

Assessment of ML Techniques for Student Dropout Prediction: A Comparative Study

Nisha Rani¹ and Pachigolla Venkata Suresh²

¹Research Scholar, School of Computer and Information Sciences (SOCIS), IGNOU, New Delhi.
Email: nishaphdignou@gmail.com

²Professor, School of Computer and Information Sciences (SOCIS), IGNOU, New Delhi.
Email: pvsuresh@ignou.ac.in

Abstract—Student dropout is a significant issue in higher education, affecting institutional efficiency and individual student outcomes. Existing research on dropout prediction methods lacks comprehensive evaluation, robust cross-validation steps and fails to evaluate various ML models on diverse metrics, necessitating a thorough comparative analysis. To this end this study aims to assess various ML techniques for predicting student dropout, comparing their effectiveness and identifying key predictors. The work analysed four widely-used ML models LR, RF, DT, and SGT. Our evaluation metrics included accuracy, precision, recall, and the area under the ROC curve to ensure a comprehensive performance. Furthermore, accuracy of the models is assessed with application of cross-validation and without application of cross-validation employed to ensure robust and reliable performance comparisons. The results indicate that LR and RF are the most effective models, achieving accuracies of 77.03% and 76.8%, respectively with highest precision and a robust recall.

Index Terms— Student dropout prediction, machine learning (ML), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Stochastic Gradient Descent (SGT).

I. INTRODUCTION

Student dropout in higher education [1], [2] represents a significant and multifaceted issue with profound implications for individuals, institutions, and society at large. This phenomenon is characterized by students discontinuing their studies before completing their degree programs, resulting in a range of negative outcomes [3]. For students, dropping out often leads to diminished career prospects, lower earning potential, and the burden of educational debt without the compensatory benefits of a degree. The decision to leave school can stem from various factors including academic difficulties, financial challenges, personal issues, and a lack of engagement or support within the educational environment [4]. Higher education institutions face considerable repercussions due to student dropout [5]; these include financial losses from decreased tuition revenue, increased costs related to recruitment and retention efforts, and potential damage to their reputation and rankings. High dropout rates can also impact the institution's ability to secure funding and grants, as these often depend on graduation rates and overall student success metrics. On a broader scale, student dropout contributes to a less educated workforce, which can hinder economic growth and innovation. It exacerbates social inequalities, as those from disadvantaged backgrounds are often disproportionately affected. The societal cost also includes lost potential in terms of human capital and productivity.

The issue of student dropout is complex and influenced by a myriad of interconnected factors. Academic preparedness, socio-economic status, mental health, and external responsibilities such as work and family all play crucial roles. Additionally, institutional factors such as the quality of academic advising, availability of support services, campus climate, and the overall inclusivity and accessibility of the educational environment significantly impact student retention [6]. Efforts to mitigate student dropout rates have included a range of strategies, from early intervention and support programs to the implementation of more flexible learning pathways and financial aid. However, identifying at-risk students and effectively addressing the diverse and dynamic causes of dropout remain significant challenges. In recent years, ML techniques have emerged as powerful tools to predict student dropouts [7], [8] by analysing large datasets encompassing academic performance, demographic information, and behavioural indicators, ML models can identify patterns and risk factors associated with dropout. These predictive insights enable institutions to proactively target interventions and support resources to those most in need, potentially mitigating the risk of dropout and improving overall retention rates. Despite the significant impact of student dropout on educational institutions and individual futures, existing studies on dropout prediction often suffer from several limitations and gaps. Many prior research efforts focus on a limited set of machine learning techniques without providing a comprehensive comparative analysis of multiple models.

