Interpreting the Machine Learning Approaches to Predict CGPA of the University Students

Ranjit Paul [1], Sadiq Hussain [2], Kamala Devi Kannan [3], Bijoy Kumar Mondal [4], Arun K. Baruah [5], Silvia Gaftandzhieva [6]

[1] North Lakhimpur College (Autonomous), Assam, India
[2] Examination Branch, Dibrugarh University, Dibrugarh, Assam, India
[3] Department of Artificial Intelligence and Data Science, MepcoSchlenk Engineering College, Sivakasi, India
[4] Sociology Discipline, Khulna University, Khulna-9208, Bangladesh
[5] Department of Mathematics, Dibrugarh University, Dibrugarh, Assam, India
[6] Faculty of Mathematics and Informatics, University of Plovdiv PaisiiHilendarski, Plovdiv, Bulgaria

***Abstract* – This study focuses on building a model to predict Cumulative Grade Point Average (CGPA) categories and identify students who might struggle academically during their time at university. The study uses student-related data to classify CGPA into specific grades (O, A+, A, B+, B, C+, C and F) and tested different machine learning methods like Naive Bayes, JRIP, J48, Random Forest, and CatBoost. Since some grades, like C and A+, had fewer students, a single-point crossover technique to balance the dataset were used. The models were assessed with ten-fold cross-validation, and CatBoost performed the best after balancing the data. To determine which factors played the largest role in the predictions, SHapley Additive exPlanations (SHAP) were used. It shows that the semester grade point average (SGPA) is one of the primary factors, especially highlighting those students in their 3rd semester who need extra support to achieve a better overall CGPA. These results show how early predictions can help universities support students and improve their academic success and retention.**

DOI: 10.18421/TEM144-50
https://doi.org/10.18421/TEM144-50

Corresponding author: Silvia Gaftandzhieva,
Faculty of Mathematics and Informatics, University of Plovdiv PaisiiHilendarski, Plovdiv, Bulgaria
Email: sissiy88@uni-plovdiv.bg

Received: 30 November 2024.
Revised: 20 June 2025.
Accepted: 27 June 2025.
Published: 27 November 2025.

***Keywords* – Machine Learning, classification, SHAP, cumulative grade point average, universities.**

1. Introduction

Entering university is a significant milestone in a student's life, offering opportunities for learning and personal growth. However, it also brings many challenges that affect their studies and well-being.

Students often find it hard to adjust to the independence of university life after the more structured school environment. Common issues include coping with academic pressure, managing time, balancing studies and social life, and dealing with personal or financial problems. Without proper support, these challenges can lead to poor academic results or even students leaving their courses, which concerns both students and universities. Besides, it is a crucial issue, but remains a persistent challenge for educational institutions. Traditional methods often struggle to identify at-risk students early and provide adequate support [1]. Many current strategies rely on manual processes and basic data analysis, which are not always reliable in identifying students who may need help [2]. Furthermore, the interventions offered are often generic rather than tailored to individual students, reducing their effectiveness [3].

Machine learning is helpful for this purpose, as it can analyse large datasets of student data to find patterns and predict who might face difficulties. It helps universities provide the proper support, such as extra resources or guidance, to help students succeed and stay in their courses. Unlike traditional approaches, this method offers real-time insights into student behaviour, participation, and performance [4]. This paper aims to predict Cumulative Grade Point Average (CGPA) using classification methods to identify struggling students and provide targeted support, such as additional classes or mentoring.

It also seeks to classify students at risk of failure or dropout using these methods, enabling timely interventions to reduce dropout rates. By evaluating and comparing these methods, the study is focused on identifying the most accurate and adaptable approach for predicting at-risk students across different university programmes. Additionally, it incorporates SHapley Additive exPlanations (SHAP) for explainability, offering insights into the key factors influencing outcomes and fostering informed decision-making aligned with eXplainable AI (XAI) principles.

The first objective is to use classification methods to predict the CGPA of students. This prediction will help identify students who might face academic difficulties early in their university life. Identifying these students can allow universities to offer extra support, such as additional classes, one-on-one mentoring, or other resources, to help them improve. Studies show that early academic performance often predicts long-term success and retention in higher education [5]. By focusing on students who may need help, universities can improve their chances of succeeding academically and reduce the risk of dropouts. Once these students are identified, institutions can provide support like counselling, tutoring, or workshops to help them overcome challenges. Early interventions have improved retention rates and created a better learning environment [6].

The second objective is to compare different methods for classifying at-risk students to find the most effective one. Universities often have diverse student populations and unique challenges, so it is essential to determine which classification method works best. By testing and comparing approaches, this study aims to identify a reliable method to be used in one institution and others with similar datasets. It ensures that the findings can be applied to various educational contexts, improving their usefulness and impact [7].

