A Comparative Study of Student Success and Retention Rates in Pursuing Higher Education

1st Tanmay Sule
Depatment of Mathematical Sciences
Stevens Institute of Technology
Jersey City, USA
tsule@stevens.edu

2nd Prerna Desai
Depatment of Mathematical Sciences
Stevens Institute of Technology
Jersey City, USA
pdesai21@stevens.edu

3rd Akanksha Wagh
Depatment of Mathematical Sciences
Stevens Institute of Technology
Hoboken, USA
awagh3@stevens.edu

***Abstract*—With the educators, institutions, and policymakers pushing towards better-skilled graduates, understanding the cause and effects of the existing education system is necessary. So is studying the impact it has made on students. So, as a conscious group of individuals, we will be studying the student success rates in their pursuit of higher education. This research project aims to enhance student success and retention by analyzing factors influencing dropout rates in diverse academic disciplines. Leveraging a comprehensive dataset encompassing demographic, socio-economic, and academic performance data, we seek to uncover the motivations driving students to persist or discontinue their educational pursuits. Our insights will enable institutions to implement targeted interventions and support mechanisms to enhance student retention and academic achievement.**

I. Introduction

One of the most important issues facing higher education is student retention. In education, retention rate is a crucial assessment parameter. High retention rates are achieved as students re-enroll from one academic year to the next. To do this, educational institutions should offer the proper guidance and instruction, in addition to other strategies to improve student performance and keep them from putting off finishing their education. Over the years, this field has received a great deal of attention in the education sector due to the curiosity of numerous scholars and organizations about what causes lead to students not finishing their degrees and what factors do not. Because of low student retention, institutions may find it difficult to maintain a high graduation rate. According to general statistics, the percentage of undergraduate college students who drop out is 40year of study. The issue of student dropouts remains a major worry, particularly for higher education institutions searching for more effective retention techniques, even though the reasons for these decisions vary greatly depending on the data analyzed and the many perspectives held on student retention in institutions. Numerous theoretical and empirical research projects aim to investigate the factors that may contribute to student dropout rates. Current research has indicated financial worries as one of the primary issues resulting from expensive schooling or inadequate support for education. Students' financial situation may be impacted by low student retention at universities. A low graduation rate in higher education institutions is caused by the requirement that students repay their education loans even if they are unable to complete their studies. Low-paying jobs may arise from not living up to students' expectations as the need for highly qualified graduates grows. Moreover, low retention rates may be a sign that an institution is not fulfilling its purpose. Universities may create fewer professionals and practical skills, which can result in a significant loss of human capital if they do not adequately educate students' mindsets for the corporate world. Parents, teachers, and children alike all benefit greatly from a high retention rate. For instance, if a student is not participating in the assigned readings or coursework, an examination of their development might yield useful information that teachers can use to improve their lesson plans and instructional resources to increase student interest and learning. "At-risk" kids are those who demonstrate poor progress or who are most likely to drop out of school for many reasons. Early identification of these individuals can yield several advantages and beneficial resources for both educational institutions and learners, including an early warning system and designed to address problems that could cause students to postpone or stop their studies. Predicting the successful students can also assist teachers in identifying the qualities of an exceptional student, which can then be codified into a set of guidelines that new university students can be given. Predictive analytics, or more precisely Learning analytics, is the term for sophisticated, high-tech solutions focused on data and analysis that are being adopted by numerous organizations. Using vast volumes of historical data, analytics are used in education to identify trends and patterns that can be used to forecast future student success. The importance of applying learning analytics to improve student retention has been demonstrated by numerous empirical research. By analyzing data to forecast students' overall achievements as early as possible before their second semester, Georgia State University and an estimated 1,400 other colleges and universities are employing learning analytics to reroute students toward a better and more successful career. Students receive recommendations based on these forecasts, such as advice to enroll in tutoring sessions, pursue particular degrees, and so forth. While it is true that "student's degree completion" is generally correlated with student retention, this review will gather and examine empirical studies about student retention and attrition in three distinct classroom environments—online, traditional, and blended learning.

