# Predicting Student Dropout and Academic Success

Yugma Patel
1230126328
ypatel42@asu.edu
SCAI, ASU

Vrukshal Patel
1230283056
vpate107@asu.edu
SCAI, ASU

Ansh Ray
1229548426
aray49@asu.edu
SCAI, ASU

Vaidahi Patel
1232209799
vpate127@asu.edu
SCAI, ASU

Vipsa Kamani
1231818590
vkamani1@asu.edu
SCAI, ASU

**Abstract**

This project aimed to address the prevalent issue of student dropout and academic success by developing predictive models using a comprehensive dataset containing information on students' backgrounds, application details, previous qualifications, family background, admission grades, financial status, and academic performance. The primary challenge was to uncover patterns within this rich dataset to accurately predict student outcomes. The importance of addressing this problem lies in enabling educational institutions, policymakers, and researchers to identify at-risk students early, facilitating timely interventions and support systems, thereby contributing to the broader understanding of educational outcomes and the formulation of effective strategies for student success. The project involved data exploration, preprocessing, feature engineering, and the development of classification models using algorithms such as logistic regression, decision trees, random forests, support vector machines, and gradient boosting. Ensemble methods and neural networks were also explored. The models were evaluated using cross-validation techniques and metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Feature importance analysis and hyperparameter tuning were employed to refine the models further.

**Index Terms**

Student Dropout, Academic Success, Predictive Models, Classification Models, Machine Learning

## I. Introduction

In contemporary education systems, student dropout stands as a formidable challenge, impacting individuals, communities, and society at large. Dropout, characterized by students prematurely disengaging from their educational pursuits, carries significant repercussions, including reduced opportunities for individuals, the perpetuation of social inequalities, and economic costs to society. Beyond hindering personal development, dropout contributes to social disintegration and economic stagnation, amplifying existing disparities and impeding progress towards inclusive growth.

Addressing student dropout requires a nuanced understanding of its multifaceted drivers, ranging from socioeconomic disadvantage to institutional barriers and inadequate support systems. Leveraging data-driven approaches is crucial for gaining insights into dropout patterns and identifying at-risk students. By harnessing advanced analytics and predictive modelling techniques, educational stakeholders can develop targeted interventions to mitigate dropout risk and foster student success. Ultimately, by empowering institutions, policymakers, and researchers with actionable insights, we can work towards building more resilient and equitable education systems that ensure every student has the opportunity to thrive.

Furthermore, as education systems evolve in response to technological advancements and shifting societal norms, new challenges and opportunities arise in the pursuit of student retention and academic success. The emergence of online learning platforms, adaptive technologies, and competency-based education models offers promising avenues for personalized learning and support interventions. However, it also introduces complexities related to digital access, online engagement, and data privacy. As such, efforts to address dropout must adapt to these changing landscapes, integrating innovative solutions that leverage technology while also safeguarding equity and inclusion in education.

### *A. Background*

In today's global landscape, the prosperity of a nation heavily relies on its youth, who play a pivotal role in propelling economic growth through heightened productivity, innovation, and competitiveness. Education serves not only to empower individuals by instilling confidence and ensuring stable incomes but also to offer enduring support and facilitate personal growth. Overall, education stands as an indispensable force for both personal and societal advancement, with the collective effort to ensure students' academic success. Nevertheless, the issue of student dropout prevails, posing significant challenges with widespread implications. Tertiary education completion rates exhibit substantial variations across OECD countries, with the average dropout rate at 31% among 19 OECD nations, resulting in an overall completion rate of 69% [1]. These findings underscore the critical need to address dropout risk factors and fortify student support mechanisms to enhance completion rates in higher education across diverse countries.

