Print ISSN: 2321-6379
Online ISSN: 2321-595X

Original Article

Predicting Students Academic Success and Dropout Using Supervised Machine Learning

Divvyam Arora

High School Senior STEM (Scholar Badge) Student and President, Physics Club (Redshift), Delhi Public School, Gurgaon, Haryana, India

Abstract

Introduction: Dropout in higher education is a phenomenon of utmost importance that must be addressed. Its' effect on the student, institution and the socio-economic growth of the country cannot be over looked. To improve the academic success rates, higher education institutions are increasingly searching for effective measures to identify the students at risk and prevent dropouts.

Purpose: The purpose of this research project was to explore multiple machine learning techniques to identify at-risk students early in their academic journey, enabling timely intervention and support.

Materials and Methods: By utilizing a comprehensive dataset from different undergraduate programs and 18 machine learning tools, a classification model was developed using python that analyzes academic and demographic features to effectively identify students in need of assistance.

Results: Tuned XGBoost and Stacking Classifier gave the best results, but Stacking Classifier was preferred due to its high precision value. The features that were found most important in the prediction process of the two selected machine learning models were; "Curricular units 2nd semester (approved)', 'Curricular units 1st semester (approved)', 'Curricular units 1st semester (enrolled)', 'Curricular units 2nd semester (grade)', 'Course' and 'Curricular units 1st semester (evaluations)'. However, 'Course' was not an important feature for Stacking Classifier.

Conclusion: The outcomes of this research demonstrate the selection of Stacking classifier as an appropriate machine learning model that can be used to reliably identify features and make predictions related to student academic success and dropout. The results showcase the potential of machine learning to contribute to proactive measures for student success.

Key words: Academic, Algorithm, Attribute, Dataset, Dropout, Feature, Machine learning, Model, Performance, Success

INTRODUCTION

Dropout in higher education is a phenomenon of utmost importance that must be addressed.[1] It leads to wastage of resources as well as decreased rates of students' satisfaction and success. This, in turn, has an adverse impact on job market and socioeconomic status in the country.[2] Its' effect on the student, institution, and the socioeconomic growth of the country cannot be overlooked.[3,4]

Dropout does not have any universally accepted definition. It is a very complex and multifactorial phenomenon. The dropout rates vary between various reported studies. This reason for this variation is accounted to; how dropout is defined, the source of data, and the method of calculation.[1,4]

Access this article online

www.ijss-sn.com

Month of Submission : 07-2023
Month of Peer Review : 07-2023
Month of Acceptance : 08-2023
Month of Publishing : 09-2023

The common definition refers to students leaving studies without completing their academic program and obtaining the graduation degree. This defines dropouts from a macro-perspective.[5] This definition leads to much lower dropout rates as compared to micro-perspective definition where even institution and program changes are included in dropouts.[1]

The dropout can be voluntary, like, if a student takes transfer to another institution or program (which are not cases of proper dropouts); or decides to leave the institution for a lucrative job offer. The dropout can also be forced due to financial crisis and personal or family-related problems. Another factor that is frequently analyzed

Corresponding Author: Divvyam Arora, H. No. 880, Sector 46, Gurgaon, Haryana, India.

in the reported researches is the timing of the dropout (early or late). Dropout has impact on society (in terms of socioeconomics), the institution (in terms of funding, performance, and academic success), and at a personal level for the student and his/her family (self-doubts and waste of time and money). At the level of nation, various countries need to widen the policies for increasing the number of highly qualified people for the society and economy. In all, there are push and pull factors that can be related to dropout. Push factors are within the chosen institution or program relating to preferences and competencies of the students; pull factors are from outside the institution or program relating to personal or family problems, financial crisis, or job offers.[6,7]

To improve academic success rates, higher education institutions are increasingly searching for effective measures to identify the students at risk and prevent dropouts.[1] Reducing academic dropout and failure in higher education is a crucial goal. Various researchers have presented machine learning models to predict student outcomes in the past decade (2010–20).[8] Early warning systems have also been developed by various research workers in the past.[9-11] More researches have been conducted in recent years to predict the student outcomes at the earliest.[12-15]

However, many studies mentioned in the literature have utilized limited dataset and machine learning models. The purpose of this research project was to explore multiple machine learning models to identify at-risk students early in their academic journey, enabling timely intervention and support.

