Katherine Kane

CS619 Data Mining

8/23/23

Student Dropout and Academic Success in Higher Education

1. Introduction

This project delves into the world of higher education data, aiming to uncover insightful patterns pertaining to student academic success and the likelihood to continue educational paths. The central objective is to predict whether undergraduate students will drop out, enroll, or graduate upon completing a full academic year. The report structure is organized into distinct sections, beginning with a detailed dataset description and attribute analysis. Subsequently, data preparation and preprocessing steps are executed to ensure robustness. An initial exploratory data analysis follows, shedding light on correlations and trends within the data. The methodology involves employing classification data mining techniques, including the Random Forest, J48, Random Tree, and IBk algorithms, to decode intricate relationships between student demographics, curricular unit performance, and enrollment statuses. These analyses cater to the classification categories: "Dropout," "Enrolled," and "Graduate." Notably, the Random Forest algorithm emerges as the top performer in accuracy. Through these findings, the report aims to uncover the most influential factors of predicting student educational trajectories.

2. Dataset Description

The dataset under analysis is titled "Predict students' dropout and academic success," and it is sourced from the UCI Machine Learning Repository. It encompasses a comprehensive range of information, including demographic data, socioeconomic indicators, enrollment details, and data at the end of both the initial and subsequent semesters. The dataset pertains to undergraduate

students enrolled during the academic years spanning from 2008/2009 through to 2018/2019. There are 35 attributes and 4424 records in the dataset. Key attributes include curricular units approved by semester, semester grades, if the tuition fees are up to date, if a scholarship is held, age at enrollment, if student is a debtor, and gender. Below is a table of all the attribute names and descriptions.

Attribute Name	Meaning
maritalstatus	Marital status of the student
applicationmode	Mode of application for the course
applicationorder	Order of application for the course
course	Course subject
daytime_eveningattendance	Course time
previousqualification	Previous educational qualification
previousqualification(grade)	Grade of the previous qualification
nacionality	Nationality of the student
mothersqualification	Education level of the student's mother
fathersqualification	Education level of the student's father
mothersoccupation	Occupation of the student's mother
fathersoccupation	Occupation of the student's father
admissiongrade	Admission grade
displaced	Whether the student is displaced
educationalspecialneeds	Educational special needs of the student
debtor	Whether the student is a debtor
tuitionfeesuptodate	Whether tuition fees are up to date
gender	Gender of the student
scholarshipholder	Whether the student holds a scholarship
ageatenrollment	Age of the student at the time of enrollment
international	Whether the student is an international student
curricularunits1stsem(credited)	Credited curricular units in the first semester

curricularunits1stsem(enrolled)	Enrolled curricular units in the first semester
curricularunits1stsem(evaluations)	Evaluated curricular units in the first semester
curricularunits1stsem(approved)	Approved curricular units in the first semester
curricularunits1stsem(grade)	Grade average in the first semester
curricularunits1stsem(withoutevaluations)	Curricular units without evaluations in the first semester
curricularunits2ndsem(credited)	Credited curricular units in the second semester
curricularunits2ndsem(enrolled)	Enrolled curricular units in the second semester
curricularunits2ndsem(evaluations)	Evaluated curricular units in the second semester
curricularunits2ndsem(approved)	Approved curricular units in the second semester
curricularunits2ndsem(grade)	Grades averages in the second semester
curricularunits2ndsem(withoutevaluations)	Curricular units without evaluations in the second semester
unemploymentrate	Unemployment rate at the time of enrollment
inflationrate	Inflation rate at the time of enrollment
gdp	Gross Domestic Product (GDP) at the time of enrollment
target	Outcome: Dropout, Enrolled, Graduate

