**Capstone Project Paper**

**“Using Machine Learning Methods to Predict Academic Success in College”**

Kamilia Adil

Capstone Advisor: Prof. Howard T. Everson

A capstone research submitted to the Graduate Faculty in Data Analysis and Visualization in partial fulfillment of the requirements for the degree of Master of Science, The City University of New York.

2024

**Acknowledgments:**

I would like to express my heartfelt gratitude to all the staff at the Graduate Center, CUNY, including the administration, professors, and support personnel. A special thank you goes to Professor Howard T. Everson for his invaluable guidance and support throughout my capstone project, as well as during the two classes I had the privilege of taking with him. I am deeply appreciative of the knowledge and wisdom imparted by all my professors, who have not only enriched my understanding of data analysis but also provided invaluable life lessons and advice. I extend special thanks to Jason Nielsen for his unwavering support and responsiveness throughout my journey in completing this master's degree, even during late nights and vacations. Their collective assistance and encouragement have been instrumental in shaping my academic and personal growth, and for that, I am truly grateful.

**Table of Contents**

1 INTRODUCTION............................................................................................................................4

2 LITERATURE REVIEW....................................................................................................................5

3 DATA DESCRIPTION......................................................................................................................6

4 METHODOLOGY……......................................................................................................................7

5 RESULTS……………………………………………………………………………………………………………………………….11

6 DISCUSSIONS..............................................................................................................................19

7 LIMITATIONS AND FUTURE RESEARCH…………………………………………………………………………….…….20

8 CONCLUSION.............................................................................................................................21

1. **Introduction:**

In the US, only 62% of students who start a postsecondary degree or certificate program finish their program within six years in 2023 (National Student Clearinghouse Research Center 2023). Additionally, 2022 saw a notable decline in undergraduate credential earners, marking the first decrease in a decade, with a reduction of 1.6% or approximately 58,800 fewer credential earners than the previous year (National Student Clearinghouse Research Center 2023).

Against this backdrop of societal and educational challenges and evolving student dynamics, I see an increasing need for a deeper understanding of the determinants of academic success in college. What are the factors and circumstances that drive postsecondary program completion? In this project, I argue that leveraging advanced machine learning methodologies offers a promising avenue for exploring this complexity and deriving insights into the factors affecting academic achievement at the postsecondary level.

The study described here focuses on unraveling the multifaceted predictors of academic success among college students, categorizing them into three key segments: dropout, enrolled, and graduate, upon completion of their standard course duration. My primary aim is to utilize machine learning techniques to uncover discernible patterns within a rich dataset encompassing vital student enrollment details, academic trajectories, demographics, and socio-economic factors, thus providing comprehensive insights into the factors shaping academic success.

Acknowledging the inherent challenges posed by the intricate nature of students' academic pathways and the constraints of available data in postsecondary education generally, I adopt a rigorous methodology to extract meaningful insights. Beginning with meticulous exploratory data analysis (EDA), I address data integrity, feature relevance, and class imbalance concerns. Techniques such as correlation analysis, outlier detection, and oversampling were employed to refine the dataset, ensuring a robust dataset for in subsequent modeling efforts.

My analysis extends to evaluating the performance of various supervised machine-learning algorithms in predicting student outcomes. Models ranging from Decision Trees to Support Vector Machines undergo scrutiny based on accuracy, precision, recall, and F1-score metrics. Additionally, I assess the discriminatory power of these models through Receiver Operating Characteristic (ROC) analysis, providing a comprehensive understanding of their classification capabilities.

Furthermore, my research delves into unsupervised machine learning methodologies, specifically employing KMeans clustering to uncover potential clusters within the dataset. K-means clustering is one of the simplest and most commonly used clustering algorithms. It tries to find cluster centers that are representative of certain regions of the data. The algorithm alternates between two steps: assigning each data point to the closest cluster center, and then setting each cluster center as the mean of the data points that are assigned to it. The algorithm is finished when the assignment of instances to clusters no longer changes (Andreas C. Müller, Sarah Guido 2016). Through rigorous evaluation of cluster purity and distance metrics, I aim to illuminate the underlying structure of student profiles.

The findings of this study hold significant implications for educational stakeholders, policymakers, and practitioners endeavoring to implement targeted interventions to bolster student success. By identifying salient predictors of academic outcomes and leveraging advanced analytical techniques, this research strives to inform evidence-based strategies that nurture student retention, mitigate dropout rates, and foster graduation.

1. **Literature review:**

Predicting students' dropout and graduation has garnered significant attention in educational research, driven by the imperative to enhance student support systems and improve academic outcomes. While traditional academic variables have long been considered in predictive modeling, recent studies underscore the importance of incorporating broader contextual factors, including personal issues, mental health challenges, and socioeconomic status, which exert substantial influence on academic success.