II. DETAILED EXPLANATION

A. Related Work

Numerous studies have consistently shown that academic performance, often measured by GPA [9], is a strong predictor of student retention. For instance, a study by Tinto [4] established the link between academic integration and student persistence, suggesting that students with higher GPAs are less likely to drop out. Similarly, a more recent study [10] confirmed that students who perform well academically are more likely to persist and graduate. Regular class attendance and active engagement in coursework are also critical predictors. In [11] study found that attendance has a significant impact on student performance and retention. Engagement in learning activities, attendance also plays a crucial role in determining whether students stay or leave. Research has highlighted the impact of socio-economic status [12] on dropout rates. Students from lower SES backgrounds face greater financial pressures and often have less access to educational resources, which increases their risk of dropping out. In recent years, the use of machine learning and predictive analytics has gained prominence in identifying at-risk students. Studies [13] demonstrated the effectiveness of ML models in predicting dropout by analysing various predictors, including academic performance, demographic factors, and behavioural data [14] Utilizing classification algorithms the study analyses predictors like past academic performance, attendance, and online participation. Another study [15] aims to predict both student dropout and academic success using a variety of ML techniques. By integrating data on academic performance, demographic information, and behavioural factors, the research constructs predictive models to identify at-risk students and those likely to succeed. The paper [16] evaluates multiple ML techniques to predict student dropout in university courses. Techniques such as LR, DTs, and Neural Networks are employed to analyse factors including academic records, engagement metrics, and socio-demographic data. Another more recent research [17] delves into the reasons behind student dropout by leveraging ML for predictive modelling and factor analysis. Using algorithms like RFs and Gradient Boosting, the study examines a wide range of predictors including academic performance, financial status, and personal circumstances. Existing studies reviewed frequently lack robust cross-validation methods, which are crucial for ensuring the reliability and generalizability of the predictive models. The datasets used in previous research are often narrow in scope, failing to incorporate a diverse range of factors such as socio-economic status, behavioural patterns, and academic performance over time. Furthermore, the interpretability and practical implementation of these models in real-world educational settings remain underexplored. Addressing these limitations is essential for developing more accurate, reliable, and actionable predictive models that can effectively identify at-risk students and inform timely intervention strategies. This study aims to fill these gaps by rigorously evaluating multiple ML techniques, employing thorough cross-validation, and utilizing a comprehensive dataset to enhance the prediction of student dropout.

B. Logistic Regression

LR is a widely used statistical method for binary classification problems and is particularly suited for predicting the probability of an event occurring [18]. In the context of predicting student dropout, LR can be employed to model the likelihood that a student will drop out based on various input features such as demographics, academic

performance, behavioural data, and support interactions. LR works by fitting a logistic function to the input data.

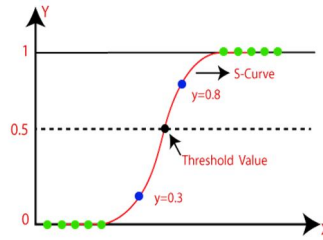


Figure 1: Logistic regression and thresholding of the s-curve [19]

The logistic function, also known as the sigmoid function, outputs a probability value between 0 and 1, which can be interpreted as the likelihood of a particular outcome—in this case, whether a student will drop out (1) or not (0). LR provides clear insights into the influence of each feature on the probability of dropout. The coefficients indicate the strength and direction of the relationship between each predictor and the dropout likelihood. It is computationally efficient and can be trained quickly even on relatively large datasets. The model outputs probabilities, which can be thresholded as shown in the figure 1 to classify students into dropout risk categories and facilitate targeted interventions.

C. SGT

SGT is an iterative method for optimizing a differentiable objective function, typically a loss function. Unlike traditional gradient descent [20], [21], which calculates the gradient of the loss function using the entire dataset, SGD updates the model parameters using only a single sample (or a small batch of samples) at each iteration. This results in faster convergence, especially for large datasets. In this study, SGD can be applied to optimize the parameters of several ML models, such as Neural Networks. By using SGD [21], these models can efficiently learn from the student dataset, to accurately predict the likelihood of student dropout. SGD is particularly advantageous for large datasets as it updates parameters more frequently and converges faster than batch gradient descent. It can handle large-scale data and complex models with numerous parameters.