### B. Problem Statement

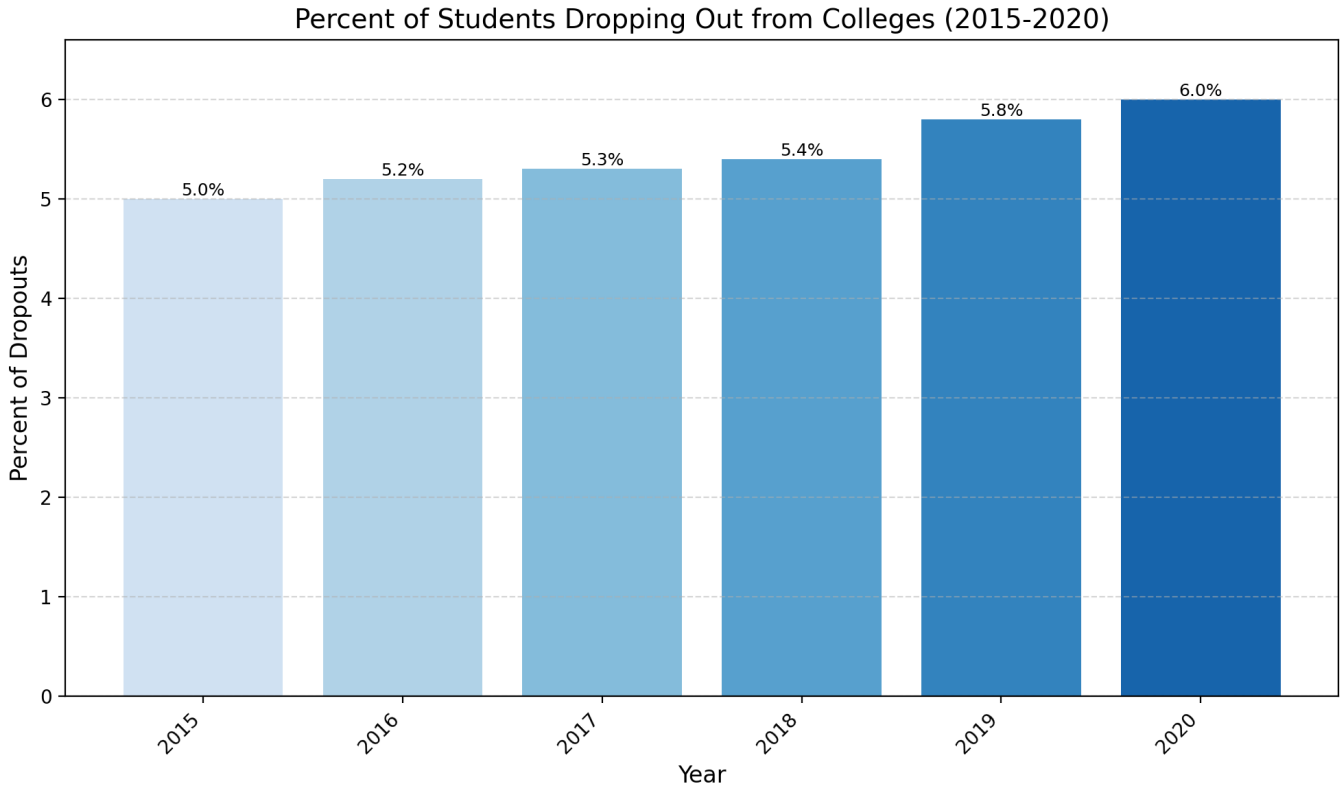


Fig. 1: Percent of students dropping out from college

In the United States, college completion remains a challenge, with dropout rates hovering around 29.2% for students entering in 2017. This means nearly a third of students never earn their degree within six years, according to the Education Data Initiative [2]. Furthermore, an estimated 35% of students change majors at least once during their college career, indicating a shift in academic focus [3]. These trends, while concerning individually, raise a combined issue: a significant number of students are either leaving college entirely or taking longer to graduate due to a shift in their academic path. This paper will investigate the factors contributing to this rise in dropouts and course changes, exploring the financial, academic, and personal reasons why students struggle to find stability in their higher education pursuits.

Figure 1 shows the number of students dropping out of college across the United States from different racial identities. As we can see the dropout rate is increasing at a slow pace even though efforts are being made to reduce this.

The project aimed to develop predictive models that could accurately classify students as likely to drop out or succeed academically, based on a comprehensive dataset containing various factors such as student backgrounds, application details, previous qualifications, family background, admission grades, financial status, and academic performance.

### C. Significance

Understanding the factors influencing dropout and academic success is crucial for educational institutions, policymakers, and researchers. By developing accurate predictive models, at-risk students can be identified early, enabling timely interventions and support systems. This contributes to the broader understanding of educational outcomes and helps in the formulation of effective strategies for student success.

