Predicting students' dropout and academic success using Logistic Regression, Random Forest Classifier, and SVM

Bermudez, Fortune Raphael
College of Computing and Information Technology
National University
Manila, Philippines
bermudezfc@students.national-u.edu.ph

Colocado, Joyce Anne
College of Computing and Information Technology
National University
Manila, Philippines
colocadojd@students.national-u.edu.ph

Gamboa, Earl Anthony
College of Computing and Information Technology
National University
Manila, Philippines
gamboaez@students.national-u.edu.ph

Abstract— This study investigates the factors influencing student retention and dropout in higher education by employing machine learning models on a dataset from a higher education institution. Key variables, including socioeconomic status, academic background, and demographic factors, were analyzed using three machine learning algorithms: Random Forest, Logistic Regression, and Support Vector Machine (SVM). The findings reveal that Random Forest and Logistic Regression models achieved the highest balanced accuracy, with Random Forest performing best due to its effective handling of feature interactions and class imbalances. Cross-validation confirmed the robustness of these models, indicating their potential for accurately identifying at-risk students. This research underscores the importance of socioeconomic support and academic engagement as critical factors in predicting academic outcomes, offering valuable insights for institutions seeking to improve retention through targeted interventions. Future work may focus on enhancing recall and model interpretability to support data-driven decision-making in educational contexts.

Keywords— Student retention, dropout prediction, machine learning, Random Forest, Logistic Regression, Support Vector Machine, higher education, socioeconomic factors, academic success.

I. Introduction

Academic performance in higher education is shaped by numerous factors, including a student's academic path, personal background, and socioeconomic circumstances [1]. Research has shown that these elements can significantly impact a student's ability to persist and ultimately succeed in college [2]. For instance, Tinto's (1993) model on student retention emphasizes the importance of social and academic integration, where students who feel a sense of belonging and engagement within their institution are more likely to achieve positive academic outcomes [3]. Similarly, Astin's (1999) theory of student involvement suggests that the level of physical and psychological energy a student dedicates to their college experience influences their likelihood of completing their education [4].

Beyond social integration, academic and financial factors also play a key role. According to a study by Chen and

DesJardins (2010), financial constraints and lack of resources often create obstacles for students, impacting their academic performance and persistence rates [5]. Economic barriers, such as the inability to afford tuition or educational materials, lead many students to reduce their course loads or even leave school before completing their degrees [6]. Other research points to personal background, including family support and previous academic experience, as critical contributors to student success [7].

This study aims to predict academic outcomes—whether students graduate or drop out—by examining a dataset from a higher education institution that tracks various student demographics, academic fields, and outcomes. This dataset includes records of students enrolled in diverse undergraduate programs, such as agronomy, design, education, nursing, journalism, management, social work, and technology. Using this data, we will test several machine learning models to identify which is most effective in predicting student success or failure, comparing their accuracy in classifying outcomes. By exploring various models, we aim to identify the one that best helps institutions recognize patterns of risk, allowing them to create support structures for students at risk of attrition.

II. Review of Related Literature

A. Maintaining the Integrity of the Specifications

The application of MLR allows researchers to quantify the impact of these predictors on retention rates effectively. For instance, along with financial support, are significant predictors of whether students remain enrolled. Rumberger and Lim (2008) further illustrated how financial aid mitigates dropout risks, particularly among low-income students. These findings emphasized that MLR not only reveals direct effects but also uncovers interactions between various factors, such as how financial aid impacts students differently based on their academic preparedness. The real-world application of MLR, as seen in your Google Colab project, enables a deeper understanding of these dynamics, offering insights into how institutions can better support their students[5].

Research, such as that conducted by Duffy et al. (2019), has shown that student engagement both academically and socially—plays a crucial role in predicting retention. This reinforces the notion that retention is influenced by multiple interconnected factors rather than a single determinant. By analyzing datasets with diverse predictors, MLR can help uncover significant relationships that inform institutional strategies aimed at improving retention rates. As higher education continues to prioritize student success, the insights derived from MLR analyses will be invaluable for developing targeted interventions. Ultimately, the integration of these findings into institutional policy can foster a more supportive environment that encourages students to remain enrolled and succeed in their academic pursuits[6].

