Predicting Student Dropout Rates Using Machine Learning Techniques

Course: CIND 820
For: Professor Tamer Abdou
By: Tiffany Okotako

Table of Contents

Predicting Student Dropout Rates Using Machine Learning Techniques

Introduction

As university tuitions increase and governments spend increasingly more on post-secondary eduction combined with the growing demand for educational attainment, it is imperative that resources are spent efficiently to have a lasting positive effect on society as a whole. A drawback to increased spending on post-secondary education is the student dropout rate. When students fail to complete their studies, they take away a space from someone who would have completed their education and they also deplete the financial resources of the government, any scholarships they would have received, and any financial investment made by themselves or their parents. Overall, students dropping out of post-secondary education is something that researchers, policy makers, and university personnel can strive to improve.

This study utilizes a dataset from researchers in Portugal to assess the efficacy of machine learning techniques in predicting dropout rate. The dataset consists of university enrollment data from the 2008/2009 school year to the 2018/2019 school year. Within their dataset, the researchers delineate dropouts as any student that changes courses, schools, or stops their studies completely (Realinho et al, 2018). Due to their definition of dropouts any student that changed their major or enrolled in a different school within the timeframe of the dataset is identified as a dropout significantly increasing the dropout rate (Realinho et al, 2018). The OECD has celebrated Portugal for their historic improvements in “access, attainment and performance” within their educational system in the 20 years preceding 2018 (Liebowitz, et al, 2018). Between 2005 and 2015, the proportion of Portuguese students who graduated from secondary school increased from 50% to 80% accompanied by significant improvements in reading, science, and math (Liebowitz, et al, 2018).

The data that will be used can be found on the UC Irvine Machine Learning Repository as dataset 697 under the name: “Predict Students' Dropout and Academic Success”. This dataset has 37 variables, one of which is called Target which is the indicator for if someone is still enrolled, dropout, or graduated from their studies. There are:

- four variables about the student’s background: age, nationality, gender, and marital status,
- three variable abouts the student’s academic standing: admission grade, previous qualification, previous qualification (grade)
- three variables about the course and application method: application mode, application order, course
- four variables for father’s and mother’s educational attainment and occupation,
- six binary (yes/no) variables: displaced, educational special needs, debtor, tuition fees up to date, scholarship holder, and international. One variable for day or night classes.
- twelve variables for educational status and results in the first and second semester,
- three variables about the economic climate: unemployment rate, inflation rate, and GDP.

Questions

This paper will answer four questions related to the dataset:

1. Is the Apriori method or the Decision Tree method most helpful in classifying dropouts?

2. Is the Random Forest method more accurate than logistic regression in predicting dropouts?
3. Overall are classification or machine learning techniques more accurate in assessing the data?
4. What are the characteristics most common amongst those who dropout?

Methodology

The paper will start with an introduction to the dataset and the defining characteristics utilizing graphs. Next three classification techniques will be assessed: Decision Tree, Apriori Algorithim and K Nearest Means. Then three machine learning techniques will be assessed: Random Forest, logistic regression, and Naïve Bayes. Finally, a conclusion will be given.

Preliminary Information

Prior to analyzing classification and machine learning techniques we will look at the current breakdown of the dropout rate by different characteristics. There were 35 variables and 4424 rows in the dataset. The demographic variables and data relating to the student's status in school such as scholarship holder or day/nighttime class attendance will be looked at in detail.

Gender

Dropout rate is highest amongst males who have a 45.1% dropout rate while the dropout rate amongst females is only 25.1%. Likewise, the graduation rate for males is lower with only 35.2% of the male students listed as graduates while 57.9% of the female students were graduates.

Marital Status

Only 4 of the 4424 students were widowers making up 0.09% of all students within the dataset meanwhile 88.58% of students were single. Only one widow was a dropout and only one widow was a graduate with a 25% rate for both dropout and graduate. Due to the size of this category, the results for widowers should not be extrapolated. Single students had the second lowest dropout rate with 30.2% being dropouts and 51.4% being graduates. Both married and divorced students had similar dropout percentages at 47% however, 39.4% of married students were graduates compared to only 35.1% of divorced students being graduates.

Course

The subject with the highest dropout rate is Biofuel Production Technologies with a dropout rate of 66 %. However, only 12 students of the 4424 are enrolled in this course. Within this group, 77.78% of dropouts were male and there were 0 male graduates whereas 33.33% of graduates and dropouts were female. The course with the lowest dropout rate was Nursing. There were 617 female students in nursing making up 80.5% of the students within the course. The female dropout rate was only 13.13% which is the lowest dropout rate of any course, and the total graduation rate for all nursing student was 71.54% while the female nursing graduation rate was 59.06% which is the highest graduation rate of any course. In general, courses that had more males than females such as Informatics Engineering and Agronomy had a high total dropout rate, 54.11% and 40.95% respectively. Informatics Engineering is the only subject where the male dropout rate was higher than the female dropout rate with 52.76% of males dropping out and 85.71% of females dropping out. Likewise, only 8.6% of males graduated from this course whereas 0% of females graduated. Courses with more females than males such as Social Service (evening attendance), Veterinary Nursing, and Social Service had lower dropout rates. These courses are all in the bottom six when ranked by dropout rate. The only outlier is Basic Education with 9 males and 183 females enrolled in the program however, a 44.27% total dropout rate. There could be other factors that explain this anomaly such as low willingness to continue in a general program. It is important to note that if students switch from Basic Education to a specific major, for example nursing, they would be considered as a dropout due to how the authors defined the term. Basic education is the only Course whereby the dropout rate for males (44.44%) and the dropout rate for females (44.26%) has less than 1% difference between them.

Age Group

Of the 4424 students, 2551 or 57.66% of them were between the ages of 17 to 20. This age group had the lowest dropout rate of 21.24% and the highest graduate rate of 60.29%. The highest dropout rate was amongst those aged 27 to 30 with 289 students in this category. The dropout rate was 61.25%. The second highest dropout rate was for those aged 31 to 39 with 414 students in this category and a dropout rate of 55.07%. The dropout rate of students 40 and above was 51.21% which is similar to the dropout rate of those aged 24-26 who had a dropout rate of 48.74%.

Nationality

By nationality, there was only 1 Turkish student so Turkish had a 100% enrolled rate. Eastern Europeans had the highest dropout rate at 45.5% and Latin Americans had the second highest dropout rate at 38.1%. The lowest dropout rate was Africans with a 19.4% dropout rate and 58.3% graduation rate and Western Europeans were second with a 20.0% dropout rate and 55% graduation rate. The dropout rate for Portuguese students was 32.2% with a 50% graduation rate.

International vs Non-International Students

Everyone who was not Portuguese was considered an international student. Overall, the results and statistics were nearly identical between both groups. The dropout rate for international students was 29.1% with a 49.1% graduation rate and the dropout rate for Portuguese student was higher at 32.2% with a greater graduation rate of 50.0%.

Financial Considerations: Scholarships, Debtors, Tuition-Up To Date

The dropout rate was 38.7% for those who did not have a scholarship and only 12.2% for those who did have a scholarship. For those who were considered “debtors” the dropout rate was 62.0% with a graduation rate of 20.1%, however, the dropout rate for non-debtors was 28.3% with a graduation rate of 53.8%. The dropout rate of those whose tuition was **not** up to date was 86.6% with a graduation rate of only 5.4%. The dropout rate of those whose tuition was up to date was 24.7% with a graduation rate of 56.0%.

Daytime vs Night Attendance

Day and night-time class attendance had stark differences in dropout and graduation rate. For those who took night classes 42.9% dropped out and 41.6% graduated. For those enrolled in day classes, only 30.8% dropped out and 51.0% graduated.

Educational Special Needs

The dropout rate for those who did and did not have special needs was nearly identical. Those **without** special needs had a 32.1% dropout rate and 50% graduation rate while those who **did have special needs** had a 33.3% dropout rate and 45.1% graduation rate.

Displaced

The dropout rate for those who were displaced was 27.6% with a graduation rate of 54.6% and the dropout rate for those who were not displaced was greater at 37.6% with a graduation rate of 44.3%.

Classification Methods

Decision Tree

The first classification method to be analyzed is the Decision Tree (DT) method. A Decision Tree is a supervised learning type of classification technique that attempts to mimic the decision-making process. With this type of classification technique, the program is taught the correct class label for a given input and once the machine learns the correct input it can be used to predict future inputs. The first step to working with the data was classifying the variables. There were 6 variables for marital status, 21 for nationality, 18 for application mode, 17 different courses, 17 variables for previous qualifications and 34 variables for mother's and father's qualifications, and 46 variables for mother's and father's occupation. These variables were put into groups based on similarity to make the analysis process easier. Additionally, there are 8 binary variables codes as either 1 or 0, one of which was gender that was transformed to be male or female, another was daytime or evening attendance which was transformed to be daytime or evening and finally six variables: displaced, educational special needs, debtor, tuition fees up to date, scholarship holder, and international that were transformed to be yes or no. For the age variable, ages were categorized into groups. Additionally, the application order variable was categorized into top three, 4th -6th choice and 7th to 9th choice likewise admission grade was placed into five groups based on percentile. The data was one-hot encoded, and a Decision Tree model was made based on the Sci-kitlearn package in Python. The data was split into two sets, 80% which was used for training and 20% which was used for testing. After the Decision Tree was created, it was reduced to 30% of its original size. Then it was pruned to have a maximum depth of 5 which refers to the number of levels beneath the initial node. This is to make the results clearer to see. The root node is chosen as the feature that is most able to categorize the data. In this case it was tuition fees up to date. In the preliminary graphs, it was determined that 86.6% of students who did not have their tuition fees up to date were dropouts, making this feature the most stratifying. The Gini index is given for each node. The Gini index varies from 0 to 1 and is a determiner of how similar the elements in a node are to one another. The Gini impurity index "measures how often a randomly chosen element of a set would be incorrectly labeled if it were labeled randomly and independently according to the distribution of labels in the set" (Decision tree learning, Wikipedia). A Gini score of 0 means that all datapoints belong to the same class and the node is a leaf node. The colors on the tree are also significant the darker the colour the closer the Gini score is to 0 and the lighter the colour the closer the Gini score is to 1. Purple was for Enrolled, Orange was for Dropouts, and Green was for Graduates.

The next level within the Decision Tree was non-scholarship holder. On the left-hand side, was those who **did not** have tuition up to date and did not hold a scholarship and on the right-hand side was student who **did** have tuition up to date but also did not hold a scholarship. As viewed in the preliminary data, scholarship holders had a 12.2% dropout rate contrasted with the 38.7% dropout rate for non-scholarship holders.

On the left-hand side of the tree at the second node for non-scholarship holders, the Gini score was 0.242 for Dropouts meaning that members of this group, who did not have their tuition up to date and were not scholarship holders were more likely to be dropouts. On the right-hand side where tuition was up to date, but the students were also not scholarship holders the Gini score was 0.588 meaning that on average 58.8% of the time a student with these two characteristics would be incorrectly mislabeled.

At the third level on the left-hand side of the DT, there were two categories, Mother's Occupation being an "Intermediate Profession" and being in Veterinary studies both of which were more likely to predict that someone was a dropout. On the right-hand side the third level had two indicators: Male and age group 17-20 which were more indicative of being enrolled in studies.

Accuracy is the proportion of correctly classified instances (both true positives and true negatives) among the total instances in the test dataset. It gives an overall measure of the classifier's performance across all classes. In this case, the overall accuracy is 0.63, meaning that the classifier correctly classified 63% of the instances in the test dataset which is not that great.

Precision is the ratio of correctly predicted positive observations to the total predicted positives. In this context, it represents the accuracy of the classifier in predicting each class. Precision for the Dropout class is 0.67, which means that out of all the instances predicted as Dropout, 67% were correctly predicted as Dropout. Precision for the Graduate class is 0.75, indicating that out of all the instances predicted as Graduate, 75% were correctly predicted as Graduate. Precision for the Enrolled class is 0.61, meaning that out of all the instances predicted as Enrolled, 61% were correctly predicted as Enrolled.

Recall, also known as sensitivity, is the ratio of correctly predicted positive observations to the actual positives. In this context, it represents the classifier's ability to correctly identify each class. Recall for the Dropout class is 0.60, indicating that out of all the actual Dropout instances, 60% were correctly identified as Dropout. Recall for the Graduate class is 0.02, meaning that only 2% of the actual Graduate instances were correctly identified as Graduate. This is an area of concern. Recall for the Enrolled class is 0.88, suggesting that 88% of the actual Enrolled instances were correctly identified as Enrolled.

The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall, especially when there is an imbalance between the number of instances in different classes. F1-score for each class is calculated as: 2 * (precision * recall) / (precision + recall). The F1-score for the Dropout class is 0.63, indicating a balance between precision and recall for this class. The F1-score for the Graduate class is 0.04, which is very low due to the low recall for this class. The F1-score for the Enrolled class is 0.72, indicating a good balance between precision and recall for this class.

Support: Support is the number of actual occurrences of each class in the test dataset. It provides insight into the distribution of classes in the dataset. Within this dataset, there were 316 instances of Dropout, 151 instances of Graduate, and 418 instances of Enrolled. Out of the 4424 instances only 151 were Graduates which translates to 3.4% which is why the prediction for Graduates was abysmal. Likewise, the precision and F1 score which depend on the algorithim' s ability to predict graduate classes was also very low.

Apriori Algorithim

The second technique to be analyzed is the Apriori Algorithim which is another classification technique. To make the comparison between the Decision Tree and Apriori fair, the exact same variable categorization was used for both. Due to the low instances of graduate a separate dataset was analyzed that only contained Enrolled and Dropouts. Despite the categorization and grouping of variables the support was exceptionally low even with choosing a low minimum support. The support chosen was 0.2. The results are as follows: For Dropouts, the result was 0.225091 for No and for Enrolled the result was 0.209109 for No. For Graduates due to the limited data the algorithm was not able to generate a data frame. The support of an itemset in an association rule mining represent the proportion of transactions in the dataset that contain that item. It can also be interpreted as the frequency of a dataset. With this data, the Apriori algorithm was not able to produce a tangible dataset of items that frequently go together. For both dropout and enrolled, it was determined that "No" appeared in approximately 20 % of the time in each itemset. A support of 0 would mean that the itemset does not appear in any transaction whereas a support of 1 means that an itemset appears in every transaction in the dataset. Even when the code was modified to parse strings, the support did not increase past 20%.

K Nearest Neighbours

The next classification technique used was the K Nearest Neighbours or KNN classification technique. K-nearest neighbours is a non-parametric supervised learning classifier that utilizes proximity to make classifications or predictions about the grouping of an individual data point (IBM, n.d.). This machine learning technique is based off the assumption that similar points can be found near one another. The KNN algorithm strives to predict the correct class for a data point by calculating the distance between a specified datapoint and the other training points. In the code, 80% of the data was used for training and 20% was used for testing. For this dataset, one to seven neighbours were assessed and the greatest accuracy of 60.90% was achieved with 5 neighbours. Surprisingly 3 and 7 neighbours both yielded an accuracy of 59% whereas 4 neighbours had a lower accuracy at 58%. Accuracy is the ratio of the correctly predicted observations to the total observations. Overall accuracy of the classifier is 61%, meaning 61% of the instances were correctly classified across all classes. The overall accuracy is brought down due to the issues with placing the enrolled class.

The precision is the ratio of correctly predicted positive observations to the total predicted positives. Dropouts had the highest precision at 66% meaning that of all the values predicted as dropout, 66% of them were dropouts. Precision is usually interpreted as a quality measure and higher precision means that an algorithm returns more relevant results than irrelevant results

(Wikipedia, Precision and Recall). Enrolled had a precision of 36% meaning that for all the instances predicted as enrolled, only 36% of them were actually enrolled.
Recall is the ratio of the correctly predicted positive observations to the total number of relevant instances. Recall is generally interpreted as the sensitivity measure or a measure of the completeness of the positive predictions. The recall was the highest with the graduate class with a score of 77% meaning that within this class, 77% were correctly classified. The recall was extremely low for the enrolled class with a score of only 26% meaning that the algorithm failed to return most of the relevant results.

Machine Learning Techniques

Random Forest

The first machine learning technique analyzed was the Random Forest technique. The Random Forest is a machine learning technique that combines the output of multiple Decision Trees to reach a single result (IBM, n.d.). The algorithm makes predictions and selects the best solution by means of voting (Bhushan et al, 2024). Unlike a Decision Tree that considers all possible outcomes and maps them with branches, the Random Forest technique only selects a subset of those features enabling the prediction techniques to be more precise. With the Random Forest algorithm, the dataset was split into 80% training and 20% testing. The overall accuracy achieved was 75 % making it the most accurate machine learning model used to analyze the data. The precision score was the highest for the dropout class with a precision of 85% and the lowest for the enrolled class with a precision of 48% meaning that for all the instances that were predicted as enrolled only 48% of them were actually enrolled and for all the instances predicted as graduates 85% of them were actually graduates. The recall was greatest for the graduate class with a value of 95% and lowest for the enrolled class with a value of 10%. This is an enormous range and signifies something deeper with regards to how the dataset was interpreted and analyzed. This means that the algorithm failed to return relevant results for the enrolled class but was able to retrieve relevant results for the graduate class.

Logistic Regression

The next machine learning technique used was the Logistic Regression technique. Logistic regression is a linear supervised learning classifier mainly used for binary classification, but it can handle more than two cases. Logistic regression is generally used to analyze datasets with one or more independent variables by utilizing a logistic function to model the probability of a datapoint belonging to a specific class (Said et al, 2023). Logistic regression can be understood as modelling the probability of an outcome, in this case the target variable, based on inputs, which would be the other 34 variables in the dataset. Logistic regression is a type of supervised machine learning that is able to classify objects. Because our target variable had three outcomes, multinomial logistic regression was used. The overall results are as follows: the accuracy of the program was 69.99% meaning almost 70% of the instances were correctly classified by the algorithm. Precision is the proportion of true positives out of all the positive predictions and can be best understood by the formula Precision = True Positives / (True Positives + False Positives). The precision for the dropout class was 80% but only 50% for the enrolled class meaning that the algorithm made a lot of false positive predictions for the enrolled class. Recall can be interpreted as the ability of the model to capture all positive instances in the dataset. A high recall means that the algorithm can successfully identify most of the positive instances and minimize false negatives. The recall for the graduate class was 95% which is remarkably high however the recall for the enrolled class was only 10% confirming our knowledge that the algorithm was not able to identify the enrolled class very effectively. Logistic regression can be sensitive to outliers which can explain for the loss of accuracy within the model (Said et al, 2023).

Naïve Bayes

The last machine learning technique utilized was the naïve bayes algorithm. This was done in two ways. First was a regular Gaussian Naïve Bayes which is a classification technique based on the probabilistic approach of a Gaussian distribution that "assumes conditional independence between every pair of features given in the value of the class variable" (Scikit-learn, n.d.). The second method utilized was a cross-validation Gaussian Naïve Bayes technique. The cross-validation method used was a 5-fold cross validation. The goal of cross-validation is to build an estimator against different cross sections of the data to develop an aggregate understanding of the algorithm's performance across all sections. The overall goal with cross-validation is to choose a model with an unbiased split. Unfortunately, the cross-validation Gaussian Naïve Bayes only outperformed the regular Naïve Bayes in two areas: for the graduate class, the cross-validation had a higher precision (73% compared with 71%) and a higher F1 score (79% vs 78%). For all other metrics the regular Gaussian Naïve Bayes was better than the cross-validation Gaussian Naïve Bayes. The regular Gaussian Naïve Bayes will be discussed. The overall accuracy was 69.94% meaning that the algorithim was able to correctly classify a point almost 70% of the time. The precision for the dropout class was 81% compared to 37% for enrolled and 71% for the graduate class. Greater precision means that the algorithim returns more relevant results than irrelevant results. This was only true for the graduate class. The recall was highest for the graduate class (86%), moderate for the dropout class (69%) and abysmal for the enrolled class (26%). Recall refers to the algorithms ability to return relevant results meaning that the algorithim struggled the most with students who were still enrolled. It is probable that the Gaussian Naïve Bayes would perform better on a binomial classification instead of a multinomial classification. The F1 score refers to the balance between both the precision and recall and it was the lowest for the enrolled class with an F1 score of 31%. The F1 score of the dropout class was 74% and the graduate class was 78%. This indicates a good balance between precision and recall for both the graduate and dropout class.

