Final Report

Evaluating the Efficacy of Various Machine Learning Architectures in Predicting Student Dropout and Academic Success

Avery Li and Rem Turatbekov

Thomas Jefferson High School for Science and Technology
Machine Learning 1 TJ AV

Submitted in Partial Completion of Q1 Project, as a Continuation of Intermediate Report Submitted Previously

Director: Selma Yilmaz

Submitted: October 21, 2024

Contents

1 Introduction:

1.1 Intellectual Merit

Our research stems from an ongoing discussion of the prevalence of dropouts in recent years, heightened by the COVID-19 pandemic. A student's success in the workforce is largely dependent on their ability to achieve a higher education, especially in our time of such an oversaturated job market. Thus, it is necessary to analyze the potential of dropouts to ensure the success of students at various levels of education. We gather data from various higher education institutions regarding the performance of students in their respective schools, taking into account absenteeism, grades, and other factors to try and predict a student's dropout and academic success.

1.2 Project Statement

It is imperative that we are able to predict student dropout to plan appropriate intervention if necessary. In this project, we aim to construct a categorical classification machine learning model that learns to classify a student as one of three statuses—dropout, enrolled, and graduate. After taking steps to clean the data and sort it down to its core attributes, we hope to construct a model to be able to map a student's statistics to their academic success at a certain institution. Several factors will be considered, including a student's grades, the days they were absent, age of enrollment, as well as external factors, including scholarship and financial status.

2 Data Report:

Data was taken from UCI, a website dedicated to various datasets. The specific dataset was funded by program SATDAP, and data was collected across various higher education institutions, which was acquired through various disjoint databases and compiled into a comprehensive dataset: https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

2.1 Description of Dataset

The dataset has a total of 36 attributes, ranging from information about the student to information about their personal background. A comprehensive list of all the attributes as well as their meaning, if necessary, follows. The dataset includes Marital Status, Application mode (type of application to the school), Application order, Course, Daytime/evening attendance, Previous qualification, Previous qualification (grade), Nationality, Mother's qualification, Father's qualification, Mother's occupation, Father's occupation, Admission grade, Displaced, Educational special needs, Debtor, Tuition fees up to date, Gender, Scholarship holder, Age at enrollment, International, Curricular units 1st sem (credited), Curricular units 1st sem (enrolled), Curricular units 1st sem (evaluations), Curricular units 1st sem (grade), Curricular units 1st sem

(approved), Curricular units 1st sem (without evaluations), Curricular units 2nd sem (credited), Curricular units 2nd sem (enrolled), Curricular units 2nd sem (evaluations), Curricular units 2nd sem (approved), Curricular units 2nd sem (grade),Curricular units 2nd sem (without evaluations), Unemployment rate, Inflation rate, GDP, and finally our class attribute, Target. A single instance can therefore be more aptly described as a student-situation pair (accounting for GDP, unemployment rate, etc…), rather than just a student. There are a total of 4424 instances of these student-situation pairs in this dataset, each representative of a student at some institution.

The distribution of the class is fairly uniform, with around 33% of each class type. There are no missing values in this dataset, but we have received approval to continue on with our project in spite of this.

2.2 Data Preprocessing

First we would need to normalize all our data points. This proves to be somewhat of a difficult task, as some data is nominal qualitative, disguised as ordinal qualitative data, and so figuring out the relative significance, or lack thereof, of each data. Another problem is figuring out which attributes relate to each other. The grades earned in a semester and the relationship between evaluations and without evaluations might be proportional, so we should figure out these relationships in order to reduce our data size. The next step would be figuring out which attributes have low correlation, and select an appropriate amount of attributes from there.

3 Data Mining:

3.1 Current Issues

Issue 1: The data was initially delimited by semicolons instead of commas, making the data unreadable in the CSV format. Code is detailed in Appendix B.1.

Issue 2: Many of the attributes in this dataset are inherently nominal qualitative (Nationality, Marital Status, Mother qualifications) with seemingly random integer numbers assigned to each unique value of each attribute, creating a false sense of order that the original data does not imply. We resolve the issue by understanding the meaning, then fixing the values as strings they corresponded to. Code is detailed in Appendix B.2.

Issue 3: Additional classification is taxing on the model itself, and is not our intended goal. The dataset has 3 classes total—dropout, graduate, enrolled—with the latter two representing a non-dropout. Our goal is not to predict the distinction between graduate and enrolled; therefore, we merge the two classes into a single “Non_Dropout” class, making our model a binary classification problem instead. The proportion is now relatively balanced, but the difference in

size between the two classes must be accounted for when splitting the dataset. Code is detailed in Appendix B.3.

3.2 Discarding Columns

The easiest columns to remove were those of derived attributes or ones that clearly had no impact on dropout rate. The attribute “International” in our case refers to “Non-portuguese”, with a one-to-one ratio to Nationality. Since the Nationality attribute provides more information, we can remove the International column.