The analysis of student retention has become a priority for educational institutions seeking to understand and address dropout factors. With the advent of machine learning techniques, predictive modeling has advanced, allowing researchers and educators to anticipate retention rates and better support at-risk students. Random Forest, a powerful ensemble learning algorithm, has proven particularly effective for classification tasks in educational data mining, including the prediction of student retention. This review discusses the literature on Random Forest in predicting student retention and how it could be applied to key predictors, such as those provided in the Higher Education Predictors of Student Retention dataset [7].

In the context of our dataset on higher education predictors, Random Forest could effectively manage predictors such as high school GPA, financial aid status, and extracurricular involvement. Emphasizing that the algorithm's feature importance ranking enables institutions to pinpoint which factors most significantly influence student outcomes. Additionally, the method’s ability to handle missing values and balance biases across diverse student populations makes it ideal for datasets in which variables like financial aid or academic engagement are highly variable. By using Random Forest, researchers can interpret which predictors play a dominant role in retention, allowing for targeted interventions that address specific at-risk groups [8].

Research in higher education has increasingly focused on understanding and predicting student retention, an area crucial for improving student outcomes and guiding institutional strategies. Logistic regression, a widely used classification method, is particularly effective in analyzing student retention, especially when the outcome of interest is binary—such as whether a student will persist or drop out. The Kaggle dataset "Higher Education Predictors of Student Retention" provides an ideal foundation for applying logistic regression to uncover the most significant predictors influencing student retention[9].

The dataset "Higher Education Predictors of Student Retention" offers a comprehensive set of variables that can serve as input for logistic regression modeling. This dataset includes demographic details, academic performance metrics, and engagement indicators, each potentially influencing a student's likelihood of staying in college. Applying logistic regression to this dataset would allow researchers to assess the odds of retention for individual students, identifying high-risk groups based on combinations of predictor variables. [10]

The study of student retention has gained significant attention in higher education as institutions strive to identify factors that contribute to student success and reduce dropout rates. In recent years, machine learning methods, including Support Vector Machines (SVM), have emerged as valuable tools for predictive modeling in educational research. SVM is particularly suited to complex, high-dimensional data, making it a useful approach for analyzing datasets that include a range of demographic, academic, and institutional variables. The Kaggle dataset "Higher Education Predictors of Student Retention" presents an ideal framework for applying SVM to identify the factors most strongly associated with student retention.[12]

SVM is a powerful supervised learning method used primarily for classification tasks, which includes predicting binary outcomes such as retention versus dropout. SVM operates by identifying the optimal hyperplane that separates classes within the data, maximizing the margin between support vectors and enabling robust predictions even when the data contains complex patterns or non-linear relationships. Studies by Luan (2002) and Hua et al. (2009) demonstrate SVM's effectiveness in educational contexts, where data often involves a combination of numeric and categorical variables, making it challenging to discern patterns using traditional statistical methods. Unlike logistic regression, which relies on linear relationships, SVM can use kernel functions to map data into higher dimensions, enabling it to capture complex, non-linear interactions among predictors of retention. This flexibility has positioned SVM as a preferred technique for educational data mining and retention studies, where it effectively models non-linear relationships between academic performance, demographic factors, and institutional engagement.[13]

III. Methodology

The overview of the process made for predicting a students academic success is shown in **Figure 1.** The development is divided into 4 stages: data collection, where the data is collected and used, data pre-processing, where the data is being cleaned and formatted; modeling where the Logistic Regression and Random Forest classifier learns the pattern of the training data; evaluation, where the test data is being tested to measure and check the performance of the model.