### D. Existing Literature

Numerous studies are currently exploring machine learning techniques and educational data mining to tackle the challenge of predicting student dropout and academic performance. However, many of these studies come with limitations regarding the scope of data utilized or the generalizability of their approaches. For example, the authors of [4] are utilizing student data from a German university to develop models for predicting dropout risk. Although their models are achieving promising results, their dataset is limited to a single institution, potentially restricting the applicability of the models across diverse educational contexts.

Similarly, the authors of [5] are employing clustering techniques and decision trees to predict student dropout at a Thai university. While their approach considers various student attributes, their study does not incorporate socioeconomic or macroeconomic factors that could significantly impact student outcomes. Furthermore, our study employs advanced feature engineering techniques, including the extraction of temporal features and the integration of economic indicators, to uncover patterns and relationships that may be overlooked by traditional approaches. This comprehensive approach, combined with the application of diverse classification algorithms, ensemble methods, and neural networks, aims to develop robust and interpretable models that can inform targeted interventions and support strategies for student success.

### *E. System Overview*

The proposed machine learning system involved data exploration, preprocessing, feature engineering, and the development of classification models. The system aimed to address the challenges of handling a large dataset spanning multiple semesters, integrating economic indicators, and addressing both classification and regression tasks within a coherent model.

### *F. Data Collection*

The dataset utilized in this study was sourced from two primary repositories: the [6] Kaggle and [7] UCI Machine Learning Repository.

### *G. Components of the ML System*

The machine learning system consists of several essential components, such as :

1) Data Exploration: Descriptive statistics, data visualization, and correlation analysis are performed to gain insights into the dataset and identify potential predictors.
2) Data Preprocessing: Handling missing values, normalization, and encoding categorical variables, followed by applying SMOTE to address the class imbalance.
3) Feature Engineering: Extracting temporal features, integrating economic indicators, and generating features based on various models.
4) Model Development: Utilizing classification algorithms such as logistic regression, decision trees, random forests, support vector machines, gradient boosting, KNN, etc.
5) Model Evaluation: Implementing cross-validation techniques and assessing performance using metrics like accuracy, precision, recall, F1-score, and Area under the Receiver Operating Characteristic Curve.
6) Iterative Refinement: Conducting feature importance analysis and hyperparameter tuning to improve model performance.

### *H. Experimental Results*

Through exploratory data analysis (EDA), valuable insights are gained into the distribution of the target variable and the relationships between the features and the target. It is observed that parameters such as Nationality, GDP, Inflation rate, and Unemployment rate do not significantly impact the Target field, leading to their exclusion from further consideration in this project. Initially, the K-Nearest Neighbors (KNN) algorithm achieves an accuracy of 85%. However, after hyper tuning, a notable improvement is observed in the KNN model's accuracy, which rises to 88%. This underscores the critical role of model tuning in enhancing predictive performance within this project. Similarly, other models also demonstrate enhanced accuracy following hyperparameter tuning. For instance, the Random Forest Classifier achieve accuracies of 92% after hyperparameter optimization, surpassing their initial performance levels. These findings highlight the effectiveness of ensemble methods such as Random Forest, as well as the significance of model tuning, in accurately predicting student dropout and academic success based on the diverse set of features present in the dataset utilized in this project.

## II. Important definitions and problem statement

### *A. Data*

The dataset used in this project contains information about students enrolled in different undergraduate degree programs at a higher education institution. The dataset includes details about the student's academic path, demographics, socio-economic factors, and their academic performance in the first and second semesters.

### *B. Prediction target*

The target variable in this problem is the 'Target' feature, which has three possible values: 'Enrolled', 'Dropout', and 'Graduate'. To simplify the task, we have categorized 'Enrolled' and 'Graduate' as 'Not Dropout'. Therefore, the main objective is to develop a classification model that accurately predicts whether a student will drop out or not.

### C. Variables

The dataset contains various features that describe the students, such as marital status, application order, course, daytime/evening attendance, application mode, previous qualification, nationality, mother's and father's qualifications, mother's occupation, admission grade, displaced status, educational special needs, debtor status, tuition fees status, gender, and scholarship holder status. There are also features related to the student's academic performance, such as the number of credited, enrolled, evaluated, and approved curricular units in the first and second semesters.