3.3 Normalizing Quantitative Data

Much of our data is nominal, rather than ordinal data, so we disregard any qualitative implications they may have otherwise had. However, other qualitative variables need to be normalized. Those attributes are Admission grade, Age at enrollment, all attributes with the prefix Curricular units, Unemployment rate, GDP, and Inflation rate. All were normalized to a scale from 0 to 5 through decimal scaling, and all the rest were normalized to the same scale through simple normalization. Code is detailed in Appendix B.4.

3.4 Weka Cross Validation

Our data is not very large, numbering just over 4,000 instances after preprocessing. Instead of splitting between training, validation, and test sets, we use Weka’s stratified 10-fold cross validation to measure the performance of our classification models by taking the average of the accuracy obtained from each fold. While this algorithm is computationally expensive, it does not waste much data, as can be the case when fixing an arbitrary validation set, and is superior to the repeated holdout method since every instance is guaranteed to be used for both training and testing.

4 Attribute Selection:

Our dataset is left with 36 attributes after preprocessing. Each of the following attribute selection algorithms are performed on Weka, with the exception of pure intuition, which has its own subsection. The chosen attribute selection algorithms are CorrelationAttributeEval, GainRatioAttributeEval, InfoGainAttributeEval, and OneRAttributeEval. A detailed description of each follows. A description of how to perform each evaluation is provided in Appendix E.1.

4.1 CorrelationAttributeEval

The Pearson Correlation Coefficient between an attribute x and target y is calculated by:

$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

The value r calculated by this formula ranges from -1 to 1, inclusive, with 0 being the worst possible outcome representing no correlation. We chose to select all attributes with an absolute r-value greater than 0.075, leaving us with 21 attributes. The full analysis is found in Appendix C.1 and selected attributes are shown below:

```
Curricular units 2nd sem (grade)
Curricular units 2nd sem (approved)
Curricular units 1st sem (grade)
Curricular units 1st sem (approved)
Tuition fees up to date
Age at enrollment
Scholarship holder
Debtor
Gender
Curricular units 2nd sem (evaluations)
Curricular units 2nd sem (enrolled)
Previous qualification
Application mode
Curricular units 1st sem (enrolled)
Marital status
Displaced
Admission grade
Curricular units 1st sem (evaluations)
Daytime/evening attendance
Curricular units 2nd sem (without evaluations)
Previous qualification (grade)
```

4.2 InfoGainAttributeEval

Information gain is a metric given by the entropy of certain subsets of data in a decision table. Entropy of any dataset is given by:

$$H = -(\sum p_i \log_2 p_i)$$

where p_i is the proportion of the label to the dataset.

Information gain is then calculated by the proportion difference in entropies:

$$IG_{split} = H - (\sum \frac{|D_j|}{|D|} \times H_j)$$

This formula implies that the information gain is the difference between the entropy of the output class and the weighted entropies of a single class decision tree; the higher the information gain, the better the predictor. We decide to choose all attributes with an IG_{split} value of greater than .025, leaving 20 attributes in the dataset. The full analysis is found in Appendix C.2 and selected attributes are shown below:

```
Curricular units 2nd sem (approved)
```

```
Curricular units 2nd sem (grade)
Curricular units 1st sem (approved)
Curricular units 1st sem (grade)
Tuition fees up to date
Curricular units 2nd sem (evaluations)
Curricular units 1st sem (evaluations)
Age at enrollment
Application mode
Course
Scholarship holder
Previous qualification (grade)
Debtor
Mother_occupation
Curricular units 2nd sem (enrolled)
Previous qualification
Gender
Father_qualification
Father_occupation
Mother_qualification
```

4.3 GainRatioAttributeEval

Gain ratio is an attempt at lessening certain biases by using the information gain metric mentioned in section 4.2; We introduce a new metric, Intrinsic Information:

$$II = -(\sum \frac{|D_j|}{|D|} \times \log_2 (\frac{|D_j|}{|D|}))$$

or, in other words, the entropy of the proportions of the subcategories themselves.

Given this metric, the gain ratio is calculated as

$$GR = \frac{IG_{split}}{II}$$

We choose all attributes with a GR value greater than .14, leaving us with 21 attributes. The full analysis is found in Appendix C.3 and selected attributes are shown below:

```
Tuition fees up to date
Curricular units 2nd sem (grade)
Curricular units 2nd sem (approved)
Curricular units 1st sem (approved)
Curricular units 1st sem (grade)
Debtor
Scholarship holder
Curricular units 2nd sem (evaluations)
Age at enrollment
Curricular units 1st sem (evaluations)
Gender
```

```
Previous qualification
Application mode
Curricular units 1st sem (without evaluations)
Curricular units 2nd sem (without evaluations)
Admission grade
Curricular units 2nd sem (enrolled)
Previous qualification (grade)
Nacionality
Curricular units 1st sem (enrolled)
Marital status
```