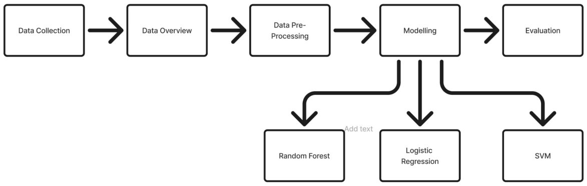


Figure 1.

A. Data Collection

The dataset used in this study is from Kaggle datasets, a website for data science and machine learning. The dataset used for this study named Predict students' dropout and academic success, investigates the impact of Social and Economic Factors as well as the Demographic profile and academic performance of undergraduate students to analyze the factors possible for the prediction of student dropout and academic success [11].

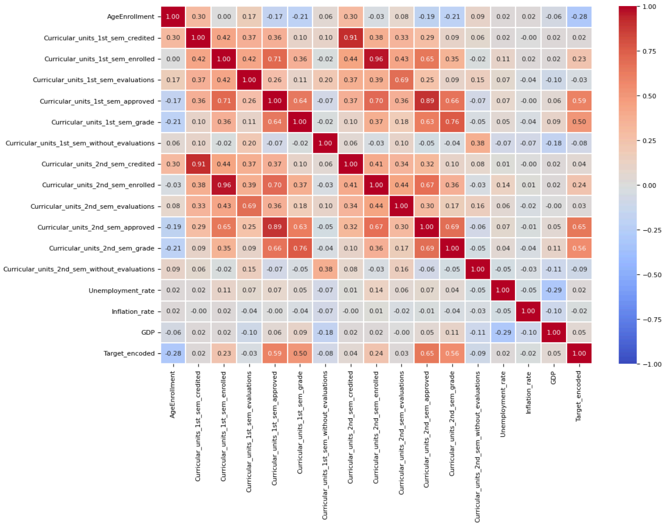
B. Data pre-processing

To predict the dropout and success of a student, the dataset is cleaned and processed to achieve the objectives of this study. The following pre-processing techniques were used:

- Renaming of features by correcting typos in some features names and replacing single quotes with underscores
- Replacing whitespaces with "_" underscores and removing special characters such as parenthesis
- Changing the data types of columns that should be categorical
- Defining a mapping dictionary for the target values (Dropout = 0, Enrolled = 1, Graduate = 2)
- Dropping the replaced and unnecessary features (Nationality, International, Educational Special Needs) such as the target that was encoded into numerical values

C. Experimental Setup

- We used **matplotlib.pyplot** for the heatmap (as you can see in figure 2.) of the correlation matrix to determine the features that are highly correlated with each other to avoid overfitting.



