Introduction

Understanding and addressing student retention is vital due to its significant impact on individuals, institutions, and society. High attrition rates in postsecondary education impose considerable financial burdens on students, with those who drop out often incurring substantial debt without gaining the long-term benefits of a degree. According to **McFarland et al. (2019)** and the **U.S. Department of Education (2015)**, dropouts earn significantly less than graduates and are far more likely to default on loans, deepening long-term economic disparities. The consequences extend beyond finances, as dropping out also adversely affects students' mental health and sense of accomplishment, as noted by **Freudenberg and Ruglis (2007)**. These individual challenges translate into broader societal costs, including lost productivity and economic potential.

For institutions, student attrition leads to billions of dollars in lost tuition revenue annually. Research by **Raisman (2013)** and **Wellman et al. (2012)** estimates that early student departures account for an annual loss of $16.5 billion in the U.S. alone. These financial inefficiencies are particularly concerning for underfunded institutions, such as community colleges, which disproportionately serve minority and low-income students. The problem is exacerbated by declining institutional resources, as described by **Bound et al. (2010)** and **Bowen et al. (2009)**, making it increasingly challenging for colleges to provide adequate support systems that encourage retention. Additionally, research by **Schneider (2010)** and **Wellman et al. (2012)** underscores the importance of investing in early interventions to retain students, particularly during their first year when dropout rates are highest.

Student attrition has significant societal implications, with low retention rates impacting education and workforce outcomes. Groups such as racial minorities and women often experience lower retention, contributing to disparities in opportunities and outcomes. Improving retention rates enhances graduation numbers and ensures better returns on educational investments. Research, such as that by **Matz et al. (2023)**, demonstrates how data-driven methods can identify at-risk students and guide effective solutions. This project focuses on understanding the reasons behind low retention rates and identifying practical strategies to improve student success.

Related Literature and Methods

Role of Machine Learning in Education

The integration of machine learning (ML) in education is revolutionizing how data is used to enhance learning experiences and decision-making. Fischer et al. (2020) emphasize that ML enables the analysis of large educational datasets to identify patterns, predict outcomes, and tailor interventions for diverse student needs. These capabilities are particularly valuable for identifying at-risk students and improving learning outcomes. However, challenges such as data quality, algorithmic bias, and the need for transparent methodologies must be addressed to ensure equitable applications.

Hilbert et al. (2021) and the National Academy of Education (2017) further highlight ML's ability to personalize learning, provide real-time feedback, and support resource optimization. ML helps model complex interactions between student behaviors, demographics, and academic performance, offering actionable insights for educators. While the benefits of ML are substantial, the National Academy stresses the importance of safeguarding student privacy and establishing governance frameworks to mitigate ethical risks. These studies collectively showcase ML's transformative potential in education, while emphasizing the need for responsible implementation.

Technical approaches in predicting student retention and our choice of models

Predicting student retention and identifying at-risk students has been a key focus of educational research, leveraging machine learning (ML) techniques to analyze diverse datasets. Ortiz-Lozano et al. use decision trees (CART and QUEST algorithms) to evaluate socio-demographic and academic predictors at different stages of a student’s academic journey. Their models, based on incremental data (e.g., admission grades, mid-term results, end-of-semester performance), achieve up to 76% accuracy for end-of-semester predictions, showcasing the importance of dynamically updated data for identifying students at risk of dropout.

Dawson et al. (2017) similarly highlight the value of predictive analytics, integrating academic and engagement-related features from learning management systems (LMS). Employing decision trees and ensemble methods, their approach analyzed login frequency, digital interaction patterns, and academic performance to provide early identification of at-risk students. Crucially, the study emphasized transitioning predictions into actionable interventions, aligning predictive insights with strategies to improve retention rates.

The "Mining University Registrar Records" study also utilized ML to predict first-year undergraduate attrition, combining demographic, geographic, and academic features, such as gender, race, parental education, and ZIP code-derived socioeconomic indicators. This study tested multiple ML models, including logistic regression, random forests, and gradient boosted trees, with logistic regression emerging as the most efficient and accurate. By evaluating subsets of data, the study highlighted the importance of demographic and academic features in predicting retention.

These studies inform our project by demonstrating the power of integrating diverse data sources, including socio-demographic, academic, and engagement metrics, to predict dropout risk. Like these approaches, our project employs ML models such as logistic regression and random forests, alongside feature engineering, to analyze comprehensive datasets. The parallels in model selection, dynamic data integration, and actionable insights underscore the potential of ML-driven strategies to guide targeted interventions and improve student outcomes.