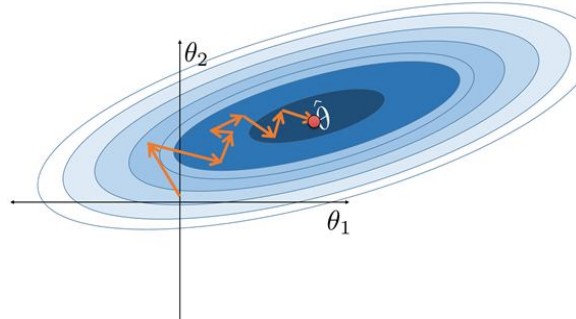


Figure 2: illustration of the Scaled gradient decent [22]

In student dropout prediction, models such as LR or neural networks can be trained using SGD. These models aim to classify whether a student will drop out or persist based on various features such as academic performance, demographic information, and engagement metrics. During training, SGD can help in identifying important features by observing the weight updates. Features with larger weight magnitudes are often more influential in the model's predictions. This insight can be used to understand which factors are most predictive of dropout.

D. Decision Tree

A DT is a popular and intuitive ML algorithm used for both classification and regression tasks [23]. In the context of predicting student dropout a DT can be employed to model the decision-making process that leads to a student either continuing their studies or dropping out. A DT algorithm builds a model in the form of a tree structure where each internal node represents a decision based on the value of a specific feature, each branch represents the outcome of that decision, and each leaf node represents a class label (in this case, dropout or no dropout).

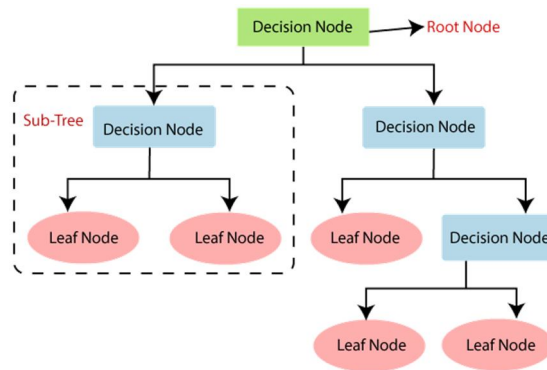


Figure 3: Process of the Decision Making Using Decision Trees [24]

The dataset is split into subsets based on an attribute value test. This process is repeated recursively, selecting the best attribute to split the data at each step. The best split is chosen using a measure of impurity or information gain. DTs are easy to interpret and visualize. The tree structure provides clear insights into the decision-making process, showing which features are most important in predicting dropout. DTs can rank features by their importance, helping identify the most significant factors contributing to student dropout. Once trained, the DT can predict the dropout probability for new students by traversing the tree based on their feature values until reaching a leaf node. The path followed through the tree provides a clear and interpretable explanation for the prediction, showing which factors contributed most to the dropout risk. The model's decisions are easy to visualize and understand as shown in figure 3, making it accessible to educators and administrators who can use the insights to develop targeted interventions. DTs can process both numerical and categorical data, making them versatile for educational datasets.

E. Random Forest

RF is an ensemble ML technique that combines multiple DTs to improve the accuracy and robustness of predictions. It is particularly effective for classification tasks such as predicting student dropout due to its ability to handle complex, nonlinear relationships and its robustness to overfitting [25], [26]. In the context of the proposed work on assessing ML techniques for student dropout prediction, RF serves as a powerful and versatile model that can leverage a wide range of predictor variables. RF works by constructing a multitude of DTs during training and outputting the class that is the mode of the classes (classification) of the individual trees. Each tree in the forest is trained on a random subset of the training data (with replacement).

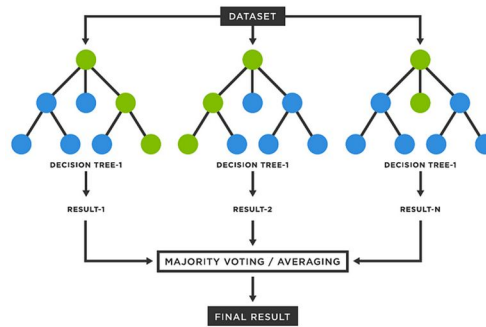


Figure 4: working of a random forest classifier [27]