The third objective is to explain how the models make predictions. This involves showing which factors, like prior grades, contributed most to the forecasts. Understanding these factors is essential for teachers, administrators, and policymakers to make informed decisions about supporting students. For instance, if grade is a key factor, universities might focus on encouraging students to attend classes regularly for poorly performed students. This objective also helps make the results more transparent and accessible to trust [8].

In summary, these objectives aim to create a system that identifies students who need help, provide insights into why they might be struggling, and offering solutions to improve their academic success and reduce dropout rates.

The research objectives are described as below:

Objective 1: Students' CGPA are predicted using classification methods. This will help identify students who struggle so it is possible to receive extra attention, such as additional classes or guidance, to improve their performance and stay on track.

Objective 2: The goal is to classify at-risk learners using diverse classification methods. By identifying these students early, universities can take steps to support them and prevent dropouts.

Objective 3: To compare different classification methods and to find the most accurate one for predicting CGPA and at-risk students. The goal is to identify which method works best and ensure it can also be applied to student data.

Objective 4: To explain the results of these predictions (SHAP) and demonstrate the essential features so stakeholders can learn how the results related to XAI are achieved. It helps teachers, policymakers, and other stakeholders understand what influences student outcomes and how to improve them.

These objectives focus on using data to support students, improve academic performance, and reduce dropout rates in higher education.

Figure 1 graphically describes the flow chart and model description.

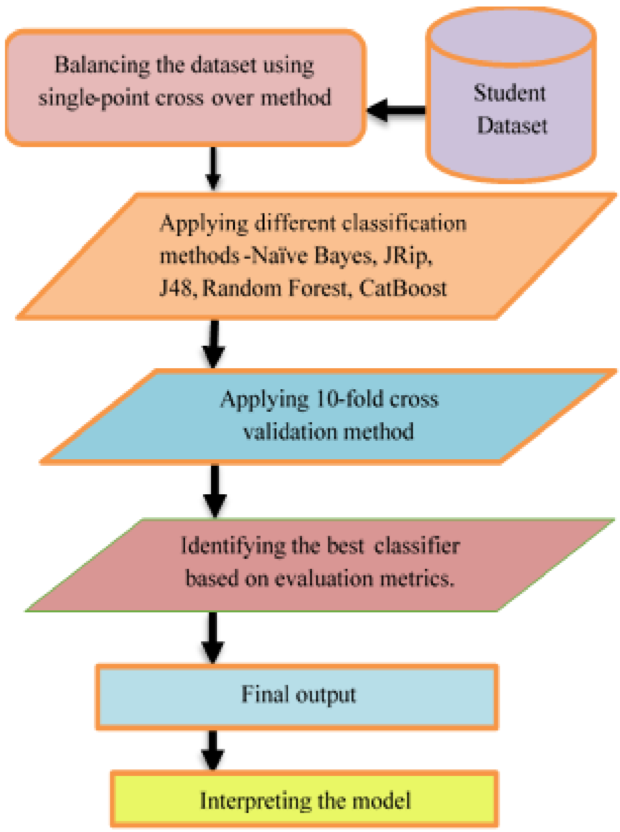

Figure 1. Flow-chart of the model

2. Literature Review

Using data mining, academic performance at a Saudi university was predicted, focusing primarily on pre-admission criteria [9]. The study highlighted the importance of Scholastic Achievement Admission Test (SAAT) scores, with Artificial Neural Networks (ANN) achieving the highest accuracy, surpassing Decision Trees, Support Vector Machines (SVMs), and Naïve Bayes.

Decision Trees, SVM, Naive Bayes, and k-Nearest Neighbour (kNN) algorithms were used to predict students' dropout rates [10]. SVM yielded 99% accuracy across critical metrics, demonstrating its effectiveness in identifying at-risk students for early intervention.

Ensemble learning techniques, particularly boosting, were investigated to predict student performance [11]. Boosting was shown to improve prediction accuracy by applying base classification techniques such as J48, kNN, Naïve Bayes, and Random Tree.

The impact of factors such as school type and study time on academic performance was examined using Decision Tree, JRip, and Naïve Bayes models [12]. The study found that Decision Tree and JRip models achieved over 80% accuracy in predicting student success.

This study [13] proposed an enhanced Random Forest model to predict academic performance, showing that the optimised classifier is better than Naïve Bayes, Bagging, and Boosting.

Feature selection methods were explored, and it was found that Decision Tree classification, combined with the principal component method, performs best in predicting student outcomes [14]. The prediction in academic procrastination in online learning was studied, and various classifiers, including ZeroR, Naïve Bayes, JRip, PART, and J48 were tested [15]. Successful predictions aid in early interventions to improve student motivation.