Questions

The first question this paper will answer is: which classification technique is better, the Decision Tree or Apriori Algorithm? Although the accuracy of the Decision Tree was only 63%, the low score can be attributed to the inclusion of the "Graduate" category which had very few instances. It was much better at classifying both Enrolled and Dropouts within the dataset. On the other hand, the Apriori algorithm was not able to generate a support greater than 22%. Overall, I would say the Decision Tree classification method is more accurate and performed better on this dataset.

Unlike the Decision Tree that had the lowest precision with the Graduate class, the KNN algorithm has the lowest precision and recall with the enrolled class and a lower overall accuracy. Of the three classification techniques utilized, the Decision Tree performed the best. The KNN method was not as accurate or precise as the Decision Tree classification technique and the Apriori Algorithm performed the worst. For the data that was analyzed, the Decision Tree classifier is a robust method which had a higher accuracy than Apriori Algorithm and KNN while also providing the visual aspect which makes it easier to see how the classification process works.

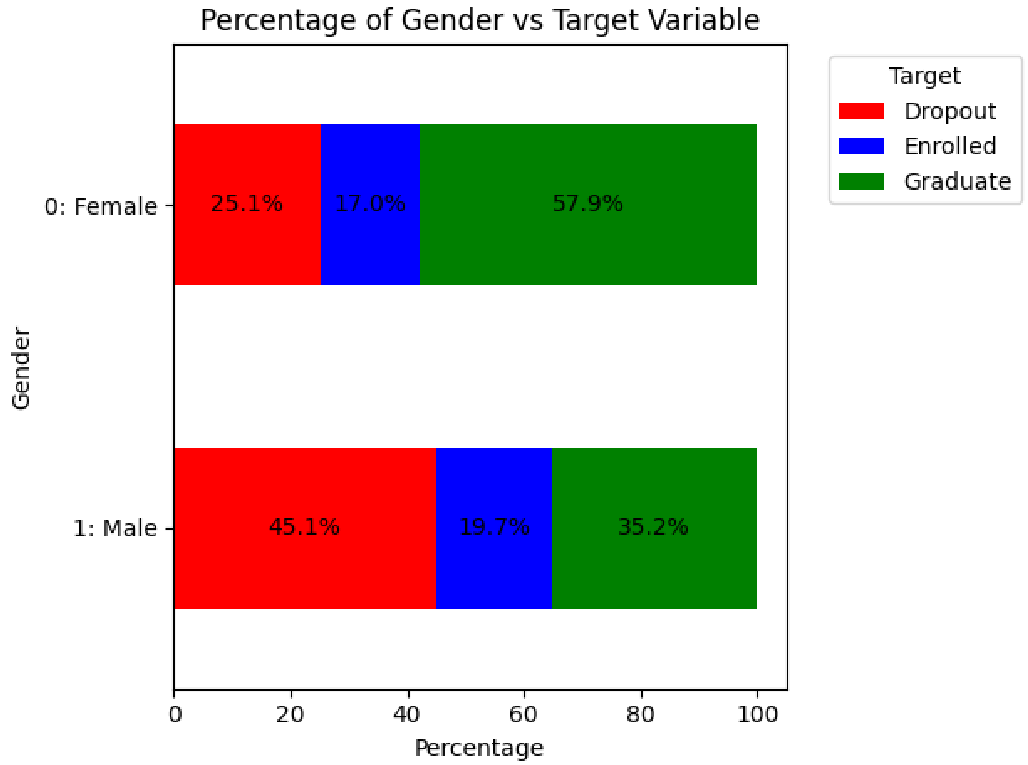
The second question to be answered in this paper is which method is more accurate in predicting dropouts: the Random Forest method or logistic regression. Based on the accuracy scores of the two methods, overall, the Random Forest method was most accurate with an accuracy score of 75.8% whereas the logistic regression had an accuracy score of 69.9%. For the Random Forest method, the precision and recall for the dropout class was 85% and 76% respectively whereas the precision and recall for the logistic regression was 80% and 66% respectively. That means that the Random Forest method outperformed the logistic regression method in both correctly identifying dropouts and retrieving relevant results. The Gaussian Naïve Bayes was slightly more precise in classification of graduates and dropouts however did not perform better in the remaining categories. The Cross-Validation Gaussian Naïve Bayes was worse than the logistic regression in every way.

The third question to be answered is whether classification or machine learning algorithms performed better with this dataset. Within the classification category, the winner was the Decision Tree method and within the machine learning category, the winner was the Random Forest method. However, there are two areas where the Logistic Regression method outperformed the Random Forest method which are: precision for the enrolled class and recall for the graduate class. When comparing the Random Forest method and the Decision Tree Method, there are two areas where the Decision Tree performed better than the Random Forest method: the precision and recall score for the enrolled class were higher. However, the recall for the graduate class was significantly lower. The Decision Tree recall value for the graduate class was 2% compared to the value of 92% for the Random Forest class. There was high variation in performance of the Decision Tree algorithim when compared to the Random Forest. Overall, based on the techniques analyzed, the Random Forest machine learning algorithm outperformed the Decision Tree classification model. The Random Forest machine learning algorithm was more consistent with its prediction of classes compared to the large jumps and discrepancies witnessed with the Decision Tree classification techniques.

The last question to be answered is what characteristics are most similar amongst those who dropout. The first aspect seemed to be financial, among the dropouts in the dataset, 86.6% of them did not have their tuition up to date which suggest some type of financial problems or issues. Likewise, of those who did not have a scholarship, 38.7% of them were dropouts compared with 12.2% of dropouts among those who did have a scholarship. Similarly, of those who were debtors, 62% were dropouts compared with 28.3% of dropouts of those who were not debtors. The next determining factors were gender and subject. 45.1% of males were dropouts which is a significant compared to only 25.1% of females being dropouts. It could be in part due to the definition of dropout potentially if males were more likely to switch subjects or majors compared with their female counterparts, but this dataset does not contain enough information to answer such a question. Certain majors like Biofuel Production Technologies, Equiniculture, and Informatic Engineering had very high dropout rates however the results are slightly skewed due to low attendance within these programs. Even within the Decision Tree, the background information about parents, their educational level and their career did not show up that much and was only relevant in the third and fourth nodes of the Decision Tree.

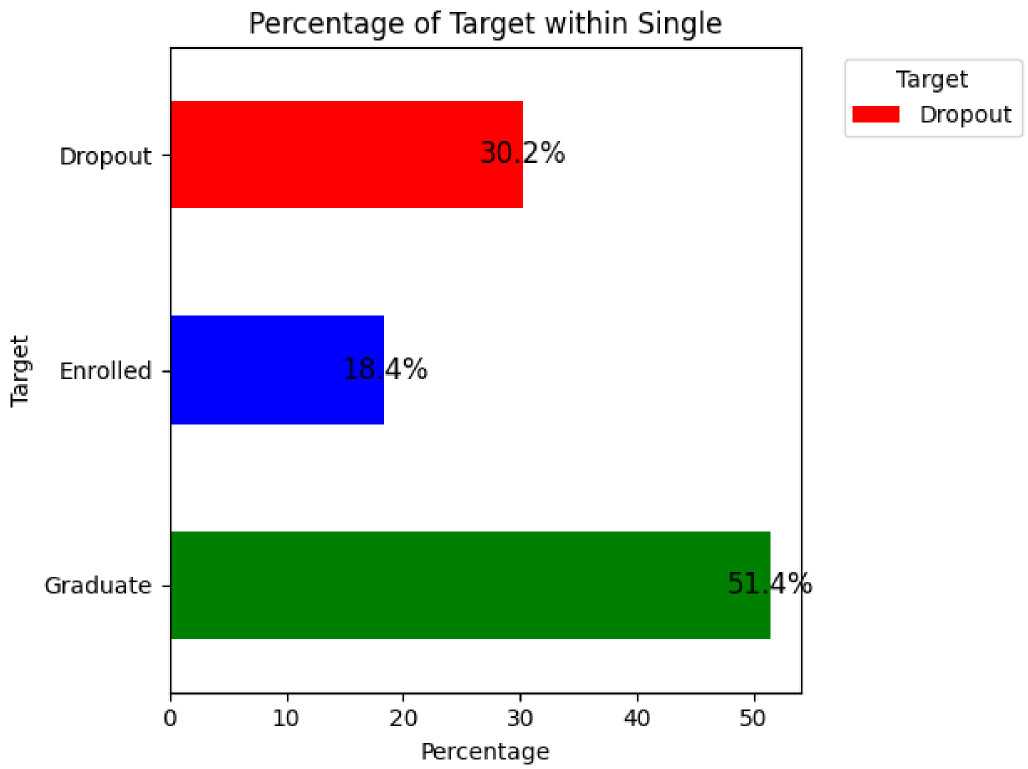
Graphs

1a- Gender vs Dropout Rate

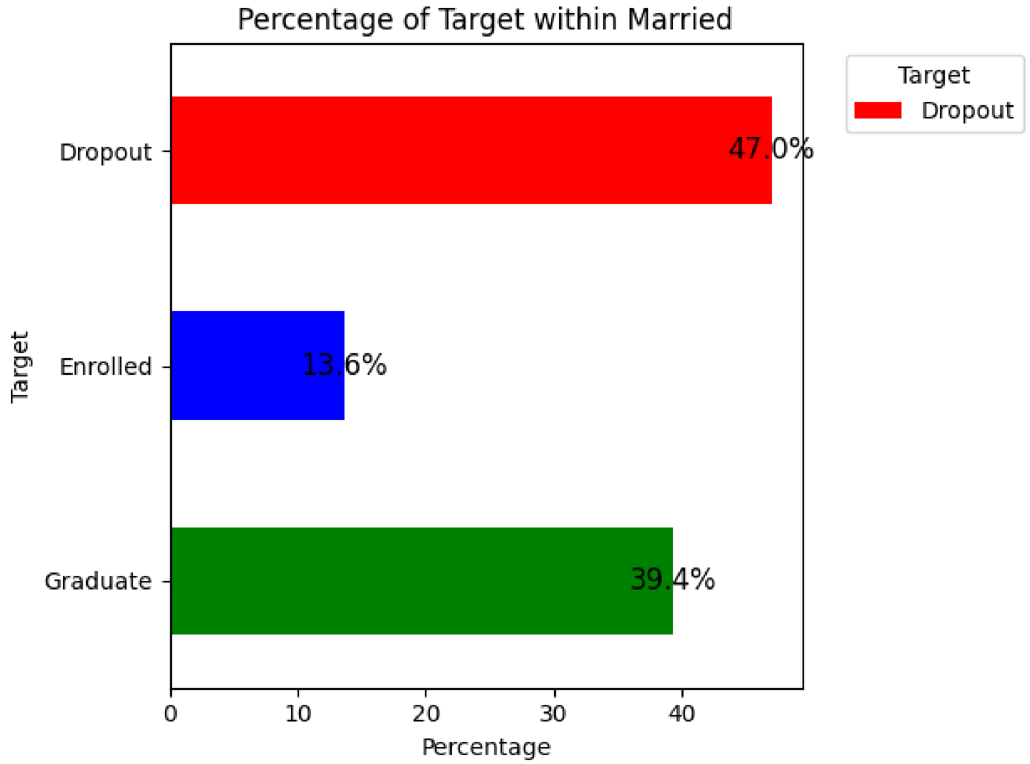


Marital Status vs Dropout Rates

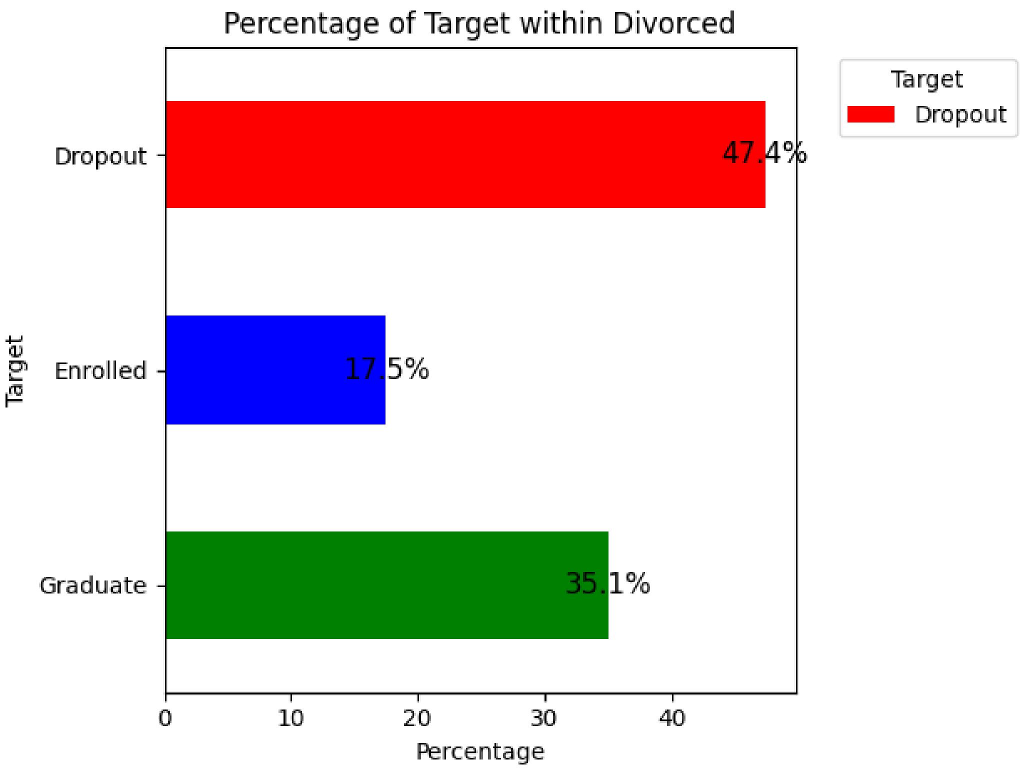
2a Single



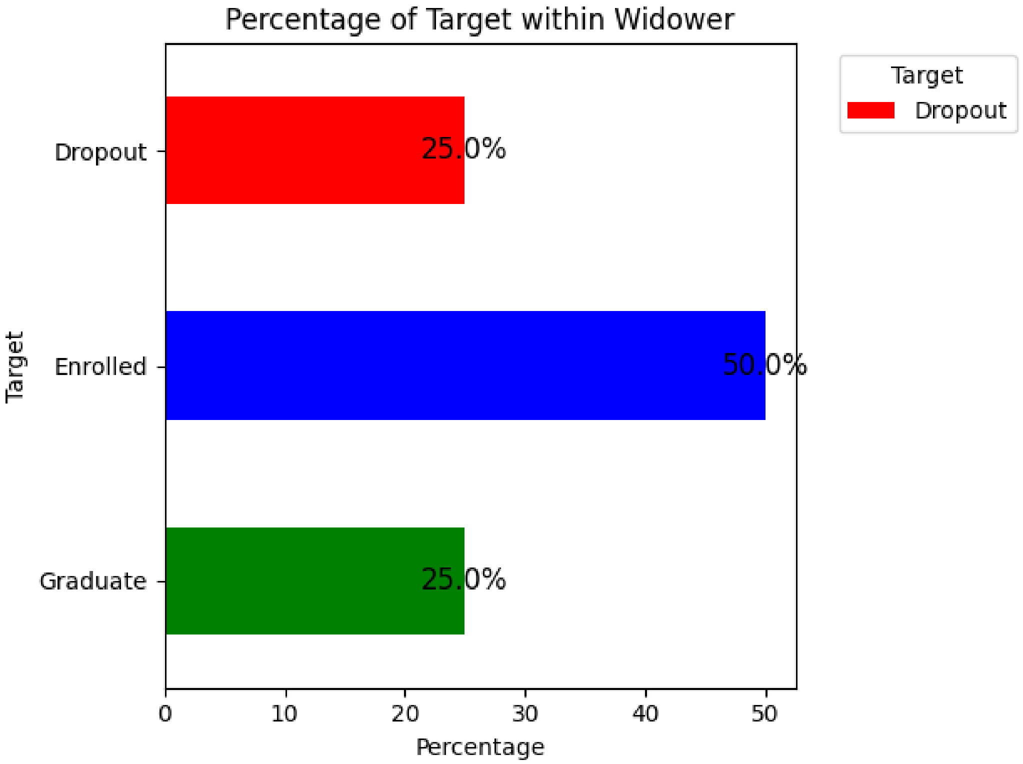
2b Married



2b Divorced



2c Widower



Marital Category Breakdown

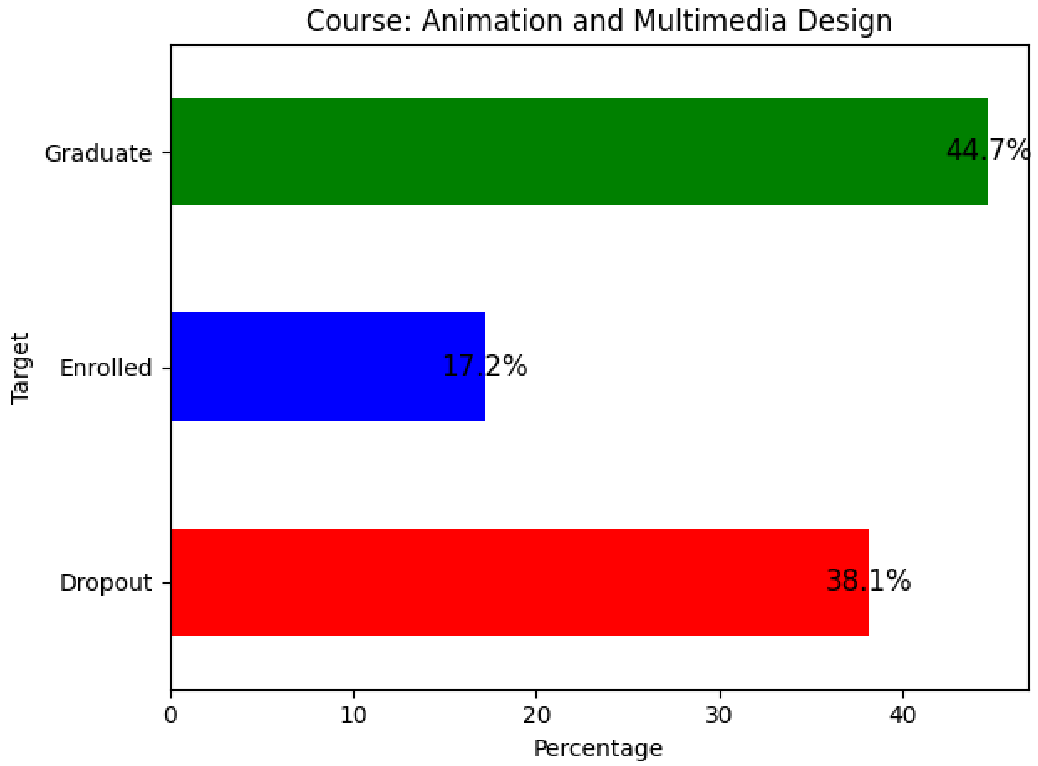
Marital Category	Dropout	Enrolled	Graduate	Total	Percentage
Single	1184	720	2015	3919	88.5800
Married	190	55	159	404	9.1300
Divorced	46	17	34	97	2.1900
Widower	1	2	1	4	0.0900