### D. Objective

The goal of this project is to develop a machine-learning model that can accurately predict whether a student will drop out or not based on the given features. The problem is formulated as a Binary-category classification task, and there is a significant imbalance in the class distribution, with one class being much more prevalent than the others.

### E. Constraints

The system is capable of handling large datasets with numerous features and instances, as educational institutions often deal with voluminous student data. Computational efficiency and scalability are crucial considerations, particularly when dealing with complex models and real-time predictions. Moreover, the system must account for potential class imbalances, a common challenge in educational data, where certain outcomes (like dropouts) may be underrepresented compared to others (non-dropouts). Appropriate techniques, such as oversampling or undersampling, are employed to ensure that the models are not biased toward the majority class. Furthermore, the system provides interpretable and actionable results to aid educational institutions, policymakers, and researchers in decision-making and intervention strategies. While achieving high predictive accuracy is important, understanding the factors contributing to dropout risk or academic success is equally crucial for developing targeted support systems and effective policies.

## III. Overview Of Proposed System

The proposed approach 2 involves the following key steps: Exploratory data analysis (EDA) to understand the distribution of the target variable and the relationships between the features and the target. Handling of the class imbalance problem using the Synthetic Minority Over-sampling Technique (SMOTE). Evaluation of various classification models, including K-Nearest Neighbors (KNN), Logistic Regression, Stochastic Gradient Descent (SGD) Classifier, Gradient Boosting Classifier, Support Vector Classifier (SVC), and Random Forest Classifier. Optimization of the models through hyperparameter tuning using GridSearchCV. Comparison of the performance of the optimized models based on evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

## IV. Experiments

### A. Data Description

The dataset comprises data available at the time of student enrollment, including academic trajectories, demographic profiles, and socio-economic indicators. Additionally, it incorporates records of students' academic performance after their first and second semesters. With a sample size of 4424 entries, the dataset encompasses 37 distinct variables, encompassing a range of factors such as marital status, previous qualifications, parental occupations, and economic circumstances. Notably, no missing values were detected within the dataset, ensuring its completeness for analysis. Ethical considerations surrounding data collection and usage, including issues of consent, privacy, and confidentiality, were duly addressed.

### B. Evaluation metrics

In our Evaluation Metrics section, we meticulously examined the performance of our predictive models using a comprehensive set of metrics. These metrics provided a quantitative assessment of how well our models generalized to unseen data and captured the nuances of predicting dropout behaviour. By employing techniques such as precision, recall, and F1-score, we gained insights into the models' abilities to correctly classify both dropout and non-dropout instances. Additionally, we utilized classification reports to present a detailed breakdown of the models' performance across different classes, shedding light on potential biases or areas for improvement.

### C. Baseline methods for comparison

The lineup of baseline models comprised KNN, Logistic Regression, SGD Classifier, Gradient Boosting Classifier, SVC, and Random Forest Classifier. These models, drawn from established methodologies in the field of dropout prediction, provided a robust foundation for assessing the efficacy of our novel approaches. By subjecting each baseline model to the same evaluation criteria as our proposed methods, we were able to discern their relative strengths and weaknesses.