In this vein, several studies have leveraged machine learning (ML) methods to forecast student outcomes. Notably, Kaggle users (competitors) have explored the predictive capabilities of logistic regression, a versatile statistical technique well-suited for binary classification tasks, using Python's Sklearn library. Their approach involved partitioning the dataset into training and test sets, followed by the application of logistic regression to make predictions on the test data. Of particular significance is their utilization of all available features to elucidate the target variable, whether indicating dropout or graduation.

The evaluation of predictive models entails a nuanced examination of various performance metrics. Competitors assessed their logistic regression model using metrics such as cross-entropy loss, accuracy, precision, and recall. Cross-entropy loss quantifies the disparity between predicted and actual outcomes, with lower values indicative of superior model performance. Accuracy gauges the proportion of correctly classified instances, while precision measures the model's ability to correctly identify positive cases. Furthermore, recall assesses the model's capacity to capture all positive instances within the dataset.

The outcomes of the competitor's analysis yielded promising results, exemplified by a cross-entropy loss of 2.80, an accuracy score of 0.92, a precision score of 0.92, and a recall score of 0.85. Despite these commendable achievements, critical reflections on the methodology are warranted. Logistic regression, while effective in many scenarios, may struggle to capture the nuanced relationships between predictors and student outcomes, particularly when assumptions of linearity are violated. Additionally, addressing class imbalances and delving into coefficient analysis could deepen the understanding of model behavior and enhance interpretability.

In summary, existing literature underscores the evolving landscape of predictive modeling in education, with ML methodologies offering promising avenues for exploring the multifaceted determinants of academic success. By embracing a holistic approach that integrates diverse factors and robust evaluation techniques, researchers can glean invaluable insights into student trajectories and inform evidence-based interventions aimed at fostering retention, mitigating dropout rates, and promoting graduation.

1. **Data description:**

The dataset utilized in this project is sourced from the University of California, Irvine Machine Learning Repository. This dataset contains multiple disjoint databases consisting of relevant information available at the time of enrollment, such as application mode, marital status, chosen course, and more. Additionally, this data can be used to estimate overall student performance at the end of each semester by assessing curricular units. It encompasses details known at the time of student enrollment, including academic path, demographics, socioeconomic factors, and academic performance at the end of the first and second semesters. The dataset follows students throughout their academic journey, providing valuable insights into their progress and outcomes.

Formatted as a CSV file, the dataset comprises 4424 instances, with each instance representing a student, and 37 columns. Of these columns, 36 serve as attributes, capturing various aspects of student information, while one column represents the target variable, indicating whether a student dropped out or graduated. The creation of this dataset was supported by the program SATDAP – Capacitacao da Administacao Publica program, funded under the grant POCI-05-5762-FSE-000191 in Portugal. However, it's important to note that while this dataset offers comprehensive insights into student demographics, performance, and outcomes, its limitation to higher education institutions may affect its generalizability to institutions in the U.S. or other countries.

* **DataFrame Information Summary:**

A screenshot of a computer program

Description automatically generated

1. **Methodology:**

To ensure the integrity of my data, I conducted an exploratory data analysis (EDA) journey and utilized various data visualization techniques, bar plots, boxplots, histograms, and correlation matrices. Initially, I meticulously checked for missing and duplicated values within the dataset, ensuring any gaps were filled using the mean value. I made a deliberate decision to exclude the "Enroll" class from the target variable, focusing solely on predicting dropout and graduation. Subsequently, I transformed the target variable, representing "Dropout" as 1 and "Graduate" as 0.

Moving forward, I delved into descriptive statistics to uncover anomalies in the dataset. This analysis highlighted variations in data point ranges across features, prompting the need for data scaling. Following this, I performed correlation analysis to comprehend the relationships between variables and their influence on the target variable.

* **Summary Statistics for Key Variables:**

A screenshot of a data sheet

Description automatically generated

Leveraging data visualization techniques, I employed the Seaborn library to craft box plots for categorical variables and cat plots to visualize numerical variables. These visual aids were instrumental in identifying outliers, assessing target variable imbalance, and grasping feature relationships with dropout or graduation.

* **Distribution of Student Graduation and Dropout:**

A graph showing a number of blue and orange squares

Description automatically generated

To address outliers within the dataset, I turned to the Interquartile Range (IQR) method. The IQR is used to measure variability by dividing a dataset into quartiles. The data is sorted in ascending order and split into 4 equal parts: Q1, Q2, and Q3, called the first, second, and third quartiles, respectively. The interquartile range is calculated as the difference between the upper and lower quartiles in the given data. It is a useful measure of variability for skewed distributions, representing the range between the first and the third quartiles, namely Q1 and Q3: IQR = Q3 – Q1 (GeeksforGeeks, Interquartile Range in Statistics 2023). Employing a threshold of 1.5 times the IQR ensured data integrity and mitigated the impact of outliers on subsequent analysis or modeling, thereby enhancing the quality and reliability of the dataset.