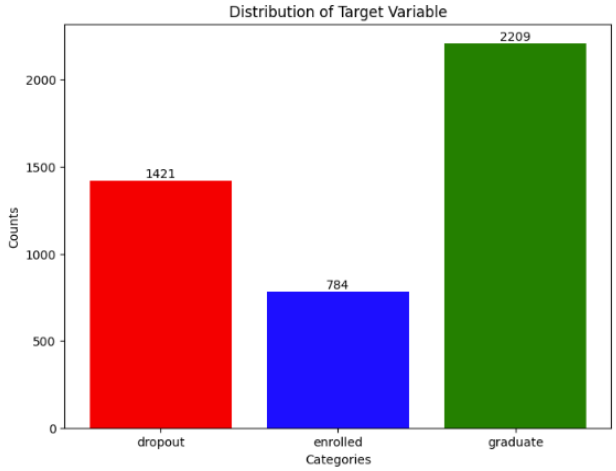
3. Data Preparation

During the process of preparing the data for analysis and mining, several steps were taken to ensure that the data was suitable for further exploration. The process started with the conversion of the initial CSV file format into the appropriate ARFF file. Upon examining the dataset, no missing values were found.

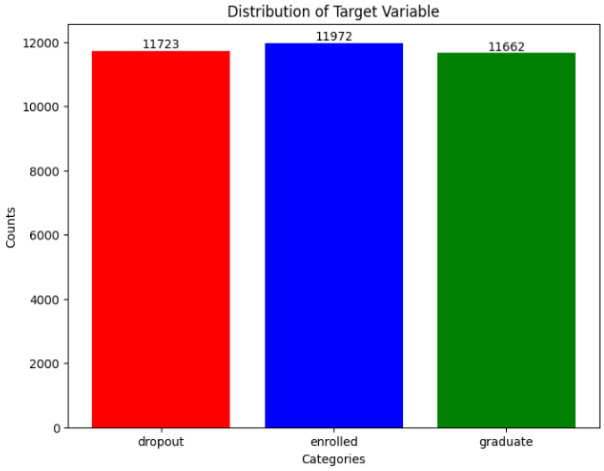
With the intention of managing the dimensionality of the dataset and reducing the risk of overfitting, a process of feature selection was initiated. Two ranking techniques, Information Gain Ranking and Correlation Ranking, were applied to the attributes. Comparing the rankings from both techniques, 9 attributes were selected for analysis as having the highest rankings of importance. The attributes are: curricularunits2ndsem(approved),

curricularunits1stsem(approved), curricularunits2ndsem(grade), curricularunits1stsem(grade), tuitionfeesuptodate, scholarshipholder, ageatenrollment, debtor, and gender.