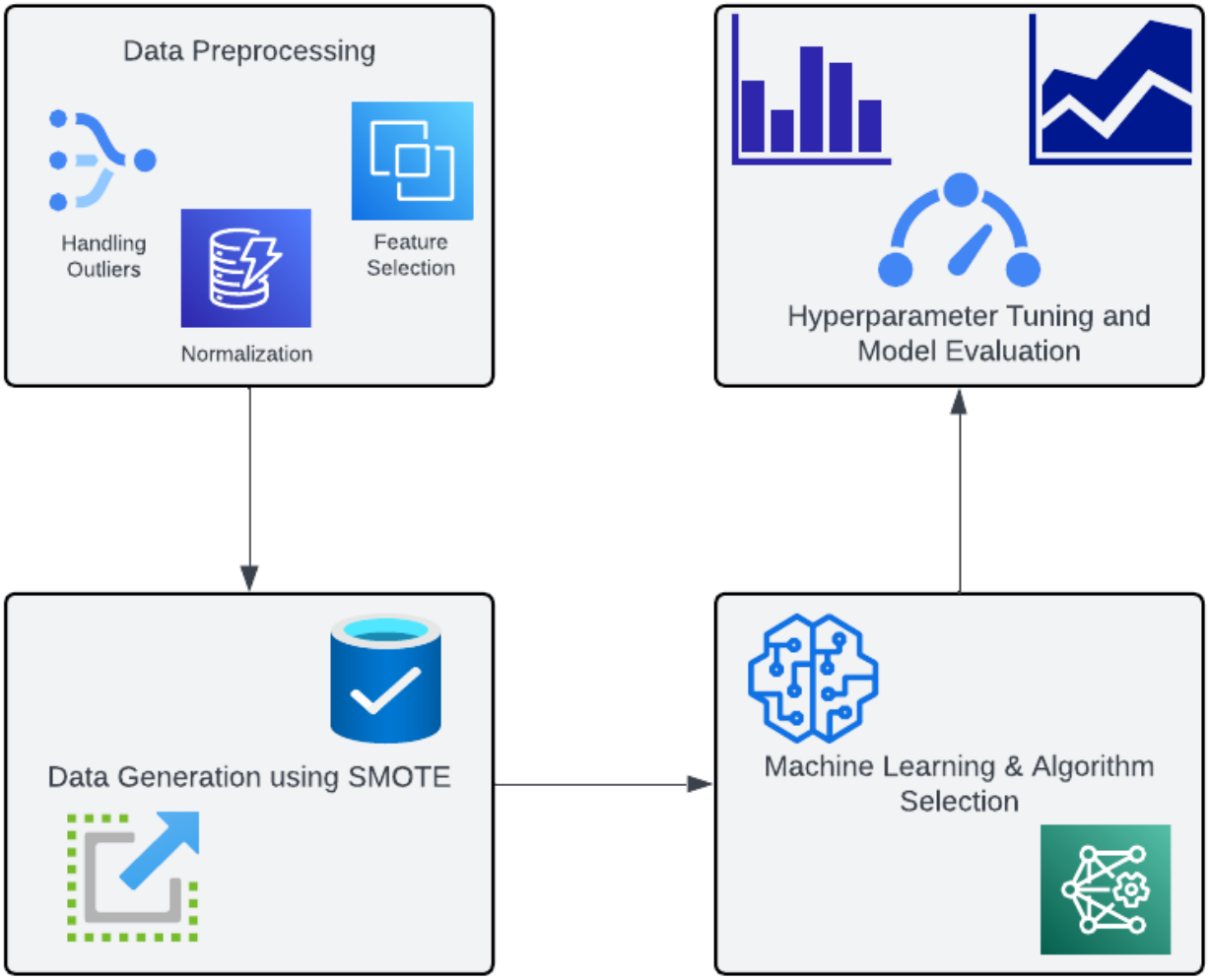


Fig. 2: ML Pipeline for the dataset

### *D. Overall performances*

We systematically analyzed all the features encompassing various aspects such as marital status, application order, course selection, attendance preferences, qualification history of students and their parents, as well as socio-economic indicators like GDP, inflation rate, and unemployment rate. Notably, after rigorous examination, we found that nationality, GDP, inflation rate, and unemployment rate had negligible impact on our target field, leading us to exclude them from further analysis.

For example, figure 3 and figure 4 represent bar charts showing the count of students in different categories (Graduate, Dropout, Enrolled) across varying levels of unemployment and inflation rates. The absence of a discernible pattern in both the count of students across categories (Graduate, Dropout, Enrolled) with fluctuating unemployment rates and the variability in counts across categories with different inflation rates suggests that neither unemployment nor inflation rates significantly influence students' decisions to graduate, drop out, or enrol.

Our primary focus centred on predicting student dropout rates, categorizing enrolled and graduated students as 'Not Dropout,' and delineating a class distribution of 3001 'Not Dropout' instances against 1421 'Dropout' instances, thereby revealing a class imbalance. To mitigate this imbalance, we employed the Synthetic Minority Over-sampling Technique (SMOTE) to ensure robustness and accuracy in our predictive modelling.