In their study, the authors [16] used various data mining methods and applied techniques like Correlation Attribute Evaluator (CAE), Information Gain Attribute Evaluator (IGAE) and Gain Ratio Attribute Evaluator (GRAE) to improve prediction accuracy by eliminating irrelevant attributes. Their research shows that the combination of classification algorithms and feature selection significantly refines prediction models.

The study [17] used SVM, Naïve Bayes, and ANN to analyse factors influencing students' academic performance and choice of major, aiming to improve prediction accuracy and inform proactive academic support measures.

In their study, authors [18] applied data mining techniques to predict the academic performance of secondary school students at an early stage, using Naïve Bayes, Random Forest, and J48.

The results showed that Naïve Bayes achieved the best accuracy, emphasising the importance of early interventions.

Early student performance prediction was studied using various classifiers, including rule-based, decision-tree, probabilistic, and multilayer perceptron methods [19]. The results showed that Random Forest produced the highest accuracy.

The study by the authors [20] compared the effectiveness of Naïve Bayes Classifier and Decision Tree J48 in predicting students' academic achievements. After pre-processing (cleaning and discretising), both algorithms achieved over 88.75% accuracy.

In this study [21], six classification algorithms were applied to build prediction models to forecast student performance, recognize at-risk individuals, and enhance academic interventions. It was found that integrated algorithms outperform single classifiers, with Logit Boost being faster than Random Forest.

Several machine learning algorithms (Naïve Bayes, Random Forest, Logistic Regression, kNN, SVM, XGBoost, and Ensemble Voting Classifier) were employed to forecast students' performance, find at-risk individuals, and enhance academic interventions [22]. The study found that ML models outperform ensemble methods, making them helpful in predicting student performance.

The authors in [23] applied Earned Value Management (EVM) and machine learning algorithms Random Forest, KA, kNN, SVM, and Naïve Bayes to aanalyze the factors influencing students' academic performance and forecast their final grades. The authors concluded that Random Forest attained the highest level of accuracy.

Several machine learning algorithms were examined to predict student dropout and academic performance, with class imbalance addressed using SMOTE resampling [24]. SMOTE enhanced model accuracy while boosting algorithms like CatBoost and LightGBM beat traditional methods such as logistic regression and decision trees, applying SHAP to improve model interpretability.

3. Methodology

In this study, machine algorithms, namely Naive Bayes, JRip, J48, Random Forest, and CatBoost were used. Each was selected for its capability to manage different types of data and its potential to uncover meaningful patterns.

3.1. Naïve Bayes

Naive Bayes is a commonly applied probabilistic algorithm in the field of machine learning, mainly employed for classification tasks [25].

Based on the observed features, this algorithm uses Bayes' theorem to estimate the probability that a particular instance falls into a certain class. A core feature of Naive Bayes is its "naive" assumption of conditional independence, meaning it considers each feature independent of the others when given the class label [26]. Despite this simplification, Naive Bayes often performs effectively in real-world applications, delivering strong results even when the independence assumption does not fully hold.

Using Bayes' theorem, Naive Bayes calculates the probability of a class C, given the set of features X. The formula for Bayes' theorem (1) is:

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)} \tag{1}$$

$P(C|X)$ represents the posterior probability of the target class, $P(X|C)$ is the likelihood (the probability of probability of the predictor given the specified class), $P(C)$ is the prior probability of the class, and $P(X)$ is the prior probability of the features.

3.2. Jrip

The JRip classifier, short for "Java RIPPER," applies the RIPPER (Repeated Incremental Pruning to Produce Error Reduction) algorithm. It is a rule-based classification algorithm. Developed by William W. Cohen, RIPPER is particularly suited for binary classification tasks [27]. The algorithm analyses each class, starting with the smallest and moving to the larger ones. For each class, it first creates a basic set of rules. Then, it carefully adjusts these rules to reduce errors one step at a time, making the rules more accurate and reliable as it progresses [28].

3.3. J48

The *Decision Tree* algorithm is applied to identify how the attribute values perform across different instances. Based on the patterns observed in the training data, it can predict the class labels for unseen new cases [29]. It generates rules that assist in predicting the target variable and clarifies the data distribution through tree structures [30]. By applying the J48 algorithm, which extends the ID3 algorithm, various applications can be classified to achieve accurate results [31].

3.4. Random Forest

Random Forest, a machine learning algorithm for classification and regression, builds multiple decision trees and integrates their predictions to enhance accuracy while minimizing overfitting.

This algorithm has three primary hyperparameters, which must be set prior to training.

Node size, the number of trees, and the number of features sampled are the key parameters that allow the classifier to handle regression and classification problems [32].

One advantage of Random Forests is showing which features (or inputs) are most important by calculating how much each affects prediction accuracy. However, training Random Forests can be slow when working with large datasets because of the built decision trees [33].