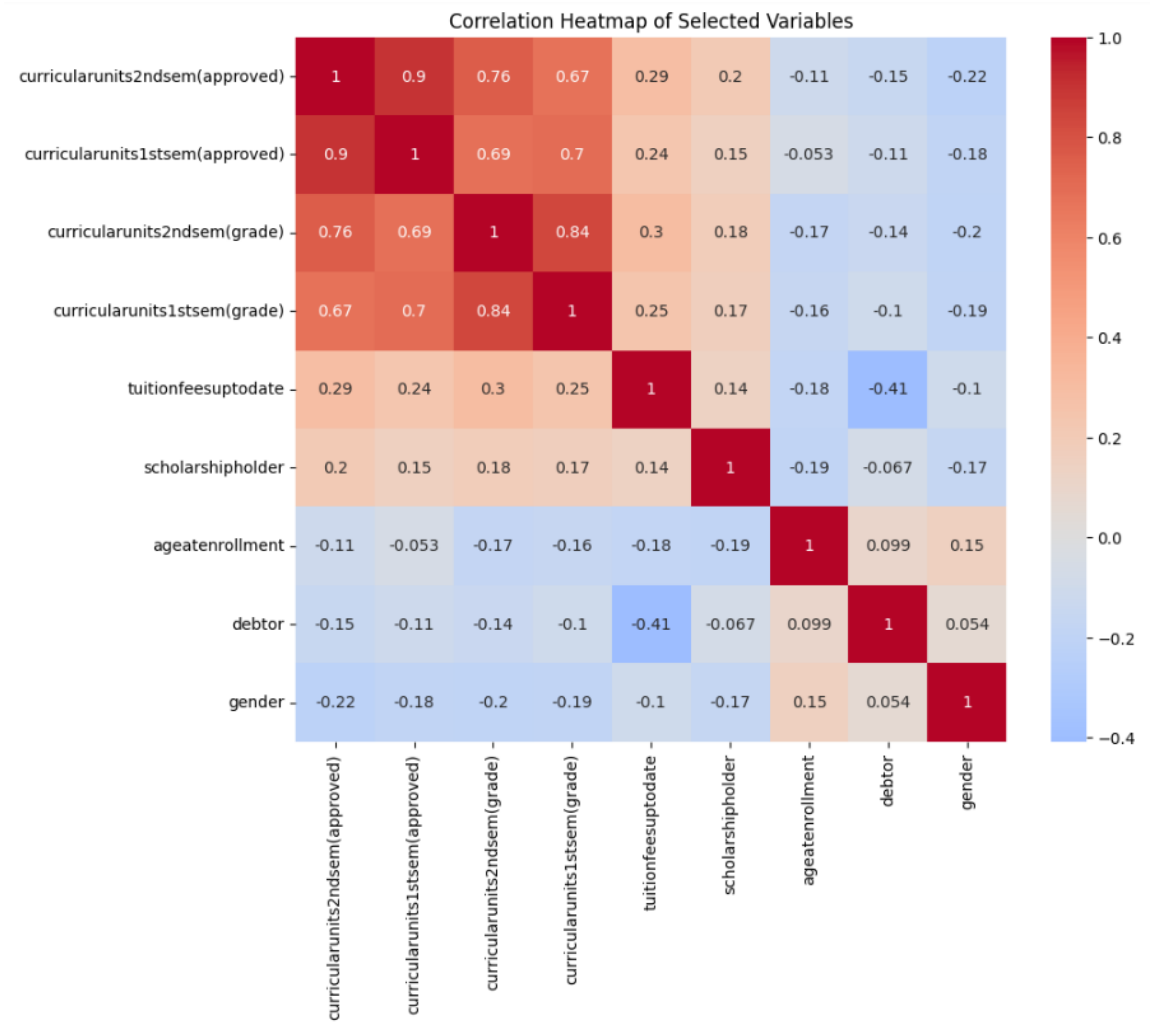
The graph below shows the distribution of the target variable. Additional histograms of the 9 attributes are located in the appendix. 32% of the students are classified as a dropout, 18% represent enrolled, and 50% in the graduate category. Looking at the distribution, there is a clear imbalance of classes.



To handle the imbalance, I used the oversampling technique to increase the number of instances in the minority classes. This generated synthetic samples to the dropout and enrolled classes to balance the distribution and allow for more accurate predictions. Below is the updated distribution of the classes.



Additionally, a heatmap was created to visually explore the relationships between the various variables in the dataset. This step was helpful in identifying potential multicollinearity that may be of concern. Particularly, the curricular units approved and curricular units grade variables indicated high correlations between the first semester and second semester records for each of those variables. In higher education, excelling in the first semester often translates to continued success in the following semesters, while poor performance is likely to persist through the progression of college.



Finally, the data was randomized and split into train and test files containing 80 percent and 20 percent of the instances, respectively.

4. Methodology

I conducted my data analysis utilizing a range of classification algorithms to derive predictions from the academic success dataset. The algorithms employed were Random Forest, J48, Random Tree, and K-Nearest Neighbors (KNN). The Random Forest algorithm merges multiple decision trees to improve predictive accuracy and mitigate overfitting. It constructs a forest of decision trees, using random subsets of features and bootstrapped datasets in order to reduce variance and enhance generalization (Source: Weka). Next, the Random Tree algorithm constructs decision tree ensembles, similar to Random Forest, to enhance predictive accuracy. Unlike Random Forest, it builds fully grown trees without bootstrapping or feature sampling (Source: Weka).

