–216. <https://doi.org/10.1109/ICAICST53116.2021.9497847>
- Samsudin NAM, Shaharudin SM, Sulaiman NAF, Smail SI, Mohamed NS, Husin NHM (2022) Prediction of student's academic performance during online learning based on regression in support vector machine. *Int J Inf Edu Technol*. 12:1431–1435
- Pathan S, Bhushan M, Bai A (2020) A study on health care using data mining techniques. *J Crit Rev* 7(19):7877–7890. <https://doi.org/10.31838/jcr.07.19.896>
- Pawar S, Bhushan M, Wagh M (2020) The plant leaf disease diagnosis and spectral data analysis using machine learning—a review. *Int J Adv Sci Technol* 29(9s):3343–3359

- Sharma D, Aggarwal D (2021) A predictive approach to academic performance analysis of students based on parental influence. *Adv Intell Syst Comput* 1165:75–84. [https://doi.org/10.1007/978-981-15-5113-0\\_6/COVER](https://doi.org/10.1007/978-981-15-5113-0_6/COVER)
- Shetu SF, Saifuzzaman M, Moon NN, Sultana S, Yousuf R (2021) Student's performance prediction using data mining technique depending on overall academic status and environmental attributes. *Adv Intell Syst Comput* 1166:757–769. [https://doi.org/10.1007/978-981-15-5148-2\\_66/COVER](https://doi.org/10.1007/978-981-15-5148-2_66/COVER)
- Staudemeyer, R.C., Morris, E.R., 2019. Understanding LSTM—a tutorial into Long Short-Term Memory Recurrent Neural Networks.
- Su X, Yan X, Tsai CL (2012) Linear regression. *Wiley Interdiscip Rev Comput Stat* 4:275–294. <https://doi.org/10.1002/WICS.1198>
- Suryadewiansyah MK, Tju TEE (2022) Naïve bayes dan confusion matrix untuk efisiensi analisa intrusion detection system alert. *Jurnal Nasional Teknologi Dan Sistem Informasi* 8(2):81–88
- Verma U, Garg C, Bhushan M, Samant P, Kumar A, Negi A (2022) Prediction of student's academic performance using machine learning techniques. *2022 Int Mob Embed Technol Conf MECON 2022*:151–156. <https://doi.org/10.1109/MECON53876.2022.9751956>
- Webb GI (2016) Naïve Bayes. *Encyclopedia of Machine Learning and Data Mining*. Springer, USA, pp 1–2
- Yakubu MN, Abubakar AM (2022) Applying machine learning approach to predict students' performance in higher educational institutions. *Kybernetes* 51:916–934. <https://doi.org/10.1108/K-12-2020-0865>
- Yulianto LD, Triayudi A, Sholihati ID (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

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.