3.5. CATBOOST

The CatBoost gradient boosting algorithm, introduced in 2018 [34], is designed to process categorical features without the necessity for typical preprocessing approaches. CatBoost represents a form of the gradient boosting algorithm which can manage numerical and categorical data without requiring additional steps to convert categorical data into numbers.

Critical features of CatBoost include:

- The algorithm efficiently handles categorical features without requiring preprocessing.
- It often delivers excellent results with default parameters, saving time on hyperparameter optimisation.
- The model can process missing values without requiring imputation.
- Unlike other models, this algorithm can scale columns automatically.
- This model implements various methods to avoid overfitting, such as robust tree boosting and ordered boosting.
- It internally performs cross-validation to choose the best hyperparameters.
- This algorithm offers a Graphics processing unit (GPU)-accelerated version, allowing users to train models faster on large datasets. This implementation boosts scalability and performance, mainly when using multiple GPU cards.

These features make CatBoost a robust, scalable, and user-friendly tool for machine-learning tasks involving complex datasets.

Due to the Decision Tree-based nature, CatBoost is particularly effective for tasks involving categorical, heterogeneous data. However, it is sensitive to hyperparameter settings, such as the highest number of iterations, tree depth, and categorical feature combinations [35].

3.6. SHAP

SHAP is an XAI technique to understand complex machine-learning models [36]. SHAP is based on principles from cooperative game theory, a mathematical framework used to study how groups of players (or features) can work together and fairly share the rewards (or influence) of their collaboration.

The technique uses explicitly Shapley values, a concept from game theory, to quantify how each player (feature) influences the total payoff (model prediction).

SHAP assigns additive importance values to individual features, ensuring properties like consistency and local accuracy, making it a reliable tool for understanding model behaviour.

4. Results and Discussion

The dataset was collected from North Lakhimpur University in Assam, India. Table 1 describes the dataset.

Table 1. Description of the educational dataset

Attribute	Description	Values
RollNo	Roll Number of a student that uniquely identifies a particular student	18BA001 etc.
Sex	Sex of a particular student	Male, Female
Programme	Programme the student was pursuing	BSC (Bachelor of Science) BA (Bachelor of Arts), BCA (Bachelor of Computer Application)
Year	Admission Year	2018, 2019, 2020
SGPA1, SGPA2, SGPA3, SGPA4, SGPA5, SGPA6	Semester Grade Point Average of 1st, 2nd, 3rd, 4th, 5th and 6th Semester	O (Percentage >=95) A+ (Percentage >=85) A (Percentage >=75) B+ (Percentage >=65) B (Percentage >=55) C+ (Percentage >=45) C (Percentage >=40) F (Percentage <=40)
Core Course	Major Course or Core Course of the student	ANT ANTHROPOLOGY ASM ASSAMESE BCA COURSES OF BCA PROGRAMME BOT BOTANY CHE CHEMISTRY ECO ECONOMICS EDU EDUCATION ELE ELECTRONICS ENG ENGLISH GEO GEOGRAPHY HIN HINDI HIS HISTORY HSC HOME SCIENCE MAT MATHEMATICS PHI PHILOSOPHY PHY PHYSICS PSC POLITICAL SCIENCE
CGPA (target variable)	Cumulative Grade Point Average	O (Percentage >=95) A+ (Percentage >=85) A (Percentage >=75) B+ (Percentage >=65) B (Percentage >=55) C+ (Percentage >=45) C (Percentage >=40) F (Percentage <=40)