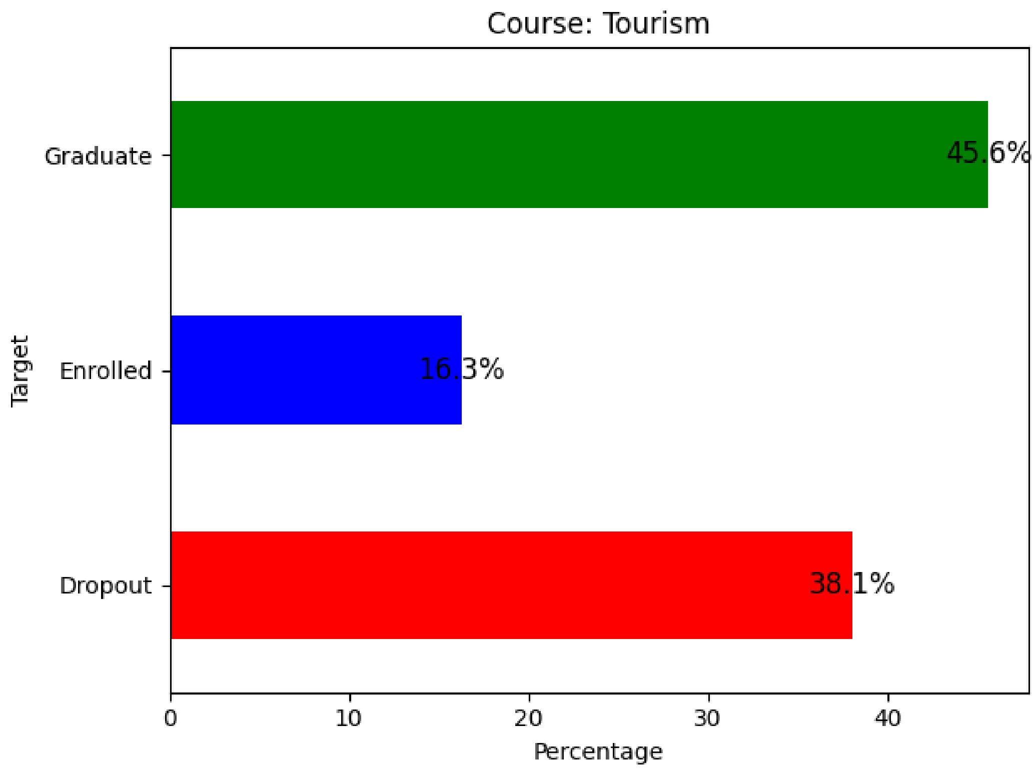
Course Category

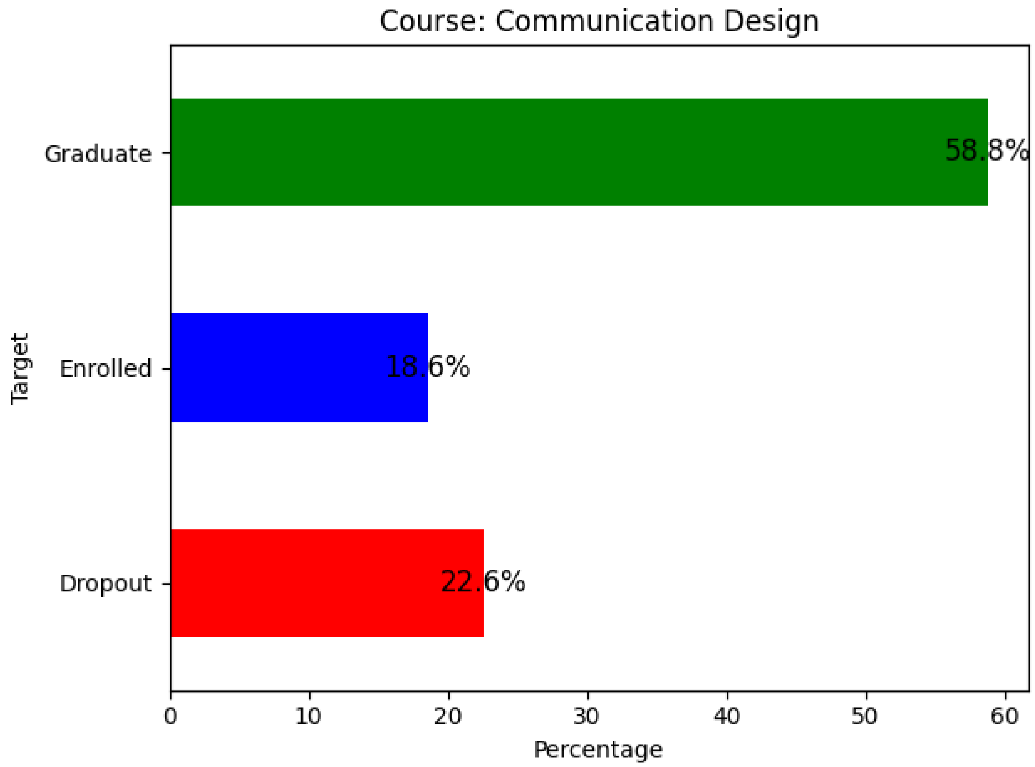
Course Category Breakdown

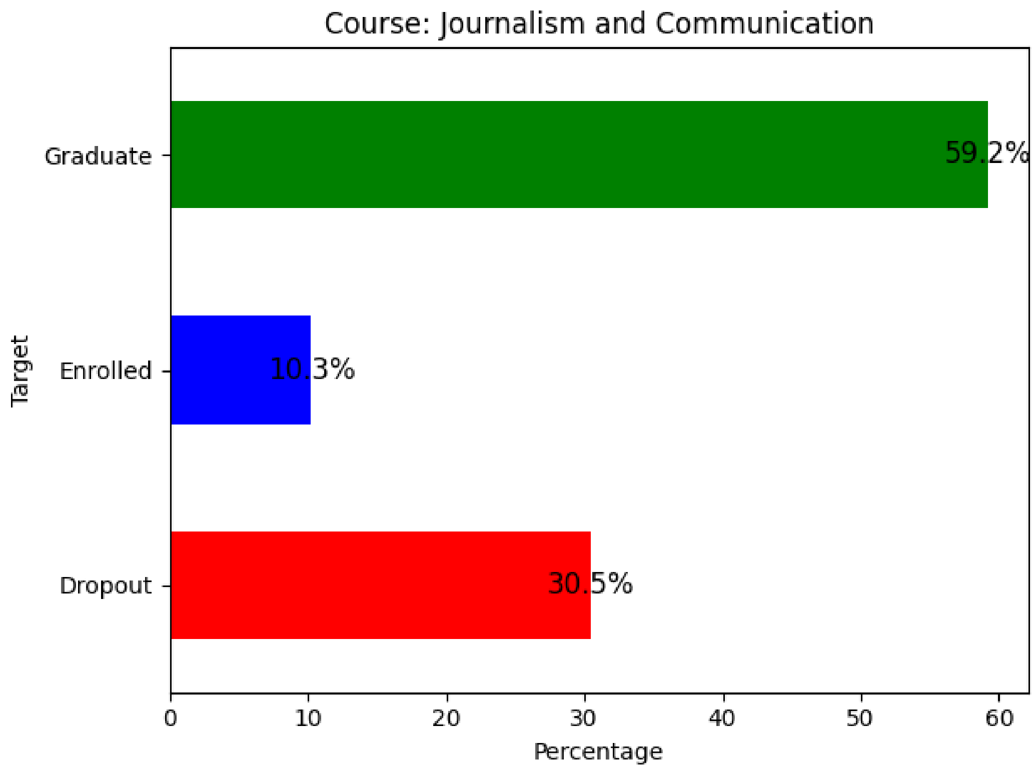
Course Category	Male	Female	Dropout Total %	Enrolled Total %	Graduate Total %	Male Dropout %	Female Dropout %	Male Graduate %	Female Graduate %
Biofuel Production Technologies	9	3	66.6667	25	8.3333	77.7778	33.3333	0	33.3333
Equiniculture	62	79	55.3191	14.8936	29.7872	64.5161	48.1013	24.1935	34.1772
Informatics Engineering	163	7	54.1176	37.6471	8.2353	52.7607	85.7143	8.589	0
Management (evening attendance)	136	132	50.7463	20.1493	29.1045	60.2941	40.9091	27.9412	30.303
Basic Education	9	183	44.2708	26.0417	29.6875	44.4444	44.2623	22.2222	30.0546
Agronomy	149	61	40.9524	17.619	41.4286	46.3087	27.8689	35.5705	55.7377
Oral Hygiene	19	67	38.3721	19.7674	41.8605	63.1579	31.3433	21.0526	47.7612
Animation and Multimedia Design	117	98	38.1395	17.2093	44.6512	39.3162	36.7347	41.8803	47.9592
Tourism	100	152	38.0952	16.2698	45.6349	44	34.2105	32	54.6053
Advertising and Marketing Management	122	146	35.4478	17.9104	46.6418	44.2623	28.0822	37.7049	54.1096
Management	162	218	35.2632	28.4211	36.3158	49.3827	24.7706	22.8395	46.3303
Social Service (evening attendance)	42	173	33.0233	9.7674	57.2093	40.4762	31.2139	47.619	59.5376
Journalism and Communication	115	216	30.5136	10.2719	59.2145	35.6522	27.7778	53.0435	62.5
Veterinary Nursing	60	277	26.7062	22.2552	51.0386	51.6667	21.2996	33.3333	54.8736

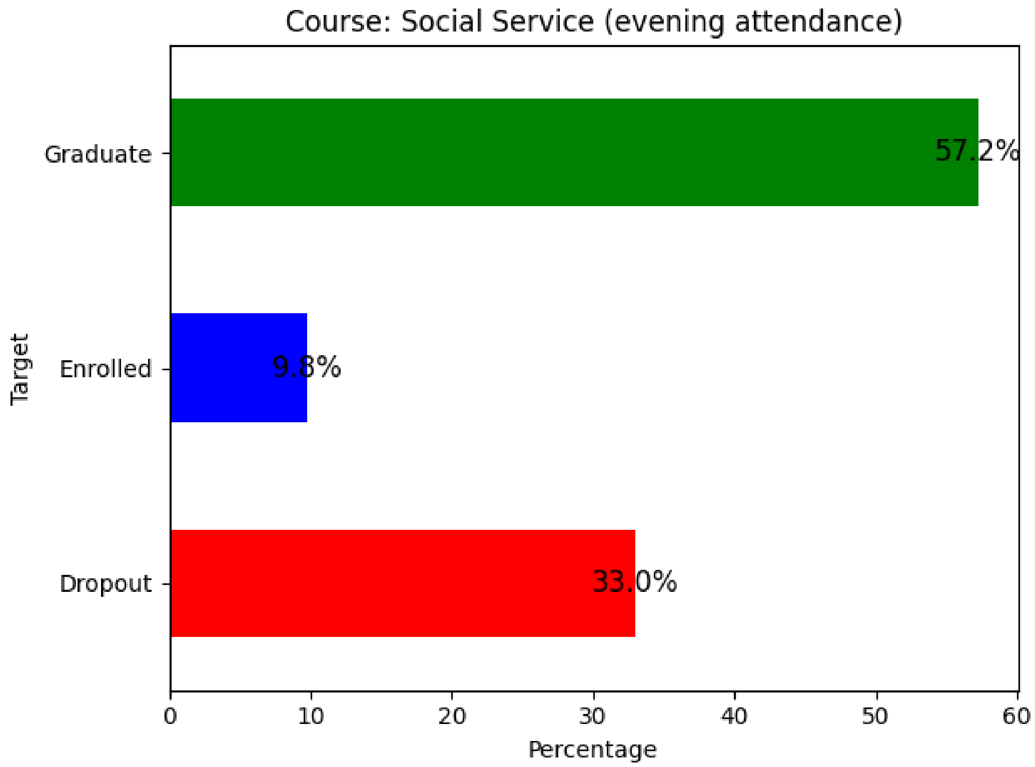
Communication Design	96	130	22.5664	18.5841	58.8496	30.2083	16.9231	50	65.3846
Social Service	46	309	18.3099	11.831	69.8592	47.8261	13.9159	45.6522	73.4628
Nursing	149	617	15.4047	13.0548	71.5405	24.8322	13.128	59.0604	74.5543

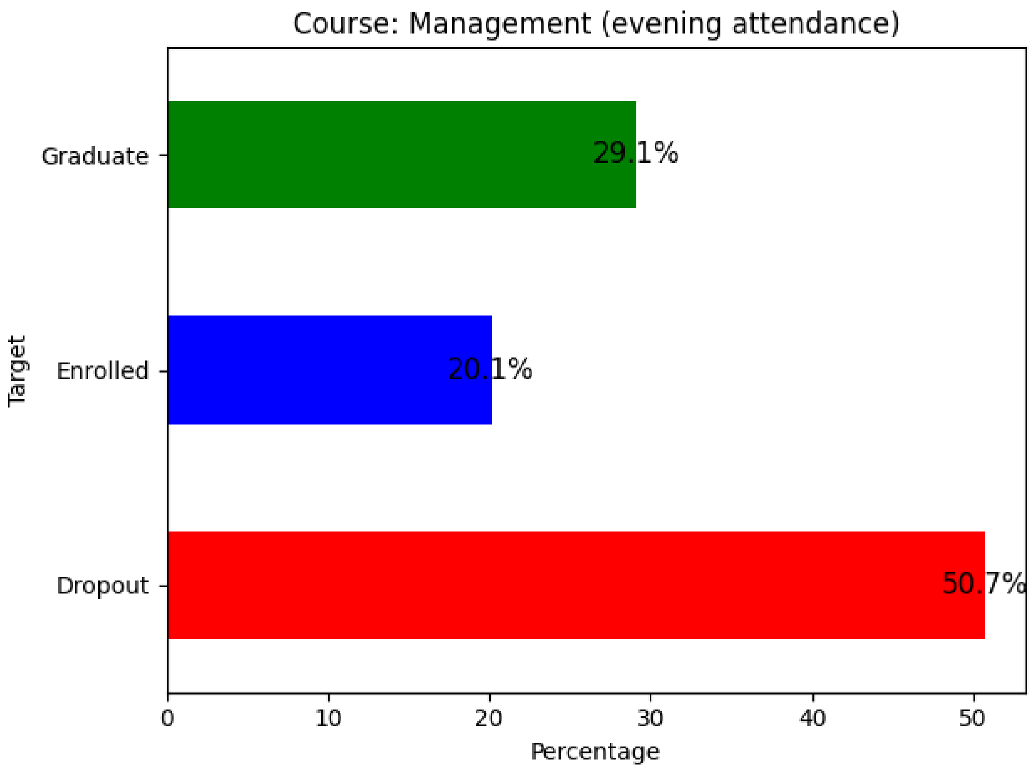


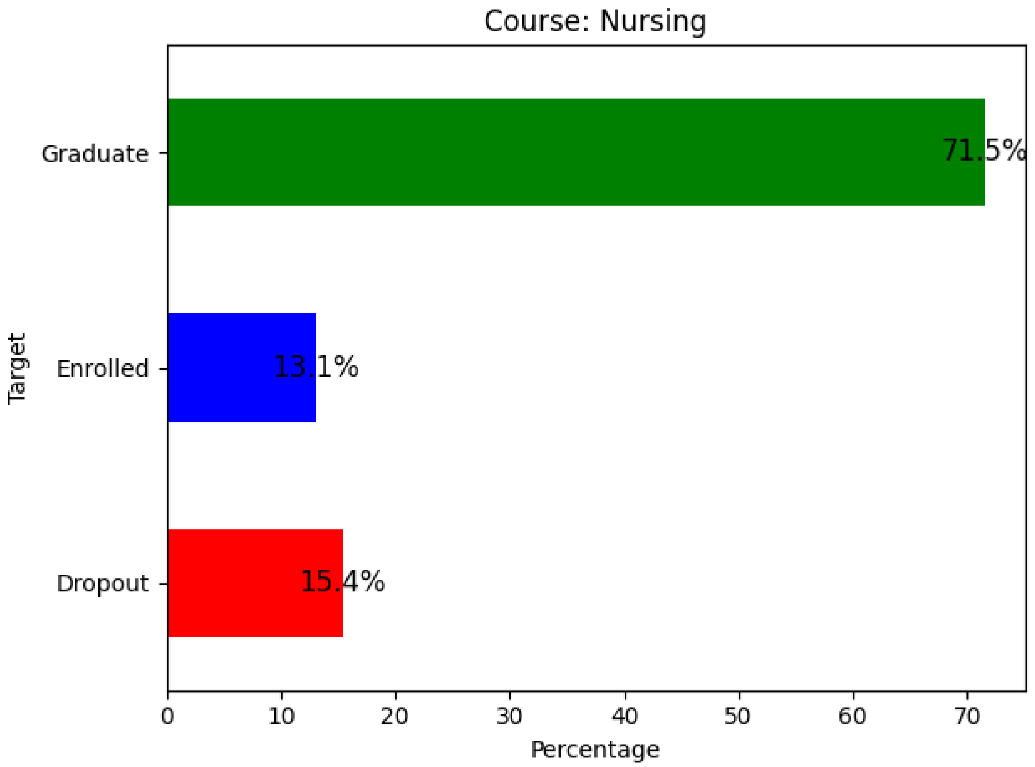

Course: Tourism
Target
Graduate
Enrolled
Dropout
45.6%
16.3%
38.1%
0
10
20
30
40
Percentage

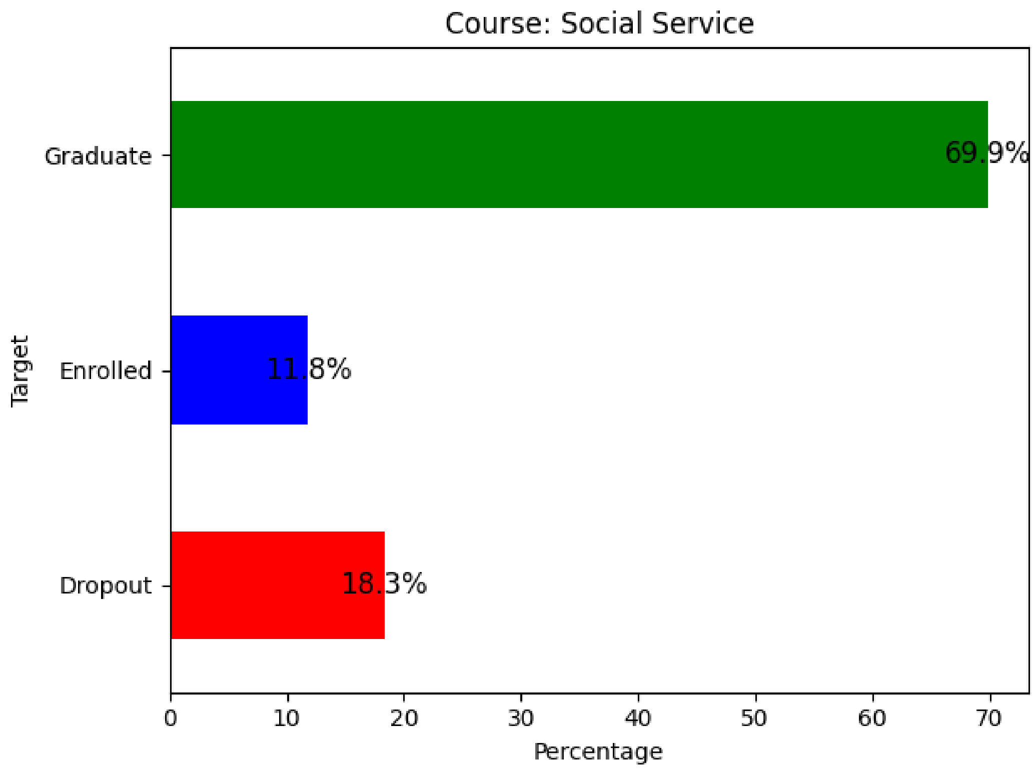

Course: Communication Design
Target
Graduate
Enrolled
Dropout
58.8%
18.6%
22.6%
0
10
20
30
40
50
60
Percentage

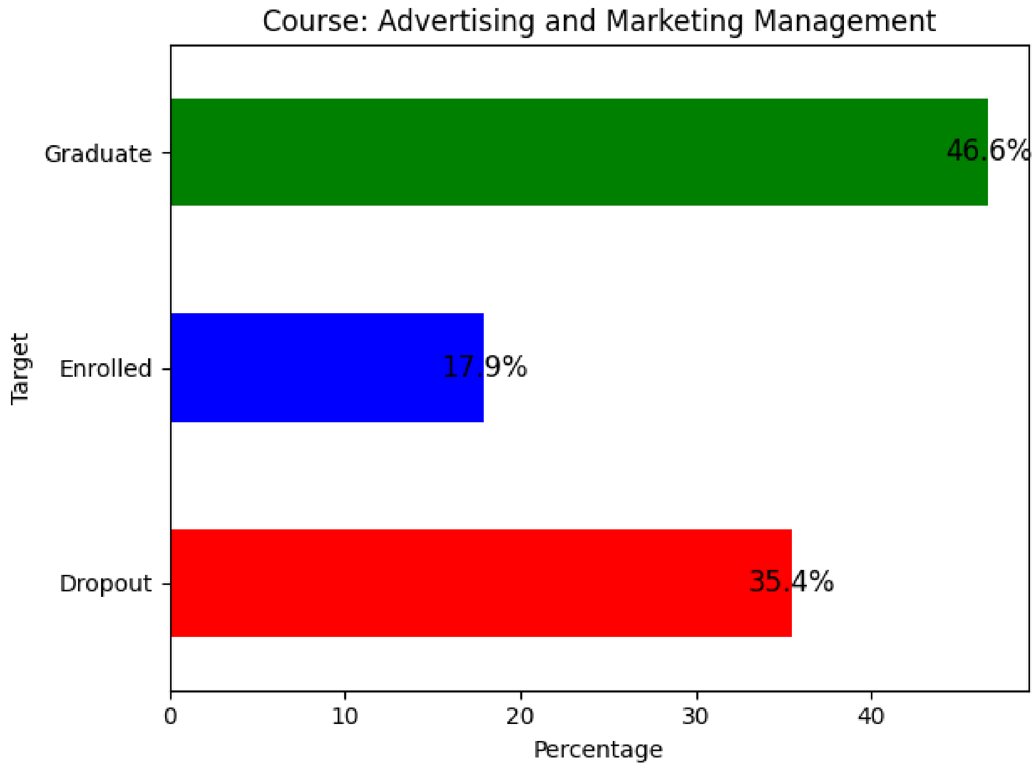

Course: Journalism and Communication
Graduate
Enrolled
Dropout
Target
59.2%
10.3%
30.5%
0
10
20
30
40
50
60
Percentage