METHODOLOGY

There are three types of methodologies that are used in analyzing the research problem where main objectives and sub-objectives are used. They are quantitative, qualitative, and mixed. As far as qualitative are concerned, the dataset involves an intuitive reasoning which cannot be translated into variables dependent or independent. The quantitative dataset is easily depicted in the form of dependent and independent variables which could be further studied in the form of correlation and regression analysis and various statistical tests to indicate the extent of dependency of the dependent variable on the independent ones. A mixed-method approach uses both quantitative and qualitative explanations to understand and predict the final result.[16] This research paper uses primarily quantitative dataset to analyze the reasons as to why there is an increasing dropout rate at the undergraduate level using sophisticated machine learning and advanced algorithms. Addressing the issues discovered would go a long way in reduction of dropout rates.

Data

The dataset used was retrieved from the University of California, Irvine's Machine Learning Repository. This was taken from various higher education institutions which were disjointed among each other. The common factor was that it was related to students who were enrolled in various undergraduate degrees. It included demographic data, socioeconomic data, and microeconomic data. It also included academic data of students at the time of enrollment as well as at the end of the first and second semesters. It contained 4424 records in total with 36 attributes. The link for dataset is mentioned below.

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Data Pre-processing

The dataset was modified to fit comma-separated values (CSV) format. This allows data to be saved in tabular format and is mostly used for importing and exporting important information. It is a plain text file that stores data by delaminating data entries with a comma. CSVs can be opened in text editors, spreadsheet programs, and specialized applications. The original repository consisted of three main dependent variables, namely "graduate," "dropout," and "enrolled." In the research paper under consideration, only "graduate" and "dropout" were considered, as these two dependent variables would be adequate to explain the reasons behind the dropout rates, indicating a binary outcome. The dataset consisted of 36 independent and one dependent variable, which was the final outcome to be predicted by the model [Table 1].

Processing Primary Data

Coding technology was performed using "Python" programming language on the "Jupyter Lab" software. The process involved in importing necessary "libraries" (a collection of precompiled codes that can be used later on in a program for some specific well-defined operations.). The analysis involved converting attribute data types in the following form, for example.

- Marital status being a discrete variable was converted into categories
- The data were then checked for the following:
  - Repeat values
  - The enrolled and graduate numbers were synchronized with each other.

Data Analysis

Variable dependency

Used Φk correlation matrix. Those variables with a correlation of <0.4 with the final outcome were removed [Figure 1].

Table 1: Dataset attributes and their brief description

S. No.	Attribute name	Role	Type	Brief description	Missing values
1	Marital status	Feature	Integer	Demographic	None
2	Nationality	Feature	Integer	Demographic	None
3	Displaced	Feature	Integer	Demographic	None
4	Gender	Feature	Integer	Demographic	None
5	Age at enrollment	Feature	Integer	Demographic	None
6	International	Feature	Integer	Demographic	None
7	Application mode	Feature	Integer	Academic data at enrollment	None
8	Application order	Feature	Integer	Academic data at enrollment	None
9	Course	Feature	Integer	Academic data at enrollment	None
10	Daytime/evening attendance	Feature	Integer	Academic data at enrollment	None
11	Previous qualification	Feature	Integer	Academic data at enrollment	None
12	Previous qualification (grade)	Feature	Continuous	Academic data at enrollment	None
13	Admission grade	Feature	Continuous	Academic data at enrollment	None
14	Mother's qualification	Feature	Integer	Socioeconomic	None
15	Father's qualification	Feature	Integer	Socioeconomic	None
16	Mother's occupation	Feature	Integer	Socioeconomic	None
17	Father's occupation	Feature	Integer	Socioeconomic	None
18	Educational special needs	Feature	Integer	Socioeconomic	None
19	Debtor	Feature	Integer	Socioeconomic	None
20	Tuition fees up to date	Feature	Integer	Socioeconomic	None
21	Scholarship holder	Feature	Integer	Socioeconomic	None
22	Curricular units 1st sem (credited)	Feature	Integer	Academic data at end of semester 1	None
23	Curricular units 1st sem (enrolled)	Feature	Integer	Academic data at end of semester 1	None
24	Curricular units 1st sem (evaluation)	Feature	Integer	Academic data at end of semester 1	None
25	Curricular units 1st sem (approved)	Feature	Integer	Academic data at end of semester 1	None
26	Curricular units 1st sem (grade)	Feature	Integer	Academic data at end of semester 1	None
27	Curricular units 1st sem (without evaluations)	Feature	Integer	Academic data at end of semester 1	None
28	Curricular units 2nd sem (credited)	Feature	Integer	Academic data at end of semester 2	None
29	Curricular units 2nd sem (enrolled)	Feature	Integer	Academic data at end of semester 2	None
30	Curricular units 2nd sem (evaluation)	Feature	Integer	Academic data at end of semester 2	None
31	Curricular units 2nd sem (approved)	Feature	Integer	Academic data at end of semester 2	None
32	Curricular units 2nd sem (grade)	Feature	Integer	Academic data at end of semester 2	None
33	Curricular units 2nd sem (without evaluations)	Feature	Integer	Academic data at end of semester 2	None
34	Unemployment rate	Feature	Continuous	Macroeconomic	None
35	Inflation rate	Feature	Continuous	Macroeconomic	None
36	GDP	Feature	Continuous	Macroeconomic	None
37	Target	Target	Categorical		None