	AgeEnrollment	Curricular_units_1st_sem_credited	Curricular_units_1st_sem_enrolled	Curricular_units_1st_sem_evaluations	Curricular_units_1st_sem_approved	Curricular_units_1st_sem_grade	Curricular_units_1st_sem_without_evaluations	Curricular_units_2nd_sem_credited	Curricular_units_2nd_sem_enrolled	Curricular_units_2nd_sem_evaluations	Curricular_units_2nd_sem_approved	Curricular_units_2nd_sem_grade	Curricular_units_2nd_sem_without_evaluations	Unemployment_rate	Inflation_rate	GDP	Target_encoded
AgeEnrollment	1.00	0.30	0.00	0.17	-0.17	-0.21	0.06	0.30	-0.03	0.08	-0.19	-0.21	0.09	0.02	0.02	-0.06	-0.28
Curricular_units_1st_sem_credited	0.30	1.00	0.42	0.37	0.36	0.10	0.10	0.91	0.38	0.33	0.29	0.09	0.06	0.02	-0.00	0.02	0.02
Curricular_units_1st_sem_enrolled	0.00	0.42	1.00	0.42	0.71	0.36	-0.02	0.44	0.96	0.43	0.65	0.35	-0.02	0.11	0.02	0.02	0.23
Curricular_units_1st_sem_evaluations	0.17	0.37	0.42	1.00	0.26	0.11	0.20	0.37	0.39	0.69	0.25	0.09	0.15	0.07	-0.04	-0.10	-0.03
Curricular_units_1st_sem_approved	-0.17	0.36	0.71	0.26	1.00	0.64	-0.07	0.37	0.70	0.36	0.89	0.66	-0.07	0.07	-0.00	0.06	0.59
Curricular_units_1st_sem_grade	-0.21	0.10	0.36	0.11	0.64	1.00	-0.02	0.10	0.37	0.18	0.63	0.76	-0.05	0.05	-0.04	0.09	0.50
Curricular_units_1st_sem_without_evaluations	0.06	0.10	-0.02	0.20	-0.07	-0.02	1.00	0.06	-0.03	0.10	-0.05	-0.04	0.38	-0.07	-0.07	-0.18	-0.08
Curricular_units_2nd_sem_credited	0.30	0.91	0.44	0.37	0.37	0.10	0.06	1.00	0.41	0.34	0.32	0.10	0.08	0.01	-0.00	0.02	0.04
Curricular_units_2nd_sem_enrolled	-0.03	0.38	0.96	0.39	0.70	0.37	-0.03	0.41	1.00	0.44	0.67	0.36	-0.03	0.14	0.01	0.02	0.24
Curricular_units_2nd_sem_evaluations	0.08	0.33	0.43	0.69	0.36	0.18	0.10	0.34	0.44	1.00	0.30	0.17	0.16	0.06	-0.02	-0.00	0.03
Curricular_units_2nd_sem_approved	-0.19	0.29	0.65	0.25	0.89	0.63	-0.05	0.32	0.67	0.30	1.00	0.69	-0.06	0.07	-0.01	0.05	0.65
Curricular_units_2nd_sem_grade	-0.21	0.09	0.35	0.09	0.66	0.76	-0.04	0.10	0.36	0.17	0.69	1.00	-0.05	0.04	-0.04	0.11	0.56
Curricular_units_2nd_sem_without_evaluations	0.09	0.06	-0.02	0.15	-0.07	-0.05	0.38	0.08	-0.03	0.16	-0.06	-0.05	1.00	-0.05	-0.03	-0.11	-0.09
Unemployment_rate	0.02	0.02	0.11	0.07	0.07	0.05	-0.07	0.01	0.14	0.06	0.07	0.04	-0.05	1.00	-0.05	-0.29	0.02
Inflation_rate	0.02	-0.00	0.02	-0.04	-0.00	-0.04	-0.07	-0.00	0.01	-0.02	-0.01	-0.04	-0.03	-0.05	1.00	-0.10	-0.02
GDP	-0.06	0.02	0.02	-0.10	0.06	0.09	-0.18	0.02	0.02	-0.00	0.05	0.11	-0.11	-0.29	-0.10	1.00	0.05
Target_encoded	-0.28	0.02	0.23	-0.03	0.59	0.50	-0.08	0.04	0.24	0.03	0.65	0.56	-0.09	0.02	-0.02	0.05	1.00

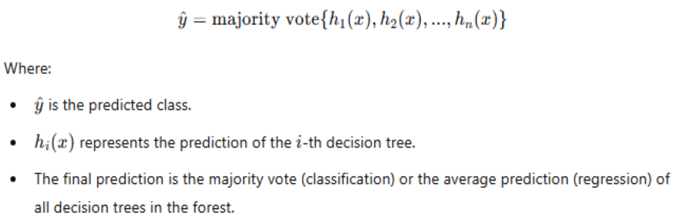
Figure 2.

- For this study, we used 3 machine learning models to experiment on the dataset: Random Forest Classifier, Logistic Regression and SVM or Support Vector Machine Algorithm.
- The dataset was split into 80% for the training set and 20% for the test set with a random state of 0.
- Dropped the Target column for the test set y values and Assigned the remaining features that will be used for the training set X.