J48 employs the C4.5 algorithm to create classification models in the form of decision trees. By recursively dividing the dataset based on the attribute with the best information gain, J48 aims to achieve insights into feature importance and decision paths (Source: Weka). Finally, KNN is an instance-based algorithm that classifies instances based on the class labels of their k-nearest neighbors in the feature space. It calculates distances between instances to identify the closest neighbors and assigns the majority class among those neighbors as the predicted class. KNN is robust against noise, and adaptable to imbalanced datasets (Source: Weka). In summary, the selection of Random Forest, J48, Random Tree, and KNN for data analysis was based on their distinct advantages and their potential to address various dataset challenges.

5. Results

The Random Forest algorithm performed the best out of all the classifiers. The model was constructed with 100 iterations, and demonstrated strong performance, achieving an accuracy of approximately 94.64% on the test set. This indicates the model's proficiency in

accurately predicting a student's class for the majority of instances in the test data. Furthermore, both the mean absolute error (0.0934) and root mean squared error (0.1827) are at low levels. The table below presents the confusion matrix, which provides an understanding of the model's performance across individual classes (Dropout, Enrolled, and Graduate). The Random Forest model excelled across all three classes, particularly in effectively predicting the Dropout and Graduate classes.

	Predicted Dropout	Predicted Enrolled	Predicted Graduate
Actual Dropout	2213	70	57
Actual Enrolled	82	2266	54
Actual Graduate	69	47	2213

Next, I used the J48 decision tree algorithm with unpruned trees. The model led to a tree with 1481 leaves and a total size of 2961 nodes. The root node chosen was if the number of approved curricular units in the second semester is equal to or less than 4, then it proceeded to new choices. The next conditions were if tuition fees are not paid, and the number of approved curricular units in the first semester is equal to or less than 5, the model proceeds to assess further conditions. If the age at enrollment is equal to or less than 18.5, the model evaluates the student's debtor status. If the student is not a debtor, the predicted class is Dropout with a confidence level of 98.0%. The decision path proceeded on with a series of conditions to contribute to the model's predictions.

The J48 model demonstrated good accuracy of approximately 89.51% on the test set. This indicates that the model accurately predicted the class labels for a majority of instances in the test data. Additionally, both the mean absolute error (0.0832) and root mean squared error

(0.2471) remained relatively low, implying low prediction errors. The J48 model's unpruned decision tree structure provides a transparent insight into the decision-making process.

Similarly, the Random Tree classifier results in a decision tree that provides insights into the classification process. The first decision made by the model is if the tuition fees are up to date. The next decision is if the number of approved curricular units in the first semester is less than 5. Other variables include gender, grade in the first semester, and age at enrollment before determining a class. The RandomTree classifier achieved an accuracy of 88.6296% on the test set, and had a low mean squared error (0.0787) and root mean squared error (0.2686). Comparing the random tree and J48 models, both determine tuition fees up to date, curricular units in the first semester, and age at enrollment as key attributes in the prediction of students' academic success.

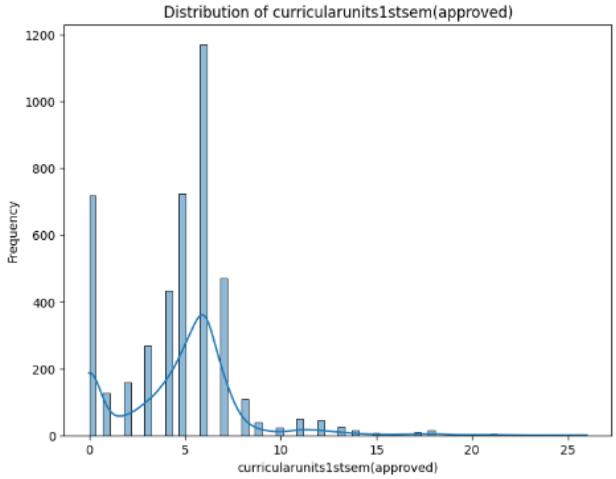
Lastly, the IBk classifier was used with k=3 and Euclidean distance. It employs a distance-based approach, classifying instances by comparing them to their k-nearest neighbors. The model achieved an accuracy of 87.1447% on the test set. Despite its reasonable performance, the IBk classifier did not achieve as high an accuracy as the other classifiers previously evaluated. This could be attributed to its sensitivity to the distribution of data. The presented confusion matrix below displays the model's classification results for each class.