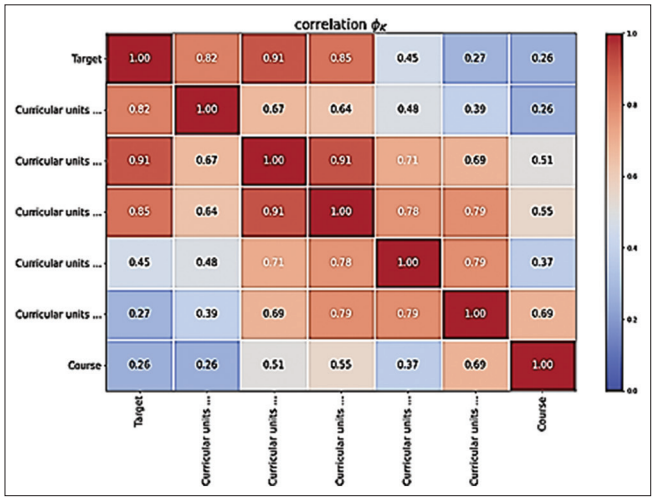

Figure 1: Φk correlation matrix

Descriptive analysis

- Histograms were plotted for each variable to authenticate data distribution.
- Outliers (data points that differ significantly from other observations) were removed and boxplots were plotted for each variable.

In this case, all values are beyond the range of (quartile 1–1.5* interquartile range [IQR] to quartile 3 +1.5*IQR).

- Bivariate analysis was done for the following variables with the target variable – curricular units 1st semester (approved), curricular units 1st semester (enrolled), curricular units 2nd semester (approved), curricular units 2nd semester (grade), and curricular units 1st semester (evaluations) [Figure 2a-e].

Machine Learning Model Training and Evaluation

- The data were further split into 30% testing set and 70% training set. A python function was created to calculate different metrics – accuracy [used to measure the model performance in terms of measuring the ratio of sum of true positive and true negatives out of all the predictions made, precision (used to measure the model performance in measuring the count of true