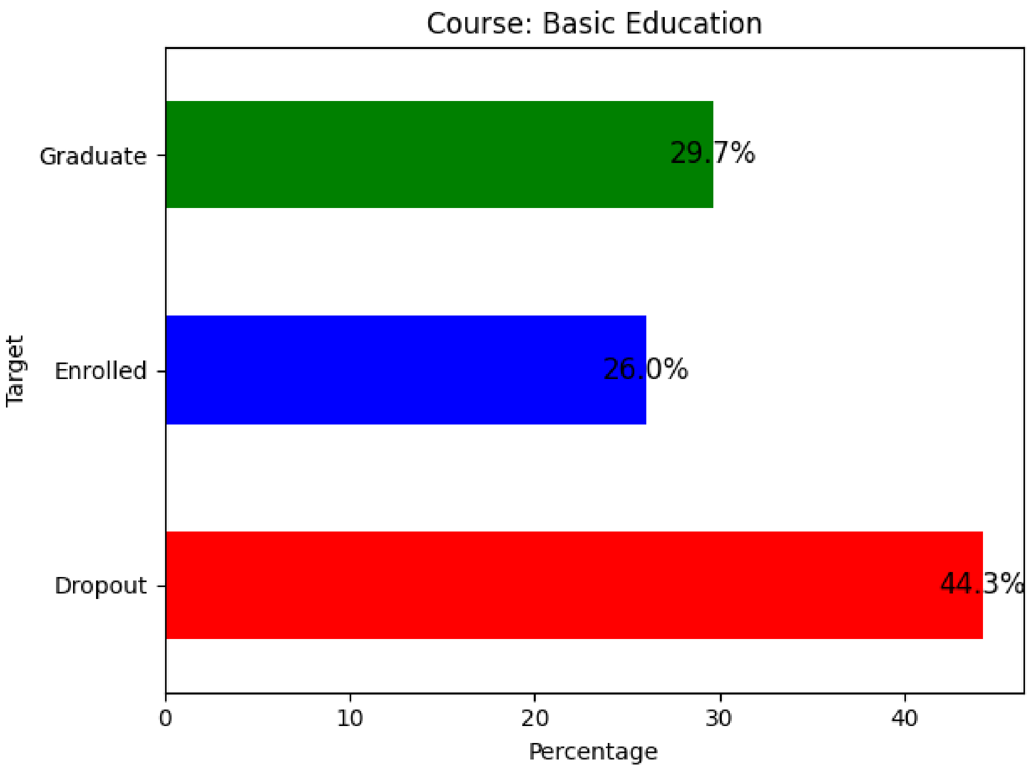

Course: Social Service (evening attendance)
Graduate
Enrolled
Dropout
Target
57.2%
9.8%
33.0%
0
10
20
30
40
50
60
Percentage

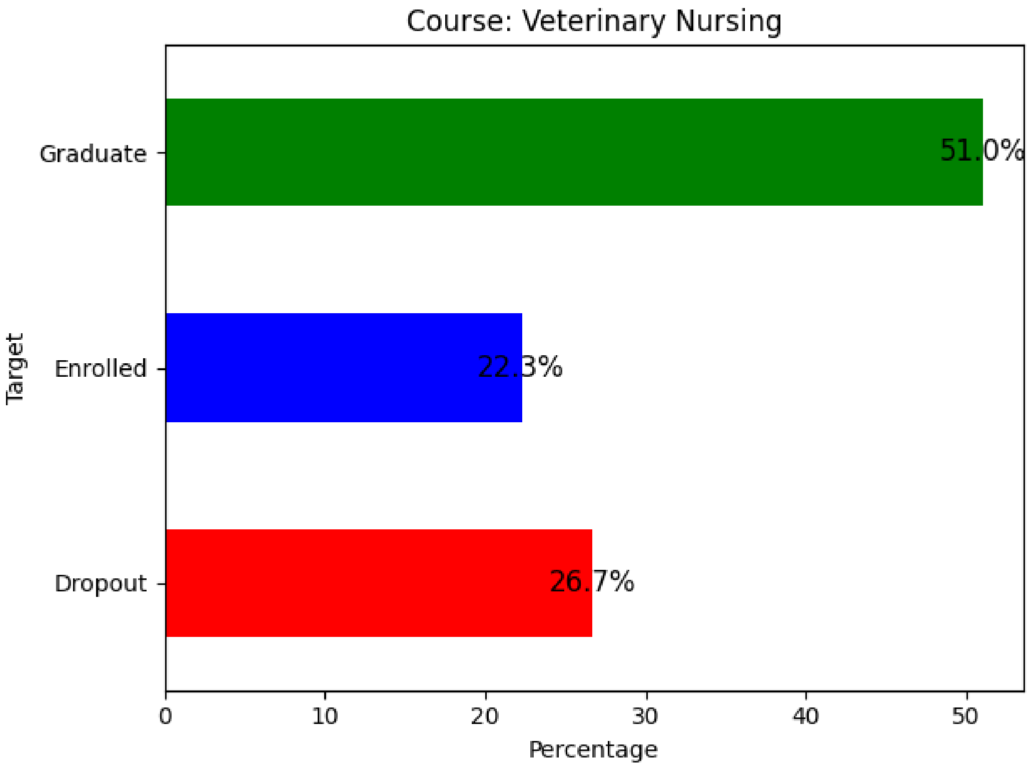

Course: Management (evening attendance)
Graduate
Enrolled
Dropout
Target
29.1%
20.1%
50.7%
0
10
20
30
40
50
Percentage

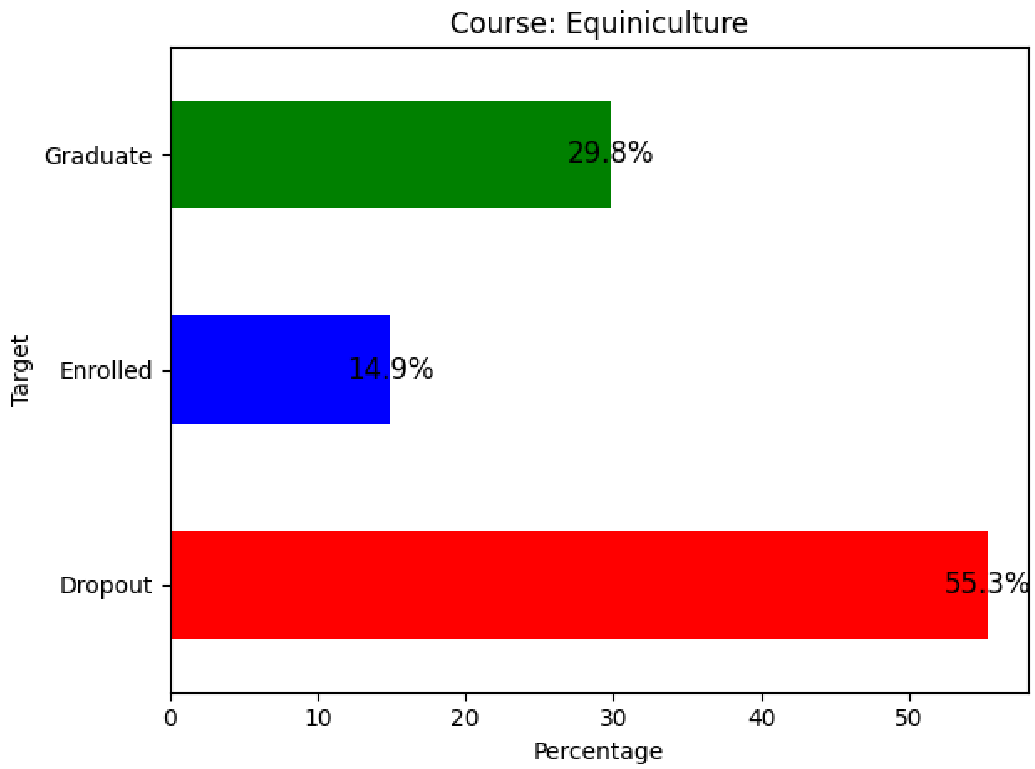

Course: Nursing
Graduate
Enrolled
Dropout
Target
71.5%
13.1%
15.4%
0
10
20
30
40
50
60
70
Percentage

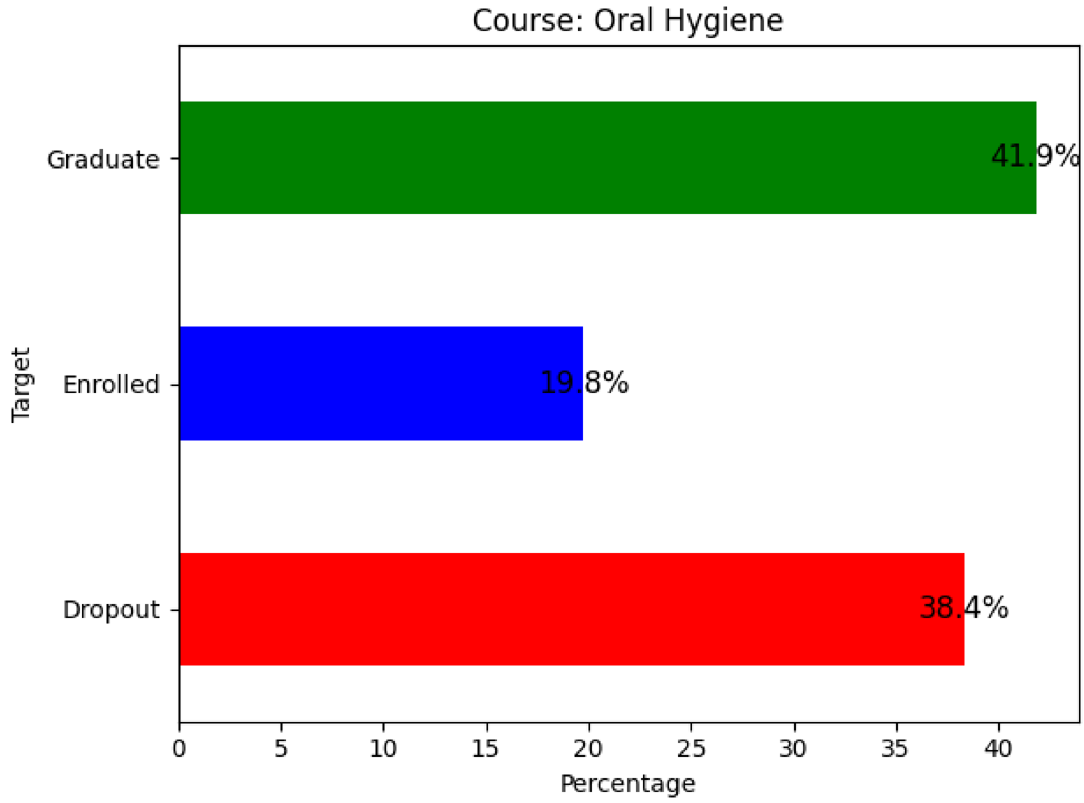

Course: Social Service
Graduate
Enrolled
Dropout
Target
69.9%
11.8%
18.3%
0
10
20
30
40
50
60
70
Percentage

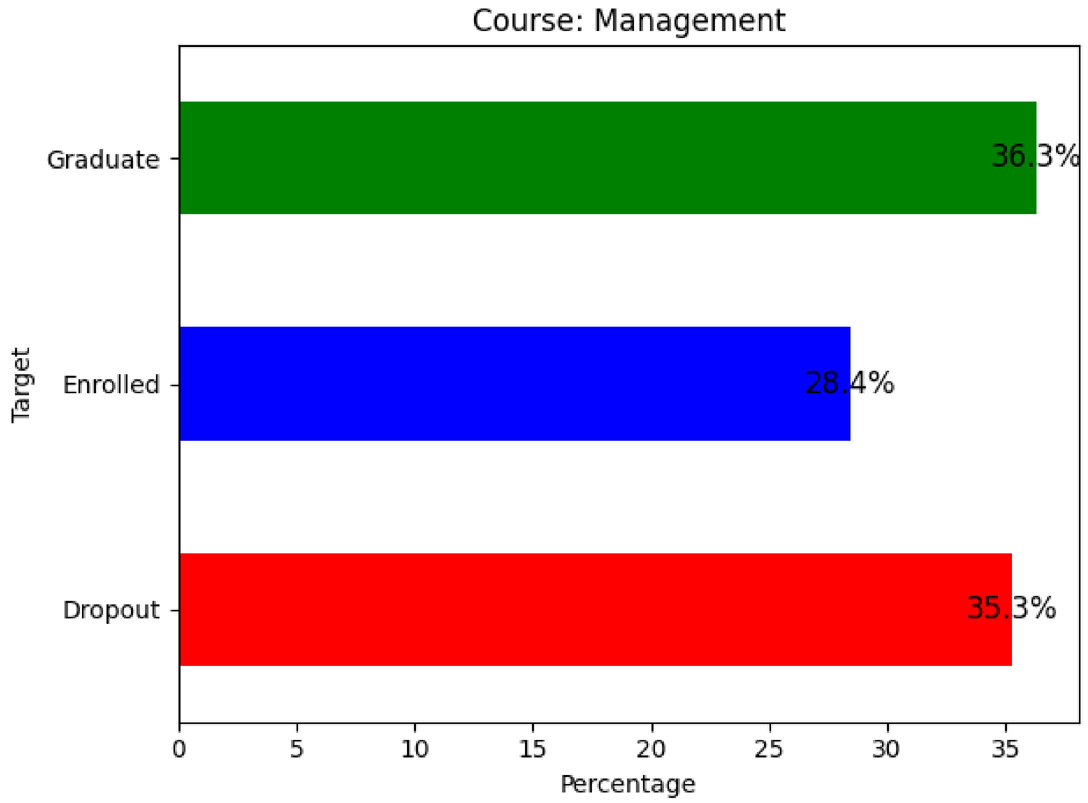

Course: Advertising and Marketing Management
Graduate
Enrolled
Dropout
Target
46.6%
17.9%
35.4%
0
10
20
30
40
Percentage

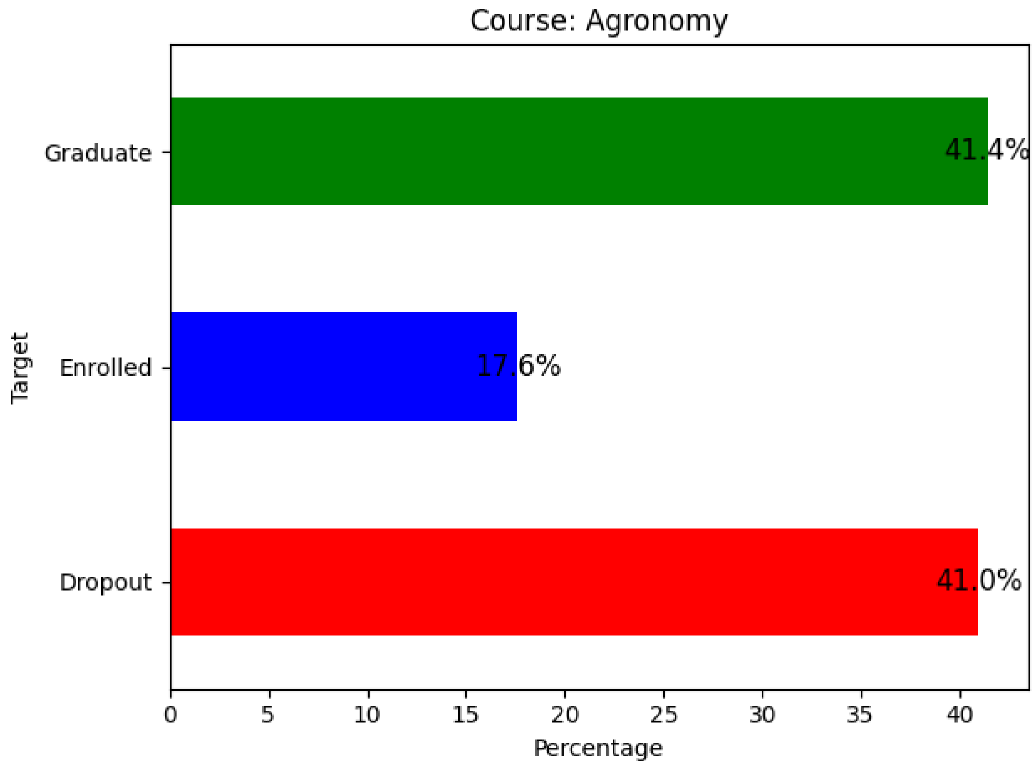

Course: Basic Education
Graduate
Enrolled
Dropout
Target
29.7%
26.0%
44.3%
0
10
20
30
40
Percentage

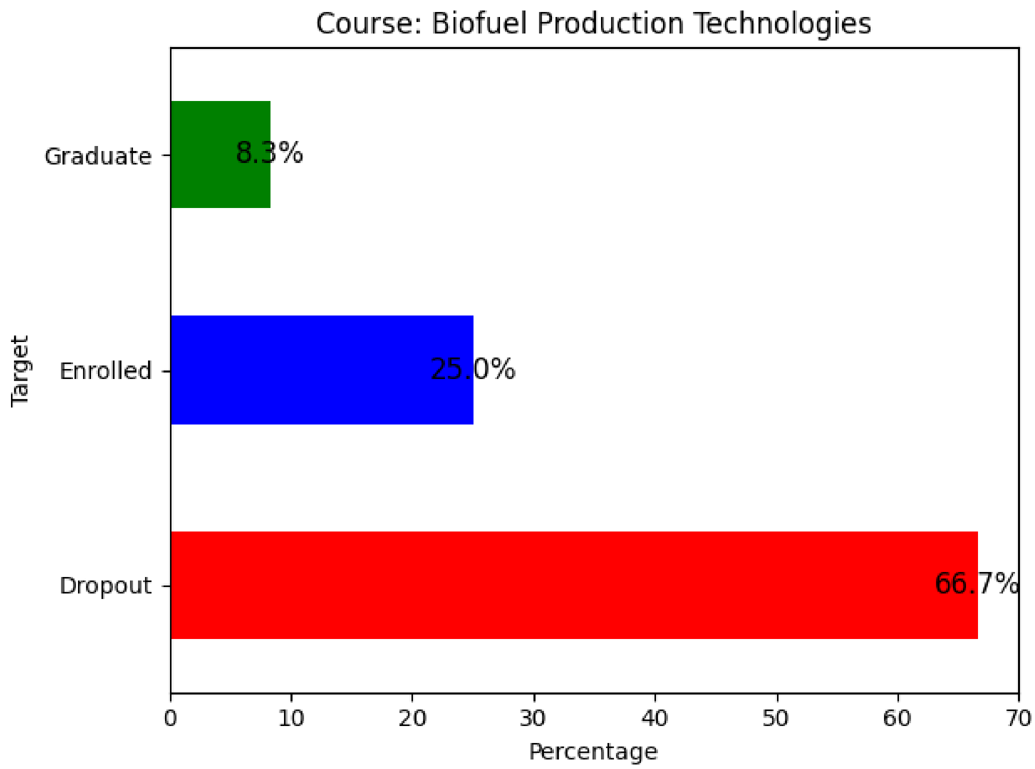

Course: Veterinary Nursing
Graduate
Enrolled
Dropout
Target
51.0%
22.3%
26.7%
0
10
20
30
40
50
Percentage