4.4 OneRAttributeEval

OneR is an algorithm designed to choose the best attribute with a one-to-one ratio of the class labels. The algorithm is as follows:

```
for each attribute:
      for each unique value in attribute:
            count labels for instances with the unique value
            determine most frequent label
            assign value to label
      compute error rate given rule chosen
choose rule with lowest error rate
```

We choose all attributes with resulting value greater than 68%, leaving us with 19 attributes. The full analysis is found in Appendix C.4 and selected attributes are shown below:

```
Curricular units 2nd sem (approved)
Curricular units 2nd sem (grade)
Curricular units 1st sem (approved)
Curricular units 1st sem (grade)
Tuition fees up to date
Curricular units 2nd sem (evaluations)
Curricular units 1st sem (evaluations)
Debtor
Application mode
Age at enrollment
Mother_occupation
Previous qualification
Previous qualification (grade)
Father_qualification
Mother_qualification
Father_occupation
Curricular units 1st sem (enrolled)
Curricular units 2nd sem (enrolled)
Course
```

4.5 Intuition

Intuitively, a student's grades in the semester is likely the largest driving factor for whether they drop out. Thus, we select 12 Curricular Unit attributes. Additionally, we consider financial and

personal factors—special needs, debtor, tuition fee up to date—all to be important. Selected attributes are shown below:

```
Curricular units 1st sem (credited)
Curricular units 1st sem (enrolled)
Curricular units 1st sem (evaluations)
Curricular units 1st sem (without evaluations)
Curricular units 1st sem (approved)
Curricular units 1st sem (grade)
Curricular units 2nd sem (credited)
Curricular units 2nd sem (enrolled)
Curricular units 2nd sem (evaluations)
Curricular units 2nd sem (without evaluations)
Curricular units 2nd sem (approved)
Curricular units 2nd sem (grade)
Educational special needs
Debtor
Tuition fees up to date
Displaced
Gender
Age at enrollment
Scholarship holder
Daytime/evening attendance
```

5 Classifier Models:

After choosing the attributes, we run various models on the dataset in Weka to test for efficiency. The four classifier models we selected are J48, NaiveBayes, Logistic, and Random Forest. A brief description of each model follows.

5.1 J48

J48 is a decision tree algorithm that uses entropy to measure information gain and subsequently construct a decision tree. At each iteration, the algorithm splits the tree using the attribute with the greatest information gain, typically finding a near-optimal classifier.

5.2 NaiveBayes

NaiveBayes is a statistical predictor operating under the Bayes' Theorem:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

The probability for a certain attribute X with unique values x_i for a class label C_i is given by

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

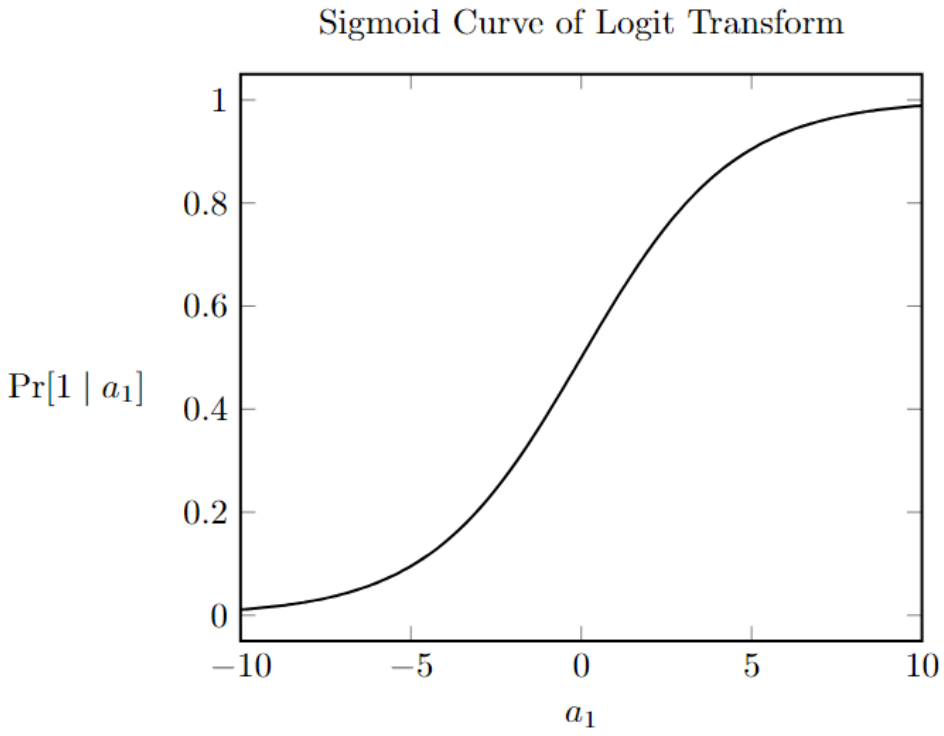
The model uses this value to determine the correct probability distribution for prediction.