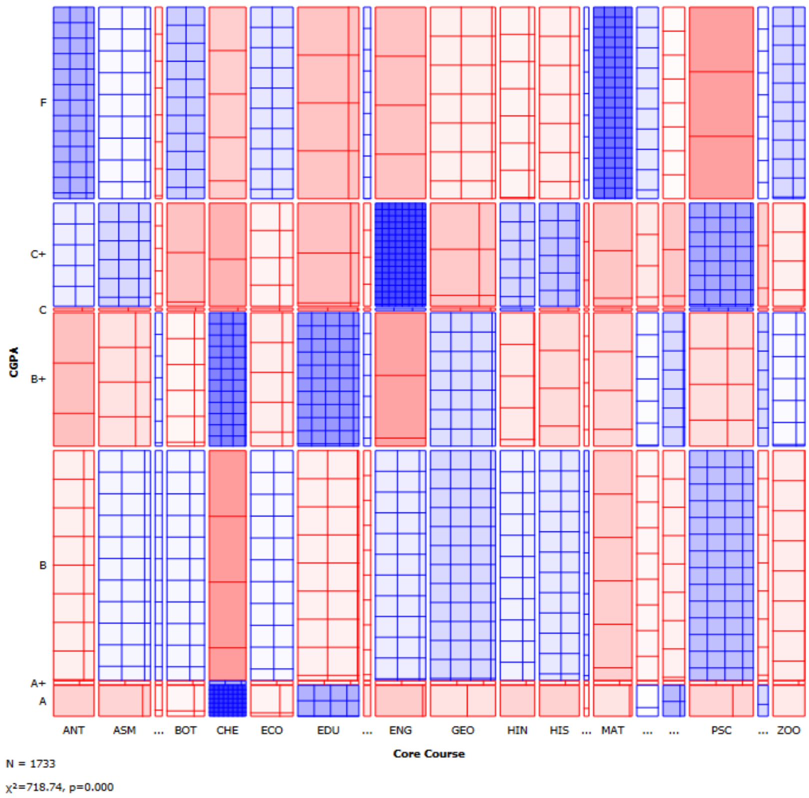

Figure 2. Sieve diagram for courses against CGPA (before balancing)

There are 12 attributes, of which CGPA is the dependent variable. There are 1733 instances, and grades C and A+ of CGPA make the dataset imbalanced. Figure 2 depicts the sieve diagram for different core courses against the CGPA obtained before balancing. A single-point crossover method was utilised to make the dataset balanced. After balancing, the number of instances was raised to 1834. Different machine-learning approaches were applied before balancing the dataset. The outcome is depicted in Table 2. Widely accepted metrics, such as recall, precision, F-measure, and accuracy were applied, to measure the efficacy of the proposed method. Higher values of these metrics represent the model's efficiency.

Table 2. Comparison of different classifiers based on various evaluation metrics (before balancing)

Classifier	Precision	Recall	F-Measure	Accuracy
Naïve Bayes	84.2	86.0	85.1	88.5
JRip	82.2	80.0	81.1	83.5
J48	83.9	83.3	83.4	83.3
Random Forest	83.7	83.0	83.4	85.2
CATBOOST	90.1	91.2	90.7	91.6

The classification technique CatBoost outperformed the other classifiers with an accuracy of 91.6%. Figure 3 demonstrates ‘CGPA’ against ‘Programme’ after balancing the dataset.

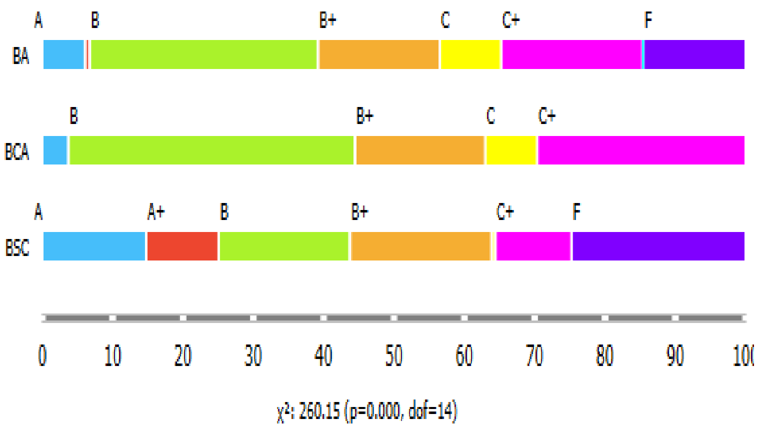

Figure 3. CGPA programme-wise after applying single-point crossover data balancing technique

After balancing the dataset, the following results are obtained and demonstrated in Table 3.

Table 3. Comparison of different classifiers based on various evaluation metrics (after balancing)

Classifier	Precision	Recall	F-Measure	Accuracy
Naïve Bayes	89.0	88.7	88.8	88.7
JRip	84.6	83.9	84.1	83.9
J48	85.5	85.1	85.1	85.1
Random Forest	87.1	86.2	86.0	86.2
CATBOOST	92.7	92.7	92.7	92.7

In both cases, a 10-fold cross-validation technique was applied. CatBoost showcased superior performance with an accuracy of 92.7%. Table 4 provides the confusion matrix for the CatBoost classifier.

Table 4. CatBoost Classifier's Confusion Matrix

a	b	c	d	e	f	g	<-- classified as
293	16	26	0	0	0	0	a = B+
14	72	0	0	0	0	0	b = A
20	0	525	24	0	0	0	c = B
0	32	229	0	0	0	0	d = C+
0	0	1	1	474	0	0	e = F
0	0	0	0	0	55	0	f = C
0	0	0	0	0	0	52	g = A+

Two feature selection attributes were applied to find the most influential characteristics of the dataset. Table 5 demonstrates the top 5 attributes using the two rank-based methods to do the task.