In addressing the issue of imbalanced data, I applied the SMOTE (Synthetic Minority Over-sampling Technique) oversampling method to the training data. Imbalanced data poses significant challenges in machine learning tasks due to the unequal distribution of classes within the dataset. When one class significantly outweighs the other, the model may exhibit bias towards the majority class during training, leading to poor performance in predicting the minority class. This imbalance can result in inaccurate model evaluations, reduced sensitivity to the minority class, and decreased overall predictive power. By oversampling the minority class using SMOTE, I effectively balanced the class distribution, thus enhancing the model's ability to generalize across both classes. Post-application of SMOTE, I converted the oversampled arrays back to DataFrames for further analysis, ensuring balanced class proportions throughout the dataset.

With preprocessing completed, I proceeded to split the dataset into training and testing sets using the train\_test\_split function, allocating 70% of the data for training and 30% for testing. This strategic split ensured reproducibility, as I set the random\_state parameter to 0.

Subsequently, I standardized the features in the dataset using StandardScaler, a method provided by the sci-kit-learn library in Python. By fitting it to the training data and transforming both the training and testing sets to have a mean of 0 and a standard deviation of 1, this meticulous standardization process guaranteed consistency in feature scaling across the dataset, ensuring that all features were on the same scale for optimal model performance.

In evaluating the models, I employed cross-validation for each model in the dataset using KFold with 5 splits. This involved calculating accuracy scores for each fold and averaging them to assess the model's performance. The results were meticulously analyzed to discern mean accuracy along with the standard deviation, providing valuable insights into the model's stability and generalization ability.

Model training commenced with a laser focus on binary classification using supervised machine learning algorithms. These algorithms are used whenever we want to predict a certain outcome from a given input, and we have examples of input/output pairs (Andreas C. Müller, Sarah Guido 2016). Specifically, I utilized:

**KNN (K-Nearest Neighbors):** This considers exactly one nearest neighbor, which is the closest training data point to the point we want to predict. The prediction is then simply the known output for this training point (Andreas C. Müller, Sarah Guido 2016).

**Gaussian Naive Bayes:** uses the probability of observing predictor values, given an outcome, to estimate the probability of observing outcome Y = I, given a set of predictor values (Peter Bruce, Andrew Bruce 2017).

**Decision Trees:** are widely used models for classification and regression tasks. Essentially, they learn a hierarchy of if/else questions, leading to a decision (Andreas C. Müller, Sarah Guido 2016). A decision tree is a flowchart-like tree structure where each internal node denotes the future, branches denote the rules and the leaf nodes denote the result of the algorithm (GeeksforGeeks, Decision Tree 2023).

**Random Forest:** A type of bagged estimate based on decision tree models (Peter Bruce, Andrew Bruce 2017). A random forest is essentially a collection of decision trees, where each tree is slightly different from the others. The idea behind random forests is that each tree might do a relatively good job of predicting but will likely be overfitting on part of the data. If we build many trees, all of which work well and overfit in different ways, we can reduce the amount of overfitting by averaging their results (Andreas C. Müller, Sarah Guido 2016).

**Logistic Regression:** is used for binary classification where we use a sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1 (GeeksforGeeks, Logistic Regression in Machine Learning 2024).

**Support Vector Machine:** is used for both classification and regression. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the future space. The hyperplane tries to ensure that the margin between the closest points of different classes should be as maximum as possible. The dimension of the hyperplane depends upon the number of features (GeeksforGeeks, Support Vector Machine (SVM) Algorithm 2023).

Evaluation of the models entailed generating comprehensive classification reports to compare accuracy, precision, recall, and F1 scores. Accuracy measures the percent (or proportion) of cases classified correctly (Peter Bruce, Andrew Bruce 2017), precision measures the accuracy of a predicted positive outcome (Peter Bruce, Andrew Bruce 2017), recall is the ratio between true positives and the some of the true positive and false positive. The recall also known as sensitivity, measures the strength of the model to predict a positive outcome—the proportion of the 1s that it correctly identifies (Peter Bruce, Andrew Bruce 2017), and the F1-score is a measure of a model’s performance that combines precision and recall. It is defined as the harmonic mean of precision and recall, where the best value is 1 and the worst value is 0 (GeeksforGeeks, Precision, Recall and F1-Score using R 2023).

Furthermore, I conducted feature importance analysis which rates how important each feature is for the decision a tree makes. This analysis assigns a value between 0 and 1 for each feature, where 0 means “not used at all” and 1 means “perfectly predicts the target (Andreas C. Müller, Sarah Guido 2016). The feature importances always sum to 1. and this analysis is performed using techniques such as Gini impurity for decision trees and random forest classifiers.