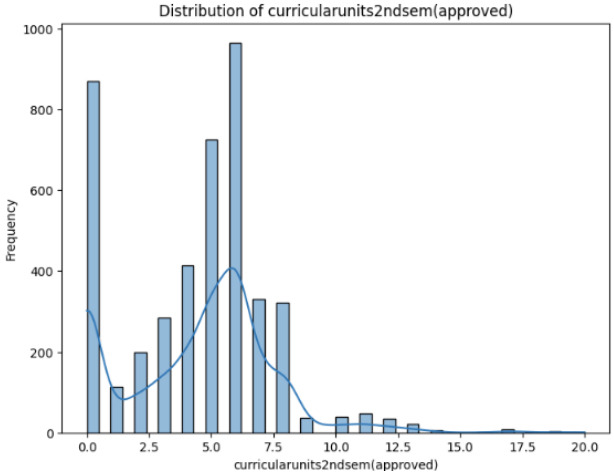
	Predicted: Dropout	Predicted: Enrolled	Predicted: Graduate
Actual: Dropout	2062	200	78
Actual: Enrolled	207	2103	92
Actual: Graduate	147	185	1997

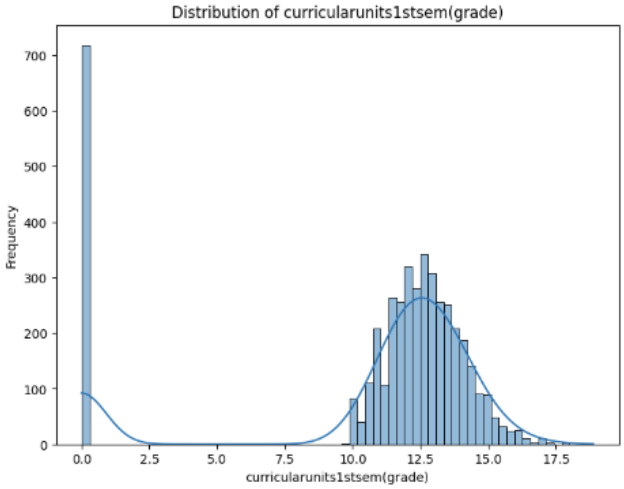
6. Conclusion

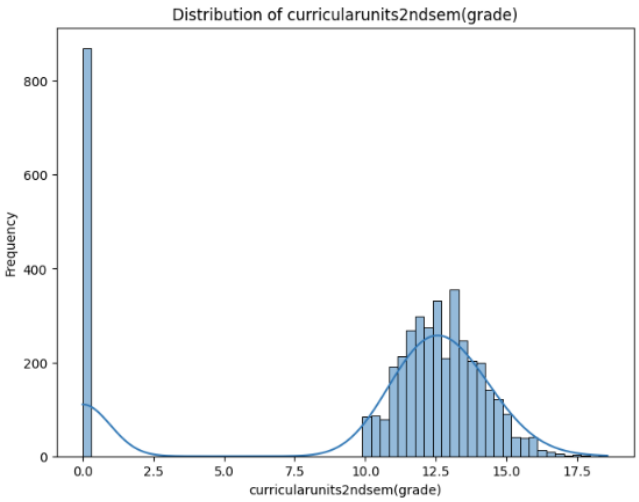
In summary, the report's analysis of various classification algorithms for predicting student outcomes yielded insightful findings that have significant implications for higher education institutions. The Random Forest algorithm emerged as the most effective, achieving a high accuracy of 94.64% on the test set. The classifier seems to predict early identification of at-risk students well, and the low mean absolute error further suggests its reliability. The J48 decision tree algorithm achieved an accuracy of 89.51%, unraveling key insights into its decision-making process. Similarly, the Random Tree classifier showcased an accuracy of 88.63%, indicating the importance of attributes such as tuition fees, curricular units, and age in forecasting academic success. Furthermore, the IBk classifier's accuracy of 87.14%, while reasonable, indicates its comparatively lower performance in relation to the other classifiers. Collectively, these findings equip higher education institutions with valuable information to create student support systems that ensure the success of a diverse range of students.