Del Bonifro et al. (2020) employ machine learning techniques to predict first-year undergraduate dropout rates at the application stage, focusing on features available prior to enrollment. Using a dataset of 15,000 students, they integrate demographic attributes (e.g., gender, age), academic metrics (e.g., high school final marks), and additional learning requirements (ALR) to build predictive models such as Linear Discriminant Analysis (LDA), Random Forests (RF), and Support Vector Machines (SVM). The study emphasizes early-stage predictions, demonstrating the utility of application-time data for identifying at-risk students and showing that ALR significantly enhances prediction accuracy. Challenges such as dataset imbalance are also addressed, with strategies to mitigate classification biases.

This study aligns with our project's focus on leveraging diverse data sources and machine learning models to predict student outcomes. Our feature engineering strategies, including interaction variables like curricular evaluations and socioeconomic predictors, parallel Del Bonifro et al.'s incremental feature evaluations. Additionally, the use of multiple models, such as logistic regression, random forests, and gradient-boosted trees, underscores shared goals of addressing class imbalance and optimizing early prediction to guide proactive interventions.

Dekker et al. (2009) highlighted the effectiveness of decision trees in predicting dropout rates among Electrical Engineering students, achieving accuracies of 75–80%. By utilizing domain-specific attributes like academic performance and enrollment details, their approach demonstrated the practical and interpretable nature of decision trees for educational interventions, aligning with our use of similar methods. However, our project faced challenges such as overfitting when applying SMOTE-generated synthetic data.

Berens et al. (2018) explored machine learning techniques for early dropout detection, with ensemble methods like AdaBoost and decision trees excelling in retention prediction. Incorporating diverse attributes such as demographics, prior education, and enrollment type, their study underscores the importance of robust classifiers and diverse data sources, paralleling our findings with Random Forests, which effectively capture complex interactions and address class imbalance.

Beaulac and Rosenthal (2019) demonstrated the strength of Random Forests in predicting student outcomes using a decade-long dataset. By capturing nuanced interactions between academic metrics and program-specific variables, Random Forests provided actionable insights and consistent performance. Similarly, our study found Random Forests to outperform other models, leveraging academic and socioeconomic attributes while maintaining robustness to noise and imbalance. Expanded discussions on these parallels will be detailed in the methods section.

Dataset description

The dataset, derived from the Polytechnic Institute of Portalegre, Portugal, aims to reduce academic dropout and failure rates through machine learning (Martins et al., 2020). It spans the academic years 2008/09 to 2018/19 and includes **3,623 records** from undergraduate programs in various disciplines such as agronomy, education, management, and technology. The dataset

comprises **25 independent variables** grouped into demographic, socioeconomic, and academic categories. Demographic features include age at enrollment, gender, marital status, nationality, and special needs, while socioeconomic variables encompass parental education and employment, student-worker status, student grants, and debt. Academic attributes focus on pre-enrollment factors like admission grades and retention years in high school.

The dataset predicts academic outcomes based on enrollment data, excluding post-registration academic assessments (Martins et al., 2020). Each record is classified into one of three outcomes: **Success** (degree completed on time), **Relative Success** (degree completed within three extra years), or **Failure** (degree not completed or exceeding three extra years). These categories correspond to low-risk, medium-risk, and high-risk students, respectively. To address **class imbalance**, with 56% labeled as "Success," 16% as "Relative Success," and 28% as "Failure," extensive preprocessing was conducted. This included handling anomalies, outliers, and excluding incomplete records from recent academic years to ensure accurate classification.

Preprocessing, oversampling

The dataset's class imbalance, with graduates dominating over dropouts and enrolled students, risks biasing models toward the majority class. To address this, **SMOTE (Synthetic Minority Oversampling Technique)** was applied. SMOTE generates synthetic samples for underrepresented classes by interpolating between existing data points, enhancing data diversity and reducing overfitting to the majority class (Chawla et al., 2002). SMOTE effectively improved class balance, particularly aiding in the prediction of dropouts and enrolled students. Models like Random Forests leveraged the added diversity to improve accuracy and robustness, while Logistic Regression achieved its highest performance (79.27%) when combined with SMOTE and scaling. However, some linear models, such as Perceptron and SGD, struggled with noise and class overlap introduced by synthetic data, emphasizing the importance of careful model selection and tuning when addressing class imbalance.