5.3 Logistic Regression

The idea is to make linear regression also produce probabilities. Logistic Regression directly estimates class probabilities for both continuous and discrete data, but is typically used for binary classifications; often, a threshold will be set where a prediction of greater than 50% probability results in a positive classification. The algorithm uses a weighted linear sum embedded in a multidimensional logit transform. The probability is given by:

$$Pr[1|a_1, a_2, \ldots, a_k] = \frac{1}{1 + \exp\left(-(w_0 + w_1 a_1 + \ldots + w_k a_k)\right)}$$

Considering only one dimension, or one variable, a_1, the resulting graph of the output $Pr[1|a_1]$ ranges from 0 to 1 and takes form of an S-shape characteristic of the sigmoid function:



The basis of logistic regression is to choose weights to maximize the log-likelihood, which measures how well the model predicts the actual outcomes in the training data. The function to be maximized is shown below:

$$\sum_{i=1}^{n} \left((1 - x^{(i)}) \log \left(1 - \Pr\left[1 \mid a_1^{(1)}, a_2^{(2)}, \ldots, a_k^{(k)} \right] \right) + x^{(i)} \log \left(\Pr\left[1 \mid a_1^{(1)}, a_2^{(2)}, \ldots, a_k^{(k)} \right] \right) \right)$$

5.4 Random Forest

Random Forest instantiates a number of classification trees and uses the output of each to determine a total classification for the model. The trees are randomly generated and equally weighted so that a single tree cannot dominate the entire forest. The trees are independent of each other, and are collectively used to evaluate the output of the model.

6 Results:

The results of our analysis are shown. Each subsection is a combination of attribute selection/classifier model pair, evaluated on accuracy and other metrics through 10-fold cross validation. Result output is pasted directly from Weka. The steps taken are detailed in section E.2.

6.1 CorrelationAttributeEval + J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3821               86.3698 %
Incorrectly Classified Instances       603               13.6302 %
Kappa statistic                          0.6745
Mean absolute error                      0.1998
Root mean squared error                  0.3433
Relative absolute error                 45.8262 %
Root relative squared error             73.5169 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area
Dropout        0.712    0.065    0.839      0.712   0.770      0.679    0.848     0.742
Non_Dropout    0.935    0.288    0.873      0.935   0.903      0.679    0.848     0.871
WeightedAvg    0.864    0.216    0.862      0.864   0.860      0.679    0.848     0.830

=== Confusion Matrix ===

    a    b   <-- classified as
 1012  409 |    a = Dropout
  194 2809 |    b = Non_Dropout
```

6.2 CorrelationAttributeEval + NaiveBayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3701               83.6573 %
```

```
Incorrectly Classified Instances       723               16.3427 %
Kappa statistic                          0.615
Mean absolute error                      0.1669
Root mean squared error                  0.3829
Relative absolute error                 38.261  %
Root relative squared error             82.0077 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.695   0.097   0.773     0.695  0.732     0.617 0.872    0.771
Non_Dropout 0.903   0.305   0.862     0.903  0.882     0.617 0.872    0.916
WeightedAvg 0.837   0.238   0.834     0.837  0.834     0.617 0.872    0.869

=== Confusion Matrix ===

    a    b   <-- classified as
  988  433 |    a = Dropout
  290 2713 |    b = Non_Dropout
```

6.3 CorrelationAttributeEval + Logistic

```
=== Stratified cross-validation ===
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3848               86.9801 %
Incorrectly Classified Instances       576               13.0199 %
Kappa statistic                          0.6868
Mean absolute error                      0.1918
Root mean squared error                  0.3127
Relative absolute error                 43.976  %
Root relative squared error             66.9732 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.709   0.054   0.861     0.709  0.778     0.693 0.909    0.856
Non_Dropout 0.946   0.291   0.873     0.946  0.908     0.693 0.909    0.933
WeightedAvg 0.870   0.215   0.869     0.870  0.866     0.693 0.909    0.908

=== Confusion Matrix ===

    a    b   <-- classified as
 1007  414 |    a = Dropout
  162 2841 |    b = Non_Dropout
```

6.4 CorrelationAttributeEval + RandomForest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3829               86.5506 %
Incorrectly Classified Instances       595               13.4494 %
Kappa statistic                          0.6784
Mean absolute error                      0.209
Root mean squared error                  0.3211
Relative absolute error                 47.926  %
Root relative squared error             68.7733 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.713   0.062   0.844     0.713  0.773     0.683 0.906    0.857
Non_Dropout    0.938   0.287   0.873     0.938  0.904     0.683 0.906    0.939
WeightedAvg    0.866   0.215   0.864     0.866  0.862     0.683 0.906    0.913