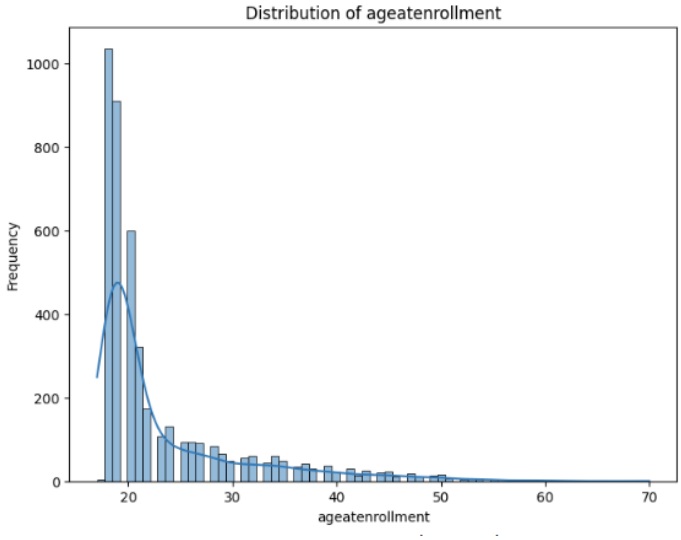
7. Appendix

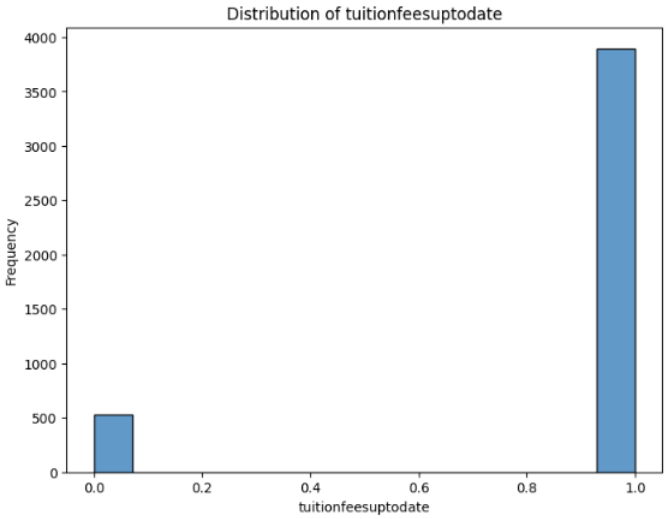

Distribution of curricularunits1stsem(approved)
Frequency
curricularunits1stsem(approved)