Categorical variables in the dataset were encoded using `LabelEncoder` to convert text-based data into numerical values for model compatibility. Outliers were detected using z-scores, where records with z-scores above 3 were removed to ensure a cleaner dataset and mitigate skewed results. Finally, a correlation matrix was generated to analyze relationships between numerical features, aiding feature selection and preprocessing decisions.

Results and Discussion

Summary of Results

The dataset with attributes ranging from academic performance to socioeconomic and demographic factors, posed challenges due to its class imbalance and diverse feature scales. Results table can be viewed in figure 2. Preprocessing techniques such as scaling and SMOTE resampling were crucial to optimize model performance. Among all models, Random Forests and Logistic Regression emerged as the top performers, achieving accuracies of 78.26% and 79.27%, respectively, after appropriate preprocessing. Random Forests excelled at capturing complex, non-linear feature interactions, such as the interplay between curricular unit grades and socioeconomic indicators like GDP or unemployment rate. Logistic Regression, in contrast, stood out for its robustness and scalability, efficiently modeling linear relationships and adapting well to resampling and scaling to address class imbalance.

Other models demonstrated varying levels of success. Decision Trees showed strong interpretability, achieving 70.31% accuracy with scaling and cross-validation but struggled with overfitting to synthetic data generated by SMOTE. The SGDClassifier and KNN benefited from scaling but were hindered by their sensitivity to noise and redundant features, with KNN achieving a peak accuracy of 67.37%. Naive Bayes models underperformed due to their independence assumptions, which were violated by interdependent features like curricular unit grades and evaluations. While these models have strengths in simplicity and efficiency, their limitations underscore the need for models like Random Forests and Logistic Regression that balance interpretability and accuracy, making them well-suited for predicting student dropout in this dataset.

model	without scaling and without CV	without scaling and with CV	with scaling and without CV	with scaling and with CV	smote resampled without scaling	smote resampled with scaling and without CV	smote resampled with scaling and with CV
logistic regression	0.709040	0.744786	0.755650	0.774115	0.689302	0.792558	0.7927
sgdclassifier	0.690678	0.693885	0.733051	0.743021	0.706215	0.652542	N/A
perceptron	0.635593	0.676217	0.721751	0.724632	0.676554	0.659605	N/A
decision tree classifier	0.676554	0.703081	0.676554	0.703081	N/A	N/A	0.663842
random forest classifier	0.755650	0.781896	0.757062	0.782603	0.754237	0.754237	N/A
gaussiannb	0.682203	0.681874	0.631356	0.670215	0.663842	0.538136	0.556497
knn	0.646893	0.672328	0.658192	0.673735	0.624294	0.632768	N/A

Figure 1. Summary of model performance

Insights from Dataset Characteristics

The dataset consists of 35 attributes capturing a mix of demographic, academic, and socioeconomic factors, making it well-suited for exploring predictive relationships related to student dropout, enrollment, and graduation. The correlation matrix (figure 2) provides a detailed view of feature relationships, which informs both preprocessing and model selection.

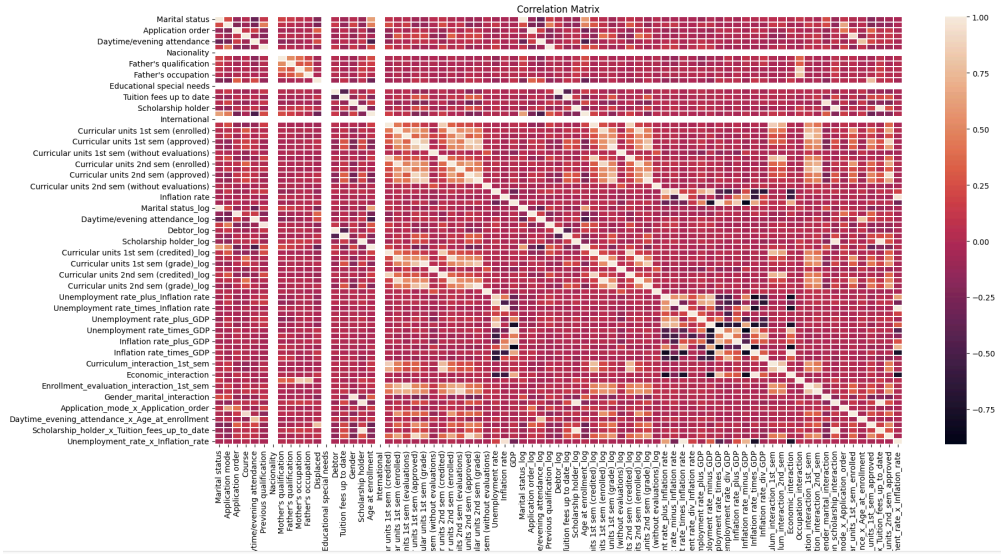

Figure 2. Correlation matrix for dataset attributes.