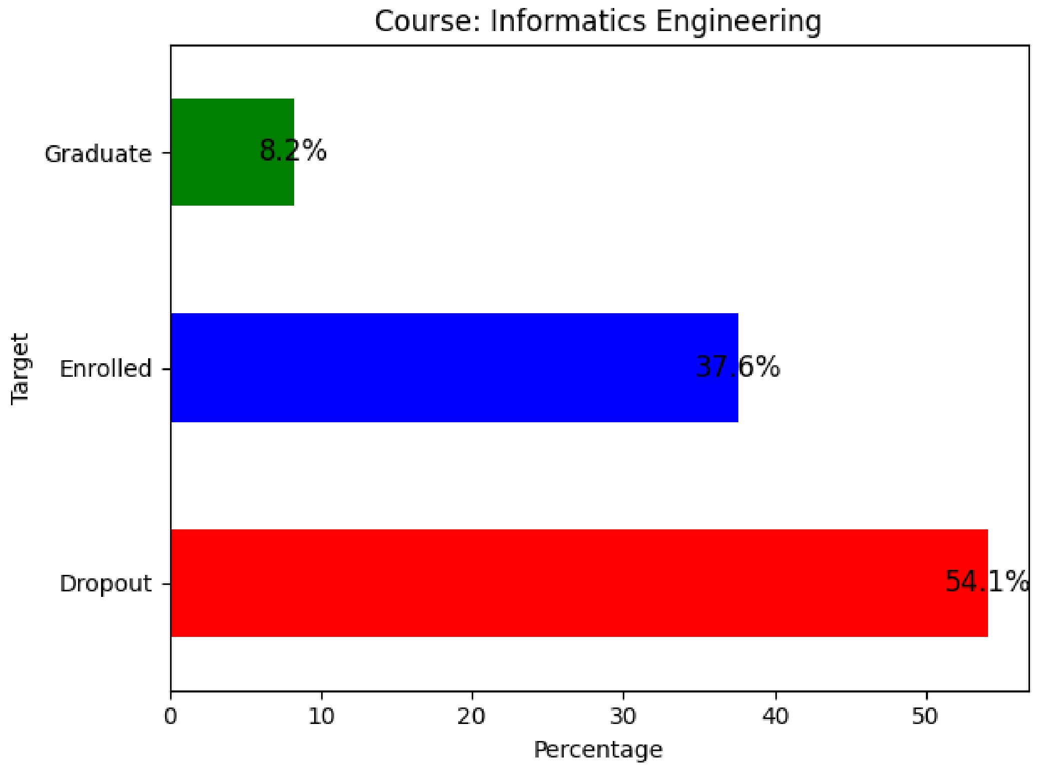

Course: Equiniculture
Graduate
Enrolled
Dropout
Target
29.8%
14.9%
55.3%
0
10
20
30
40
50
Percentage


Course: Oral Hygiene
Graduate
Enrolled
Dropout
Target
41.9%
19.8%
38.4%
0
5
10
15
20
25
30
35
40
Percentage


Course: Management
Graduate
Enrolled
Dropout
Target
36.3%
28.4%
35.3%
0
5
10
15
20
25
30
35
Percentage


Course: Agronomy
Graduate
Enrolled
Dropout
Target
41.4%
17.6%
41.0%
0
5
10
15
20
25
30
35
40
Percentage

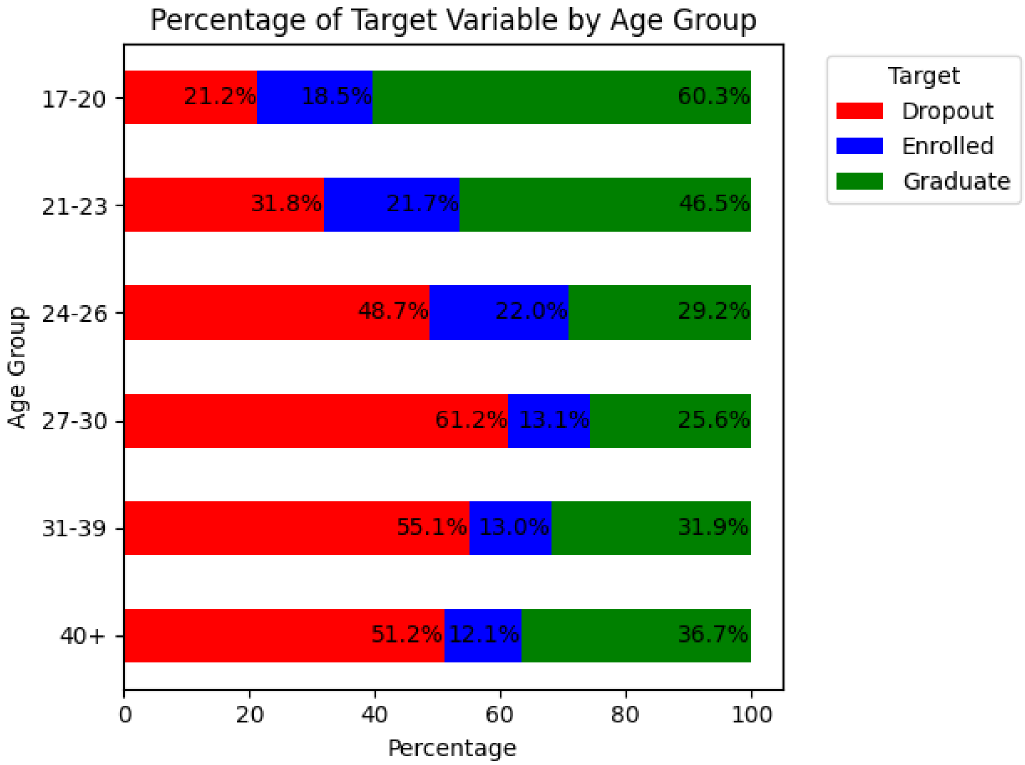

Course: Biofuel Production Technologies
Graduate
8.3%
Enrolled
25.0%
Dropout
66.7%
Target
0
10
20
30
40
50
60
70
Percentage

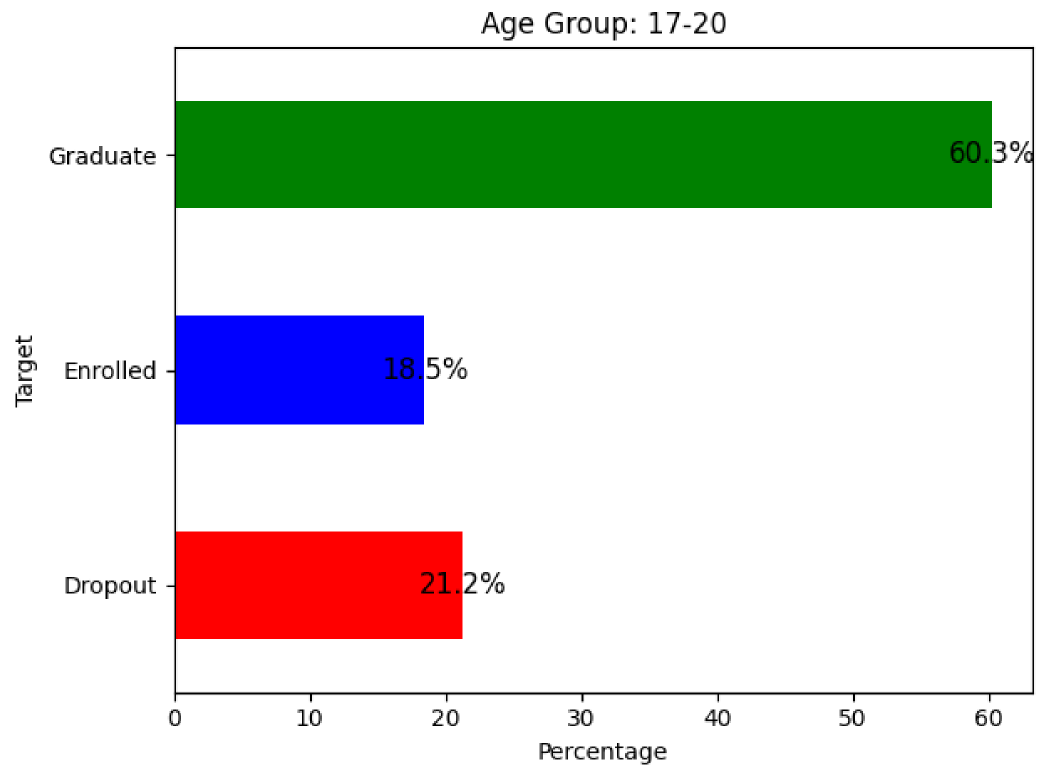

Course: Informatics Engineering
Graduate
8.2%
Enrolled
37.6%
Dropout
54.1%
Target
0
10
20
30
40
50
Percentage

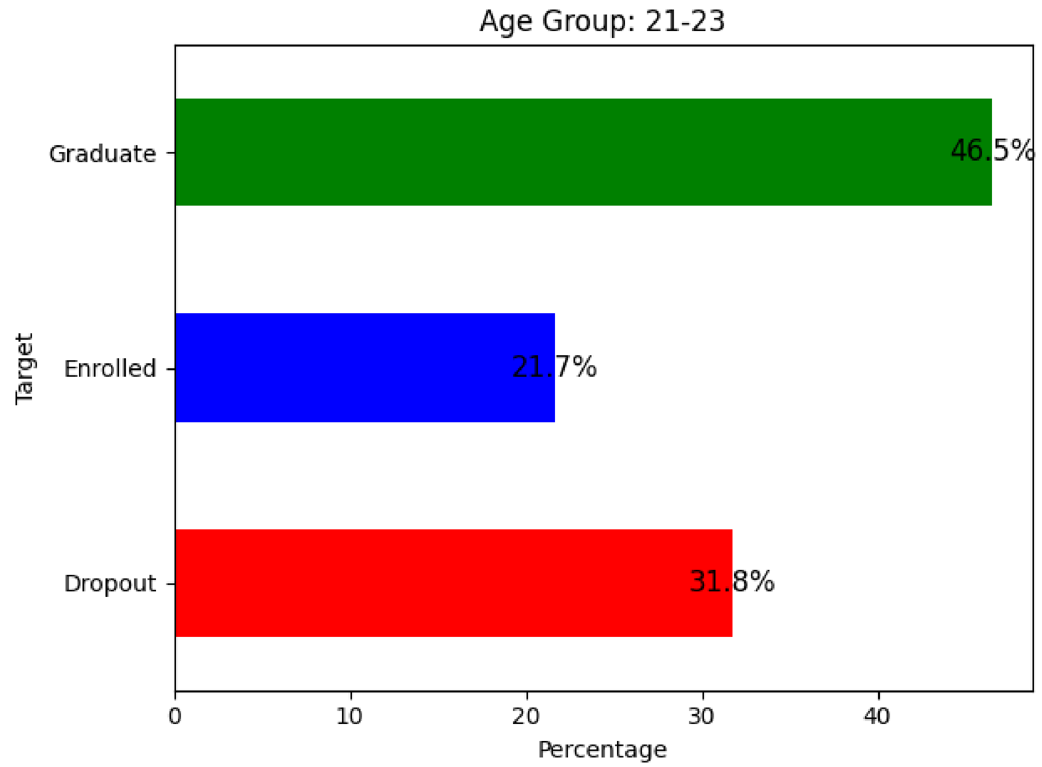
Age

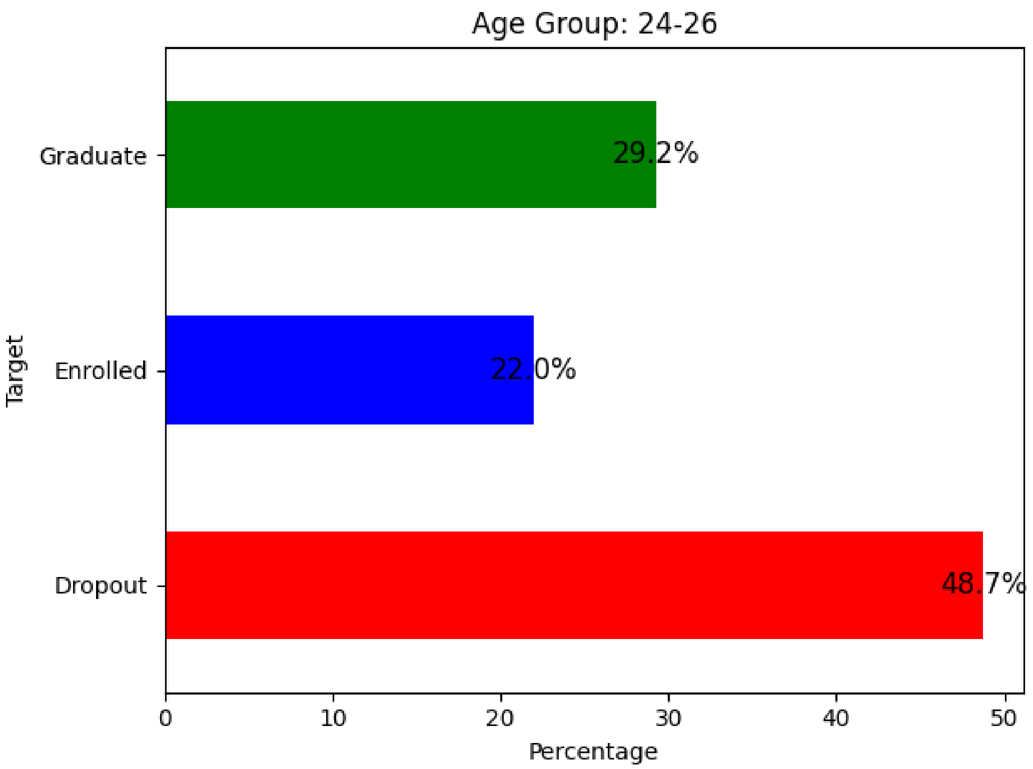
Age Group Breakdown

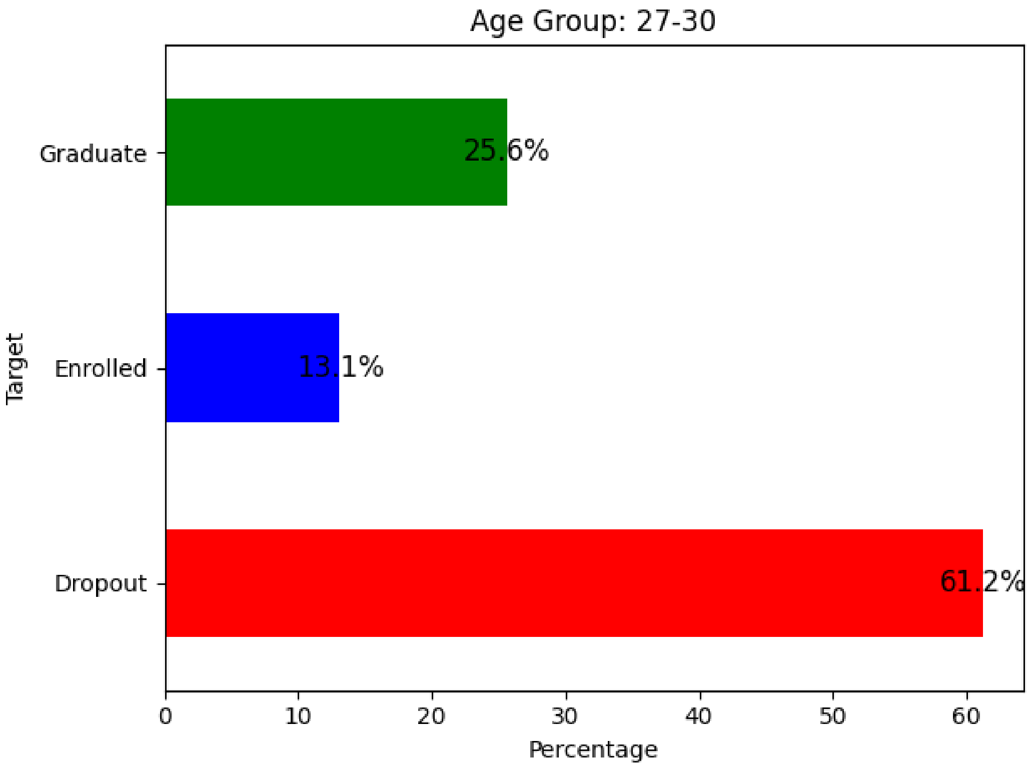
Age Group	Dropout	Enrolled	Graduate	Total by Age Group	Total by Percentage
17-20	21.24657	18.46335	60.29008	2551	57.66274864
21-23	31.78808	21.68874	46.52318	604	13.65280289
24-26	48.74214	22.01258	29.24528	318	7.188065099
27-30	61.24567	13.14879	25.60554	289	6.532549729
31-39	55.07246	13.04348	31.88406	414	9.358047016
40+	51.20968	12.09677	36.69355	248	5.605786618

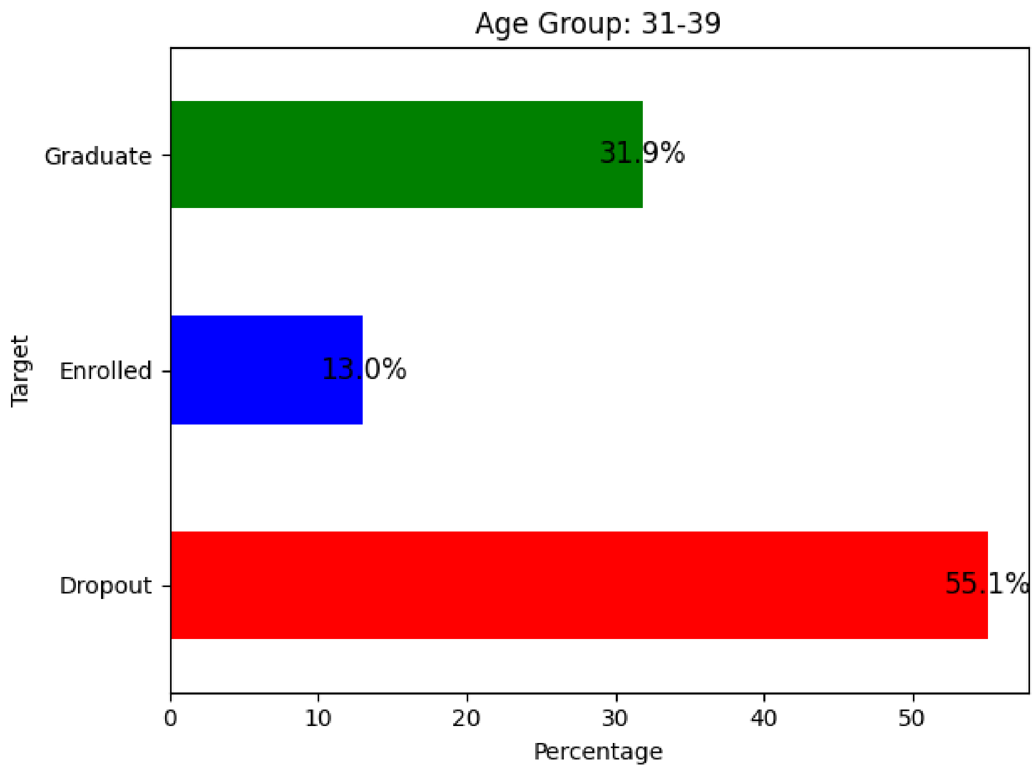


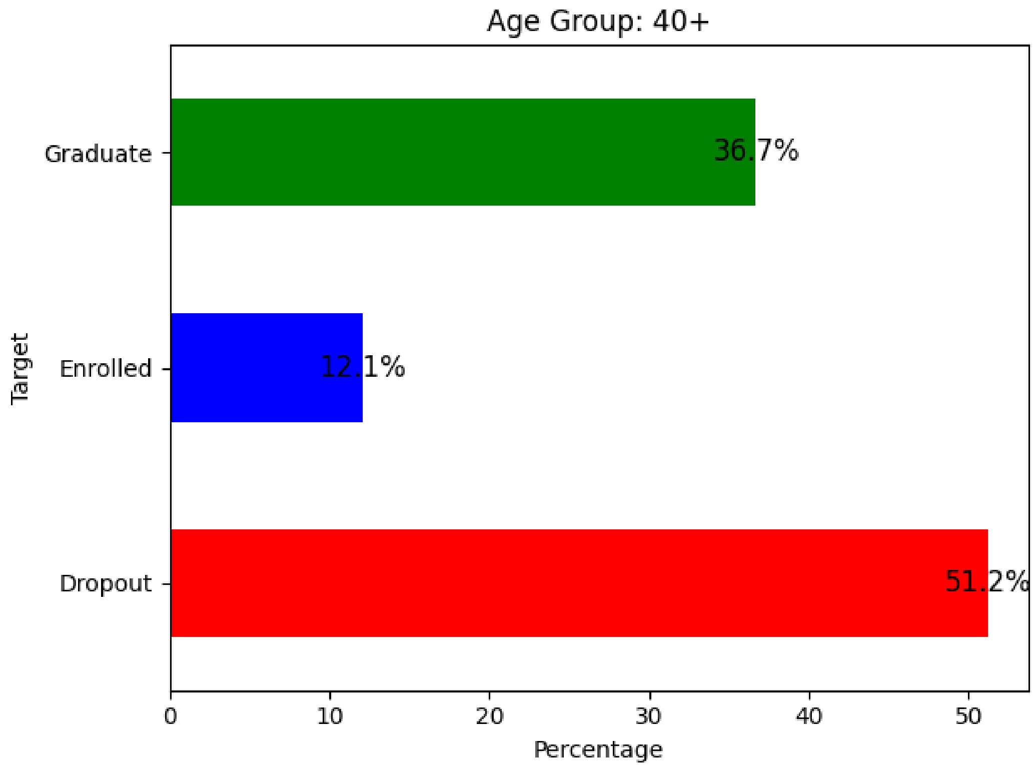

Age Group: 17-20
Target
Graduate
Enrolled
Dropout
60.3%
18.5%
21.2%
0
10
20
30
40
50
60
Percentage