After conducting the feature importance analysis, I reassessed and recalibrated my model's performance. Initially, I evaluated the model using all predictors. Subsequently, I identified and removed features with negligible importance. This method facilitated a thorough evaluation of the model's performance both before and after eliminating non-significant predictors.

Additionally, I employed ROC analysis to assess discriminatory ability, with AUC serving as a pivotal performance metric throughout the evaluation process. The AUC metric captures the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate (Matz, S.C, Bukow, C.S, Peters, H. et al. 2023). When the AUC is 0.5, the model’s predictive performance is equal to chance or a coin flip. The closer to 1, the higher the model’s predictive performance in distinguishing between students who drop out and those who graduate.

Furthermore, I delved into unsupervised machine learning, which refers to statistical methods that extract meaning from data without training a model on labeled data (data where an outcome of interest is known) (Peter Bruce, Andrew Bruce 2017).  Via K-means clustering, I utilized clustering as a technique to divide data into different groups, where the records in each group are similar to one another. The goal of clustering is to identify significant groups of data. The groups can be used directly, analyzed in more depth, or passed as a feature or an outcome to a predictive regression or classification model (Peter Bruce, Andrew Bruce 2017).

**Deployment:**

In deploying my Dash application, I meticulously followed a multi-step deployment process to ensure its successful integration and functionality within a production environment. Initially, I dedicated considerable effort to developing the application interface using a combination of HTML, Dash Core Components, and sophisticated callback functions to enable dynamic user interactions. Once the application was fully developed and thoroughly tested, I transitioned to the deployment phase, where I employed a combination of deployment tools and methodologies. Leveraging Git version control, I established a streamlined deployment pipeline to automate the deployment process, ensuring rapid and consistent deployments. This involved configuring deployment settings, managing dependencies, and executing deployment scripts to prepare the application for production. Additionally, I implemented continuous integration and continuous deployment (CI/CD) practices to automate testing and deployment tasks, further enhancing deployment efficiency and reliability. Once deployed, the application was made accessible to users via a unique URL provided by Render, enabling real-time interaction and utilization of its features. Render's robust infrastructure played a pivotal role in ensuring the reliability, security, and scalability of the deployed application. Features such as free TLS certificates, a global CDN, and private networks bolstered the application's security posture and facilitated efficient content delivery to users worldwide. Throughout the deployment process, I remained vigilant, conducting thorough monitoring and maintenance to address any issues promptly and ensure optimal performance. By meticulously following this deployment process, I was able to seamlessly integrate my Dash application into a production environment, providing users with a reliable and feature-rich application experience.

1. **Results:**

This section offers an in-depth exploration of the outcomes of my study concerning the factors influencing graduation outcomes among students. I begin by introducing the various analyses conducted, including correlation analysis, supervised machine learning, unsupervised machine learning (specifically KMeans clustering), and Receiver Operating Characteristics (ROC) analysis. Each analysis is meticulously examined to unveil significant findings and insights. The section concludes with a discussion of the implications derived from feature analyses, shedding light on the utility of predictors in the dataset.

**Results of Correlation Analysis:**

The results of the correlation analysis offer valuable insights into the relationships between different variables and their impact on graduation outcomes. The heatmap of correlations reveals significant associations between various curricular units from the 1st and 2nd semesters. Notably, "Curricular units 2nd sem (credited)" and "Curricular units 1st sem (credited)" exhibit the highest correlation at 0.95, while "Curricular units 1st sem (enrolled)" and "Curricular units 2nd sem (enrolled)" display a strong correlation of 0.94.

Regarding correlations with the target variable, all show positive relationships, indicating that higher values of the target variable (representing graduation) correspond to higher values of the correlated features. Among these, "Curricular units 2nd sem (approved)" displays the highest positive correlation (correlation coefficient = 0.654), suggesting a strong association with the likelihood of graduation. Similarly, "Curricular units 2nd sem (grade)" shows a substantial positive correlation (correlation coefficient = 0.605), indicating that higher grades in the second semester are associated with a greater likelihood of graduation.

Conversely, certain features exhibit weaker positive correlations with the target variable. For instance, "Mother's occupation" demonstrates a very low correlation coefficient of 0.001, suggesting a minimal relationship with graduation compared to other features. Despite variations in correlation strength, all features display positive associations with graduation, providing valuable insights into the factors influencing student success and the likelihood of graduating.