Table 5. Most influential attributes based on rank-based feature selection techniques

Feature Selection Method	High Influence Attributes
ReliefF Ranking Filter	SGPA5 (Rank: 0.46315), SGPA3 (0.377221), SGPA6 (0.32659), SGPA4 (0.288187), SGPA1 (0.275048)
Gain Ratio feature evaluator	SGPA5 (Rank: 0.4339), SGPA3 (0.3831), SGPA4 (0.3059), SGPA6 (0.2917), SGPA1 (0.2697)

The results presented in the above table show that SGPA5 and SGPA3 are the most influential attributes. For the 2018 admitted batch, SGPA6 was obtained by appearing in the examination online during the COVID period. Additionally, SGPA6 is compared as a response variable with the SGPA of other semesters.

To understand the independent variable (gender, SGPA1, SGPA2, SGPA3, SGPA4, SGPA5 and program) related to the dependent variable (SGPA6 based on 2018 online results), Pearson's chi-square test was executed, along with Fishers Exact test when appropriate. The Table 6 showed that student gender was significant with SGPA6 ($p < 0.048$). Further, the table data shows that SGPA1 and SGPA2 were significantly associated with SGPA6 ($P<0.000$ and $P<0.000$) respectively. Moreover, SGPA3, SGPA4 and SGPA5 were statistically significant with SGPA6 ($p<0.000$, $p<0.000$ and $p<0.000$) correspondingly. The study also found that the program available in the institution was also significantly related to SGPA6 ($p<0.000$) based on online examination data from 2018.

Table 6. Results of Chi-square test based on 2018 data

Independent Variables	Dependent Variable (SGPA6)	
	$\chi 2$(df)	Fisher's Exact Test
Sex	12.713(6)*	12.228
SGPA1	260.065(36)**	207.017
SGPA2	448.541(36)**	200.893
SGPA3	438.635(42)**	309.877
SGPA4	412.744(36)**	299.210
SGPA5	364.855(42)**	326.768
Program	67.914(12)**	61.869

*** significant at 1% level * significant at 5% level*

Most machine learning models follow the black-box oriented approach. To make the results interpretable and explainable, CatBoost with the SHAP method was applied.

Figure 4 showing feature importance, ranks attributes based on their impact. Here, semester grades like those from the fifth and third semesters stand out as the most influential, meaning playing a major role in determining a student's overall CGPA. Other factors, such as earlier semester grades, also contributed but to a lesser extent.

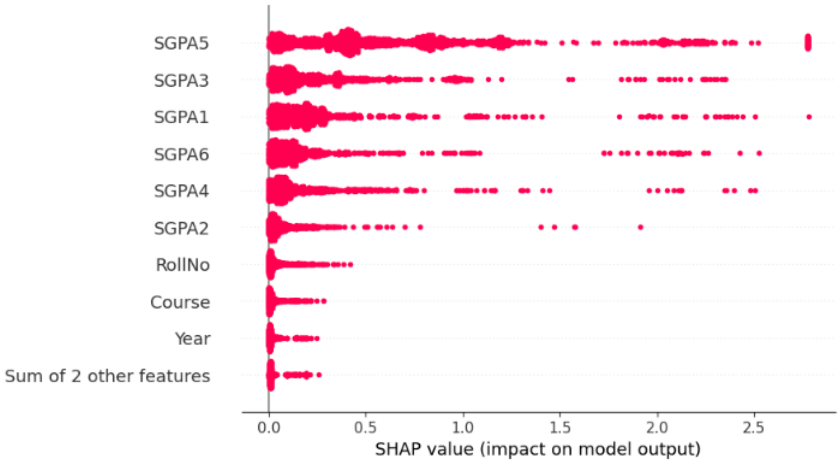

Figure 4. SHAP values having features of importance

Figure 5 provides a detailed view of how individual students' performance trends affect predictions. Warmer colors represent higher grades, while cooler colors indicate lower performance.

This visualization helps identify patterns—for instance, students with consistently high grades in later semesters tend to have better predicted CGPAs.

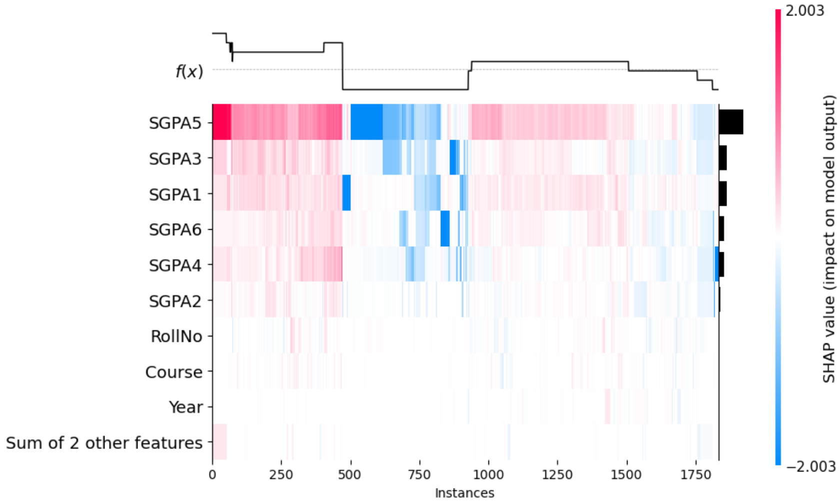

Figure 5. Heatmap based on SHAP values

Figure 6 displays the distribution of each feature's influence across all predictions. Wider sections indicate where most students fall, revealing that higher grades in key semesters, like the fifth, generally push CGPA predictions upward.