II. Related Work

Undergraduates continue to face ongoing issues with poor academic performance and high dropout rates despite consistently rising enrolment rates at U.S. postsecondary institutions. High attrition rates make it more difficult for academic institutions to prepare for enrolment and add to the

workload associated with recruiting new students. Students who leave college before receiving a terminal degree see it as a waste of their time and money, as well as an unrealized human potential. Poor academic performance increases the likelihood of dropping out of college and is frequently a sign of adjustment issues. The number of students who drop out of a university has historically been used to define student attrition at that particular institution. According to studies, more students drop out of college during their first year than the remainder of their time in school. Many research on student retention and attrition, like this one, have concentrated on first-year dropouts, or the proportion of students who do not return for their second year, since the majority of student dropouts happen at the conclusion of the first year, or the freshmen year. There is no distinction made in this definition of attrition between students who may have transferred to other colleges and earned their degrees there. Not by academic dismissal, but simply by the students leaving on their own volition at the end of their first year. Traditionally, studies on retention have been survey-based (e.g., assessing a cohort of students and tracking them for a predetermined amount of time to see if they continue their education). Researchers used this type of design to construct and validate theoretical models, such as Tinto's well-known student integration model. Others have expanded on Tinto's notion by utilizing survey-based research to create student attrition models. These survey-based research studies have been criticized for not being broadly applicable to other institutions and for being expensive and difficult to administer, while having set the groundwork for the field. An analytical technique using data frequently found in institutional databases can be employed as an alternate (and/or a supplementary) method to the usual survey-based retention study. Academic institutions regularly gather a lot of data about their students, including socioeconomic status, social media activity, educational history, and academic achievement. Survey-based and data-driven retention research are at best equivalent, according to a comparison, and data-driven research outperforms survey-based research in terms of developing a parsimonious logistic regression model. However, in practice, these two research approaches—one based on surveys, the other on theories and institutional data as well as analytical techniques—complement and support one another. In other words, while analytical studies may uncover novel associations among the variables that could inspire the creation of new theories or the improvement of preexisting ones, theoretical research may assist in identifying significant predictor variables to be used in such studies. Attrition is linked to several academic, economical, and other relevant issues. Universities with more open admission policies, no significant waiting list for candidates, and no transfer policy, according to Wetzell et al., have more severe student attrition issues than universities with an excess of applications. However, Hermanowicz discovered that more selective colleges do not always have higher graduation rates; rather, there is a chance that other variables that aren't directly related to selectivity could be at play. Higher rates of retention are frequently attained when students find that their university's environment is highly correlated with their interests. For this reason, in addition to the structural aspects of universities (such as admission and prestige of the school), the cultural aspects (such as norm and values that guide communities) should also receive equal attention. According to related research, Astin found that students' relationships with teachers, staff, and classmates have a significant impact on their persistence, or retention rate. Academic difficulty, adjustment issues, a lack of defined academic and vocational goals, ambiguity, a lack of commitment, a poor integration with the college community, incongruence, and isolation are among the variables Tinto lists as contributing to students dropping out of school. Consequently, improving studentstaff contact on campus can have a significant impact on retention. Making good connections with college staff during their first term of enrolment and successfully navigating the transition to college, supported by both initial and extended orientation and advisement programs, are crucial factors, especially for first-generation college students, in their decision to stay enrolled until their goals are achieved.

III. Our Solution

A. Description of Dataset

This dataset offers a thorough picture of the students enrolled in different undergraduate programs at a university. It contains information on social-economic characteristics, academic achievement, and demographics that can be utilized to examine potential determinants of academic success and student dropout. This dataset includes several separate databases with pertinent data that was accessible at the time of enrollment, including the kind of application, marital status, course selection, and more. Additionally, by evaluating the curricular units credited/enrolled/evaluated/approved together with their corresponding grades, this data can be used to estimate total student performance at the end of each semester. Lastly, the region's GDP, unemployment rate, and inflation rate can all provide additional insight into the role that economic considerations have in either academic success or student dropout rates. This potent analytical tool will offer insightful information about what drives students to pursue a variety of fields, including agronomy, design, education, and nursing, or to drop out of school. technology, social services, or journalism management.

To make the most of this dataset, scientists we became familiar with all of the variables it contains. These included numerical variables like the number of curricular units at the start of a semester or the age at enrollment; ordinal data measurement type variables like marital status; studied trends over time like GDP or inflation rate; frequency measurements variables like percentage of scholarship holders; etc. Furthermore, we ensured that of any potential bias present in the data. For example, if one population is underrepresented in comparison to another, as this phenomenon may produce unexpected results if ignored when conducting research with this data set. Lastly, it is crucial that we understand that the data in this Kaggle dataset only spans one semester for each admission intake; further research conducted over a longer period may be able to provide more accurate results related to a particular topic area due to further deterioration retention achievement coefficients obtained from those gradually accurate experiments unfolding various admissions seasons throughout the year.