* **Correlation with Student Dropout or Graduation:**

A graph with numbers and a number in the middle

Description automatically generated with medium confidence

1. **Results of Supervised Machine Learning:**

Before delving into the evaluation of various machine learning classifiers and their performance in predicting student outcomes regarding dropout and graduation, it is essential to understand the rationale behind the chosen ensemble approach. The decision to utilize this ensemble methodology stems from the need to comprehensively assess each model's performance individually. Rather than aggregating predictions, this approach allows for a detailed examination of the strengths and weaknesses of each algorithm in predicting student outcomes related to dropout and graduation. By evaluating algorithms such as KNN, Gaussian Naive Bayes, Decision Trees, Random Forest, Logistic Regression, and Support Vector Machine independently, I aim to identify the most suitable model for the dataset and task at hand. This methodological choice facilitates a rigorous evaluation of each model's efficacy and enables the selection of the best-performing algorithm for predicting student outcomes.

Following this rationale, the evaluation included cross-validation to assess the models' robustness and classification reports to analyze precision, recall, and F1-score. Precision measures the proportion of true positive predictions out of all positive predictions made by the model, indicating the accuracy of the positive predictions. Recall, on the other hand, measures the proportion of true positive predictions out of all actual positive instances in the dataset, indicating the model's ability to capture all positive instances. The F1-score, a metric derived from the harmonic mean of precision and recall, serves as a comprehensive measure of a model's predictive performance. It considers both precision and recall, making it particularly valuable in assessing the model's ability to balance between correctly identifying positive instances and minimizing false predictions. Therefore, incorporating the F1-score into the evaluation provides a holistic understanding of the model's reliability and effectiveness in predicting student outcomes.

* **K-Nearest Neighbors (KNN):**

K-Nearest Neighbors (KNN): Achieved a cross-validation accuracy of 90.55%, with precision, recall, and F1-score of 89%, indicating proficiency in predicting student outcomes. However, it showed potential challenges in achieving a balance between precision and recall, particularly in classifying dropout instances.

* **Decision Trees (DT):**

The Decision Trees model achieved a cross-validation accuracy of 87.36%, with precision, recall, and F1-score of 0.83, 0.84, and 0.84, respectively, and an accuracy of 85%. This model demonstrated moderate accuracy and balanced performance in predicting both dropout and graduation instances.

* **Random Forest Classifier (RFC):**

The Random Forest Classifier demonstrated strong performance with a cross-validation accuracy of 91.74% and an accuracy score of 91%. It achieved balanced precision and recall values of 0.91 and an F1-score of 0.88, indicating robustness in classifying both dropout and graduation instances.

* **Gaussian Naive Bayes (GNB):**

The Gaussian Naive Bayes model initially displayed a lower cross-validation accuracy of 20.3% and an accuracy of 22%, with precision and recall scores of 0.81 and 0.22, respectively. The F1-score was also relatively low at 0.22, indicating challenges in accurately predicting both dropout and graduation cases.

However, following feature importance analysis, where non-significant features were removed from the predictor variables (X), significant improvements were observed in the performance of the Gaussian Naive Bayes model. The cross-validation accuracy surged from 20.3% to 83%, while accuracy increased from 22% to 81%. Although precision remained consistent at 0.81, recall significantly improved from 0.22 to 0.81. Additionally, the F1 score saw a substantial increase from 0.22 to 0.83.

* **Logistic Regression (LR):**

Similarly, Logistic Regression achieved a high cross-validation accuracy of 91.04% and an accuracy score of 90%. It showed precision and recall values of 0.89 and 0.90, respectively, with an F1-score of 0.88, highlighting its effectiveness in identifying both dropout and graduation cases.

* **Linear Support Vector Machine (SVM):**

The Linear SVM model displayed comparable performance to Logistic Regression, with a cross-validation accuracy of 91.44%, an accuracy score of 90%, precision, and recall values of 0.89 and 0.90, respectively, and an F1-score of 0.88, indicating consistent performance in classifying both dropout and graduation instances.

* **Radial Basis Function Support Vector Machine (RBF SVM):**

However, the RBF SVM model showed a slightly lower cross-validation accuracy of 89.05%, an accuracy score of 88%, and precision, and recall values of 0.77 and 0.88, respectively. The F1-score was 0.82, suggesting some challenges in correctly classifying dropout instances.

* **Polynomial Support Vector Machine (Poly SVM):**

The Polynomial SVM model exhibited a high cross-validation accuracy of 91.84% and an accuracy score of 89%. It showed precision and recall values of 0.88 and 0.89, respectively, with an F1-score of 0.86, indicating robust performance in predicting both dropout and graduation instances.

* **Sigmoid Support Vector Machine (Sigmoid SVM):**

Finally, the Sigmoid SVM model achieved a cross-validation accuracy of 89.25% and an accuracy score of 87%, with precision and recall values of 0.83 and 0.87, respectively. The F1-score was 0.84, suggesting reliable performance in classifying both dropout and graduation cases.