=== Confusion Matrix ===

    a    b   <-- classified as
 1013  408 |    a = Dropout
  187 2816 |    b = Non_Dropout
```

6.5 InfoGainAttributeEval + J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3811               86.1438 %
Incorrectly Classified Instances       613               13.8562 %
Kappa statistic                          0.666
Mean absolute error                      0.2099
Root mean squared error                  0.3369
Relative absolute error                 48.1292 %
Root relative squared error             72.1505 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.692   0.058   0.849     0.692   0.762    0.673 0.865    0.788
Non_Dropout    0.942   0.308   0.866     0.942   0.902    0.673 0.865    0.892
WeightedAvg    0.861   0.228   0.860     0.861   0.857    0.673 0.865    0.859

=== Confusion Matrix ===

    a    b   <-- classified as
```

```
  983  438 |    a = Dropout
  175 2828 |    b = Non_Dropout
```

6.6 InfoGainAttributeEval + NaiveBayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3746               84.6745 %
Incorrectly Classified Instances       678               15.3255 %
Kappa statistic                          0.6404
Mean absolute error                      0.1615
Root mean squared error                  0.3734
Relative absolute error                 37.0444 %
Root relative squared error             79.9595 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.719   0.093   0.786     0.719  0.751     0.642 0.884    0.795
Non_Dropout 0.907   0.281   0.872     0.907  0.889     0.642 0.884    0.930
WeightedAvg 0.847   0.221   0.844     0.847  0.845     0.642 0.884    0.887

=== Confusion Matrix ===

    a    b   <-- classified as
 1021  400 |    a = Dropout
  278 2725 |    b = Non_Dropout
```

6.7 InfoGainAttributeEval + Logistic

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3847               86.9575 %
Incorrectly Classified Instances       577               13.0425 %
Kappa statistic                          0.6891
Mean absolute error                      0.1877
Root mean squared error                  0.3196
Relative absolute error                 43.0331 %
Root relative squared error             68.4516 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.725   0.062   0.847     0.725  0.781     0.693 0.900    0.818
Non_Dropout 0.938   0.275   0.878     0.938  0.907     0.693 0.900    0.922
WeightedAvg 0.870   0.207   0.868     0.870  0.867     0.693 0.900    0.889
```

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1030  391 |    a = Dropout
  186 2817 |    b = Non_Dropout
```

6.8 InfoGainAttributeEval + Random Forest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3832               86.6184 %
Incorrectly Classified Instances       592               13.3816 %
Kappa statistic                          0.6777
Mean absolute error                      0.2235
Root mean squared error                  0.3201
Relative absolute error                 51.2404 %
Root relative squared error             68.5588 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.701   0.056   0.856     0.701  0.771     0.685 0.908    0.866
Non_Dropout 0.944   0.299   0.870     0.944  0.905     0.685 0.908    0.935
WeightedAvg 0.866   0.221   0.865     0.866  0.862     0.685 0.908    0.913

=== Confusion Matrix ===

    a    b   <-- classified as
  996  425 |    a = Dropout
  167 2836 |    b = Non_Dropout
```

6.9 GainRatioAttributeEval + J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3810               86.1212 %
Incorrectly Classified Instances       614               13.8788 %
Kappa statistic                          0.6681
Mean absolute error                      0.2032
Root mean squared error                  0.3438
Relative absolute error                 46.5987 %
Root relative squared error             73.6234 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.706   0.065   0.837     0.706  0.766     0.673 0.849    0.751
```

```
Non_Dropout 0.935   0.294   0.870     0.935  0.901     0.673 0.849    0.873
WeightedAvg 0.861   0.221   0.860     0.861  0.858     0.673 0.849    0.834

=== Confusion Matrix ===

    a    b   <-- classified as
 1003  418 |    a = Dropout
  196 2807 |    b = Non_Dropout
```

6.10 GainRatioAttributeEval + NaiveBayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3715               83.9738 %
Incorrectly Classified Instances       709               16.0262 %
Kappa statistic                          0.6212
Mean absolute error                      0.1649
Root mean squared error                  0.3813
Relative absolute error                 37.8097 %
Root relative squared error             81.6571 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

            TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout     0.694   0.091   0.783     0.694  0.736     0.623 0.873    0.768
Non_Dropout 0.909   0.306   0.863     0.909  0.885     0.623 0.873    0.916
WeightedAvg 0.840   0.237   0.837     0.840  0.837     0.623 0.873    0.869

=== Confusion Matrix ===

    a    b   <-- classified as
  986  435 |    a = Dropout
  274 2729 |    b = Non_Dropout
```

6.11 GainRatioAttributeEval + Logistic

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3844               86.8897 %
Incorrectly Classified Instances       580               13.1103 %
Kappa statistic                          0.6842
Mean absolute error                      0.1927
Root mean squared error                  0.316
Relative absolute error                 44.1853 %
Root relative squared error             67.6687 %
Total Number of Instances             4424
```