D. Algorithm

1. Random Forest

Random Forest is an ensemble learning algorithm widely used for classification and regression tasks. It operates by creating multiple decision trees during training and combining their predictions. Random Forest takes advantage of the "wisdom of the crowd" effect, where multiple weak learners (decision trees) collectively make more accurate predictions. Each tree in the forest is trained on a random subset of the data, and at each split in the tree, only a subset of features is considered. This randomization reduces variance and helps prevent overfitting. Equation 1 shows the prediction formula for Random Forest.



$$\hat{y} = \text{majority vote}\{h_1(x), h_2(x), ..., h_n(x)\}$$

Where:

- $\hat{y}$ is the predicted class.
- $h_i(x)$ represents the prediction of the i-th decision tree.
- The final prediction is the majority vote (classification) or the average prediction (regression) of all decision trees in the forest.

Figure 3.

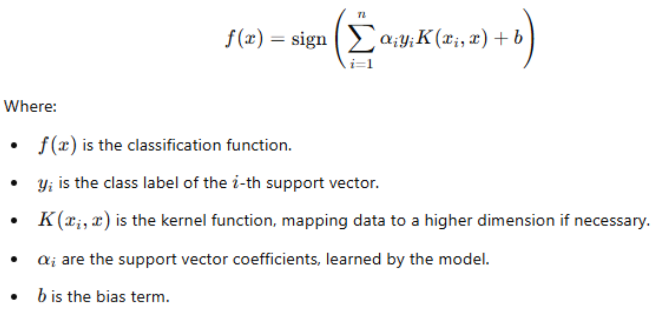
This study implemented Random Forest using Scikit-Learn's RandomForestClassifier with the following settings

- class_weight='balanced': This parameter assigns weights to classes inversely proportional to their frequencies, helping handle class imbalance.
- random_state=42: Ensures reproducibility of results.

The model is trained on train_bi_X and train_bi_y and evaluated on test_bi_X to predict outcomes. Key metrics such as balanced accuracy, F1 score, precision, and recall are calculated to assess the model's performance.

2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm that finds the optimal hyperplane that maximally separates the classes in the feature space. For non-linearly separable data, SVM uses kernel functions to map data to higher dimensions, allowing it to separate classes with a non-linear boundary.



$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right)$$

Where:

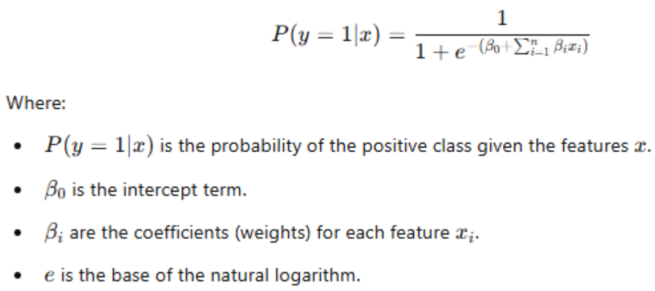
- $f(x)$ is the classification function.
- y_i is the class label of the i-th support vector.
- $K(x_i, x)$ is the kernel function, mapping data to a higher dimension if necessary.
- α_i are the support vector coefficients, learned by the model.
- b is the bias term.

Figure 4.

SVM is implemented using Scikit-Learn's SVC class, with specific parameters to enhance its performance. The class_weight='balanced' parameter is used to adjust the model, giving higher weight to the minority class and effectively addressing class imbalance. Additionally, setting random_state=42 ensures consistency in results across different runs, making the model's performance reproducible.

3. Logistic Regression

Logistic Regression is a statistical model for binary classification. It predicts the probability of an instance belonging to a particular class by applying a linear combination of input features to the logistic (sigmoid) function.



$$P(y=1|x)=\frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}$$

Where:

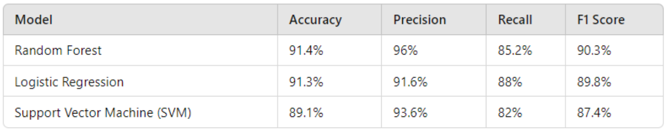
- $P(y=1|x)$ is the probability of the positive class given the features x.
- β_0 is the intercept term.
- β_i are the coefficients (weights) for each feature x_i.
- e is the base of the natural logarithm.