**Cross-Validation Scores for Student Academic Outcomes:**

**A graph of a bar graph

Description automatically generated**

* **Weighted Average of Various Metrics for Student Academic Outcomes before Conducting Feature Importance Analysis:**

**A graph of different colored bars

Description automatically generated**

* **Weighted Average of Various Metrics for Student Academic Outcomes after Conducting Feature Importance Analysis:**

**A graph of different colored bars

Description automatically generated**

1. **Results of Receiver Operating Characteristics (ROC):**

Upon assessing the AUC values of the classifiers, it's crucial to contextualize their performance within the Receiver Operating Characteristics (ROC). The AUC metric measures the model's ability to discriminate between positive and negative instances across all possible threshold values.

The Polynomial Support Vector Machine (Poly SVM) achieved the highest AUC of 0.83, indicating strong discriminatory power and effective separation of the Graduate and Dropout classes with minimal overlap. Following closely, the Logistic Regression model demonstrated robust discriminatory ability with an AUC of 0.81, signifying strong performance in classification. The Linear Support Vector Machine (Linear SVM) also displayed strong discriminatory capability, achieving an AUC of 0.80.

The Random Forest model yielded a moderate discriminatory ability with an AUC of 0.79, suggesting reasonable performance in class separation.

Moving on, the K-Nearest Neighbors (KNN) model exhibited an AUC of 0.72, indicating moderate discriminatory ability. Similarly, the Gaussian Naïve Bayes classifier achieved an AUC of 0.74, demonstrating moderate discriminatory power. Additionally, the RBF SVM, Sigmoid SVM, and KNN models obtained AUC values of 0.75, 0.73, and 0.72, respectively, indicating moderate differentiation between the classes.

Conversely, the Decision Trees model achieved the lowest AUC of 0.63, suggesting a relatively weaker ability to discriminate between positive and negative instances compared to the other models.

* **Receiver Operating Characteristic (ROC) for Different Classifiers:**

A graph of different colored lines

Description automatically generated

1. **Results of the KMeans unsupervised Machine learning:**

In evaluating the outcomes of the KMeans unsupervised machine learning approach, I utilized the elbow method to identify the optimal number of clusters, which was determined to be k = 4. This indicates that four clusters could provide an effective segmentation for the dataset. Analysis of purity scores across various distance metrics, including Euclidean, squared Euclidean, Manhattan, Chebyshev, Canberra, and chi-square distances, consistently revealed a high purity score of 88.66%. This consistent score suggests that the clusters are relatively pure. However, upon visually inspecting the clusters, I noted a vertical spread with nearby clusters, indicating potential challenges in achieving clear separation. This observation suggests that the clusters might not be distinctly separate, and centroids may be positioned far from where data points are gathered, highlighting potential limitations in accurately capturing complex data patterns.

* **Elbow Method for Optimal K in Keans Clustering:**

A graph with a line and a line

Description automatically generated with medium confidence

* **Visualization of Clusters and Centroids:**

A white background with black text

Description automatically generated

* **Performance Metrics of Supervised Learning Models:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Cross-Validation (%)** | **Accuracy**  **(%)** | **Precision**  **(%)** | **Recall**  **(%)** | **F1-score**  **(%)** | **AUC** |
| **Decision Trees** | 87 | 84 | 83 | 84 | 84 | 0.63 |
| **K-Nearest Neighbors** | 90 | 89 | 89 | 89 | 85 | 0.72 |
| **Gaussian Naïve Bayes**  **(All predictors)** | 20 | 22 | 81 | 22 | 22 | 0.74 |
| **Gaussian Naïve Bayes**  **(Non-Significant       Predictors Removed)** | 83 | 86 | 81 | 81 | 83 | 0.73 |
| **Random Forest** | 91 | 91 | 91 | 91 | 88 | 0.79 |
| **Logistic Regression** | 91 | 90 | 89 | 90 | 88 | 0.81 |
| **Linear SVM** | 91 | 90 | 89 | 90 | 87 | 0.80 |
| **RBF SVM** | 89 | 88 | 77 | 88 | 82 | 0.75 |
| **Poly SVM** | 91 | 89 | 88 | 89 | 86 | 0.83 |
| **Sigmoid SVM** | 89 | 87 | 83 | 87 | 84 | 0.73 |

1. **Results of the Feature Importance Analysis:**

Both the decision trees model and the random forest classifier highlighted "Curricular units 2nd sem (approved)" and "Curricular units 1st sem (grade)" as the most influential variables, with importance scores of 0.17 and 0.12, respectively. Conversely, other features such as 'Curricular units 2nd sem (enrolled)', 'Daytime/evening attendance', 'Previous qualification', 'Nationality', 'Educational special needs', 'Curricular units 2nd sem (credited)', 'Curricular units 1st sem (without evaluations)', 'Debtor', 'Curricular units 1st sem (enrolled)', 'Curricular units 1st sem (credited)', 'International', and 'Marital status' were deemed to have negligible importance, each scoring 0. These analyses aimed to identify the most influential variables impacting student outcomes, underscoring the significance of "Curricular units 2nd sem (approved)" and "Curricular units 1st sem (grade)" as crucial predictors of graduation.