```
=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.705   0.054   0.862     0.705  0.776     0.691 0.903    0.844
Non_Dropout    0.946   0.295   0.872     0.946  0.907     0.691 0.903    0.923
WeightedAvg    0.869   0.217   0.868     0.869  0.865     0.691 0.903    0.898

=== Confusion Matrix ===

    a    b   <-- classified as
 1002  419 |    a = Dropout
  161 2842 |    b = Non_Dropout
```

6.12 GainRatioAttributeEval + Random Forest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3831               86.5958 %
Incorrectly Classified Instances       593               13.4042 %
Kappa statistic                          0.6787
Mean absolute error                      0.2082
Root mean squared error                  0.3208
Relative absolute error                 47.7497 %
Root relative squared error             68.7113 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.709   0.060   0.848     0.709  0.773     0.684 0.904    0.857
Non_Dropout    0.940   0.291   0.872     0.940  0.905     0.684 0.904    0.937
WeightedAvg    0.866   0.217   0.865     0.866  0.862     0.684 0.904    0.9118


=== Confusion Matrix ===

    a    b   <-- classified as
 1008  413 |    a = Dropout
  180 2823 |    b = Non_Dropout
```

6.13 OneRAttributeEval + J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3803               85.9629 %
Incorrectly Classified Instances       621               14.0371 %
Kappa statistic                          0.6609
Mean absolute error                      0.2125
Root mean squared error                  0.3382
```

```
Relative absolute error                 48.7244 %
Root relative squared error             72.4251 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.685   0.058   0.848     0.685  0.758     0.668 0.862    0.792
Non_Dropout 0.942   0.315   0.864     0.942  0.901     0.668 0.862    0.890
WeightedAvg 0.860   0.232   0.859     0.860  0.855     0.668 0.862    0.859

=== Confusion Matrix ===

    a    b   <-- classified as
  974  447 |    a = Dropout
  174 2829 |    b = Non_Dropout
```

6.14 OneRAttributeEval + NaiveBayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3742               84.5841 %
Incorrectly Classified Instances       682               15.4159 %
Kappa statistic                          0.6373
Mean absolute error                      0.1628
Root mean squared error                  0.3755
Relative absolute error                 37.3264 %
Root relative squared error             80.4178 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.712   0.091   0.788     0.712   0.748    0.639 0.878    0.788
Non_Dropout 0.909   0.288   0.870     0.909   0.889    0.639 0.878    0.920
WeightedAvg 0.846   0.225   0.843     0.846   0.844    0.639 0.878    0.878

=== Confusion Matrix ===

    a    b   <-- classified as
 1012  409 |    a = Dropout
  273 2730 |    b = Non_Dropout
```

6.15 OneRAttributeEval + Logistic

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3845               86.9123 %
```

```
Incorrectly Classified Instances       579               13.0877 %
Kappa statistic                          0.687
Mean absolute error                      0.1893
Root mean squared error                  0.3208
Relative absolute error                 43.397  %
Root relative squared error             68.7127 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.719   0.060   0.851     0.719  0.779     0.692 0.899    0.813
Non_Dropout 0.940   0.281   0.876     0.940   0.907    0.692 0.899    0.920
WeightedAvg 0.869   0.210   0.868     0.869   0.866    0.692 0.899    0.886

=== Confusion Matrix ===

    a    b   <-- classified as
 1021  400 |    a = Dropout
  179 2824 |    b = Non_Dropout
```

6.16 OneRAttributeEval + Random Forest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3831               86.5958 %
Incorrectly Classified Instances       593               13.4042 %
Kappa statistic                          0.6761
Mean absolute error                      0.2225
Root mean squared error                  0.3213
Relative absolute error                 51.0158 %
Root relative squared error             68.8185 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.695   0.053   0.861     0.695  0.769     0.684 0.906    0.864
Non_Dropout 0.947   0.305   0.868     0.947  0.906     0.684 0.906    0.933
WeightedAvg 0.866   0.224   0.866     0.866  0.862     0.684 0.906    0.910

=== Confusion Matrix ===

    a    b   <-- classified as
  987  434 |    a = Dropout
  159 2844 |    b = Non_Dropout
```

6.17 Intuition + J48

```
=== Stratified cross-validation ===
```

```
=== Summary ===

Correctly Classified Instances        3778               85.3978 %
Incorrectly Classified Instances       646               14.6022 %
Kappa statistic                          0.6546
Mean absolute error                      0.1944
Root mean squared error                  0.353
Relative absolute error                 44.5829 %
Root relative squared error             75.592  %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.715   0.080   0.808     0.715  0.759     0.657 0.845    0.721
Non_Dropout    0.920   0.285   0.872     0.920  0.895     0.657 0.845    0.869
WeightedAvg    0.854   0.219   0.852     0.854  0.851     0.657 0.845    0.822

=== Confusion Matrix ===

    a    b   <-- classified as
 1016  405 |    a = Dropout
  241 2762 |    b = Non_Dropout
```