Table I presents the enhanced accuracy scores obtained post-application of hyperparameter tuning across a spectrum of models. Notably, Hypertuned XGB and Hypertuned Random Forest showcased peak accuracy and f1-score of 92%, followed closely by Hypertuned StackingClassifier and Hypertuned Bagging at 91%. However, Hypertuned Adaboost and Hypertuned MLP displayed slightly lower accuracy rates of 89% and 91% respectively, while Hypertuned KNN and Hypertuned Logistic Regression yielded accuracies of 88% and 85% correspondingly. Conversely, in Table II, the prior accuracy standings of distinct

models are delineated, revealing RandomForestClassifier and GradientBoostingClassifier leading at 91% accuracy and f1-score. Notably, KNN (Optimised) and SGDClassifier achieved 88% and 86% accuracies, respectively, whereas the basic KNN and Logistic Regression models recorded accuracies of 85%.

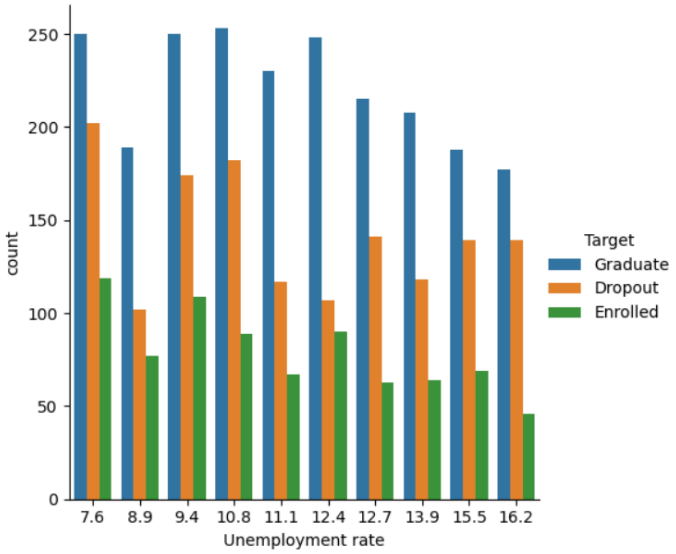


Fig. 3: Unemployement rate for each and every target attribute

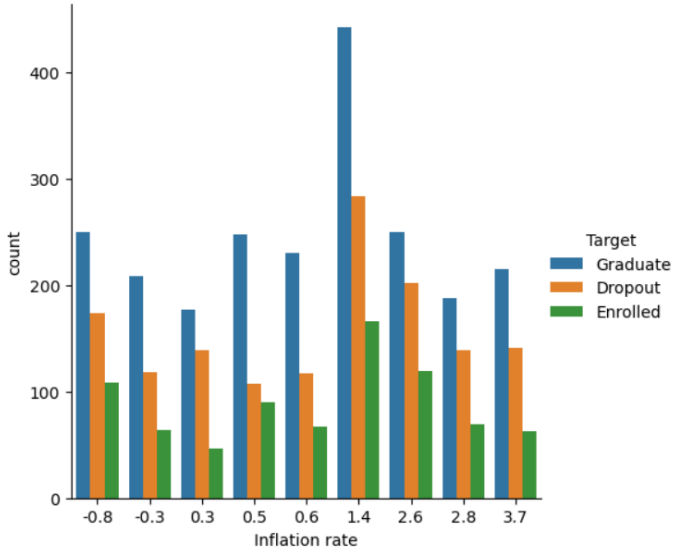


Fig. 4: People having rate of Inflation

## V. Related works

Student attrition poses a significant challenge for higher education institutions globally, prompting researchers to develop predictive models to identify at-risk students and intervene effectively. Several studies have explored the application of machine learning algorithms to leverage administrative data and improve the accuracy of dropout predictions. Understanding the determinants of attrition and implementing targeted interventions is crucial. By utilizing administrative student data from German universities and employing the AdaBoost Algorithm, they achieved prediction accuracies of up to 95% for identifying potential dropouts. This multi-method approach enhances prediction robustness, considering the heterogeneity across study programs and student demographics [8].

Researchers utilized a machine learning approach to predict academic success indicators such as Cumulative Grade Point Average (CGPA) using data from a university. Their findings revealed the significance of factors such as gender, high school exam scores, and regional demographics in determining academic outcomes. This model not only serves as a classifier but also provides actionable insights for mitigating failure and optimizing resource allocation in tertiary institutions [9].