1. **Discussion:**

In my study, I've uncovered that a student's likelihood of dropping out or graduating can be foreseen by examining various factors such as their academic journey, demographic background, socioeconomic status, and performance at the end of each semester. Interestingly, I found that elements like "Curricular units 2nd sem (approved)" play a pivotal role in determining the probability of graduation, emphasizing the significance of academic progression, particularly in the final stages of education. By "Curricular units 2nd sem (approved)," I mean the number of courses or subjects that a student completed and received approval for during the second semester. Essentially, it indicates how well a student is progressing academically during this period. My machine learning results suggest that students who complete more curricular units in the second semester are more likely to graduate. This finding highlights the importance of maintaining a strong academic performance, particularly in the later stages of a student's college journey. It implies that tracking and supporting students' progress in their coursework during critical periods like the second semester can be valuable in predicting and promoting college graduation.

Moving forward, let's delve into the performance of the predictive models I employed. We observed out-of-sample accuracies ranging up to an impressive 91%. Notably, the Random Forest classifier, Logistic Regression classifier, and Linear SVM stood out with accuracies of 91%, 90%, and 90%, respectively. However, the Poly SVM stole the show with the highest ROC-AUC performance of 0.83. This metric indicates its exceptional ability to differentiate between dropout and graduation classes. Conversely, the decision trees model exhibited a more moderate performance, clocking in with an accuracy of 84% and the lowest AUC of 0.63, implying some difficulty in accurately discerning between the two classes.

Another noteworthy finding emerged with the GNB model. Initially, it yielded a disappointing accuracy of 22%. However, upon removing non-significant predictors, the accuracy dramatically improved to 86%. This improvement can be attributed to addressing the naive assumption of the Naïve Bayes model, which operates under the premise that features are independent given the class label. When this assumption is violated, such as in cases where features are correlated or dependent, the model's performance can suffer.

Moving beyond individual model performance, our exploration revealed critical insights into the importance of academic path, demographics, and socioeconomic factors in predicting student outcomes. However, it's important to acknowledge the challenges encountered in our unsupervised machine-learning endeavors. These challenges, including the dataset's complexity, multidimensionality, and potential overlap among data points, pose significant hurdles. Factors such as the choice of distance metric, initialization of cluster centroids, and the presence of outliers or noise all contribute to these complexities.

Regarding deployment strategy, I opted to thoroughly test multiple machine learning models before selecting the most appropriate one for online deployment. Ultimately, the Poly SVM emerged as the top choice due to its superior ability to distinguish between dropout and graduation cases.

1. **Limitations and future research**

Our study identifies several limitations and suggests avenues for future research. Firstly, it's crucial to acknowledge the inherent limitations of our project. The dataset's size, comprising 4424 instances, may constrain the complexity and reliability of the developed machine learning models. While our analysis aims to offer valuable insights, a larger and more diverse dataset could lead to more robust and generalizable results. Future research endeavors should prioritize the inclusion of larger and more varied datasets.

Secondly, our dataset originates from Portugal, thereby limiting the generalizability of our findings to other countries with differing cultural backgrounds and educational systems, particularly the U.S. Future investigations should assess the extent to which our models can be applied across various cultural contexts and identify universally valid features of student dropout across diverse settings.

Lastly, accurately predicting student dropouts and academic success presents challenges due to inherent uncertainties. The dataset may not encompass all factors influencing student performance, such as personal issues or mental health challenges like anxiety or depression, which can significantly impact a student's academic engagement and outcomes. Subsequent research should delve into these nuanced aspects to enhance the predictive accuracy of dropout models.

1. **Conclusion:**

In conclusion, this study represents a comprehensive exploration into the multifaceted predictors of academic success among college students, employing advanced machine learning methodologies and rigorous analytical techniques. Through meticulous data preprocessing, model training, and evaluation, we have unearthed valuable insights into the factors shaping student outcomes, shedding light on the complexities inherent in predicting dropout and graduation. Our findings underscore the pivotal role of academic progression, demographic background, and socioeconomic factors in determining student success. While certain models, such as the Random Forest classifier and Logistic Regression, demonstrated robust performance, challenges persist, particularly in unsupervised machine learning endeavors. Despite these limitations, our research is a foundation for future investigations to enhance predictive accuracy and generalizability across diverse cultural contexts and educational systems. This study strives to inform evidence-based strategies that empower educational stakeholders to foster student retention, mitigate dropout rates, and ultimately cultivate a culture of academic excellence by bridging the gap between theory and practice.