In the initial phases of data preprocessing, the process commences with the collection of raw data, followed by a meticulous cleaning procedure aimed at detecting and addressing issues like missing values, outliers, and inconsistencies.

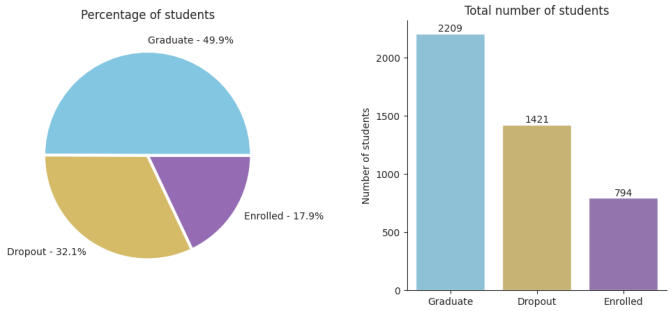

Fig. 1. Dataset Description

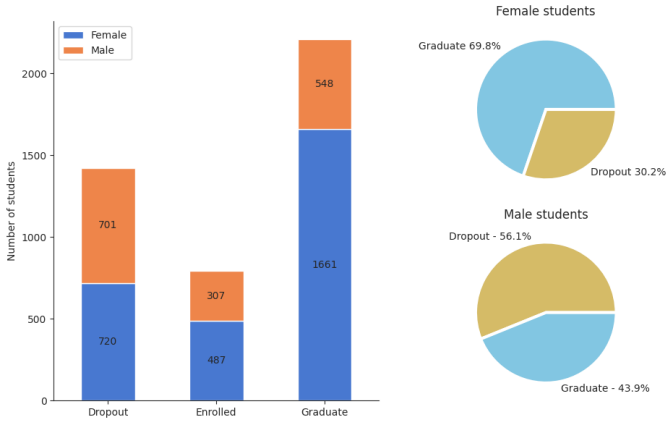

Fig. 2. Dropout Rates By Gender

Furthermore, an examination of the relationships among different features, including the target column, is conducted using techniques like correlation matrix analysis. Subsequently, a range of algorithms, including gradient boosting, random forest classifier, and decision tree, is employed on the preprocessed data. The objective is to implement the algorithm that exhibits the highest accuracy, thereby preparing the data optimally for subsequent analysis and modeling in the project.