In contrast, other factors, such as the academic program or gender, show more varied effects.

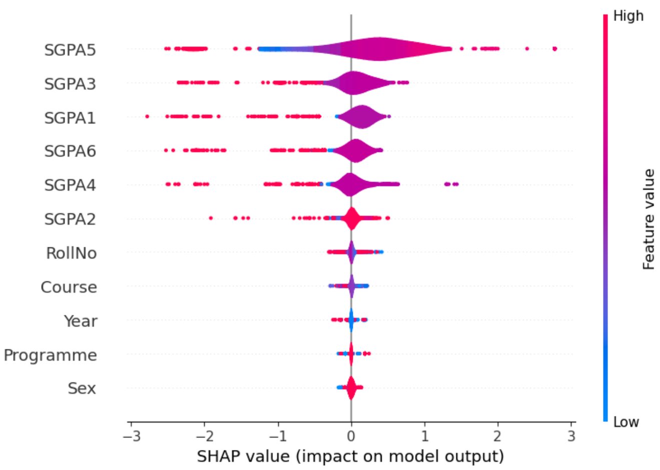

Figure 6. Violin plot using SHAP

Figure 7 breaks down the prediction for a single student, showing how each feature adjusts the final CGPA estimate. Starting from an average baseline, the plot illustrates whether a student's grades in specific semesters raise or lower their predicted outcome.

This helps educators understand why the model predicts a certain CGPA for an individual, making the decision-making process transparent.

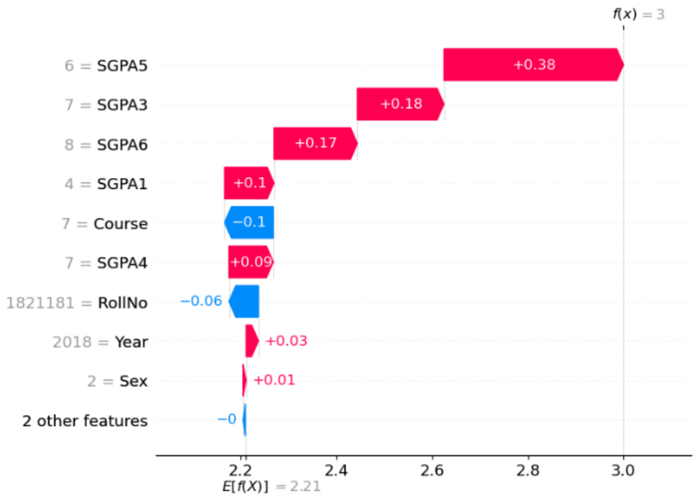

Figure 7. Waterfall plot for a single data point

Analysing these diagrams and ranking important features leads to the conclusion that SGPA5 and SGPA3 are the most influential features and contribute immensely to obtaining the evaluation metrics. These figures also provide clear, interpretable insights, ensuring that educators can trust and act on the model's predictions.

5. Conclusion

In this study, CatBoost yielded superior performance among the other classifiers for predicting CGPA. As SGPA5, and SGPA3 are the key attributes in the process, the low-performing students after obtaining the 3rd semester results may be given extra attention to avoid dropouts, which fulfils the first objective. Although the 5th semester results strongly influence overall CGPA, till the 5th semester it will be too late to handle such low-performing students. Hence, after 3rd semester, it will be the prime time to focus on such students.

The study employed five classification techniques (Naïve Bayes, JRip, J48, Random Forest, and CatBoost). CatBoost demonstrated superior performance, achieving the highest accuracy and F1-score. Future plans include testing the process on other university datasets to validate its efficacy, as this has not yet been undertaken. This fulfils the second objective of the study. The third and final objective was implemented by devising SHAP as part of the XAI framework.

It was a retrospective study which may be considered one of the study's limitations. Three programs from North Lakhimpur University, which are still active, were chosen for this study.

The results are believed to significantly assist the university's teachers and policymakers, as well as other counterparts to a great extent. The size of the dataset is another limitation. Deep learning strategies were applied to predict the CGPA. As deep learning techniques perform best with numerous instances, the collection of 2021 to 2024 datasets is planned for future research goals, enabling the employment of these methods.