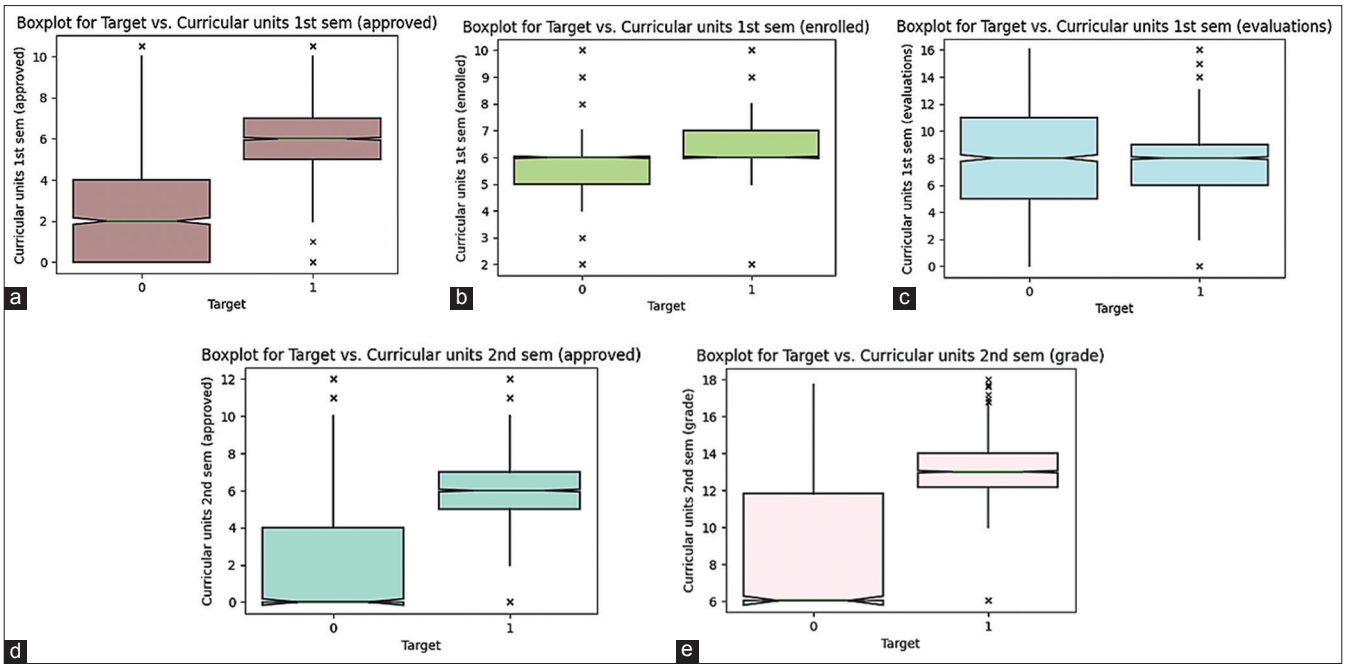

Figure 2: (a-e) Bivariate data analysis

positives in the correct manner out of all positive predictions made, recall (used to measure the model performance in terms of measuring the count of true positives in a correct manner out of all the actual positive values), F1 score (harmonic mean of precision and recall scores) – for each model. A function was created to make a confusion matrix, which is used to map the relationship between predicted and actual values, for each model.

- A total of 18 machine learning models (Neural Network, Logistic Regression, K Nearest Neighbor, AdaBoost, Gradient Boost, XGBOOST, Tuned Random Forest (RF), Support Vector Machine, Tuned Bagging Classifier, Decision Tree (GridSearch), Stacking Classifier, Combo Stacking Classifier, Tuned XGBOOST, Bagging Classifier, Tuned Bagging Classifier, Decision Tree, Tuned Decision Tree, and RF) were worked upon to try and reach a best-fit solution.
- Greater emphasis was given to "precision" than "recall." The meaning of high precision is that the model has minimum false-positive values while a higher recall score means that the model has minimum false-negative values. For the purpose of this study, precision is preferred as predicting that a student will graduate if, in reality, they would not is worse than predicting that a student will dropout if in reality, they will not.
- Compared test metrics (accuracy, precision, recall, and F1 score) of all models by first sorting them based on their accuracy and then on their precision score. Ultimately, the models with an optimal value of accuracy and precision score were chosen [Figure 3].