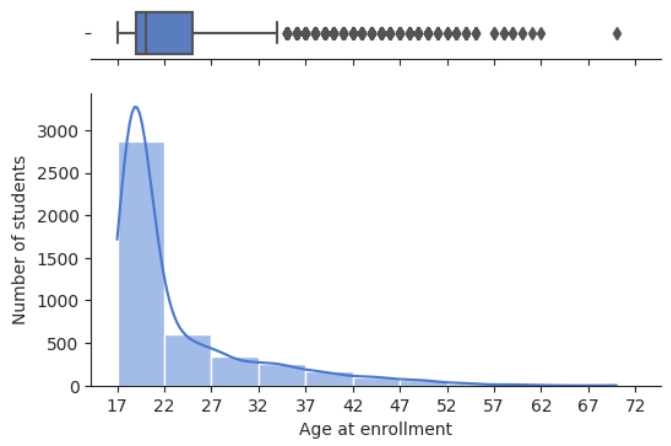

Fig. 3. Number of Students By Age

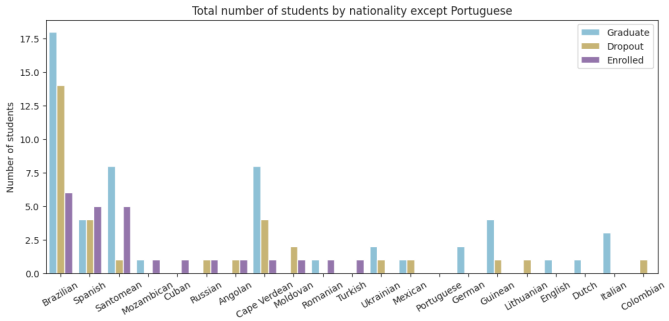

Fig. 4. Enrolment Status By Nationality

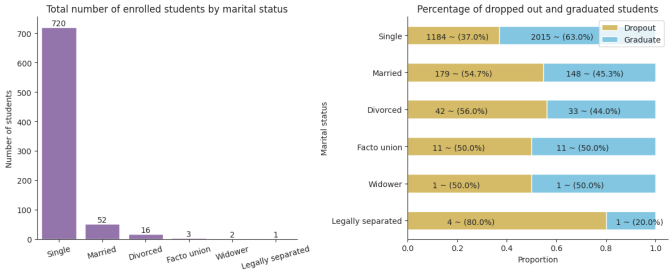

Fig. 5. Marital Status of Enrolled Students

B. *Machine Learning Algorithms*

Artificial Neural Networks (ANN) are analytical methods with biological inspiration that can represent incredibly complicated non-linear functions. We employed a backpropagation, supervised learning technique with the well-known multilayer perceptron (MLP) neural network architecture in this work. Arguably the most popular and extensively researched ANN architecture is MLP, a robust function approximator for prediction and classification tasks. Hornik et al. conducted an empirical study which shown that MLP can learn arbitrarily complex nonlinear functions to an arbitrary accuracy level, if it is given the appropriate size and structure. An illustration of the ANN architecture that was employed in this investigation. Decision trees are strong classification algorithms that are gaining popularity because of their qualities of intuitive explainability.