Age Group: 21-23
Target
Graduate
Enrolled
Dropout
46.5%
21.7%
31.8%
0
10
20
30
40
Percentage


Age Group: 24-26
Target
Graduate
Enrolled
Dropout
29.2%
22.0%
48.7%
0
10
20
30
40
50
Percentage


Age Group: 27-30
Target
Graduate
Enrolled
Dropout
25.6%
13.1%
61.2%
0
10
20
30
40
50
60
Percentage


Age Group: 31-39
Target
Graduate
Enrolled
Dropout
31.9%
13.0%
55.1%
0
10
20
30
40
50
Percentage


Age Group: 40+
Target
Graduate
Enrolled
Dropout
36.7%
12.1%
51.2%
0
10
20
30
40
50
Percentage

Ethnicity

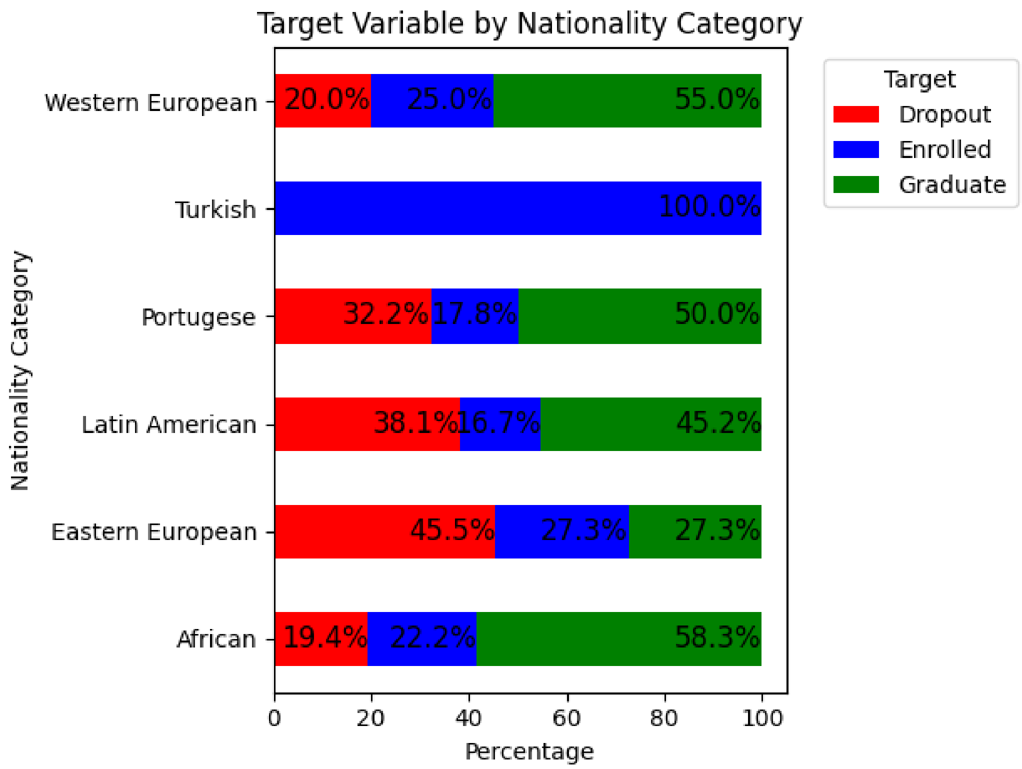
There were 21 different ethnicities listed in the dataset however, 4314 of the students were Portuguese in origin which accounted for 97.5 % of the students in the dataset. The countries were categorized into: Turkish, African, Latin American, Eastern European, and Western European.

Ethnicity Breakdown

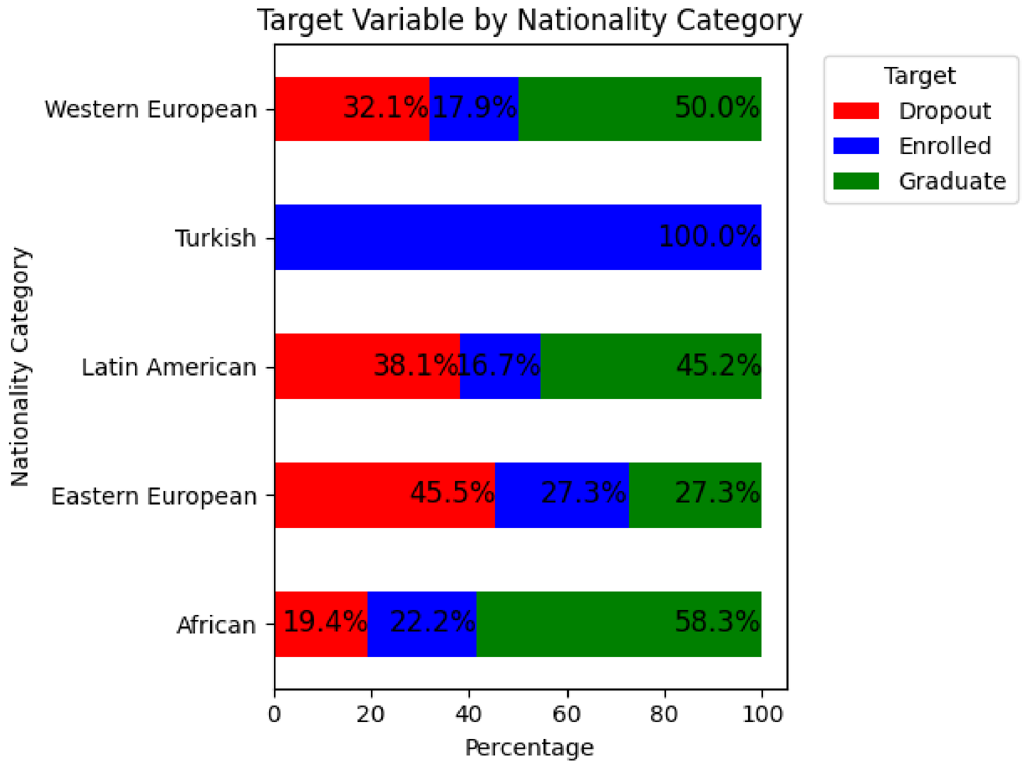
Nationality Category	Ethnicity	Total Count	Running Total
Turkish	Turkish	1	1
African	Angolan	2	3
	Cape Verdean	13	16
	Guinean	5	21
	Mozambican	2	23
	Santomean	14	37
Latin American	Brazilian	38	75
	Colombian	0	75
	Cuban	1	76
	Mexican	2	78
Eastern European	Lithuanian 1 79	1	79
	Moldova (Republic of)	3	82
	Romanian	2	84
	Russian	2	86
	Ukrainian	3	89
Western European	Dutch	1	90
	English	1	91
	German	2	93
	Italian	3	96
	Spanish	13	109
	Portuguese	4314	4423

Nationality Category	Total
Turkish	1
African	36
Latin American	42
Eastern European	11
Western European	20
Portuguese	4314

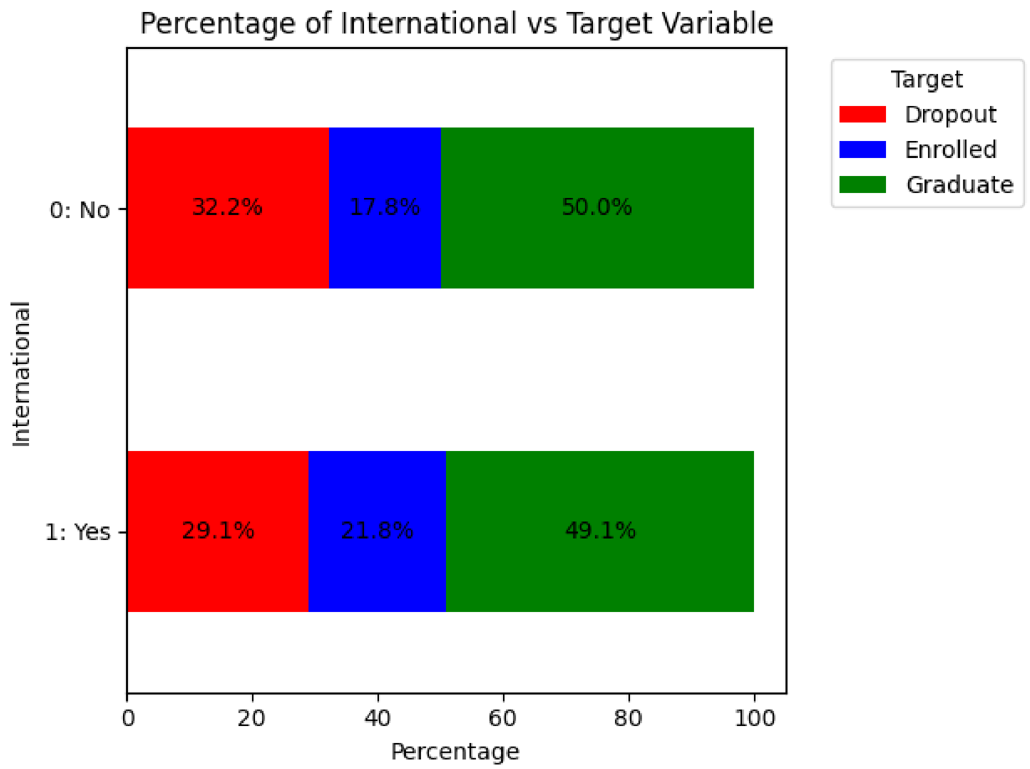
Nationality by Category with Portuguese separate from Western European



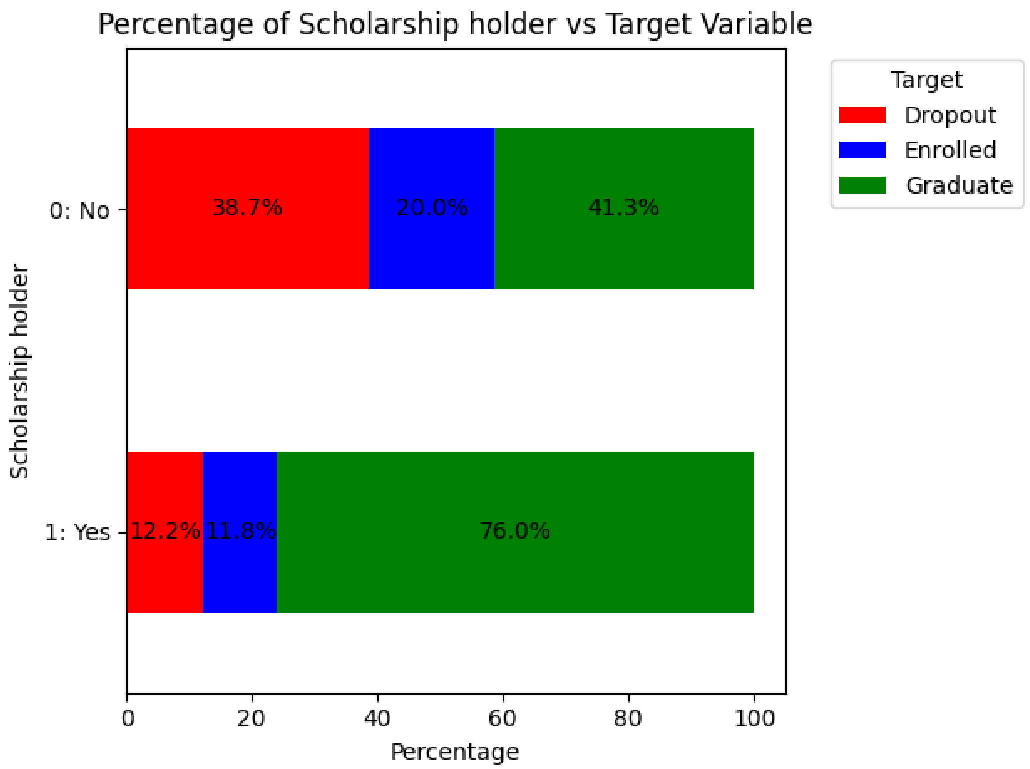
Nationality by Category with Portuguese part of Western European



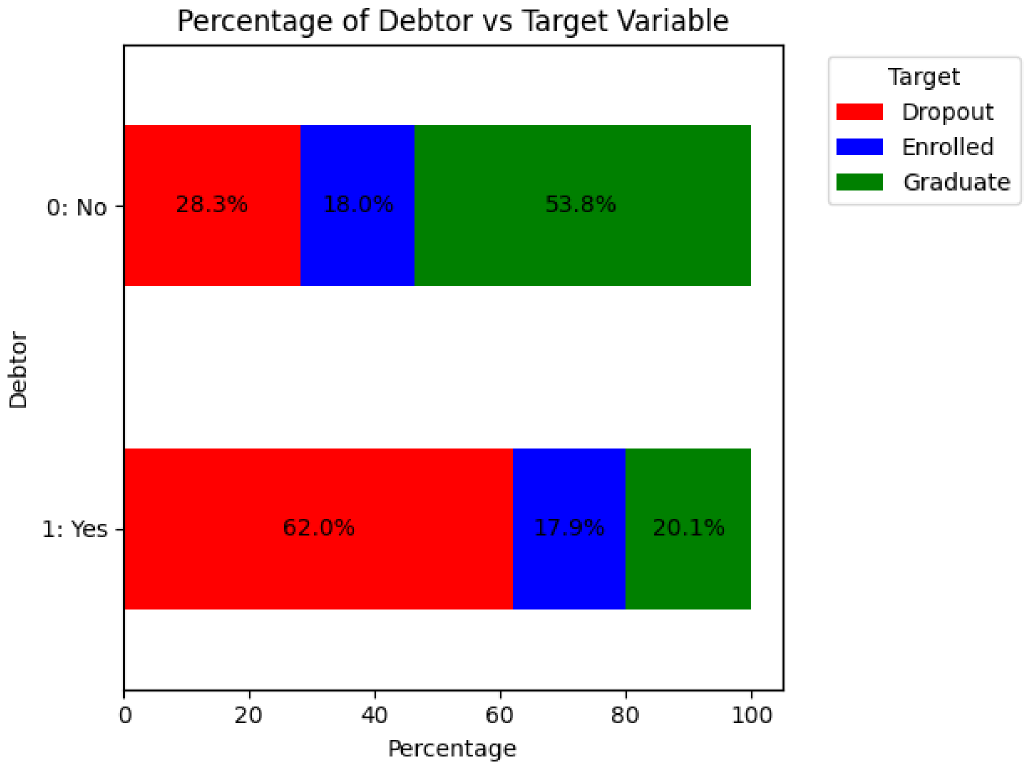
International vs Non-International Students