Attributes representing **academic performance**, such as *Curricular units 1st sem (credited, enrolled, evaluations, approved, grade)* and their second-semester counterparts, exhibit moderate to strong correlations with the target variable. These features capture critical aspects of student engagement and outcomes, with clear progression paths, such as "Curricular units (evaluations)" predicting "Curricular units (approved)" and final grades. Models like Random Forests excel in capturing these interdependencies without requiring explicit feature engineering, while Logistic Regression performs effectively when linear relationships dominate.

Socioeconomic factors, including *Tuition fees up to date*, *Scholarship holder*, and macroeconomic indicators like *Unemployment rate*, *Inflation rate*, and *GDP*, show weaker correlations but still serve as important secondary predictors. For example, "Tuition fees up to date" reflects financial stability, a known risk factor for dropout. Logistic Regression models these linear relationships well when features are appropriately scaled, whereas Random Forests integrate such features seamlessly without requiring significant preprocessing, due to their robustness to noise.

Demographic attributes, such as *Gender*, *Age at enrollment*, and *Marital status*, demonstrate weak direct correlations with the target variable but gain importance when interacting with academic and socioeconomic predictors. For instance, "Age at enrollment" can influence academic performance. Ensemble methods like Random Forests effectively leverage these interactions while de-emphasizing weaker predictors. In contrast, simpler models like Naive Bayes struggle due to their independence assumptions.

Finally, certain features, such as *Mother's qualification* and *Father's qualification*, show moderate correlations with each other but weak direct relationships with the target variable. These attributes hold predictive value when combined with academic metrics. Random Forests adeptly handle such interdependencies, while Naive Bayes models underperform, highlighting the importance of selecting models capable of managing feature correlations in dropout prediction tasks.

Model suitability for predicting student retention

1. Random Forests: Insights from Literature and Application

The findings from **Beaulac and Rosenthal (2019)** align strongly with the performance of Random Forests in this project. Their study employed Random Forests to predict academic success and major selection using a decade of course registration data from a Canadian university. Random Forests excelled in modeling feature interdependencies and delivering interpretable insights through variable importance analysis. Early academic performance metrics, such as grades in foundational courses, emerged as significant predictors. Similarly, in our project, *Curricular units (evaluations)* and *Tuition fees up to date* were identified as critical features, demonstrating Random Forests' ability to capture nuanced patterns in complex educational datasets.

Hutt et al. (2019) also support the applicability of Random Forests, showcasing their effectiveness in predicting 4-year college graduation rates using diverse features like sociodemographics, standardized test scores, and extracurricular engagement. Random Forests consistently outperformed models such as Logistic Regression and Decision Trees, with strengths in identifying key predictors through implicit feature selection and maintaining fairness in predictions across subpopulations. This robustness mirrors our project, where Random Forests handle both high-impact predictors and noisy or weakly correlated features like *Gender* or *Nationality*, ensuring reliable predictions.

Random Forests are particularly well-suited for our dataset due to their ability to capture **non-linear relationships** and interactions between features. For example, the interplay between *Curricular units (enrolled)* and *Curricular units (evaluations)* is critical for dropout prediction, and Random Forests excel at uncovering such dependencies. Additionally, their ensemble averaging reduces the impact of noise, ensuring predictors with stronger explanatory power dominate model decisions. This robustness is reflected in the consistently high accuracy of Random Forests across preprocessing configurations in our study. Both Beaulac and Rosenthal's and Hutt et al.'s findings emphasize Random Forests' adaptability and effectiveness in educational analytics. Variable importance analyses in these studies provided actionable insights—highlighting key academic and demographic features—paralleling our use of Random Forests to identify pivotal predictors like *Scholarship holder* and *Age at enrollment*.

Building on these insights, our project underscores Random Forests' role as a cornerstone of predictive analytics in education. Their ability to generalize across diverse datasets, capture feature interactions, and maintain interpretability makes them an optimal choice for dropout prediction. By integrating robust ensemble techniques with actionable insights, Random Forests provide a powerful tool for early intervention strategies to improve student retention outcomes.