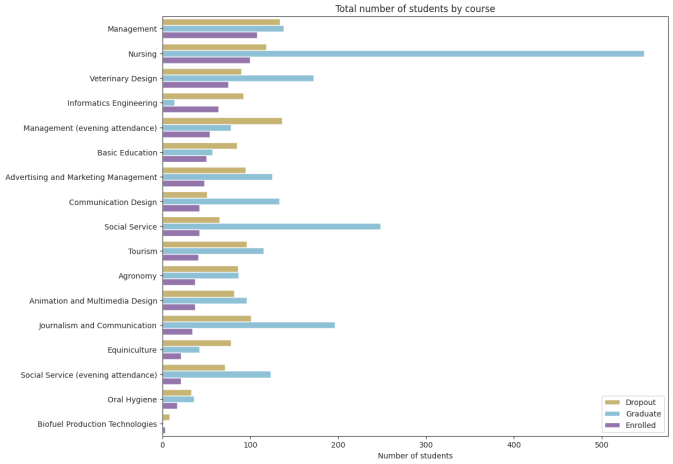

Fig. 6. Enrolment Status By Course

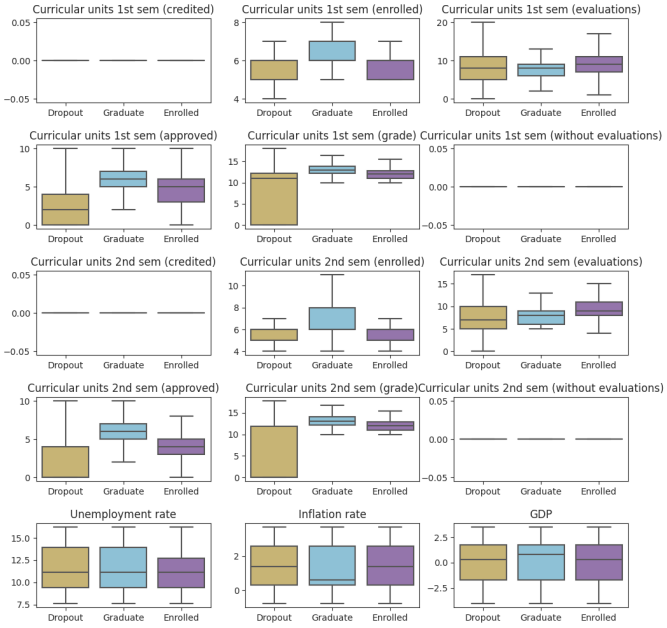

Fig. 7. Enrolment Status By Cirriculum

Breiman et al.'s CART (Classification and Regression Trees) and CHAID (CHi-squared Automatic Interaction Detector) are two well-known decision tree algorithms, as are Quinlan's ID3, C4.5, and C5. The C5 algorithm, an enhanced variant of the C4.5 and ID3 algorithms, was employed in this investigation. A more comprehensive version of linear regression is called logistic regression. Binary or multi-class dependent variable prediction is its main application. However, linear regression cannot directly represent the response variable since it is discrete. It thus develops the model to predict the probability of the event occurring rather than a precise estimate of the event itself. The usage of machine learning approaches for realworld prediction issues has expanded despite the widespread use of logistic regression as a statistical tool for classification problems due to its restrictive assumptions on normality and independence. The family of generalized linear models that includes Support Vector Machines (SVMs) makes a classification or regression decision based on the linear combination of features' value. In Support Vector Machines (SVMs), the mapping function can be either a regression function (used to estimate the numerical value of the desired output) or a classification function (used to categorize the data, as is the case in this study). In order to transform the input data—which by nature represents extremely complex nonlinear relationships—into a high-dimensional feature space for classification, nonlinear kernel functions are frequently used. In this feature space, the input data becomes more separable—that is, linearly separable—than it was in the original input space. The classes in the training data are then optimally separated by the construction of maximum-margin hyperplanes. By maximizing the distance between the two parallel hyperplanes, two parallel hyperplanes are built on either side of the hyperplane that divides the data. It is assumed that the classifier's generalization error will be better the greater the margin or separation between these parallel hyperplanes. Ensembles/Bagging (Random Forest): A random forest is a type of classifier that produces a class that is the mean of the classes produced by each individual decision tree. It is made up of several decision trees. A random forest is essentially a group (ensemble) of seemingly straightforward decision trees that can all generate a response when given a set of predictor values. It has been demonstrated that a random forest operates quite effectively on big datasets with plenty of variables. Breiman created the initial random forest induction algorithm. Ensembles/Boosted Trees: The basic idea behind boosted trees is to calculate a series of extremely basic trees, with each new tree being constructed using the prediction residuals of the one before it. To put it simply, it builds the next tree with lessons from the preceding one to reduce the number of cases that are incorrectly classified. Hastie et al. provide thorough technical descriptions of these techniques. Ensembles/Information Fusion: The process of "intelligently" merging data from two or more information sources (prediction models) is known as information fusion. In this example, the data is predictions. There is continuous discussion on the degree of sophistication of fusion approaches, but fusion (combining forecasts) yields more reliable and accurate prediction results.

C. Implementation Details

In the implementation of our machine learning project, we meticulously followed a structured approach. The journey commenced with the collection of raw data, a crucial step that laid the foundation for subsequent analyses. Following data collection, an extensive dataset-cleaning process was undertaken to ensure data integrity and reliability. Subsequently, we conducted an Exploratory Data Analysis (EDA) to gain valuable insights into the inherent patterns and characteristics of the dataset. Moving forward, the model-building phase involved the utilization of various algorithms, including Decision Tree, Logistic Regression, Support Vector Machines, Random Forest, and Gradient Boosting. Among these, Random Forest emerged as the most accurate, surpassing 91Performance metrics such as precision, recall, and F1 Score were used to evaluate and validate the robustness of the models. As a refinement step, parameters with negative correlations were identified, leading to the strategic decision to drop these columns for improved model accuracy. This structured implementation ensures a comprehensive and effective approach to understanding and predicting outcomes in our machine learning project.

IV. Comparison

In our project, we evaluated multiple machine learning models to ascertain their efficacy in predicting student dropout rates. The models encompassed Decision Tree, Logistic Regression, Support Vector Machines, Random Forest, and Gradient Boosting. Among these, Random Forest achieved an impressive accuracy exceeding 91The precision, recall, and F1 Score metrics further confirmed its robustness, boasting values of 0.889, 0.970, and 0.928, respectively. A notable aspect of our analysis revealed the presence of parameters with negative correlations in the dataset.

Random Forest (RF) is considered a powerful and versatile algorithm, and it often outperforms other algorithms in various scenarios. Here are some reasons why Random Forest might be preferred over other classifiers:

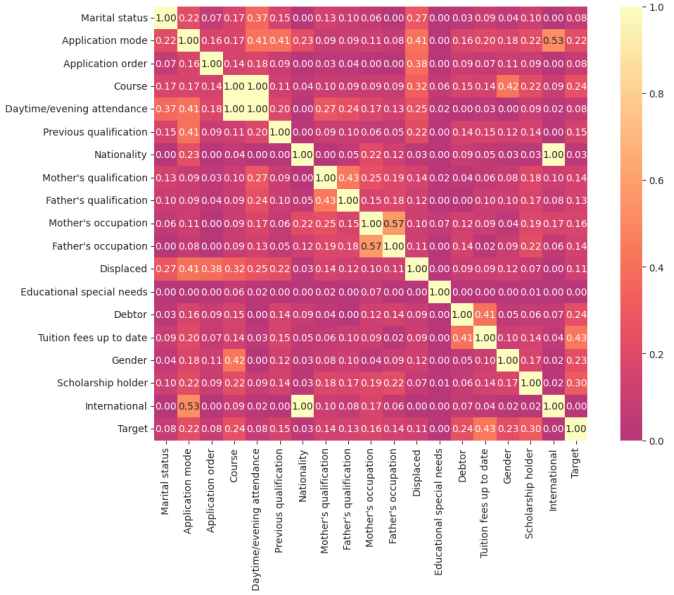

Fig. 8. Cramers V Association Plot

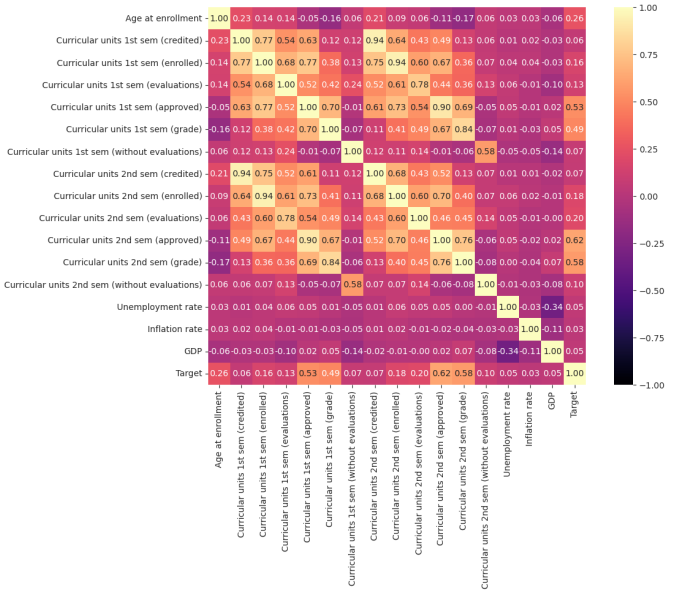

Fig. 9. Corelation and Association Plot

Ensemble Method: Random Forest is an ensemble learning method, which means it combines the predictions of multiple individual models (decision trees in this case). The ensemble approach helps to reduce overfitting and improve generalization, making it more robust.

Handling Non-Linearity: Decision Trees, including those in a Random Forest, can model complex, non-linear relationships in the data. This can be an advantage when dealing with data that doesn't follow a linear pattern, which might be a limitation of Logistic Regression.

Robust to Overfitting: Random Forest is less prone to overfitting compared to a single Decision Tree. By building multiple trees and averaging their predictions, RF reduces the

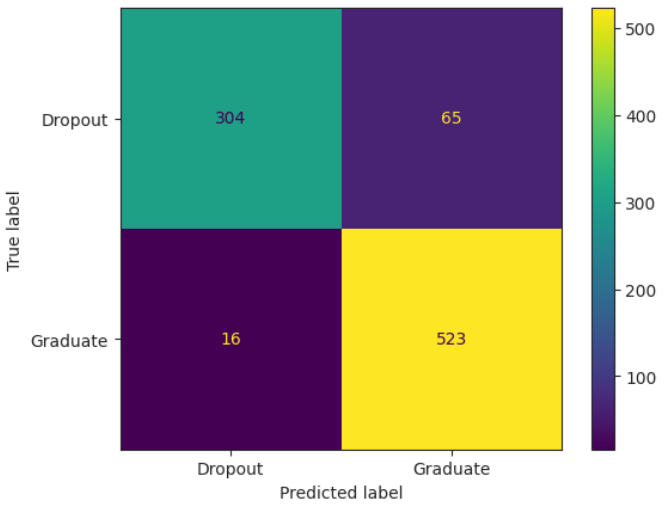

Fig. 10. Confusion Matrix

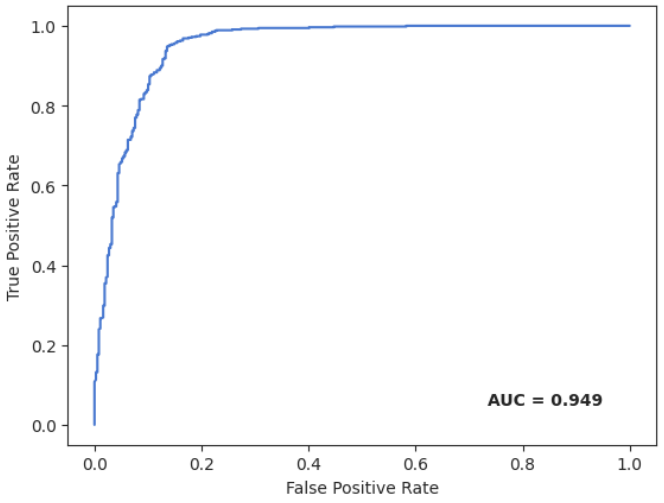

Fig. 11. Predicting Student Dropout Rates

risk of fitting the noise in the data.