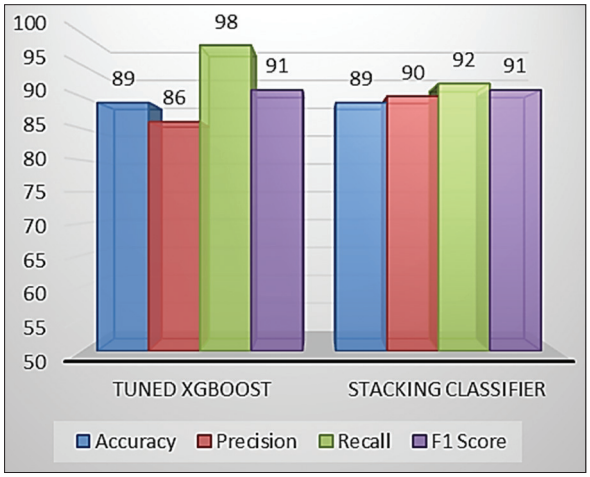

Figure 3: Comparison of test metrics for selected models

- A confusion matrix [Figure 4] was plotted for the selected models and area under the curve of a receiver operating characteristic (AUC ROC) graph was calculated to further gauge the model's predicting ability and efficiency [Figures 5 and 6].

Feature Importance

- Feature importance is important to understand the data and also plays a role in the improvement and interpretation of the models.[17] A feature importance graph was created for the selected models to understand the value of each independent attribute in the decision-making process of the selected models [Figures 7 and 8].

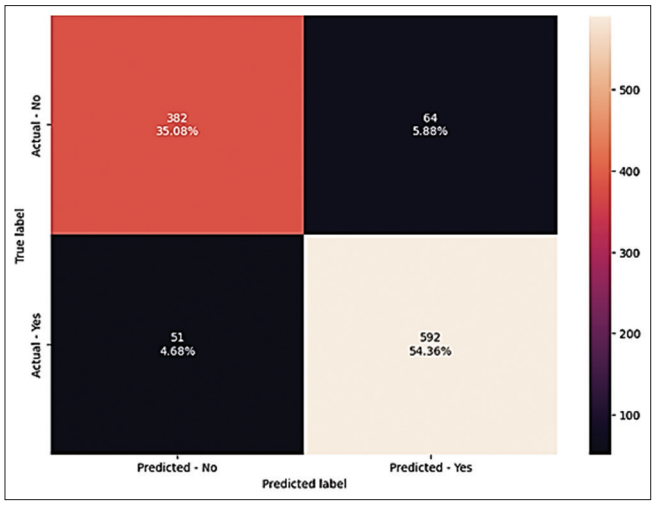

Figure 4: Stacking classifier confusion matrix

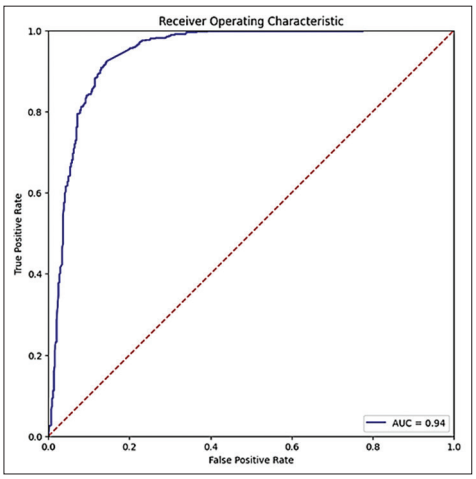

Figure 5: Receiver operating characteristic curve for stacking classifier

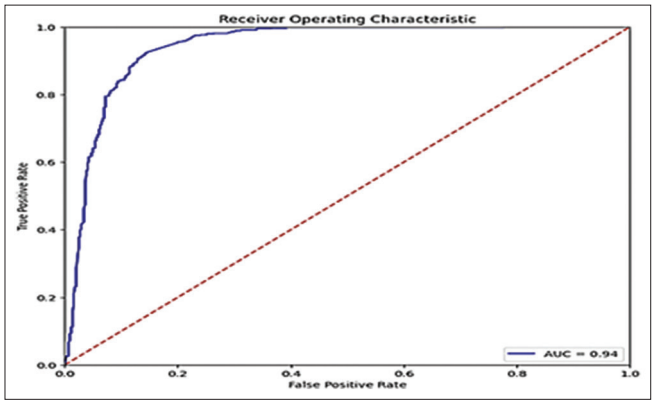

Figure 6: Receiver operating characteristic curve for tuned XGBoost

RESULTS

First, variable dependency was found with Φk correlation matrix. Those variables with a correlation of <0.4 with the final outcome were removed. Eventually, only 7 out of 35 independent variables were used for training the machine models, which were "Course," "Tuition fees up to date," "Curricular units 1st semester (enrolled)," "Curricular units 1st semester (evaluations)," "Curricular units 1st semester (approved)," "Curricular units 2nd semester (approved)," and "Curricular units 2nd semester (grade)."