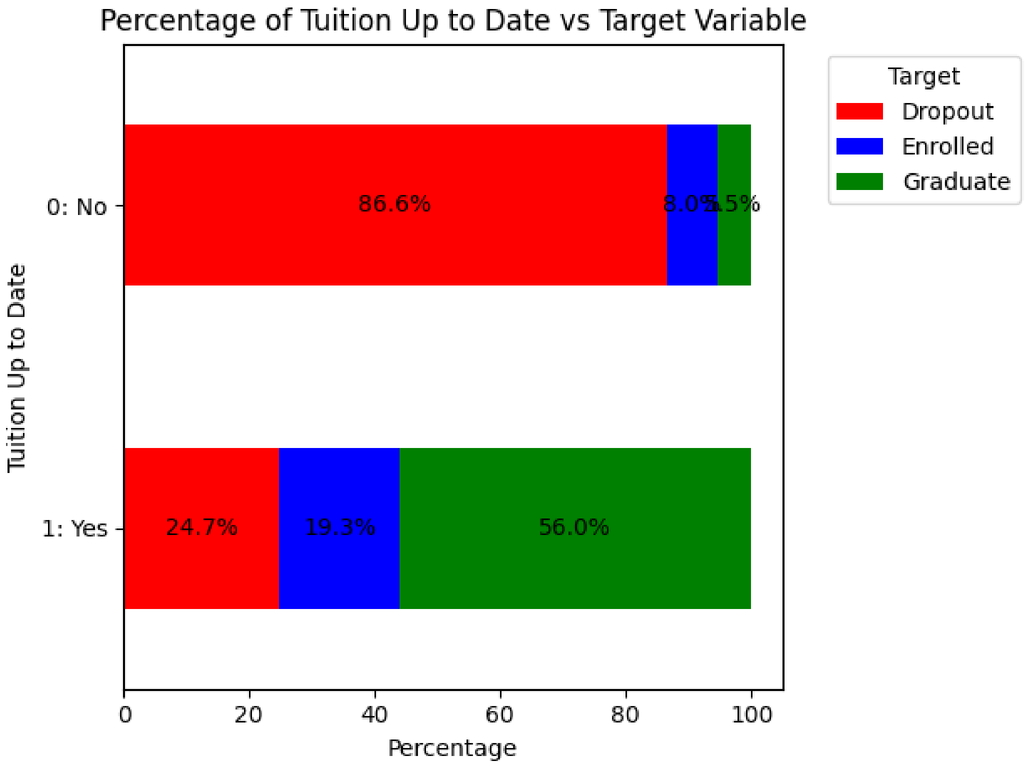
Scholarship Holders



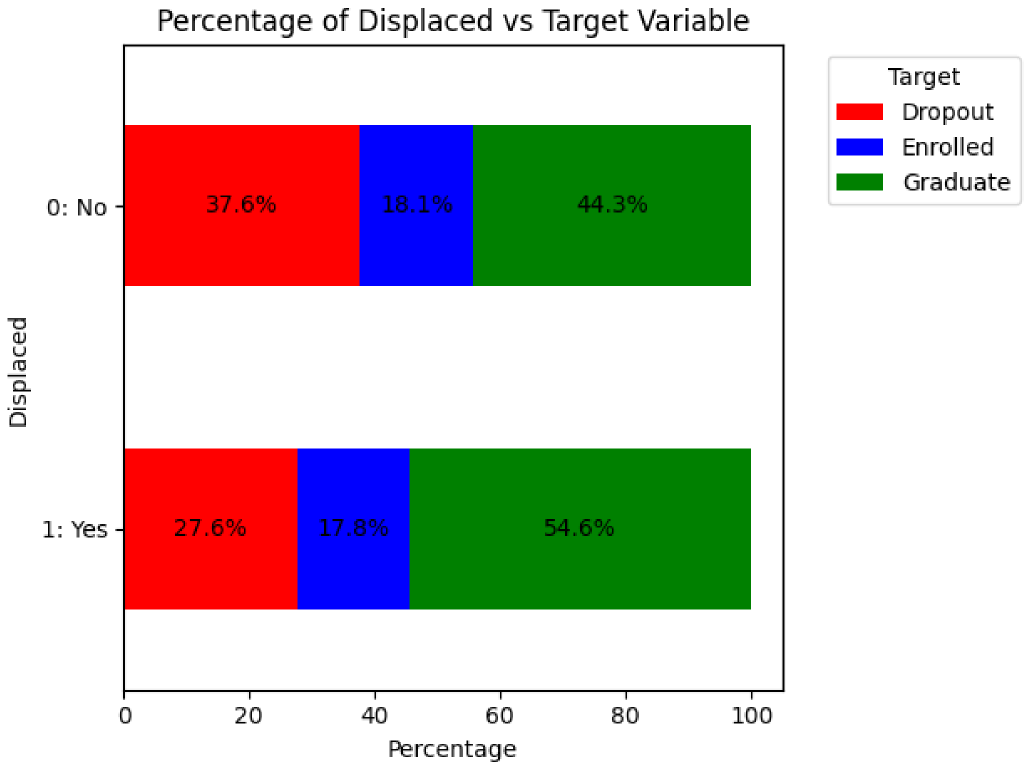
Debtors



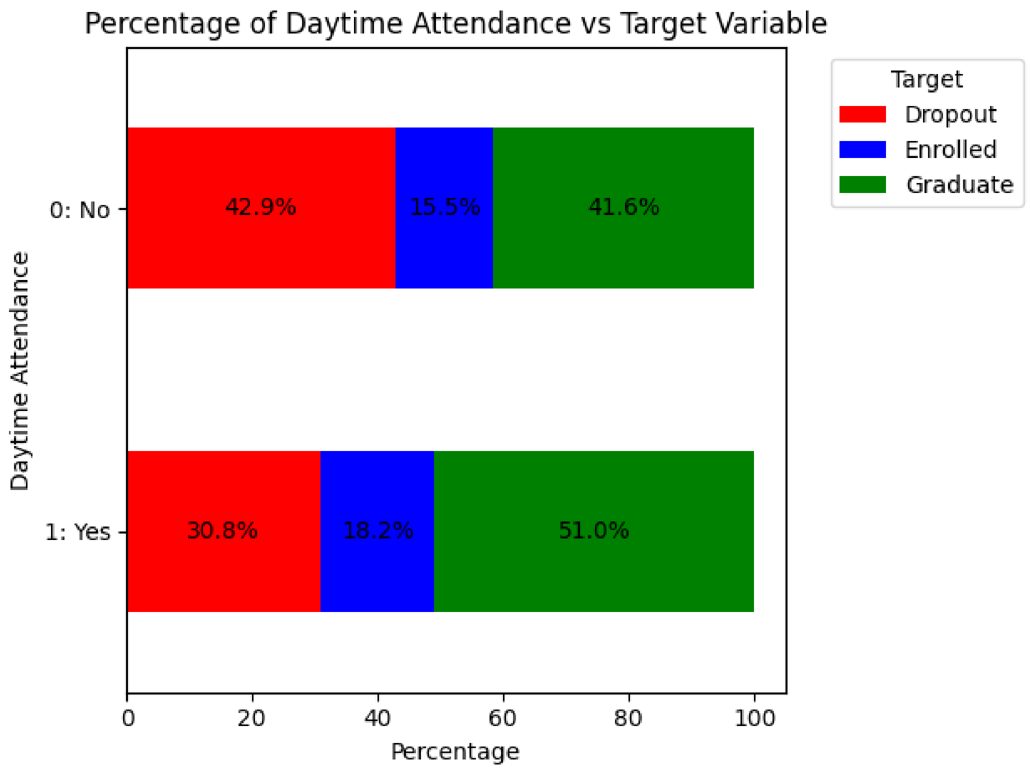
Tuition Up To Date



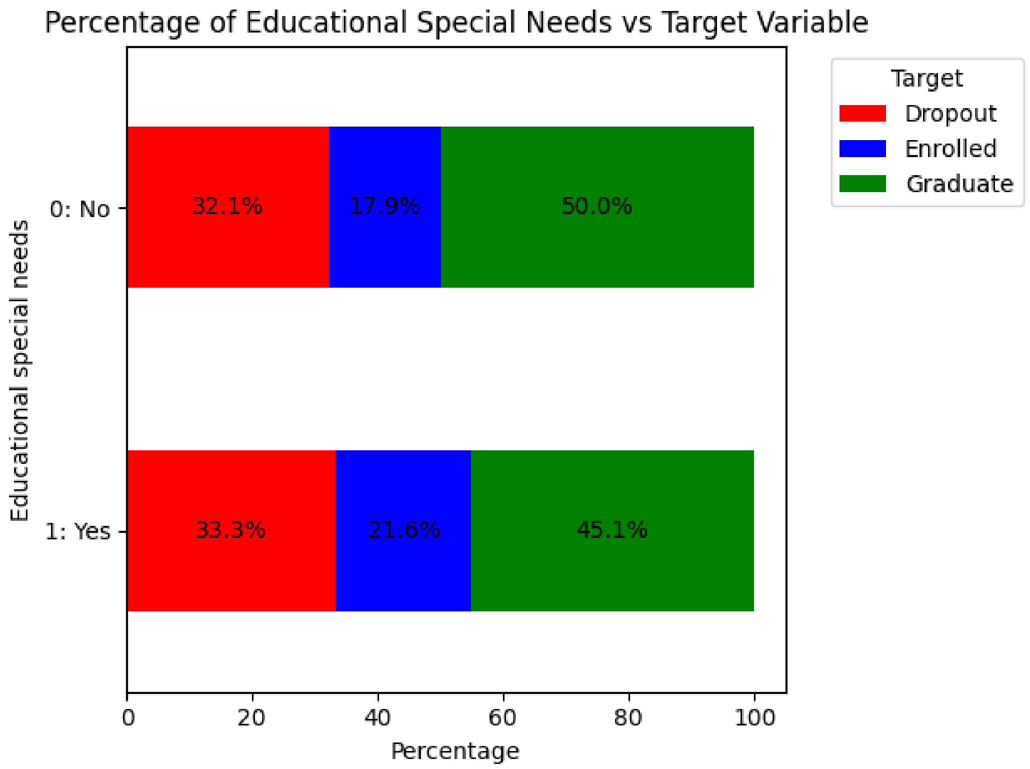
Displaced



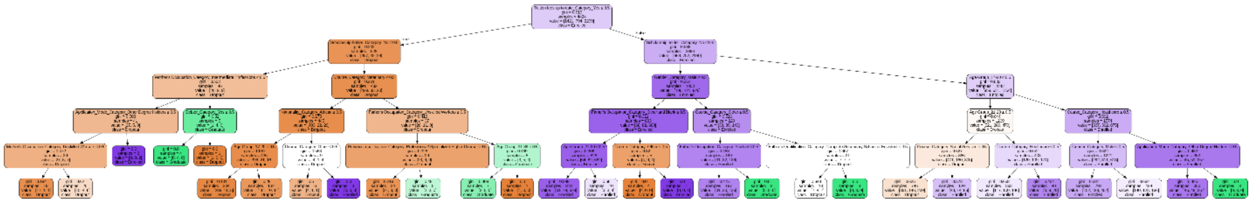
Daytime vs Night Attendance



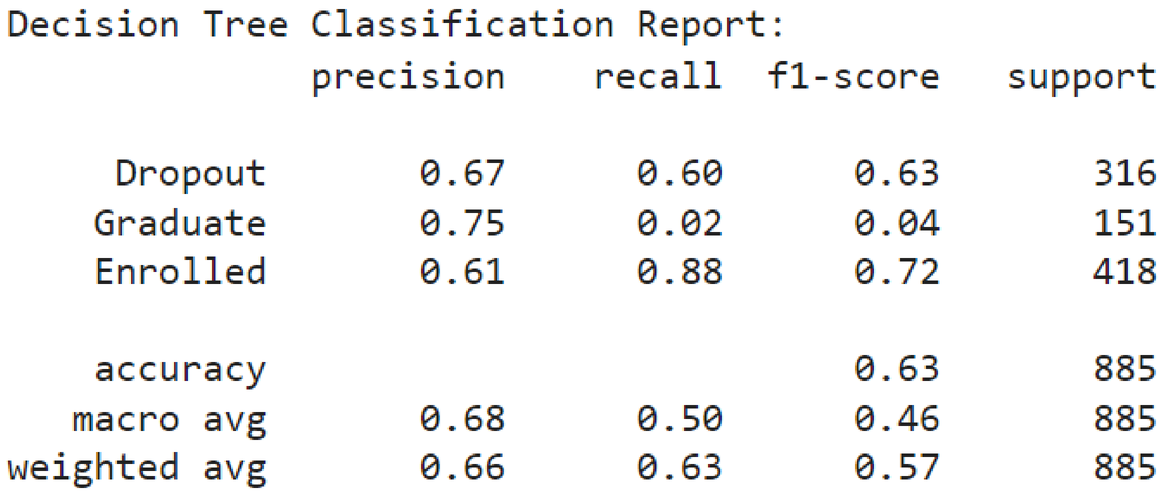
Educational Special Needs



Decision Tree

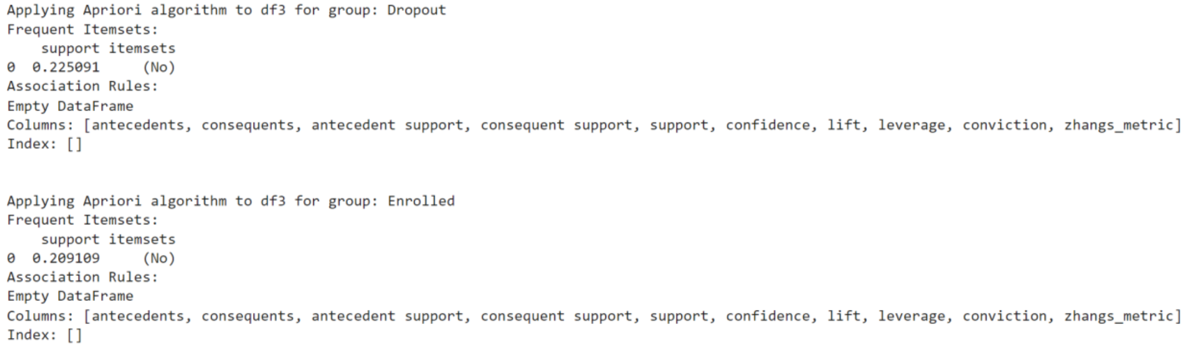


Decision Tree Classification Report



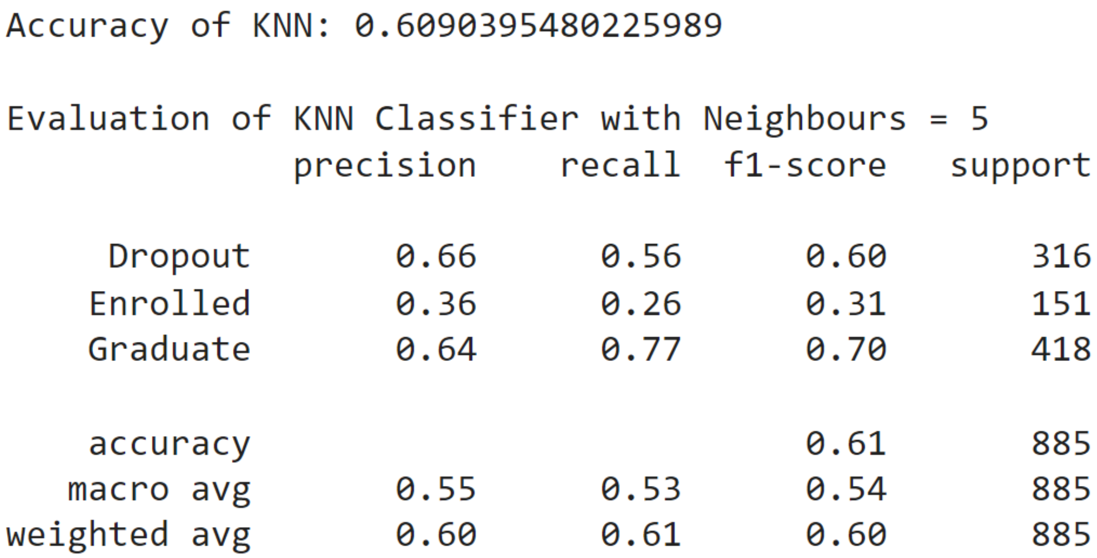
```
Decision Tree Classification Report:
              precision    recall  f1-score   support

     Dropout       0.67      0.60      0.63       316
    Graduate       0.75      0.02      0.04       151
    Enrolled       0.61      0.88      0.72       418

    accuracy                           0.63       885
   macro avg       0.68      0.50      0.46       885
weighted avg       0.66      0.63      0.57       885
```

Apriori Algorithim



```
Applying Apriori algorithm to df3 for group: Dropout
Frequent Itemsets:
    support itemsets
0  0.225091     (No)
Association Rules:
Empty DataFrame
Columns: [antecedents, consequents, antecedent support, consequent support, support, confidence, lift, leverage, conviction, zhangs_metric]
Index: []


Applying Apriori algorithm to df3 for group: Enrolled
Frequent Itemsets:
    support itemsets
0  0.209109     (No)
Association Rules:
Empty DataFrame
Columns: [antecedents, consequents, antecedent support, consequent support, support, confidence, lift, leverage, conviction, zhangs_metric]
Index: []
```

K Nearest Means Classification

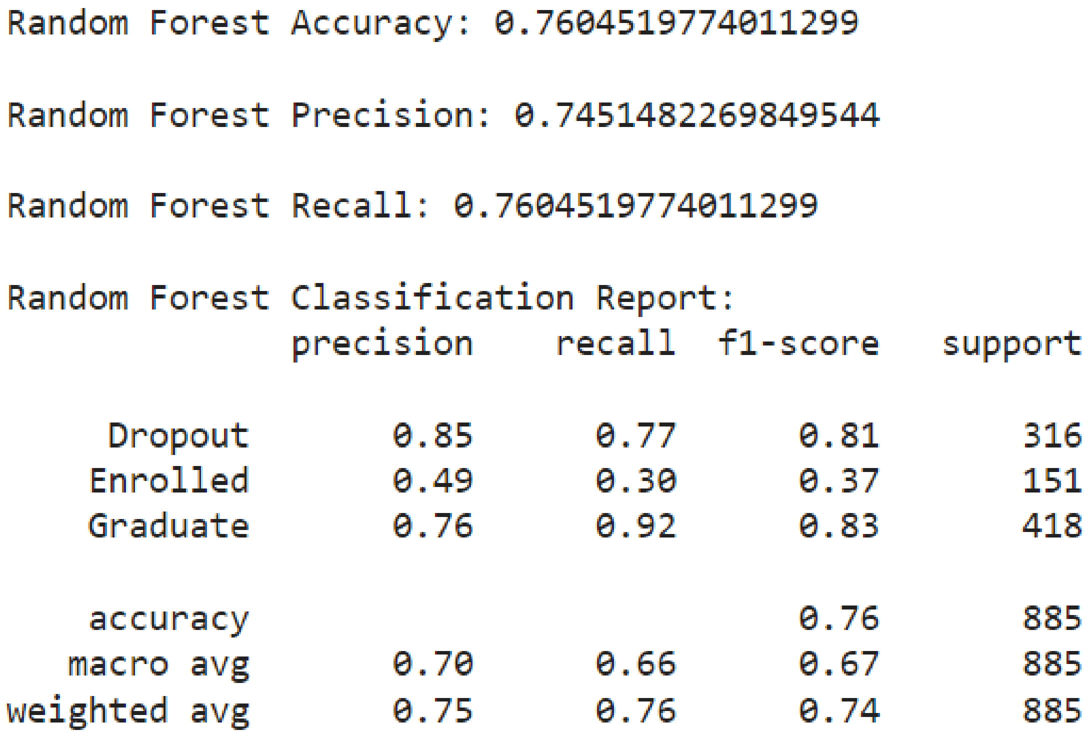


```
Accuracy of KNN: 0.6090395480225989

Evaluation of KNN Classifier with Neighbours = 5
              precision    recall  f1-score   support

     Dropout       0.66      0.56      0.60       316
    Enrolled       0.36      0.26      0.31       151
    Graduate       0.64      0.77      0.70       418

    accuracy                           0.61       885
   macro avg       0.55      0.53      0.54       885
weighted avg       0.60      0.61      0.60       885
```

K Nearest Mean Neighbour Comparison

Neighbour	KNN Value
1	KNN: 0.5649717514124294
2	KNN: 0.5288135593220339
3	KNN: 0.5988700564971752
4	KNN: 0.5853107344632769
5	KNN: 0.6090395480225989
6	KNN: 0.6033898305084746
7	KNN: 0.5966101694915255

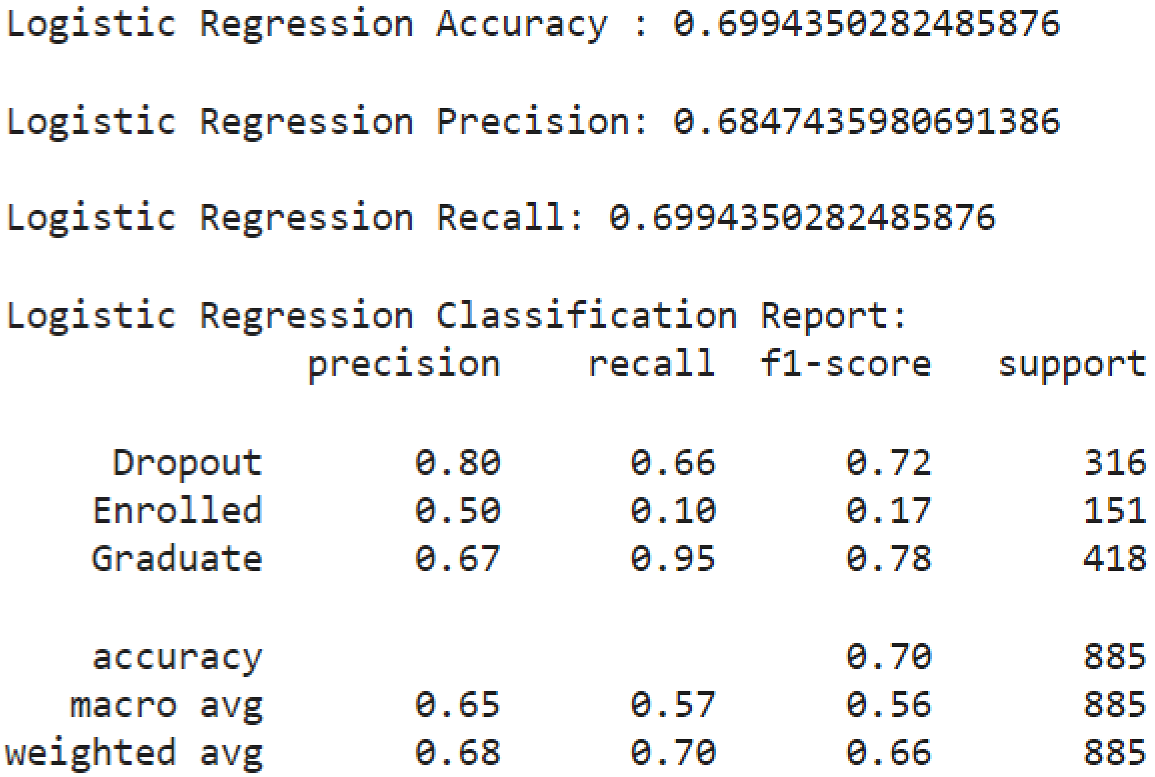
Random Forest Classifier



```
Random Forest Accuracy: 0.7604519774011299

Random Forest Precision: 0.7451482269849544

Random Forest Recall: 0.7604519774011299

Random Forest Classification Report:
              precision    recall  f1-score   support

     Dropout       0.85      0.77      0.81       316
    Enrolled       0.49      0.30      0.37       151
    Graduate       0.76      0.92      0.83       418

    accuracy                           0.76       885
   macro avg       0.70      0.66      0.67       885
weighted avg       0.75      0.76      0.74       885
```