Here are links to my notebooks for further exploration:

Notebook1**:**<https://colab.research.google.com/drive/1B8BkEA95mlHTrI_NmbMEicKY-TOKZ0R5?usp=sharing>

Notebook2**:**<https://colab.research.google.com/drive/1fQV1DMpYfxCK1pMAoQ8T2NeTHwMUijvU?usp=sharing>

Additionally, you can explore my deployed model through the following link:

<https://predict-student-dropout.onrender.com/>

# Bibliography

Andreas C. Müller, Sarah Guido. 2016. *Introduction to Machine Learning\_A Guide for Data Scientists.* United States of America: O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Claudio Brasca, Nikhil Kaithwal, Charag Krishnan, Monatrice Lam, Jonathan Law, and Varun Marya. 2022. *Using machine learning to improve student success in higher education.* April 07. https://www.mckinsey.com/industries/education/our-insights/using-machine-learning-to-improve-student-success-in-higher-education.

Elaine Brzycki, Henry Brzycki. 2016. *Student Success in Higher Education: Developing the Whole Person through High-Impact Practices.* ‎ BG Publishing.

GeeksforGeeks. 2023. *Decision Tree.* Augest 20. Accessed Augest 20, 2023. https://www.geeksforgeeks.org/decision-tree/.

—. 2023. *Support Vector Machine (SVM) Algorithm.* Jun 10. https://www.geeksforgeeks.org/support-vector-machine-algorithm/.

George D. Kuh, Jillian Kinzie, John H. Schuh, Elizabeth J. Whitt. 2005. *Student Success in College: Creating Conditions That Matter.* Jossey-Bass.

Helen Crompton, Diane Burke. 2023. "Artificial intelligence in higher education: the state of the field." *International Journal of Educational Technology in Higher Education* 22.

Hussain S, Khan MQ. 2021. *Student-Performulator: Predicting Students’ Academic Performance at Secondary and Intermediate Level Using Machine Learning.* Jun 03. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8173503/.

João Gabriel Corrêa Krüger, Alceu de Souza Britto Jr., Jean Paul Barddal. 2023. *An explainable machine learning approach for student dropout prediction.* December 15. https://www.sciencedirect.com/science/article/abs/pii/S0957417423014355.

Joshi, Naveen. 2022. "Understanding Education 4.0: The Machine Learning-Driven Future Of Learning." *Forbes.* https://www.forbes.com/sites/naveenjoshi/2022/03/31/understanding-education-40-the-machine-learning-driven-future-of-learning/?sh=47a2f4055bc2.

KATHERINE SCHAEFFER. 2022. *10 facts about today’s college graduates.* April 12. https://www.pewresearch.org/short-reads/2022/04/12/10-facts-about-todays-college-graduates/.

Lasse Rouhiainen. 2019. *How AI and Data Could Personalize Higher Education.* October 14. https://hbr.org/2019/10/how-ai-and-data-could-personalize-higher-education.

Matz, S.C, Bukow, C.S, Peters, H. et al. 2023. *Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics.* April 07. https://www.nature.com/articles/s41598-023-32484-w.

Mduma, Neema. 2023. *Data Balancing Techniques for Predicting Student Dropout Using Machine Learning.* February 27. https://www.mdpi.com/2306-5729/8/3/49.

National Student Clearinghouse Research Center. 2023. *Completing College.* November 30. https://nscresearchcenter.org/completing-college/.

—. 2023. *Undergraduate Degree Earners.* March 16. https://nscresearchcenter.org/undergraduate-degree-earners/.

Neema Mduma, Khamisi Kalegele, Dina Machuve. 2019. "A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction." *DATA SCIENCE JOURNAL* 14.

Peter Bruce, Andrew Bruce. 2017. *Practical statistics for data scientists\_ 50 essential concepts.* United States of America: O’Reilly Media, Inc., 1005 Gravenstein Highway North,.

Raghul Gandhi Venkatesan, Dhivya Karmegam, Bagavandas Mappillairaju. 2023. "Exploring statistical approaches for predicting student dropout in education: a systematic review and meta-analysis." *Springer Link.* November 29. https://link.springer.com/article/10.1007/s42001-023-00231-w.

Sean Burns. 2021. *Machine Learning's Growing Role in Research.* March 29. https://www.educause.edu/ecar/research-publications/2021/machine-learnings-growing-role-in-research/introduction-and-key-findings.

Yağcı, M. 2022. *Educational data mining: prediction of students' academic performance using machine learning algorithms.* March 03. https://slejournal.springeropen.com/articles/10.1186/s40561-022-00192-z.

Yu-Sheng Su, Yu-Da Lin, Tai-Quan Liu. 2022. "Applying machine learning technologies to explore students’ learning features and performance prediction." *frontiers.* December 21. https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.1018005/full.