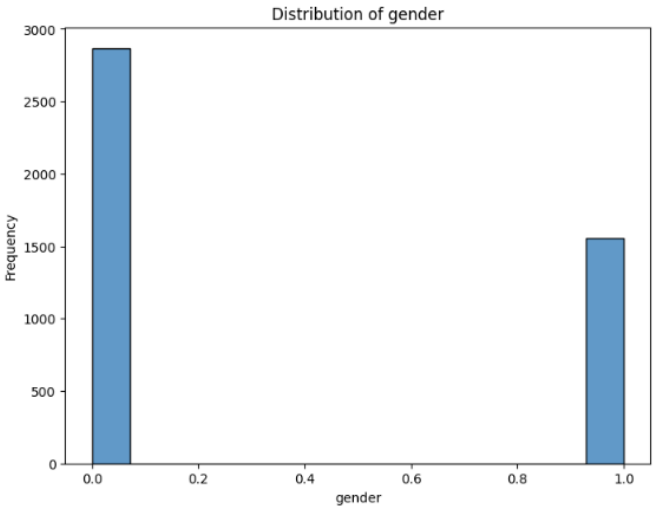

Distribution of curricularunits2ndsem(approved)
Frequency
curricularunits2ndsem(approved)

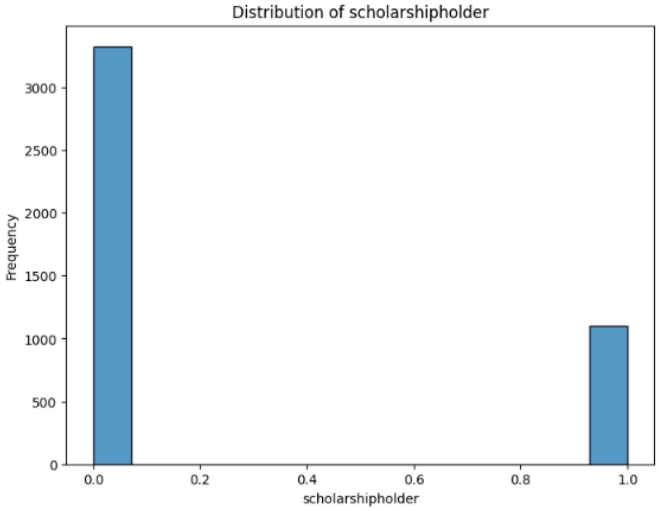

Distribution of curricularunits1stsem(grade)
Frequency
curricularunits1stsem(grade)

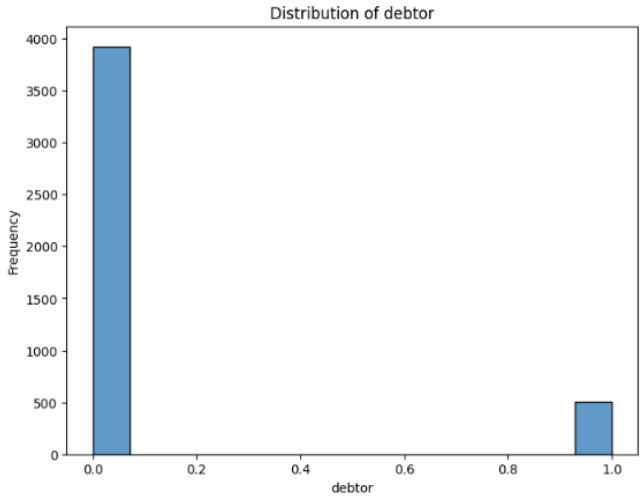

Distribution of curricularunits2ndsem(grade)
Frequency
curricularunits2ndsem(grade)


Distribution of ageatenrollment
Frequency
ageatenrollment


Distribution of tuitionfeesuptodate
Frequency
tuitionfeesuptodate
0
500
1000
1500
2000
2500
3000
3500
4000
0.0
0.2
0.4
0.6
0.8
1.0


Distribution of gender
Frequency
gender
0
500
1000
1500
2000
2500
3000
0.0
0.2
0.4
0.6
0.8
1.0


Distribution of scholarshipholder
Frequency
scholarshipholder
0
500
1000
1500
2000
2500
3000
0.0
0.2
0.4
0.6
0.8
1.0


Distribution of debtor
Frequency
debtor
0
500
1000
1500
2000
2500
3000
3500
4000
0.0
0.2
0.4
0.6
0.8
1.0