Feature Importance: Random Forest provides a feature importance score, which indicates the contribution of each feature to the model's predictions. This can be valuable for feature selection and understanding the importance of different variables in the dataset.

Versatility: Random Forest can handle both classification and regression tasks. It can be used for a wide range of problems without extensive hyperparameter tuning.

The following tables illustrates the comparison of all the algorithms with the parameters like accuracy, precision, recall and F1 Score:

V. Future Directions

The future scope of our project holds immense potential to significantly improving the educational systems by delving deeper into the intricacies of student success and retention. Despite the valuable insights gained from our machine learning

	Precision	Recall	F1-Score	Accuracy
Random Forest	88%	96%	92%	91.10%
Gradient Boosting	90%	95%	92%	91%
Logistic Regression	89%	97%	93%	90%
Support Vector Machines	86%	98%	92%	89.50%
Decision Tree	90%	89%	89%	87%

Fig. 12. Comparison Between Different Models

analysis, the constraints posed by the quality and quantity of available data present substantial avenues for improvement. The limited dataset compelled strategic decisions, such as excluding crucial demographic features and omitting columns to prevent overfitting in our models. A more extensive and robust dataset could have allowed us to harness the full power of machine learning to uncover deeper patterns and relationships. Our aim was to make a substantial difference in understanding and addressing the complex challenges associated with student success. While numerous projects have tackled this topic, we sought to bring a fresh perspective and contribute meaningful insights. However, the project reveals a stark reality — there is still much to be uncovered. The primary drawback of our model and dataset lies in its limitation to fully account for the intricate web of human factors influencing educational outcomes. Life's unpredictability and the profound impact of human emotions introduce complexities that our current model cannot entirely capture. For instance, unforeseen medical challenges or struggles with cultural adjustment by international students may lead to dropout decisions that transcend purely academic considerations. The crux of our point lies in acknowledging the multifaceted nature of human experiences, encompassing relationships and emotional well-being — all factors influencing a student's decision to succeed or drop out. While recognizing the inherent limitations in quantifying such nuanced aspects, our dataset offers valuable insights that can guide recommendations for universities and educational institutions. By understanding the broader context and the human dimensions at play, our findings can contribute to the development of targeted strategies that go beyond traditional academic support, addressing the holistic well-being of students and fostering an environment conducive to their overall success and retention.

VI. Conclusion

We trained our model using Decision Tree, Logistic Regression, Support Vector Machines, Random and Gradient Boosting. Random Forest gives better accuracy as compared to any other algorithm with little over 91These are the values for following parameters: Precision: 0.889, Recall: 0.970, F1 Score: 0.928. Few parameters in the dataset had negative Correlation. When two variables have an inverse correlation, or negative correlation, one variable rises as the other falls, and vice versa. This relationship describes an observable pattern; it may or may not indicate that the two variables are causally related. So, we dropped those columns to get highly correlated data which will increase the accuracy of our model.

The application of machine learning algorithms in our study has provided valuable insights into the factors influencing student success and retention rates across diverse academic disciplines. The models were trained to uncover patterns and relationships contributing to student outcomes. Here are some key insights derived from our analysis:

1. Predictive Factors for Success: Our machine-learning models identified specific factors that contribute significantly to student success. These include, academic performance indicators. Understanding these predictors allows for targeted interventions to support students at risk of academic challenges.

2. Discipline-Specific Trends: The analysis revealed discipline-specific trends in student success and retention. Different academic disciplines exhibit unique patterns, indicating that interventions and support mechanisms should be tailored to the specific needs of students in each discipline.

3. Early Warning Signs: Machine learning algorithms enabled the identification of early warning signs for students at risk of dropping out. Early identification allows for targeted support to address issues before they escalate.

4. Comparative Analysis of Learning Environments: Our study specifically compared student retention and attrition in three distinct classroom environments: online, traditional, and blended learning.

5. Continuous Monitoring and Adaptation: The dynamic nature of machine learning models allows for continuous monitoring and adaptation. As new data becomes available, the models can be updated to reflect evolving trends and challenges, ensuring that interventions remain relevant and effective over time.

References

1] Burke, Adam. "Student retention models in higher education: A literature review." College and University 94.2 (2019): 12-21.

2] Radovan, Marko. "Should I stay, or should I go? Revisiting student retention models in distance education." Turkish Online Journal of Distance Education 20.3 (2019): 29-40.