6.18 Intuition + NaiveBayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3715               83.9738 %
Incorrectly Classified Instances       709               16.0262 %
Kappa statistic                          0.6194
Mean absolute error                      0.1655
Root mean squared error                  0.3803
Relative absolute error                 37.9508 %
Root relative squared error             81.4399 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.685   0.087   0.788     0.685  0.733     0.623 0.860    0.755
Non_Dropout    0.913   0.315   0.860     0.913  0.885     0.623 0.860    0.895
WeightedAvg    0.840   0.242   0.837     0.840  0.837     0.623 0.860    0.850

=== Confusion Matrix ===

    a    b   <-- classified as
  974  447 |    a = Dropout
  262 2741 |    b = Non_Dropout
```

6.19 Intuition + Logistic

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3872               87.5226 %
Incorrectly Classified Instances       552               12.4774 %
Kappa statistic                          0.7011
Mean absolute error                      0.1913
Root mean squared error                  0.3097
Relative absolute error                 43.8767 %
Root relative squared error             66.3223 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.725   0.054   0.865     0.725  0.789     0.707 0.914    0.87
Non_Dropout 0.946   0.275   0.879     0.946  0.911     0.707 0.914    0.945
WeightedAvg 0.875   0.204   0.874     0.875  0.872     0.707 0.914    0.921

=== Confusion Matrix ===

    a    b   <-- classified as
 1030  391 |    a = Dropout
  161 2842 |    b = Non_Dropout
```

6.20 Intuition + Random Forest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3846               86.9349 %
Incorrectly Classified Instances       578               13.0651 %
Kappa statistic                          0.6878
Mean absolute error                      0.1969
Root mean squared error                  0.3149
Relative absolute error                 45.1578 %
Root relative squared error             67.4462 %
Total Number of Instances             4424

=== Detailed Accuracy By Class ===

               TP_Rate FP_Rate Precision Recall F-Measure MCC   ROC_Area PRC_Area
Dropout        0.720   0.060   0.850     0.720  0.780     0.693 0.911    0.864
Non_Dropout 0.940   0.280   0.876     0.940  0.907     0.693 0.911    0.940
WeightedAvg 0.869   0.209   0.868     0.869  0.866     0.693 0.911    0.916

=== Confusion Matrix ===
```

```
    a    b   <-- classified as
 1023  398 |    a = Dropout
  180 2823 |    b = Non_Dropout
```

7 Analysis:

A detailed sorting of the performance of each model by certain metrics is found in Appendix D, along with all results mentioned in this section.

7.1 General Analysis

After running 20 different models on our dataset, we found that the five best-performing models by accuracy are:

1. Intuition + Logistic
2. Correlation + Logistic
3. InfoGain + Logistic
4. Intuition + Random Forest
5. OneR + Logistic

These models achieved a minimum accuracy of 86.9%, with Intuition + Logistic performing best at 87.5% accuracy. Although we cannot say that this model is most optimal before observing other metrics, we can assume these five models are acceptable to use.

Interestingly enough, it turned out that a model's accuracy was largely dependent on the type of classification model used, rather than the attribute selection algorithm. Logistic models tended to perform best, followed by Random Forest, J48, and Naive Bayes. Therefore, Logistic models should be used if aiming for highest accuracy.

7.2 TP/FP Rate

It is not wise to only consider accuracy as the metric in choosing the best model. Our model aims to predict student dropout, not success, which means that a better focus metric would be the true positive rate of dropout. It is better to incorrectly predict a non-dropout student as dropout, rather than the opposite, as the goal of the project is toy. Hence, the true positive rate is the most important metric, taking into account all instances labeled as positive. The five best-performing models by TPR are:

1. Intuition + Logistic
2. Intuition + Random Forest
3. OneR + Logistic
4. InfoGain + Naive Bayes
5. Intuition + J48

These models achieve a minimum TPR of .710, with the highest being Intuition + Logistic at .875 TPR. Overall, the list is more closely associated with the attribute selection algorithm than

to the classifier itself. Intuitively selected attributes tend to perform best, but the trend is scattered and weak, if present at all.

7.3 Area Under Curve

Since both FPs and FNs are important, area under the ROC curve is a satisfactory measure of the model's performance. The top five models by the area under curve (AUC) metric are:

1. Intuition + Logistic
2. Intuition + Random Forest
3. Correlation + Logistic
4. InfoGain + Random Forest
5. Correlation + Random Forest

The top five models achieve a minimum AUC of .906, with Intuition + Logistic performing highest at .914 AUC.

7.4 Best Model

After examining three different metrics, we conclude that the Intuition + Logistic model is best as it outperformed all other models in every metric, achieving 87.5% accuracy, .875 TPR, and .914 AUC. Most importantly, it achieved the highest performance for TPR, our primary metric of interest for dropout prediction.

8 Conclusion:

8.1 Insights

This project gave us firsthand experience of the machine learning workflow for supervised learning. During preprocessing, we faced numerous difficulties and had to learn to troubleshoot errors in our data along with issues from the Weka software. For data analysis, we learned the importance of attribute selection through various selection algorithms paired with classifier models. Despite the apparent ease in choosing accuracy as the primary metric, we had to observe various metrics to evaluate model performance pertinent to the context and choose the best model more assuredly; we used Accuracy, True Positive Rate, and Area Under Curve of ROC to evaluate 20 models and compare them holistically in terms of student dropout.

8.2 Future Research

We only used four of the eleven available selection algorithms in WEKA, but there are many other attribute selection algorithms not included, such as Minimum Redundancy Maximum Relevance and Recursive Feature Elimination. This leads to many more model possibilities, so our chosen model is by no means the definitive best model. Future research can explore these other combinations, and with additional data, like mental health and extracurricular engagement, the predictors for student dropout and academic success can be examined more comprehensively.

Beyond the current methodology employed, the nature of the study can be altered, as longitudinal studies could provide valuable insight into how the predictive factors evolve over time in our rapidly changing world.

8.3 Division of Labor

Rem Turatbekov:

1. Finding data
2. Code for preprocessing
3. Editing reports
4. Presentation

Avery Li:

1. Drafting Reports
2. Running Models on Weka
3. Formatting code in paper
4. Finding External Resources

A Appendix A: Class Ordering:

Following are the conversion of integer value to the actual value they represent in the data table, formatted as a python dictionary. The contents of the dictionaries are used in Appendix B.3, as detailed in section 3.1.

A.1 Application Mode