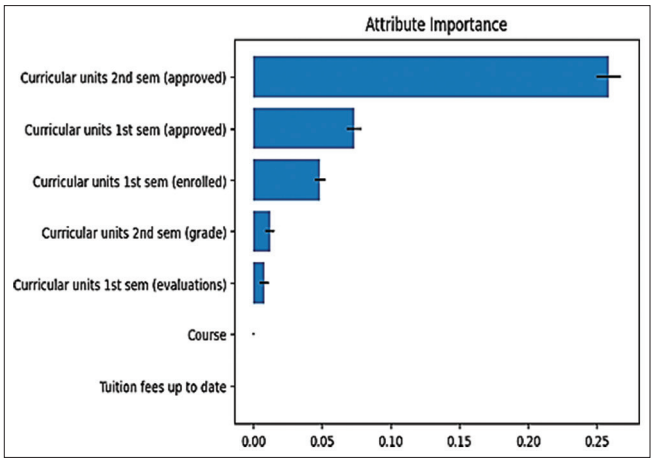

Figure 7: Feature importance for stacking classifier

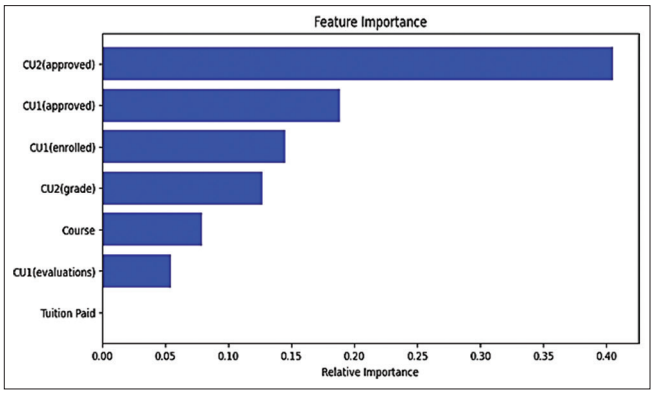

Figure 8: Feature importance for tuned XGBoost

The features that were found most important in the prediction process of the two best machine learning models were "Curricular units 2nd semester (approved)," "Curricular units 1st semester (approved)," "Curricular units 1st semester (enrolled)," "Curricular units 2nd semester (grade)," "Course," and "Curricular units 1st semester (evaluations)." However, "Course" was not an important feature for Stacking Classifier.

Out of 18 machine learning models used, ultimately two models offered the best results. One of them was Tuned XGBOOST (boosting method) and the other one was Stacking Classifier (stacking method) (preferred due to higher precision). Neural Networks gave the highest recall score, but their other metrics were low similar to other advanced models. The model that was eventually used

was stacking classifier (stacking method) model as it was equipped with AdaBoost (a statistical classification meta-algorithm), tuned AdaBoost (hyperparameter tuning), and Gradient Boost (a ML technique used in regression).

DISCUSSION

Out of all machine learning models used, Ultimately Tuned XGBOOST (boosting method) and the Stacking Classifier (stacking method) (preferred due to higher precision) offered the best results. The model that was eventually used was stacking classifier (stacking method) model.

The features that were found most important in the prediction process of the two best machine learning models were "Curricular units 2nd semester (approved)," "Curricular units 1st semester (approved)," "Curricular units 1st semester (enrolled)," "Curricular units 2nd semester (grade)," "Course," and "Curricular units 1st semester (evaluations)." However, "Course" was not an important feature for Stacking Classifier.