2. Logistic Regression.

Logistic Regression demonstrates strong performance in dropout prediction tasks, particularly when combined with data preprocessing techniques like scaling and Synthetic Minority Oversampling Technique (SMOTE). This combination yields an accuracy of **79.27%** in our study, emphasizing the model's ability to capture linear trends effectively. Key predictive features include **Curricular unit grades**, **Tuition fees up to date**, and **average academic performance metrics**, which align with the model's linear assumptions. Logistic Regression ensures equitable predictions across all target classes, making it a practical choice for systems requiring real-time and interpretable outputs.

The efficacy of Logistic Regression in educational prediction tasks is corroborated by studies such as **Aulck et al.** and **Berens et al. (2018)**. Aulck et al. highlight the model's robustness in handling structured datasets, particularly those involving demographic and academic features. Similarly, Berens et al. developed an Early Detection System (EDS) for predicting student attrition using Logistic Regression alongside other machine learning methods. Their study achieved an accuracy of **79%** at the end of the first semester, demonstrating Logistic Regression's effectiveness in leveraging administrative and academic performance data for early intervention systems.

The use of logistic regression models aligns well with datasets characterized by high-dimensional structured variables. These models provide interpretable coefficients, enabling educational institutions to identify actionable factors, such as financial instability or academic underperformance, to reduce dropout rates. By applying feature scaling and resampling techniques, logistic regression further addresses challenges like class imbalance and heterogeneous feature scales, as evidenced by its performance improvements in our study and Berens et al.'s findings. This alignment between literature insights and empirical results underscores Logistic Regression's relevance and utility in the predictive modeling of student dropout, supporting its selection as a robust baseline model in similar educational contexts.

3. **Naive Bayes: Limitations and Insights from Literature**

Naive Bayes models, particularly GaussianNB, struggle in dropout prediction tasks due to their reliance on the independence assumption, which is frequently violated in educational datasets. Features like *Curricular units (evaluations)*, *Curricular units (approved)*, and *Curricular units (grades)* are interdependent, and Naive Bayes fails to capture the relationships among them. This limitation is further compounded when synthetic data from SMOTE resampling is introduced, as Naive Bayes models are highly sensitive to changes in feature distributions caused by synthetic samples.

Pal (2012) highlights Naive Bayes' challenges in handling interdependent features, emphasizing its inability to model feature relationships effectively in educational datasets. This study found that decision tree algorithms, such as ID3, outperformed Naive Bayes by capturing dependencies critical for accurate classification. Similarly, **Beaulac and Rosenthal (2019)** and **Hutt et al. (2019)** demonstrated the advantages of ensemble methods over simpler probabilistic approaches for datasets with complex feature interdependencies, further underscoring the limitations of Naive Bayes in modeling such relationships. **Del Bonifro et al. (2020)** also reported that Naive Bayes struggled in educational contexts where demographic and academic variables were interdependent, particularly when compared to models like Random Forests or Logistic Regression, which excel at handling nuanced feature interactions.

Implications for Dropout Prediction

Naive Bayes' reliance on marginal probabilities limits its ability to account for critical predictors, such as *Curricular units (evaluations)* and *Scholarship holder*, which interact with other features to influence dropout risk. While Naive Bayes offers simplicity and computational efficiency, its inability to handle correlated features makes it less suited for dropout prediction tasks, as supported by the literature. Models like Random Forests and Logistic Regression, which are better equipped to model complex interactions, consistently outperform Naive Bayes in this domain..

4. Decision trees

Decision trees, known for their simplicity and interpretability, performed moderately well in our project but fell short compared to ensemble methods like Random Forests. Similarly, Dekker et al. (2009) reported decision tree accuracies of 75–80% for predicting Electrical Engineering student dropouts. Their success illustrates the strengths of decision trees in handling categorical and continuous features without scaling and modeling intuitive, hierarchical decision paths that align with domain-specific attributes. In our project, decision trees captured key relationships between features like *Curricular units credited* and *Grades*, strong predictors of dropout. However, overfitting was evident when synthetic data from SMOTE was introduced, particularly for weakly correlated features such as *Marital status*. Dekker et al. addressed such issues with cost-sensitive learning and focused on domain-specific attributes, improving decision tree performance without additional data.

The contrast highlights the potential for domain-driven feature selection and strategies like cost-sensitive learning to enhance decision tree performance. While decision trees provide a strong baseline for dropout prediction, their vulnerability to

overfitting and sensitivity to noise suggest the need for careful tuning or transitioning to ensemble methods like Random Forests for better accuracy and robustness.