This process, known as bagging, helps in reducing variance and improving the model's generalizability. During the construction of each tree, RF randomly selects a subset of features to consider for splitting at each node. This ensures that the trees are diverse and reduces the likelihood of overfitting. For classification tasks, the final prediction is based on the majority vote of all the individual trees. This ensemble approach helps in improving the accuracy and robustness of the model. RF tends to achieve high accuracy due to its ensemble nature, which reduces overfitting and variance compared to individual DTs. The algorithm can handle a wide range of data types and is effective with both numerical and categorical variables. By providing measures of feature importance, RF helps educators and administrators understand the key drivers of student dropout, facilitating

targeted interventions. By using RF, the university can accurately identify at-risk students and understand the underlying factors contributing to dropout. This enables the institution to design and implement targeted support programs to help these students, thereby improving retention rates and overall student success.

F. Experimental Analysis

The experimental analysis involves training and evaluating several ML algorithms to predict student dropout using the dataset from the Polytechnic Institute of Portalegre [15]. Each algorithm is assessed both with and without k-fold cross-validation ($k = 10$) using Python and the scikit-learn library. The performance of each model is compared based on accuracy, precision, recall, and ROC curve metrics. The initial train-test split results provide a baseline for comparison. Each algorithm's accuracy, precision, recall, and AUC are recorded. This method is straightforward but may suffer from high variance and may not fully represent the model's performance on unseen data. The 10-fold cross-validation results are averaged across the folds to yield a more reliable and generalizable performance measure. This method reduces the risk of overfitting and ensures that the model's performance is consistent across different subsets of the data. The results of both training methods are compared to determine the impact of cross-validation on model performance.

G. Dataset Used

The dataset used in this study is available as a CSV file, the Student dropout dataset of a higher education institution having 35 features and 4424 rows [28].

It contains comprehensive data on students enrolled between the academic years 2008-2009 at the Polytechnic Institute of Portalegre [15]. The dataset includes various features such as student demographics, academic records socio-economic and macro-economic, records, and notably contains no missing values. This rich and complete dataset serves as a robust foundation for building and evaluating ML models aimed at predicting academic performance and student dropout, thereby aiding the tutoring team in identifying at-risk students and implementing timely interventions.

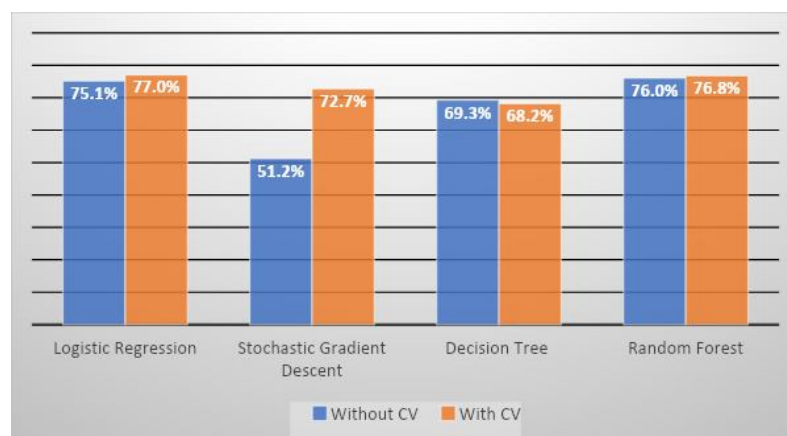


Figure 5: Accuracy of the LR, SGD, DT, and RF algorithms with using CV and without CV

The Figure 5 compares the accuracy of four ML models—LR, SGD, DT, and RF in predicting student dropout, with and without cross-validation (CV). Without CV, RF achieves the highest accuracy (76.0%), closely followed by LR (75.1%), DT (69.3%), and SGD (51.2%). With CV, all models show slight improvements except for DT, which slightly decreases from 69.3% to 68.2%. LR improves to 77.0%, SGD significantly boosts to 72.7%, and RF marginally increases to 76.8%. The enhancements with CV highlight its effectiveness in providing more reliable and robust accuracy estimates, especially for SGD, which shows the most notable improvement. RF and LR emerge as the most accurate models overall, both with and without CV, indicating their strong performance and stability in predicting student dropout.