Logistic Regression

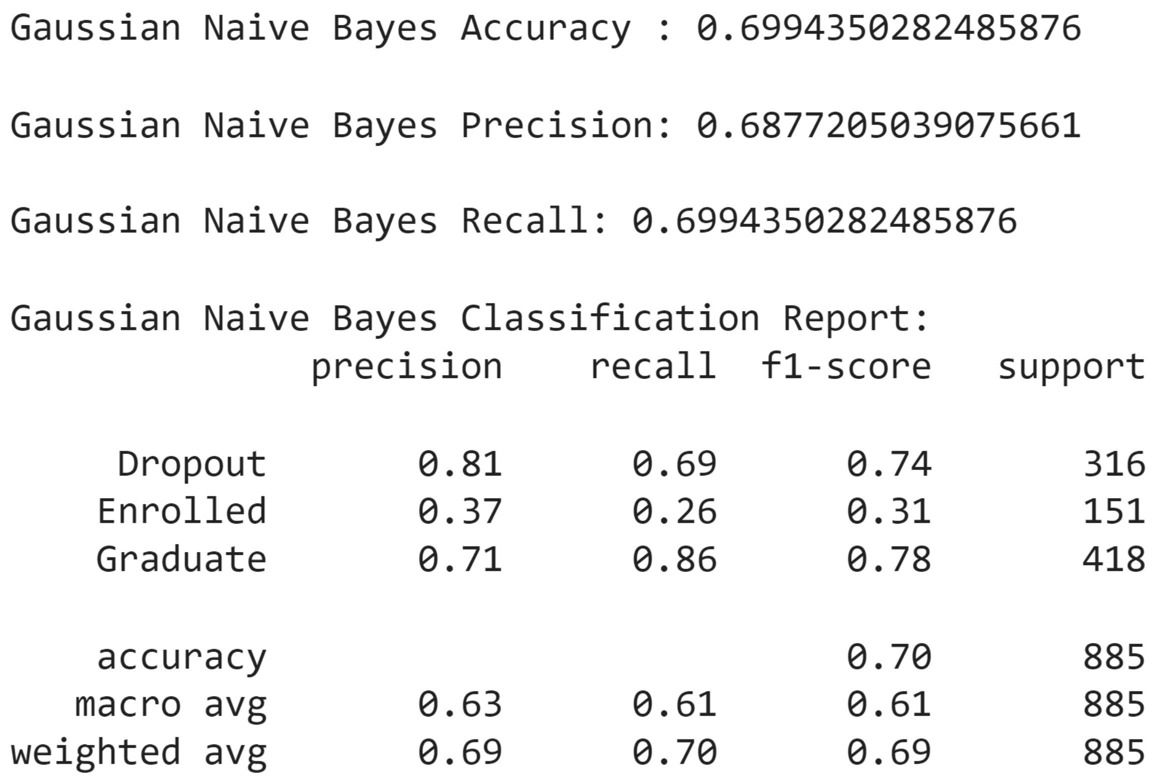
Logistic Regression Classifier



```
Logistic Regression Accuracy : 0.6994350282485876

Logistic Regression Precision: 0.6847435980691386

Logistic Regression Recall: 0.6994350282485876

Logistic Regression Classification Report:
              precision    recall  f1-score   support

     Dropout       0.80      0.66      0.72       316
    Enrolled       0.50      0.10      0.17       151
    Graduate       0.67      0.95      0.78       418

    accuracy                           0.70       885
   macro avg       0.65      0.57      0.56       885
weighted avg       0.68      0.70      0.66       885
```

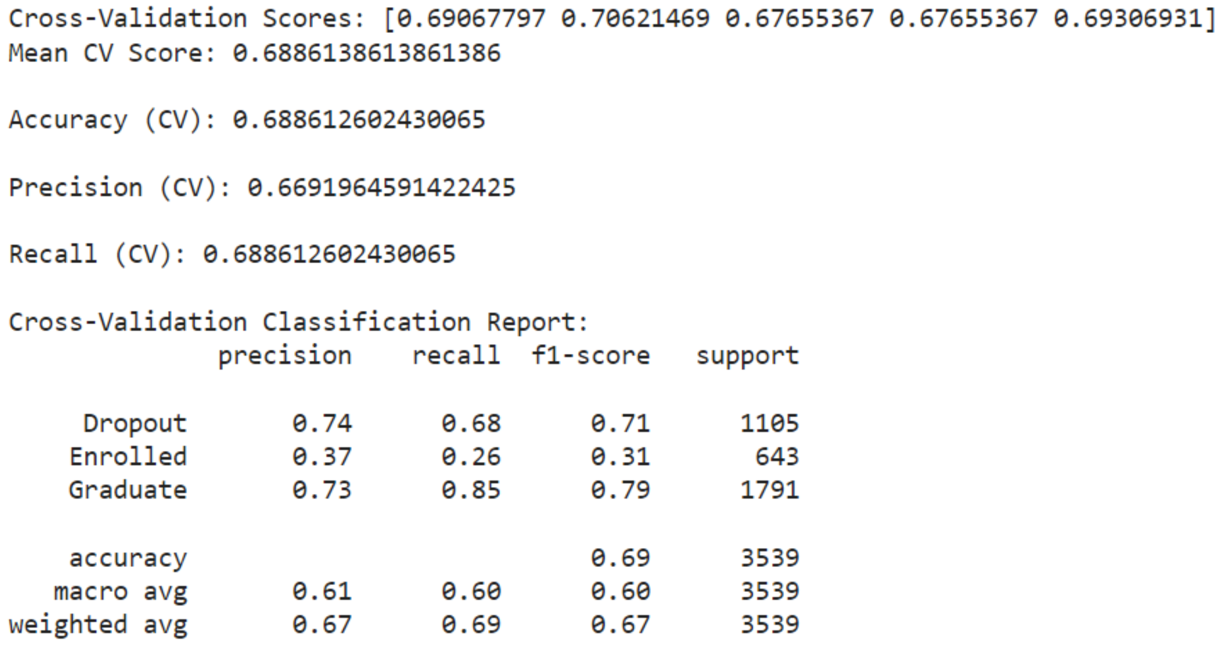
Gaussian Naïve Bayes

Gaussian Naïve Bayes



```
Gaussian Naive Bayes Accuracy : 0.6994350282485876

Gaussian Naive Bayes Precision: 0.6877205039075661

Gaussian Naive Bayes Recall: 0.6994350282485876

Gaussian Naive Bayes Classification Report:
              precision    recall  f1-score   support

     Dropout       0.81      0.69      0.74       316
    Enrolled       0.37      0.26      0.31       151
    Graduate       0.71      0.86      0.78       418

    accuracy                           0.70       885
   macro avg       0.63      0.61      0.61       885
weighted avg       0.69      0.70      0.69       885
```

Cross Validation Gaussian Naïve Bayes



```
Cross-Validation Scores: [0.69067797 0.70621469 0.67655367 0.67655367 0.69306931]
Mean CV Score: 0.6886138613861386

Accuracy (CV): 0.688612602430065

Precision (CV): 0.6691964591422425

Recall (CV): 0.688612602430065

Cross-Validation Classification Report:
              precision    recall  f1-score   support

     Dropout       0.74      0.68      0.71      1105
    Enrolled       0.37      0.26      0.31       643
    Graduate       0.73      0.85      0.79      1791

    accuracy                           0.69      3539
   macro avg       0.61      0.60      0.60      3539
weighted avg       0.67      0.69      0.67      3539
```

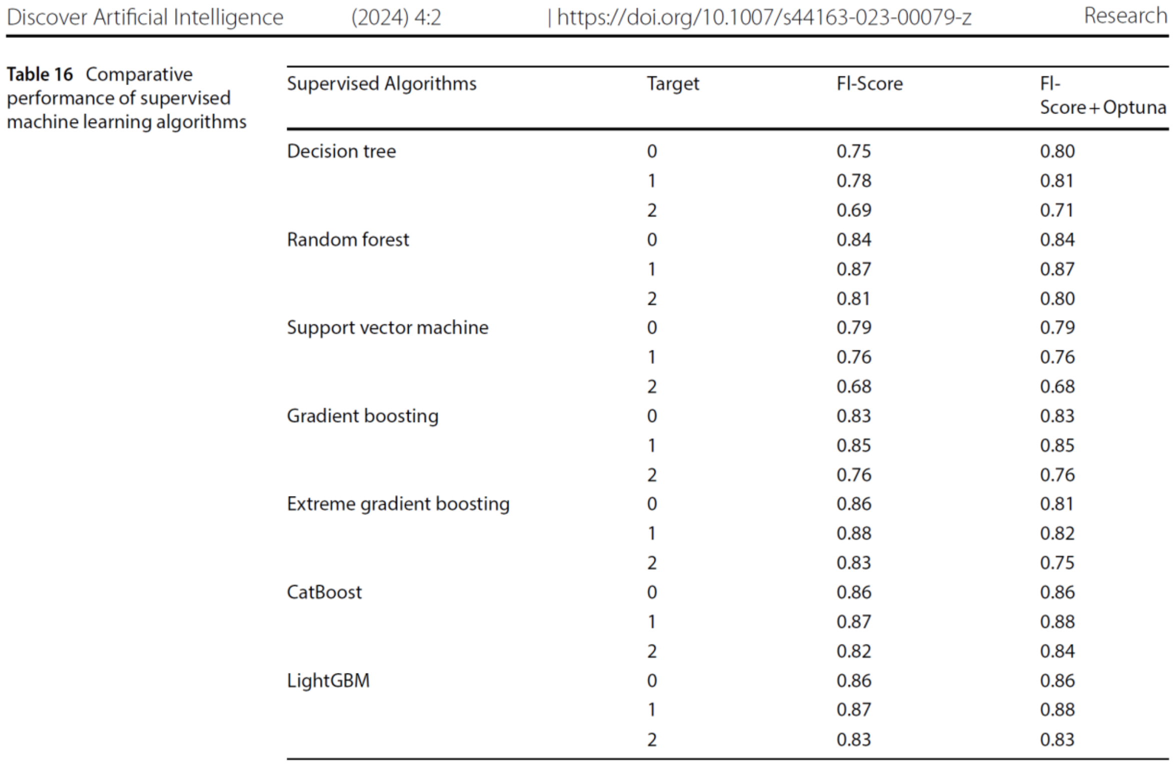
Results

Classification Results Comparison

Target	Decision Tree	KNN	Random Forest	Logistic Regression	Gaussian Naive Bayes	Cross-Validation Gaussian Naive Bayes
Precision						
Dropout	0.67	0.66	0.85	0.8	0.81	0.74
Enrolled	0.61	0.36	0.48	0.5	0.37	0.37
Graduate	0.75	0.64	0.76	0.67	0.71	0.73
Recall						
Dropout	0.6	0.56	0.76	0.66	0.69	0.68
Enrolled	0.88	0.26	0.3	0.1	0.26	0.26
Graduate	0.02	0.77	0.92	0.95	0.86	0.85
Accuracy	0.632	0.609	0.758	0.699	0.699	0.688

Results Comparison

Figure from Villar & de Andrade



Discover Artificial Intelligence (2024) 4:2 | https://doi.org/10.1007/s44163-023-00079-z Research

Table 16 Comparative performance of supervised machine learning algorithms

Supervised Algorithms	Target	FI-Score	FI-Score + Optuna
Decision tree	0	0.75	0.80
	1	0.78	0.81
	2	0.69	0.71
Random forest	0	0.84	0.84
	1	0.87	0.87
	2	0.81	0.80
Support vector machine	0	0.79	0.79
	1	0.76	0.76
	2	0.68	0.68
Gradient boosting	0	0.83	0.83
	1	0.85	0.85
	2	0.76	0.76
Extreme gradient boosting	0	0.86	0.81
	1	0.88	0.82
	2	0.83	0.75
CatBoost	0	0.86	0.86
	1	0.87	0.88
	2	0.82	0.84
LightGBM	0	0.86	0.86
	1	0.87	0.88
	2	0.83	0.83

The authors of the original dataset found that “circular units 2nd sem approved” was the strongest feature in determining wheter a student would dropout from their program (Realinho et al, 2022). The paper which used the same dataset but focused on SMOTE and gradient boosting techniques also found the same results (Villar & de Andrade, 2024). Their results as measured by the F1 score were higher than the results I obtained in my study. This is largely due in part to the lack of consistency with the precision and recall that I obtained for some of the methods used. Overall, it can be ascertained that both SMOTE and gradient boosting techniques helped address the size differences between the different classes resulting in an algorithm that was better equipped to parse the data.

Future Considerations

The mapping codes for nationality, occupation and course codes were completely wrong based on what was included in the paper and the numbering techniques used in the dataset. Although the authors have a paper that explains the codes and their meanings, it is different from what is contained in the dataset. Either they should update their dataset or update their paper so that the mapping codes make sense. It took a lot of deep diving and printing the unique values per column to classify variables into groups as I based my data and rational from the paper of the authors. The authors of the dataset also recommended utilizing feature importance techniques such as Permutation Feature Importance (Realinho et al, 2022). While I did not follow this exactly, for the Decision Tree and Apriori Method, groups such as course and parent's occupation were classified into smaller groups to make our understanding of the data more robust.

Another issue with the dataset is that the authors provided very limited explanations for the variables. For example, they fail to explain in detail the variables pertaining to a student's academic performance in their first and second semester. The maximum number of credits for the variable curricular units 1st sem enrolled is 26 but the mean is 6.270 and the median is 6. These values are so different from one another that it warrants an actual explanation. Likewise, no explanation is given for the grades in either semester with the maximum being 18.875 and the mean being 10.641 but no grading scale was provided. For other variables the authors do not explain what anything means. It is possible that some of the students applying from Portugal were not Portuguese by ethnicity and actually had a different ethnic origin. The authors do not explain dual citizenship or what categorized someone as Portuguese. More emphasis needs to be spent explaining the categories. For example, there was a stark difference in the percentages for the debtor category and tuition-up-to date category and the authors did not go into much detail. Another issue is the educational special needs category. The authors did not go into detail to explain what classified as special needs. This is important since the enrolled and graduation rate for those with and without special needs were nearly identical counterintuitive to what the average person would expect.

Classification Technique Issues

For the Decision Tree perhaps removing graduates from the dataset will aid in the prediction process and enable the classifier to be more accurate. For the Apriori algorithim, changing the classification and grouping categories to be significantly less for example grouping more occupations, or previous qualifications together might aid the classification to be more accurate. Another issue is the inclusion of graduate, dropout, and enrolled within the dataset. In future a dataset that studies students over a 5-year period and has only two variables dropout or graduate might fare significantly better with classification techniques. Such a dataset would best capture the essence of this study which is predicting students who will dropout of university. While this dataset was comprehensive, only gathering two years of data from students and not being able to determine if a student graduated in another program or university affects one's ability to extrapolate results.

With the techniques utilized each one fared differently with precision and recall of the three classes. For all techniques the precision was the lowest for the enrolled class. However, for the Decision Tree, the graduate precision was greater than the dropout precision; the opposite was true for all other techniques used. For the recall, the results varied greatly. For the Decision Tree, the recall for the graduate class was the lowest and the enrolled class was the highest. For all other classifiers, the recall was the greatest for the graduate class and lowest for the enrolled class. Had the target value been binary only having graduate or dropout, it could shed more light on the characteristics that are most similar amongst dropouts.

Suggestions

The first issue with the dataset is that the term dropout is too narrow and does not account for normal changes such as major or school changes that might occur within the average student's academic career. The other issue is that students were not followed for very long, only the grades from a student's first and second semester were calculated some students take longer to complete their studies and this is not sufficient time to collect and extrapolate results from students. Other issues with this dataset include the size of the classes and categorization of the data.

Future researchers should consider utilizing Synthetic Minority Oversampling Technique or SMOTE for short. SMOTE is a type of machine learning technique that aims to combat the issues that machine learning techniques face with imbalanced datasets. When datasets are imbalanced, some machine learning algorithms attempt to either under-sample the majority class or over-sample the minority class (Chawla et al, 2002). The SMOTE technique utilizes both features to improve performance on imbalanced datasets which aids in the prediction and classification of the minority class (Brownlee, 2021). The SMOTE technique works by synthesizing new examples from the minority class which is more helpful than simply oversampling the minority class as oversampling minority classes does not add any new information. The SMOTE technique developed by (Chawla et al, 2002) is the most widely used approach to synthesize new samples (Brownlee, 2021). Due to the underrepresentation of the graduate class, some machine learning models faced difficulty with predications and classification of this class. Other techniques that can develop predictive models and classification for datasets with imbalanced classes should be considered and utilized to have a more robust understanding of the data. The authors recommended using Balanced Random Forest, Roughly Balanced Baggin, Over-Bagging, or SMOTE-Bagging techniques (Realinho et al, 2022).

The other issue with this dataset is the lack of distinction across classes, for a student to be either a dropout or graduate they had to first be enrolled in classes. This made it difficult for certain algorithms to predict either the enrolled or dropout class. Recommendations include following students over n+1 years wehre n years is either the length of the program or the typical number of years it takes for students to complete the program. After n+1 years each student should be categorized as either a graduate or dropout and the results should be analyzed to determine the features and characteristics common to those who graduated or dropped out. Such a dataset would benefit greatly from following students over their entire academic career irrespective of if they changed majors or schools. By giving the students n+1 years and redefining the dropout to only include students who left school entirely, it would be significantly easier to make better predictions that can translate to changes in government spending or university funding and the likes.

Overall, there is importance in utilizing machine learning techniques to predict dropout rates amongst students. These results can help theorize a better study that integrates the aforementioned solutions to better model student attainment. In the future these results can be extrapolated to aid policymakers and government spending in a way that maximizes student graduation.

References

Addison, L., & Williams, D. (2023). Predicting student retention in higher education institutions (HEIs). *Higher Education, Skills and Work-Based Learning*, *13*(5), 865-885.
Aquines Gutiérrez, O., Hernández Taylor, D. M., Santos-Guevara, A., Chavarría-Garza, W. X., Martínez-Huerta, H., & Galloway, R. K. (2022). How the Entry Profiles and Early Study Habits Are Related to First-Year Academic Performance in Engineering Programs. *Sustainability*, *14*(22), 15400.

Ben Said, M., Hadj Kacem, Y., Algarni, A., & Masmoudi, A. (2023). Early prediction of Student academic performance based on Machine Learning algorithms: A case study of bachelor's degree students in KSA. *Education and Information Technologies*, 1-24.
Bhushan, M., Verma, U., Garg, C., & Negi, A. (2024). Machine Learning-Based Academic Result Prediction System. *International Journal of Software Innovation (IJSI)*, *12*(1), 1-14.
Brownlee, J. (2021, March 17). SMOTE Oversampling for Imbalanced Classification. *Machine Learning Mastery*. Retrieved from https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
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. Retrieved from: https://arxiv.org/abs/1106.1813.
IBM. (n.d.). k-nearest neighbors (KNN). Retrieved from https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20(KNN)%20algorithm%20is%20a%20non,used%20in%20machine%20learning%20today.
IBM. (n.d). Random Forest. Retrieved from https://www.ibm.com/topics/random-forest .
Jordan, L., Kostandini, G., & Mykerezi, E. (2012). Rural and urban high school dropout rates: Are they different?. *Journal of Research in Rural Education (Online)*, *27*(12), 1.
Liebowitz, D., Gonzalez, P., Hooge, E. & Lima, G. (2018). OECD Reviews of School Resources: Portugal 2018. *OECD*. Retrieved from https://www.oecd.org/education/school-resources-review/CombinedSummaryPortugal.pdf
Martins, M. V., Baptista, L., Machado, J., & Realinho, V. (2023). Multi-class phased prediction of academic performance and dropout in higher education. *Applied Sciences*, *13*(8), 4702.
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*(1), 5705.
Monteverde-Suárez, D., González-Flores, P., Santos-Solórzano, R., García-Minjares, M., Zavala-Sierra, I., de la Luz, V. L., & Sánchez-Mendiola, M. (2024). Predicting students' academic progress and related attributes in first-year medical students: an analysis with artificial neural networks and Naïve Bayes. *BMC Medical Education*, *24*(1), 74.
Paura, L., & Arhipova, I. (2014). Cause analysis of students' dropout rate in higher education study program. *Procedia-social and behavioral sciences*, *109*, 1282-1286.
Picot, G., & Hou, F. (2011). Preparing for success in Canada and the United States: The determinants of educational attainment among the children of immigrants. *Statistics Canada Analytical Branch Studies Working Paper*, (332).
Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. *Data*, *7*(11), 146.

Sandoval-Palis, I., Naranjo, D., Vidal, J., & Gilar-Corbi, R. (2020). Early dropout prediction model: A case study of university leveling course students. *Sustainability*, *12*(22), 9314.

Savona, L. A. (2010). *Predicting Student Success for Community College Students over a Ten-Year Period*. ProQuest LLC. 789 East Eisenhower Parkway, PO Box 1346, Ann Arbor, MI 48106.

Scikit-learn. (n.d.). Naive Bayes. In Scikit-learn: Machine Learning in Python (Version 0.24.2). Retrieved from https://scikit-learn.org/stable/modules/naive_bayes.html

Tamada, M. M., Giusti, R., & Netto, J. F. D. M. (2022). Predicting students at risk of dropout in technical course using LMS logs. *Electronics*, *11*(3), 468.

Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, *4*(1), 1-24.

Wikipedia contributors. (2022, January 18). Decision tree learning. In Wikipedia. Retrieved April 14, 2024, from https://en.wikipedia.org/wiki/Decision_tree_learning#Gini_impurity

Wikipedia contributors. (2024, March 20). Precision and recall. In Wikipedia. Retrieved April 14, 2024, from https://en.wikipedia.org/wiki/Precision_and_recall#:~:text=Precision%20can%20be%20seen%20as,irrelevant%20ones%20are%20also%20returned.