Popular, but lower-performing models

Despite their limited performance in our project, models like K-Nearest Neighbors (KNN), SGDClassifier, and Perceptron remain popular due to their simplicity, ease of implementation, and historical relevance in machine learning.

5. KNN

KNN's reliance on distance metrics makes it sensitive to feature scaling and susceptible to the curse of dimensionality. High-dimensional attributes in our dataset, such as academic and demographic variables, and redundant features like *Curricular units enrolled* and *Curricular units credited*, exacerbate these challenges. Scaling mitigates some issues, but SMOTE resampling introduces noise, reducing KNN's accuracy (63.27% with SMOTE and scaling). Literature supports these findings; Pal (2012) highlights KNN's effectiveness in low-dimensional, well-separated datasets but notes its limitations with imbalanced or noisy data. Dekker et al. (2009) emphasize the need for domain-specific preprocessing, which could improve KNN's performance but is not inherently facilitated by the model. Additionally, Kadar et al. (2018) illustrate KNN's potential in controlled environments with augmented data, such as emotion recognition, where the dataset characteristics align well with KNN's strengths.

6. SGDClassifier and Perceptron

Both SGDClassifier and Perceptron rely on linear boundaries, limiting their ability to model non-linear relationships inherent in features like *Curricular units evaluations* and *Grades*. While scaling and SMOTE resampling offer marginal improvements, synthetic samples disrupt their optimization, leading to suboptimal results. The popularity of these models stems from their computational efficiency and ease of implementation, but as Aulck et al. (2016) and Del Bonifro et al. (2020) demonstrate, datasets with interdependent socio-demographic and academic features often require more robust models like Logistic Regression or Random Forests. These studies further highlight the limitations of purely linear approaches in capturing the nuanced relationships required for accurate student retention predictions.

Broader Implications

The modest performance of KNN, SGDClassifier, and Perceptron highlights the importance of aligning model selection with dataset characteristics. While effective for low-dimensional, balanced datasets, these models struggle in complex domains like student retention prediction. Literature consistently shows that models like Logistic Regression and Random Forests, which handle feature interactions and class imbalances effectively, are better suited for such tasks. The findings underscore the need to pair simple models with preprocessing techniques tailored to the dataset or to opt for inherently robust algorithms capable of managing noise, redundancy, and complex feature interactions.

Bibliography

Bowen, W. G., Chingos, M. M., & McPherson, M. S. (2009). *Crossing the Finish Line: Completing College at America's Public Universities*. Princeton University Press.

Bound, J., Lovenheim, M. F., & Turner, S. (2010). Why have college completion rates declined? An analysis of changing student preparation and collegiate resources. *American Economic Journal: Applied Economics, 2*(3), 129–157.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Dawson, S., Jovanovic, J., Gašević, D., & Pardo, A. (2017). From Prediction to Impact: Evaluation of a Learning Analytics Retention Program. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 474–478.

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *International Workshop on Educational Data Mining*. Retrieved from https://www.educationaldatamining.org/EDM2009/uploads/proceedings/dekker.pdf

Fischer, C., Fishman, B., Barowy, W., Reich, J., & Turner, A. (2020). Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education, 44*(1), 130–160.

Freudenberg, N., & Ruglis, J. (2007). Reframing school dropout as a public health issue. *Preventing Chronic Disease, 4*(4).

Hilbert, S., Laubenbacher, R., & Singh, V. (2021). Machine Learning for the Educational Sciences. *Review of Education, 9*(2), e3310.

Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M. T., & Realinho, V. (2021). Early Prediction of Student's Performance in Higher Education: A Case Study. *Polytechnic Institute of Portalegre*. Retrieved from https://doi.org/10.1007/978-3-030-63099-4_9

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports, 13*, 5705. https://doi.org/10.1038/s41598-023-32484-w

Mining University Registrar Records to Predict First-Year Undergraduate Attrition. (2016). ERIC Reports. Retrieved from https://files.eric.ed.gov/fulltext/ED599235.pdf

Raisman, N. (2013). The cost of college attrition at four-year colleges & universities: An analysis of 1669 U.S. institutions. *Policy Perspectives*.

Schneider, M. (2010). Finishing the First Lap: The Cost of First-Year Student Attrition in America's Four-Year Colleges and Universities. *American Institutes for Research*.