TABLE I. EVALUATION RESULTS OF THE LR, SGD, DT, AND RF USING ACCURACY, PRECISION AND RECALL

Metric	LR	SGT	DT	RF
Accuracy	77.0%	72.7%	68.2%	76.8%
Precision	80.0%	75.7%	70.2%	78.8%
Recall	75.0%	69.7%	67.2%	76.0%

The table I presents performance metrics accuracy, precision, and recall of four ML models used for student dropout prediction, with CV applied. LR shows the highest precision at 80.0%, meaning when it predicts a student will drop out, it is correct 80% of the time. This indicates a low false positive rate. RF also shows . high precision at 78.8%, showing a similar ability to correctly identify true dropout cases. SGD shows moderate precision at 75.7%, indicating a relatively balanced prediction capability. Whereas the DT shows a Lower precision at 70.2%, implying a higher false positive rate compared to the other models. RF also has the highest recall at 76.0%, meaning it correctly identifies 76% of all actual dropout cases, indicating a low false negative rate, Figure 7. slightly lower recall (LR:) at 75.0%, still effectively identifying most dropout cases. SGT shows moderate recall at 69.7%, indicating some missed true dropout cases. DT possess the Lowest recall at 67.2%, suggesting it misses a higher proportion of actual dropout cases. LR and RF are the most effective models overall, balancing high accuracy, precision, and recall. SGT (SGD) performs moderately well but lags behind LR and RF in all three metrics. DT has the lowest performance across all metrics, indicating it is the least reliable model for predicting student dropout in this context. This analysis suggests that LR and RF are preferable choices for predicting student dropout due to their superior performance in terms of accuracy, precision, and recall.

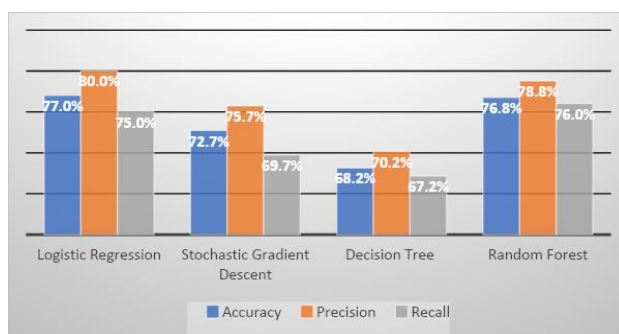


Figure 7: Comparative Visualization of the four classifiers using accuracy, precision and recall

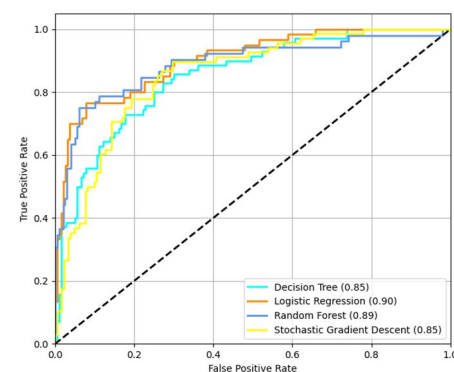


Figure 8: ROC and AUC of selected ml models for dropout prediction

The ROC curve is a graphical representation of a classifier's performance across different threshold values. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR), providing insight into the trade-off between sensitivity and specificity for the classifier. The figure compares the ROC curves of four ML algorithms LR, SGT (SGD), DT, and RF for predicting student dropout rates. A higher AUC signifies better performance, with a perfect classifier achieving an AUC of 1.0. LR achieves the highest area under the curve (AUC) 0.90 indicating that LR is the most effective model at distinguishing between students who will drop out and those who will not. RF has an AUC of 0.89, slightly lower than LR but still high showing that RF is also very effective in predicting student dropout, with strong discriminative ability. Maintains a high TPR similar to LR, but with a slight dip in some regions, indicating very competitive but marginally less robust performance compared to LR. While slightly lower than LR and RF, SGD still performs well in distinguishing between dropout and non-dropout students. Reasonably high, but the curve shows more variability, indicating that its performance is less stable across different thresholds. DT also has an AUC of 0.85, equal to SGD Indicating DT is effective, though not as strong as LR and RF. The curve is slightly lower in some regions compared to LR and RF, showing that DT is less consistent in identifying true positives across different thresholds. The ROC curves and AUC values suggest that while all four models are competent for predicting student dropout.