Similar study was conducted by researchers like Realinho *et al*.[2] They utilized the dataset from the Polytechnic Institute of Portalegre that was created by collecting information from various disjointed sources from students enrolled in different undergraduate programs, such as education, nursing, management, design, technologies, agronomy, journalism, and social service. The data included socioeconomic, demographic, macroeconomic, academic data on enrollment, and academic performance at the end of each of the two semesters with 35 attributes. The dataset was utilized to build machine learning models that could predict academic performance and dropout. The researchers utilized the RF,[18] Catboost (CATBOOST),[19] light gradient boosting machine,[20] and extreme gradient boosting (XGBOOST).[21]

The most important feature was determined using the Permutation Feature Importance technic, using F1 as the error metric. The features that were considered important in the all algorithms used were "Curricular units sem 2 (approved)," "Curricular units sem 1 (approved)," "Curricular units sem 2 (grade)," "Course," and "Tuition fees up to date." The features "Curricular units sem 1 (enrolled)," "Curricular units sem 1 (evaluations)," "Curricular units sem 2 (enrolled)," and "Curricular units sem 2 (evaluations)" were important in three of the algorithms. The authors concluded that the dataset was useful for conducting comparative studies on student success rates and also for further training in the area of machine learning tools. The machine learning tools are of great help in predicting the risks of dropout and failures.

Authors like Saa *et al*.[12] studied the dataset of a private university in the United Arab Emirates. The dataset of students' information consisted of 56,000 records with 34 attributes. The workers recommended the RF algorithm for data mining used to predict the academic performance of the students. The study also recognized four main categories to which the most important attributes affecting the student performance belonged. These included information of course and instructor, information of students' demographics, general information of student, and information related to previous performance of student. The investigators concluded that results of this study can help the higher education universities to identify the features that affect the academic performance of the students to build early warning system for predicting the low performance or failure of students. The authors also commented that educational data mining is an upcoming field that can aid in mining and identifying the important attributes in any available educational data.

Workers like Martins *et al*.[15] used a dataset from a higher education university to build classification models that can help predict performance of students. The dataset included information recorded at the time of student's registration such as the socioeconomics, demographics, and the academic path. Both standard and boosting algorithms were tested, and classification models were trained and evaluated. The results demonstrated that boosting algorithms respond better than the standard the algorithms to any classification task. However, these boosting algorithms failed to appropriately identify the cases in one of the minority classes. Additional information related to 1st-year performance of the students can also be included for further studies. These researchers stated that the machine learning techniques can help in early identification of the students at risk of failure of their academic path and help to devise strategies in advance to support them.

Many studies mentioned in the literature have utilized limited dataset and machine learning models. The purpose of this research project was to explore multiple machine learning models to identify at-risk students early in their academic journey, enabling timely intervention and support. By utilizing a comprehensive dataset from different undergraduate programs, a classification model was developed using python that analyzes academic and demographic features to effectively identify students in need of assistance. The outcomes of this research demonstrate the selection of an appropriate machine learning model, thereby showcasing the potential of machine learning to contribute to proactive measures for student success.

However, the dataset has to be continuously updated with latest student data to make reliable predictions in the long

term. Various studies in the literature have demonstrated that early prediction of students at risk is possible; however, it is difficult to generalize the utility of prediction models into the different programs.[13] We might have to consider more attributes related to region, instruction mode, cultural factors, and the details about dropout and transfer to make the prediction model more generalized.

CONCLUSION

Within the limitations of the present research, it can be concluded that

1. Tuned XGBoost and Stacking Classifier gave the best results, but Stacking Classifier was preferred due to its high precision value (XGBoost has higher recall). The accuracy, F1 score, and AUC ROC were same for both models. Stacking Classifier can be used to reliably make predictions related to student academic success and dropout.
2. The features that were found most important in the prediction process of the two selected machine learning models (Tuned XGBoost and Stacking Classifier) were "Curricular units 2nd semester (approved)," "Curricular units 1st semester (approved)," "Curricular units 1st semester (enrolled)," "Curricular units 2nd semester (grade)," "Course," and "Curricular units 1st semester (evaluations)." However, "Course" was not an important feature for Stacking Classifier.

ACKNOWLEDGMENT

I would like to extend my gratitude for Ms Aditi Misra, Director-principal of our school and my teachers, mentors, and parents for motivating me and guiding me in this project and all my scholastic activities.