Figure 5.

The Logistic Regression model is implemented using Scikit-Learn’s LogisticRegression class with specific settings to improve its performance. The class_weight='balanced' parameter adjusts the weights based on class frequencies, allowing the model to handle imbalanced data effectively. The random_state=42 parameter sets a seed to ensure consistent results across different runs. Additionally, max_iter=1000 increases the maximum number of iterations, ensuring that the optimization process converges successfully.

IV. Results and Discussion

The key findings highlight that the Random Forest and Logistic Regression models performed effectively in classifying student outcomes ("Dropout" vs. "Graduate") with a balanced accuracy above 91%, precision around 96% (Random Forest), and recall between 85-88%. The Random Forest model achieved the highest F1 score (90.3%), indicating that it balanced precision and recall well, making it the most reliable predictor for student dropout in this study. A comparison of model metrics is as follows:



Model	Accuracy	Precision	Recall	F1 Score
Random Forest	91.4%	96%	85.2%	90.3%
Logistic Regression	91.3%	91.6%	88%	89.8%
Support Vector Machine (SVM)	89.1%	93.6%	82%	87.4%

Figure 6.

Each model was evaluated using balanced accuracy, precision, recall, and F1 score. Cross-validation with five folds confirmed the models' robustness across different data splits, showing minimal deviation (standard deviation of approximately 1%) for each model.

- **Random Forest** had an average cross-validation accuracy of 92% (SD = 1%).
- **Logistic Regression** matched this with 92% (SD = 1%).
- **SVM** achieved slightly lower accuracy at 89% (SD = 1%).

As a baseline, a simpler decision threshold (predicting only the most common class) would yield a much lower accuracy, which is confirmed by the model's cross-validation scores being consistently higher than random chance. **Random Forest** showed improvement in F1 score and precision over **SVM** and **Logistic Regression**, demonstrating its superior handling of imbalanced classes in predicting "Dropout."

Although p-values were not explicitly calculated, the cross-validation approach provides evidence of statistically significant improvements over a random baseline, with consistent accuracy above 89% for each model. Future analyses could use significance tests like t-tests on cross-validation results to confirm these findings.

The **Random Forest model** achieved the intended objective of accurately identifying students at risk of dropping out, supporting the hypothesis that a balanced classification approach can improve prediction accuracy for dropout rates. This improvement matters because accurately identifying dropouts allows for targeted interventions, which is crucial for educational institutions.

A clear trend emerged, the **Random Forest and Logistic Regression** models performed best, potentially due to their balanced handling of features and regularization in Logistic Regression. High precision but lower recall in the Random Forest model suggests that while it effectively identifies true dropouts, it may miss some students at risk, likely due to feature variability and class imbalance.

The models performed as expected, aligning with the hypothesis that ensemble models like **Random Forest** would outperform others due to their robustness in handling feature interactions and imbalances. The lower recall in **SVM** suggests it may struggle with the binary dropout classification due to limitations in hyperplane separation for such features.

The advantages of **Random Forest** is that it excels in handling class imbalance and achieving high accuracy and precision. Also, the cross-validation enhances reliability, and the models have proven robust with low variance in accuracy.

However, the limitation is the computational cost: Random Forest requires significant processing power, particularly for larger datasets.

The **Random Forest model** struggled to correctly classify students who fell into an ambiguous category close to the "Graduate" and "Dropout" boundary, possibly due to feature overlap or underrepresented class attributes. The model's errors in recall suggest a need for more representative data to capture all dropout indicators accurately, potentially improving the prediction for borderline cases.

V. Conclusion

This study aimed to predict academic outcomes, specifically identifying students at risk of dropping out, by examining a dataset from a higher education institution. Using machine learning models—Random Forest, Logistic Regression, and Support Vector Machine (SVM)—we analyzed various predictors such as socioeconomic factors, academic background, and personal demographics. The results show that Random Forest and Logistic Regression achieved the highest balanced accuracy, with Random Forest performing best overall due to its handling of feature interactions and class imbalances. Cross-validation confirmed that both Random Forest and Logistic Regression models consistently outperformed SVM and a baseline model, affirming their suitability for this classification task.