3] Simpson, Ormond. "Student retention in distance education: are we failing our students?." Open learning: The Journal of Open, Distance and e-learning 28.2 (2013): 105-119.

4] Blue, Andrea. Exploring mentoring strategies needed by higher educational managers to increase student retention rates and decrease dropout rates in a higher education organization. Diss. Colorado Technical University, 2018.

5] Berge, Zane L., and Yi-Ping Huang. "A Model for Sustainable Student Retention: A Holistic Perspective on the Student Dropout Problem with Special Attention to e." Learning 13.5 (2004): 97-108.

6] McCroskey, James C., Steven Booth-Butterfield, and Steven K. Payne. "The impact of communication apprehension on college student retention and success." Communication Quarterly 37.2 (1989): 100-107.

7] Sydow, Debbie L., and Robert H. Sandel. "Making student retention an institutional priority." Community College Journal of Research and Practice 22.7 (1998): 635-643.

8] Delen, Dursun. "A comparative analysis of machine learning techniques for student retention management." Decision Support Systems 49.4 (2010): 498-506.

9] Cardona, Tatiana, et al. "Data mining and machine learning retention models in higher education." Journal of College Student Retention: Research, Theory Practice 25.1 (2023): 51-75.

10] Trivedi, Sandeep. "Improving students' retention using machine learning: Impacts and implications." ScienceOpen Preprints (2022).

11] Huo, Huade, et al. "Predicting dropout for nontraditional undergraduate students: A machine learning approach." Journal of College Student Retention: Research, Theory Practice 24.4 (2023): 1054-1077.

12] Kemper, Lorenz, Gerrit Vorhoff, and Berthold U. Wigger. "Predicting student dropout: A machine learning approach." European Journal of Higher Education 10.1 (2020): 28-47.

13] Utomo, Andy Prasetyo, Purwanto Purwanto, and Bayu Surarso. "Latest Algorithms in Machine and Deep Learning Methods to Predict Retention Rates and Dropout in Higher Education: A Literature Review." E3S Web of Conferences. Vol. 448. EDP Sciences, 2023.

14] Mduma, Neema, Khamisi Kalegele, and Dina Machuve. "A survey of machine learning approaches and techniques for student dropout prediction." (2019).

15] Shilbayeh, Samar, and Abdullah Abonamah. "Predicting student enrollments and attrition patterns in higher educational institutions using machine learning." Int. Arab J. Inf. Technol. 18.4 (2021): 562-567.

16] Bello, Felipe A., et al. "Using machine learning methods to identify significant variables for the prediction of first-year Informatics Engineering students dropout." 2020 39th International Conference of the Chilean Computer Science Society (SCCC). IEEE, 2020.

17] Lykourentzou, Ioanna, et al. "Dropout prediction in e-learning courses through the combination of machine learning techniques." Computers Education 53.3 (2009): 950-965.

18] Dake, Delali Kwasi, and Charles Buabeng-Andoh. "Using Machine Learning Techniques to Predict Learner Drop-out Rate in Higher Educational Institutions." Mobile Information Systems 2022 (2022).

19] Kiss, Viktor, Edgar Maldonado, and Mark Segall. "The Use of Semester Course Data for Machine Learning Prediction of College Dropout Rates." Journal of Higher Education Theory and Practice 22.4 (2022): 64-74.

20] Alkhasawneh, Ruba, and Rosalyn H. Hargraves. "Developing a hybrid model to predict student first year retention in STEM disciplines using machine learning techniques." Journal of STEM Education: Innovations and Research 15.3 (2014).

21] Yadav, Surjeet Kumar, Brijesh Bharadwaj, and Saurabh Pal. "Mining Education data to predict student's retention: a comparative study." arXiv preprint arXiv:1203.2987 (2012).

22] Shafiq, Dalia Abdulkareem, et al. "Student retention using educational data mining and predictive analytics: a systematic literature review." IEEE Access (2022).

23] Suhaimi, Nur Amalina Diyana, et al. "Classification of Learner Retention using Machine Learning Approaches." 2021 7th International Conference on Research and Innovation in Information Systems (ICRIIS). IEEE, 2021.

24] Opazo, Diego, et al. "Analysis of first-year university student dropout through machine learning models: A comparison between universities." Mathematics 9.20 (2021): 2599.

25] Solis, Martin, et al. "Perspectives to predict dropout in university students with machine learning." 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI). IEEE, 2018.