TABLE I: Accuracy scores achieved after implementing hyperparameter tuning on various models.

| | Model | Accuracy | f1-score |
|---|---|---|---|
| 1 | Hypertuned XGB | 92 | 92 |
| 2 | Hypertuned Random Forest | 92 | 92 |
| 3 | Hypertuned StackingClassifier | 91 | 91 |
| 4 | Hypertuned Bagging | 91 | 91 |
| 5 | Hypertuned MLP | 91 | 91 |
| 6 | Hypertuned Adaboost | 89 | 89 |
| 7 | Hypertuned KNN | 88 | 88 |
| 8 | Hypertuned Logistic Regression | 85 | 85 |
| 7 | Hypertuned Perceptron | 83 | 83 |

TABLE II: Previous accuracy scores for various models.

| | Model | Accuracy | f-1 score |
|---|---|---|---|
| 1 | RandomForestClassifier | 91 | 91 |
| 2 | GradientBoostingClassifier | 91 | 91 |
| 3 | KNN (Optimised) | 88 | 88 |
| 4 | SGDClassifier | 86 | 86 |
| 5 | KNN | 85 | 85 |
| 6 | Logistic Regression | 85 | 85 |

[10] presents a comprehensive dataset encompassing demographic, socioeconomic, and academic performance data from a higher education institution. Leveraging this dataset, classification models were developed to predict student dropout and success, facilitating targeted interventions by the tutoring team. The integration of predictive analytics into educational practices enables proactive support for students at risk of dropout or academic underachievement. By leveraging diverse datasets and advanced analytical techniques, educational institutions can implement proactive strategies to support student retention and achievement.

## VI. Conclusions

In conclusion, this project represents a significant step towards addressing the pervasive issue of student dropout and fostering academic success. By leveraging a comprehensive dataset encompassing various factors such as student backgrounds, application details, academic performance, and socio-economic indicators, predictive models were developed to classify students at risk of dropout accurately. Through extensive data exploration, preprocessing, and feature engineering, coupled with the utilization of diverse classification algorithms, including logistic regression, decision trees, random forests, support vector machines, and gradient boosting, robust models were constructed and rigorously evaluated.

The significance of this endeavor lies in its potential to enable early identification of at-risk students, thereby facilitating timely interventions and support systems. By providing educational institutions, policymakers, and researchers with actionable insights into dropout risk factors, these models contribute to a broader understanding of educational outcomes and the formulation of effective strategies for student success. Moreover, the systematic experimentation and evaluation of various models, alongside comparisons with baseline methods, offer valuable insights into the efficacy of different approaches for dropout prediction.

Moving forward, the insights gleaned from this project can inform the development of proactive strategies and interventions aimed at improving student retention and academic achievement. By leveraging advanced analytical techniques and incorporating predictive analytics into educational practices, institutions can enhance their ability to support students effectively, ultimately fostering a more inclusive and successful learning environment. Through ongoing research and collaboration, we can continue to refine and improve these models, ultimately contributing to the advancement of educational outcomes on a broader scale.

## References


[1] OECD. *Education at a Glance 2009: OECD Indicators*. OECD Publishing, Paris, 2009.

[2] Education Data Initiative. College dropout rate [2023]: by year + demographics, 2023.

[3] National Center for Education Statistics. Dropping out of college and returning: Trends in bachelor's degree attainment, May 2018.

[4] Johannes Berens, Kerstin Schneider, Simon Gortz, Simon Oster, and Julian Burghoff. Early detection of students at risk - predicting student dropouts using administrative student data from german universities and machine learning methods. *Journal of Educational Data Mining*, 11(3):1–41, Dec. 2019.

[5] Natthakan Iam-On and Tossapon Boongoen. Improved student dropout prediction in thai university using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics*, 8, 02 2015.

[6] Kaggle dataset. Online. Retrieved from https://www.kaggle.com/datasets/missionjee/students-dropout-and-academic-success-dataset.

[7] UCI machine learning repository. Online. Retrieved from https://archive.ics.uci.edu/dataset/697/predict+students+dropout.

[8] J. Berens, K. Schneider, S. Görtz, S. Oster, and J. Burghoff. Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods. *Unknown*, 2018.

[9] M. N. Yakubu and A. M. Abubakar. Applying machine learning approach to predict students' performance in higher educational institutions. *Kybernetes*, 51(2):916–934, 2022.

[10] V. Realinho, J. Machado, L. Baptista, and M. V. Martins. Predicting student dropout and academic success. *Data*, 7(11):146, 2022.