III. CONCLUSIONS

The issue of student dropout is complex and influenced by a myriad of interconnected factors. Academic preparedness, socio-economic status, mental health, and external responsibilities such as work and family all play crucial roles. However, identifying at-risk students and effectively addressing the diverse and dynamic causes of dropout remain significant challenges. By leveraging advanced ML techniques, educational institutions can gain a deeper understanding of the dropout phenomenon, tailor their support systems more effectively, and ultimately foster a more supportive and successful educational environment for all students. To this end this study aims to assess various ML (ML) techniques for predicting student dropout, comparing their effectiveness

and identifying key predictors. Our results indicate that LR and RF are the most effective models, achieving accuracies of 77.0% and 76.8%, respectively. LR displayed the highest precision (80.0%) and a robust recall (75.0%), while RF demonstrated strong performance with a precision of 78.8% and the highest recall of 76.0%. The ROC analysis further confirmed the superiority of these models, with LR achieving the highest AUC (0.90) followed by RF (0.89), indicating their excellent ability to distinguish between students who drop out and those who do not. SGT and DT, although effective, showed comparatively lower performance, with accuracies of 72.7% and 68.2% respectively, and AUC values of 0.85. The DT's lower precision (70.2%) and recall (67.2%) suggest higher variability and less consistency in predicting student dropout. Future research should explore several directions to build upon these findings. Even though selected ML models provided good results, however, use of deep learning may further improve the predictive performance and coupling deep learning model with better pre-processing steps such as clustering and feature selection may further enhance the generalization capabilities. Future research should also focus on the interpretability and transparency of predictive models to ensure that educators and policymakers can understand and trust the insights generated.