This research project was done under mentorship and esteemed guidance of Prof Jamie Fairclough as a requirement for Stanford Pre-Collegiate Summer Program course entitled "Introduction to Machine Learning" conducted at Stanford University, USA.

REFERENCES

1. Behr A, Giese M, Kamdjou HD, Theune K. Motives for dropping out from higher education-an analysis of bachelor's degree students in Germany. Eur J Educ 2021;56:325-43.
2. Realinho V, Machado J, Baptista L, Martins MV. Predicting student dropout and academic success. Data 2022;7:146.
3. Nurmalitasari N, Long ZA, Noor FM. Factors influencing dropout students in higher education. Educ Res Int 2023;2:1-13.
4. Kehm BM, Larsen MR, Sommersel HB. Student dropout from universities in Europe: A review of empirical literature. Hungarian Educ Res J 2020;9:147-64.
5. Larsen MR, Sommersel HB, Larsen MS. Evidence on Dropout Phenomena at Universities. Brief Version. Copenhagen, Denmark: Danish Clearinghouse for Educational Research; 2013.
6. Bound J, Turner S. Dropouts and diplomas: The divergence in collegiate outcomes. In: Hanushek EA, Machin S, Woessmann L. Handbook of the Economics of Education. Vol. 4. Netherlands: Elsevier; 2011. p. 573-613.
7. Ulriksen L, Madsen LM, Holmegaard HT. What do we know about explanations for drop out/opt out among young people from STM higher education programmes? Stud Sci Educ 2010;462:209-44.
8. Namoun A, Alshanqiti A. Predicting student performance using data mining and learning analytics techniques: A systematic review. Appl Sci 2020;11:1-28.
9. Arnold KE. Signals: Applying academic analytics. Educause Quarterly 2010;33. 33(1):1-10. Available from: https://er.educause.edu/articles/2010/3/educause-quarterly-magazine-volume-33-number-1-2010
10. Costa EB, Fonseca B, Santana MA, De Araújo FF, Rego J. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Comput Hum Behav 2017;73:247-56.
11. Hu YH, Lo CL, Shih SP. Developing early warning systems to predict students' online learning performance. Comput Hum Behav 2014;36:469-78.
12. Saa AA, Al-Emran M, Shaalan K. Mining student information system records to predict students' academic performance. Adv Intell Syst Comput 2020;921:229-39.
13. Akçapinar G, Altun A, Askar P. Using learning analytics to develop early-warning system for at-risk students. Int J Educ Technol High Educ 2019;16:1-20.
14. Daud A, Lytras MD, Aljohani NR, Abbas F, Abbasi RA, Alowibdi JS. Predicting student performance using advanced learning analytics. In: WWW 2017 Companion: Proceedings of the 26th International World Wide Web Conference 2017. Vol. 3-7. Perth, Australia; 2017. p. 415-21.
15. Martins MV, Tolledo D, Machado J, Baptista LM, Realinho V. Early prediction of student's performance in higher education: A case study. In: Advances in Intelligent Systems and Computing. Vol. 1365. Germany: Springer; 2021. p. 166-75.
16. Edmonds WA, Kennedy TD. An applied guide to research designs: Quantitative, qualitative, and mixed methods. 2nd ed. Thousand Oaks, CA. Sage Publications; 2016.
17. Saarela M, Jauhiainen S. Comparison of feature importance measures as explanations for Classification models. SN Appl Sci 2021;3:1-12.
18. Ho TK. Random Decision Forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada. Vol. 1; 1995. p. 278-2.
19. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: Unbiased Boosting with Categorical Features. arXiv 2019; 1-23. Available from: https://arxiv.org/pdf/1706.09516v5.pdf
20. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al*. LightGBM: A highly efficient gradient boosting Decision Tree. In: Advances in Neural Information Processing Systems. Vol. 30. United States: MIT Press; 2017. p. 3147-55.
21. Chen T and Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference, San Francisco, CA, USA; 2016. p. 13-7.

How to cite this article: Arora D. Predicting Students Academic Success and Dropout Using Supervised Machine Learning. Int J Sci Stud 2023;11(6):72-78.

Source of Support: Nil, **Conflicts of Interest:** None declared.