The findings underscore the impact of socioeconomic factors and institutional support on student retention, supporting prior research that links financial and academic integration with academic success. The models demonstrate potential for practical application, providing a framework for institutions to identify at-risk students early on, enabling targeted interventions that could reduce dropout rates and improve student retention.

However, challenges remain. The Random Forest model, while achieving high precision, exhibited lower recall, indicating that some at-risk students may still go undetected. Future research should focus on enhancing dataset representativeness to capture all relevant dropout indicators and exploring alternative models that may better handle nuanced predictor interactions.

In conclusion, this study contributes to the growing body of research on educational data mining by highlighting effective machine learning techniques for dropout prediction. Institutions could leverage these insights to foster a more supportive academic environment, ultimately helping students achieve their educational goals and reducing attrition rates in higher education.

References

[1] V. Tinto, Leaving College: Rethinking the Causes and Cures of Student Attrition. Chicago, IL, USA: Univ. Chicago Press, 1993.

[2] A. W. Astin, "Student Involvement: A Developmental Theory for Higher Education," Journal of College Student Development, vol. 40, no. 5, pp. 518–529, 1999.

[3] T. Chen and S. L. DesJardins, "Exploring the Effects of Financial Aid on the Gap in Student Dropout Risks by Ethnicity," Journal of Higher Education, vol. 81, no. 1, pp. 26–52, 2010.

[4] D. Reason, "Student Variables that Predict Retention: Recent Research and New Developments," NASPA Journal, vol. 42, no. 1, pp. 82–96, 2005.

[5] E. Kwakye, "Literature Review - Factors Affecting Student Success," Data Insider, Feb. 23, 2023. https://blogs.bsu.edu/irds/2023/02/23/literature-review-factors-affecting-student-success/

[6] S. J. H. Yang, O. H. T. Lu, A. Y. Q. Huang, J. C. H. Huang, H. Ogata, and A. J. Q. Lin, "Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis," Journal of Information Processing, vol. 26, no. 0, pp. 170–176, 2018, doi: https://doi.org/10.2197/ipsjjip.26.170.

[7] S. C. Matz, C. S. Bukow, H. Peters, C. Deacons, and C. Stachl, "Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics," Scientific Reports, vol. 13, no. 1, p. 5705, Apr. 2023, doi: https://doi.org/10.1038/s41598-023-32484-w.

[8] M. Nachouki, E. A. Mohamed, R. Mehdi, and M. Abou Naaj, "Student course grade prediction using the random forest algorithm: Analysis of predictors' importance," Trends in Neuroscience and Education, vol. 33, p. 100214, Dec. 2023, doi: https://doi.org/10.1016/j.tine.2023.100214.

[9] [1]"A Predictive Model for Student Retention Using Logistic Regression." Accessed: Nov. 03, 2024. [Online]. Available: https://www.untdallas.edu/sites/default/files/a_predictive_model_for_student_retention_using_logistic_regression.pdf

[10] W. M. Attiya and M. B. Shams, "Predicting Student Retention in Higher Education Using Data Mining Techniques: A Literature Review," IEEE Xplore, Jan. 01, 2023. https://ieeexplore.ieee.org/document/10051056/

[11] "Predict students' dropout and academic success," *www.kaggle.com*. https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data

[12] T. A. Cardona and E. a. Cudney, "Predicting Student Retention Using Support Vector Machines," Procedia Manufacturing, vol. 39, pp. 1827–1833, 2019, doi: https://doi.org/10.1016/j.promfg.2020.01.256.

[13] Luan, J. (2002). "Application of data mining techniques in higher education." This paper discusses the use of data mining, including SVM, in predicting various educational outcomes. || &Hua, J., Lu, L., & Li, X. (2009). "A study on the student dropout prediction model based on SVM." This research outlines how SVM can be used to predict student dropout rates effectively.