Acknowledgements

This paper is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project № BG-RRP-2.004-0001-C01. The paper reflects only the author's view and the EU is not responsible for any use that may be made of the information it contains.

References:

[1]. Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews)*, *40*(6), 601-618.

[2]. Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. *Learning Analytics*. Springer. Doi: 10.1007/978-1-4614-3305-7_4

[3]. Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing personalized education: A dynamic framework. *Educational Psychology Review*, *33*(3), 863-882.

[4]. Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498-506.

[5]. Herzog, S. (2005). Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. *Research in higher education*, *46*, 883-928.

[6]. Cele, N. (2021). Big data-driven early alert systems as means of enhancing university student retention and success. *South African Journal of Higher Education*, *35*(2), 56-72.

[7]. Abu-Oda, G. S., & El-Halees, A. M. (2015). Data mining in higher education: university student dropout case study. *International Journal of Data Mining & Knowledge Management Process*, *5*(1), 15. Doi: 10.5121/ijdkp.2015.5102

[8]. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*. Doi: 10.48550/arXiv.1705.07874

[9]. Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *Ieee Access*, *8*, 55462-55470.

[10]. Lottering, R., Hans, R., & Lall, M. (2020). A model for the identification of students at risk of dropout at a university of technology. *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 1-8.

[11]. Ashraf, M., Zaman, M., & Ahmed, M. (2020). An intelligent prediction system for educational data mining based on ensemble and filtering approaches. *Procedia Computer Science*, *167*, 1471-1483.

[12]. Walia, N., et al. (2020). Student's academic performance prediction in academic using data mining techniques. *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*.

[13]. Jayaprakash, S., Krishnan, S., & Jaiganesh, V. (2020). Predicting students academic performance using an improved random forest classifier. In *2020 international conference on emerging smart computing and informatics (ESCI)*, 238-243.

[14]. Triayudi, A., & Fitri, I. (2021). Comparison of the feature selection algorithm in educational data mining. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *19*(6), 1865-1871.

[15]. Xu, H., et al. (2021). Prediction and visualization of academic procrastination in online learning. *Proceedings of the 2021 6th International Conference on Distance Education and Learning*, 133-139.

[16]. Nidhi, N., et al. (2021). Predictive Model for Students' Academic Performance Using Classification and Feature Selection Techniques. *2021 2nd International Conference on Computational Methods in Science & Technology (ICCMST)*, 106-111. IEEE.

[17]. Veluri, R. K., et al. (2022). Learning analytics using deep learning techniques for efficiently managing educational institutes. *Materials Today Proceedings*, *51*(8), 2317-2320.

[18]. Alghamdi, A. S., & Rahman, A. (2023). Data mining approach to predict success of secondary school students: A Saudi Arabian case study. *Education Sciences*, *13*(3), 293.

[19]. Meghji, A. F., et al. (2023). Using educational data mining to predict student academic performance. *VFAST Transactions on Software Engineering*, *11*(2), 43-49.

[20]. Tasfia, S., et al. (2024). Exploring Weka and Python for Educational Data Mining: Naïve Bayes vs. J48. *2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, 470-475.

[21]. Feng, G., & Fan, M. (2024). Research on learning behavior patterns from the perspective of educational data mining: Evaluation, prediction and visualization. *Expert Systems with Applications*, *237*, 121555.

[22]. Bellaj, M., et al. (2024). Educational Data Mining: Employing Machine Learning Techniques and Hyperparameter Optimization to Improve Students' Academic Performance. *International Journal of Online & Biomedical Engineering*, *20*(3).

[23]. Aydogan, S. K., Pura, T., & Bingül, F. (2024). Predicting Students' Academic Performances Using Machine Learning Algorithms in Educational Data Mining. *Malaysian Online Journal of Educational Technology*, *12*(4), 131-153.

[24]. Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, *4*(1), 2.

[25]. Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. Wiley.

[26]. Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, *29*, 131-163.

[27]. Cohen, W. W. (1995). Fast effective rule induction. *Proceedings of the twelfth international conference on machine learning*, 115-123.

[28]. Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, *13*(1), 61-72.

[29]. Korting, T. S. (2006). C4. 5 algorithm and multivariate decision trees. *Image processing division, national institute for space research–INPE.*

[30]. Nadali, A., Kakhky, E. N., & Nosratabadi, H. E. (2011). Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system. *2011 3rd International Conference on Electronics Computer Technology*, *6*, 161-165.

[31]. Lee, L. E., et al. (2020). Evaluation of prediction algorithms in the student dropout problem. *Journal of Computer and Communications*, *8*(3), 20-27.

[32]. Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

[33]. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). Springer.

[34]. Prokhorenkova, L., et al. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

[35]. Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. *Journal of big data*, *7*(1), 94.

[36]. Kalita, E., et al. (2025). Educational data mining: a 10-year review. *Discover Computing*, *28*(1), 1-25.