REFERENCES

- [1] G. Eason, [1] B. M. Kehm, M. R. Larsen, and H. B. Sommersel, "Student dropout from universities in Europe: A review of empirical literature," *Hung. Educ. Res. J.*, vol. 9, no. 2, pp. 147–164, 2019.
- [2] J. M. Ortiz-Lozano, A. Rua-Vieites, P. Bilbao-Calabuig, and M. Casadesús-Fa, "University student retention: Best time and data to identify undergraduate students at risk of dropout," *Innov. Educ. Teach. Int.*, 2018.
- [3] A. Behr, M. Giese, H. D. Tegum Kamdjou, and K. Theune, "Dropping out of university: a literature review," *Rev. Educ.*, vol. 8, no. 2, pp. 614–652, 2020.
- [4] M. do C. Nicoletti, "Revisiting the Tinto's theoretical dropout model," *High. Educ. Stud.*, vol. 9, no. 3, pp. 52–64, 2019.
- [5] A. Burke, "Student retention models in higher education: A literature review," *Coll. Univ.*, vol. 94, no. 2, pp. 12–21, 2019.
- [6] M. Tight, "Student retention and engagement in higher education," *J. Furth. High. Educ.*, vol. 44, no. 5, pp. 689–704, 2020.
- [7] B. Prenkaj, P. Velardi, G. Stilo, D. Distant, and S. Faralli, "A survey of machine learning approaches for student dropout prediction in online courses," *ACM Comput. Surv. CSUR*, vol. 53, no. 3, pp. 1–34, 2020.
- [8] A. Alhothali, M. Albsisi, H. Assalahi, and T. Aldosemani, "Predicting student outcomes in online courses using machine learning techniques: A review," *Sustainability*, vol. 14, no. 10, p. 6199, 2022.
- [9] M. G. Gallego, A. P. Perez de los Cobos, and J. C. G. Gallego, "Identifying students at risk to academic dropout in higher education," *Educ. Sci.*, vol. 11, no. 8, p. 427, 2021.
- [10] J. Kinzie and G. Kuh, "Reframing student success in college: Advancing know-what and know-how," *Change Mag. High. Learn.*, vol. 49, no. 3, pp. 19–27, 2017.
- [11] A. P. Miller, *Examining The Efficacy of Attendance as A Predictor of Academic Performance*. Sam Houston State University, 2019.
- [12] C. Aina, E. Baici, G. Casalone, and F. Pastore, "The determinants of university dropout: A review of the socio-economic literature," *Socioecon. Plann. Sci.*, vol. 79, p. 101102, 2022.
- [13] S. Jayaprakash and V. Jaiganesh, "A survey on academic progression of students in tertiary education using classification algorithms," *Int. J. Eng. Technol. Sci. Res. IJETS*, vol. 5, no. 2, pp. 136–142, 2018.
- [14] A. Sarra, L. Fontanella, and S. Di Zio, "Identifying students at risk of academic failure within the educational data mining framework," *Soc. Indic. Res.*, vol. 146, pp. 41–60, 2019.
- [15] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, p. 146, 2022.
- [16] J. Kabathova and M. Drlik, "Towards predicting student's dropout in university courses using different machine learning techniques," *Appl. Sci.*, vol. 11, no. 7, p. 3130, 2021.
- [17] S. Kim, E. Yoo, and S. Kim, "Why Do Students Drop Out? University Dropout Prediction and Associated Factor Analysis Using Machine Learning Techniques." *arXiv*, Oct. 17, 2023. doi: 10.48550/arXiv.2310.10987.
- [18] A. Das, "Logistic Regression," in *Encyclopedia of Quality of Life and Well-Being Research*, F. Maggino, Ed., Cham: Springer International Publishing, 2023, pp. 3985–3986. doi: 10.1007/978-3-031-17299-1_1689.
- [19] "Logistic Regression in Machine Learning - Javatpoint," *www.javatpoint.com*. Accessed: May 28, 2024. [Online]. Available: <https://www.javatpoint.com/logistic-regression-in-machine-learning>
- [20] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford, "Accelerating Stochastic Gradient Descent for Least Squares Regression," in *Proceedings of the 31st Conference On Learning Theory*, PMLR, Jul. 2018, pp. 545–604. Accessed: May 28, 2024. [Online]. Available: <https://proceedings.mlr.press/v75/jain18a.html>
- [21] Y. Liu, Y. Gao, and W. Yin, "An improved analysis of stochastic gradient descent with momentum," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 18261–18271, 2020.

- [22] “Stochastic gradient descent - Cornell University Computational Optimization Open Textbook - Optimization Wiki.” Accessed: May 28, 2024. [Online]. Available: https://optimization.cbe.cornell.edu/index.php?title=Stochastic_gradient_descent
- [23] M. Batra and R. Agrawal, “Comparative Analysis of Decision Tree Algorithms,” in *Nature Inspired Computing*, B. K. Panigrahi, M. N. Hoda, V. Sharma, and S. Goel, Eds., Singapore: Springer, 2018, pp. 31–36. doi: 10.1007/978-981-10-6747-1_4.
- [24] B. Charbuty and A. Abdulazeez, “Classification Based on Decision Tree Algorithm for Machine Learning,” *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, Art. no. 01, Mar. 2021, doi: 10.38094/jastt20165.
- [25] IBM, “What is Random Forest?” Accessed: Aug. 31, 2023. [Online]. Available: <https://www.ibm.com/topics/random-forest>
- [26] A. Parmar, R. Katariya, and V. Patel, “A Review on Random Forest: An Ensemble Classifier,” in *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, J. Hemanth, X. Fernando, P. Lafata, and Z. Baig, Eds., Cham: Springer International Publishing, 2019, pp. 758–763. doi: 10.1007/978-3-030-03146-6_86.
- [27] D. Gunay, “Random Forest,” Medium. Accessed: May 28, 2024. [Online]. Available: <https://medium.com/@denizgunay/random-forest-af5bde5d7e1e>
- [28] Valentim Realinho, Jorge Machado, Luís Baptista, & Mónica V. Martins. (2021). Predict students' dropout and academic success (1.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.5777340>