```
data = {
    "1": "1st phase - general contingent",
    "2": "Ordinance No. 612/93",
    "5": "1st phase - special contingent (Azores Island)",
    "7": "Holders of other higher courses",
    "10": "Ordinance No. 854-B/99",
    "15": "International student (bachelor)",
    "16": "1st phase - special contingent (Madeira Island)",
    "17": "2nd phase - general contingent",
    "18": "3rd phase - general contingent",
    "26": "Ordinance No. 533-A/99, item b2) (Different Plan)",
    "27": "Ordinance No. 533-A/99, item b3 (Other Institution)",
    "39": "Over 23 years old",
    "42": "Transfer",
    "43": "Change of course",
    "44": "Technological specialization diploma holders",
    "51": "Change of institution/course",
    "53": "Short cycle diploma holders",
    "57": "Change of institution/course (International)"
}
```

A.2 Course

```
data = {
    "33": "Biofuel Production Technologies",
    "171": "Animation and Multimedia Design",
    "8014": "Social Service (evening attendance)",
    "9003": "Agronomy",
    "9070": "Communication Design",
    "9085": "Veterinary Nursing",
    "9119": "Informatics Engineering",
    "9130": "Equinculture",
    "9147": "Management",
    "9238": "Social Service",
    "9254": "Tourism",
    "9500": "Nursing",
    "9556": "Oral Hygiene",
    "9670": "Advertising and Marketing Management",
    "9773": "Journalism and Communication",
    "9853": "Basic Education",
    "9991": "Management (evening attendance)"
```

```
}
```

A.3 Previous Qualifications

```
data = {
    "1": "Secondary education",
    "2": "Higher education - bachelor's degree",
    "3": "Higher education - degree",
    "4": "Higher education - master's",
    "5": "Higher education - doctorate",
    "6": "Frequency of higher education",
    "9": "12th year of schooling - not completed",
    "10": "11th year of schooling - not completed",
    "12": "Other - 11th year of schooling",
    "14": "10th year of schooling",
    "15": "10th year of schooling - not completed",
    "19": "Basic education 3rd cycle (9th/10th/11th year) or equiv.",
    "38": "Basic education 2nd cycle (6th/7th/8th year) or equiv.",
    "39": "Technological specialization course",
    "40": "Higher education - degree (1st cycle)",
    "42": "Professional higher technical course",
    "43": "Higher education - master (2nd cycle)"
}
```

A.4 Nationality

```
data = {
    "1": "Portuguese",
    "2": "German",
    "6": "Spanish",
    "11": "Italian",
    "13": "Dutch",
    "14": "English",
    "17": "Lithuanian",
    "21": "Angolan",
    "22": "Cape Verdean",
    "24": "Guinean",
    "25": "Mozambican",
    "26": "Santomean",
    "32": "Turkish",
    "41": "Brazilian",
    "62": "Romanian",
    "100": "Moldova (Republic of)",
    "101": "Mexican",
    "103": "Ukrainian",
    "105": "Russian",
    "108": "Cuban",
    "109": "Colombian"
}
```

A.5 Mother's/Father's Qualifications

```
data = {
    "1": "Secondary Education - 12th Year of Schooling or Eq.",
    "2": "Higher Education - Bachelor's Degree",
    "3": "Higher Education - Degree",
    "4": "Higher Education - Master's",
    "5": "Higher Education - Doctorate",
    "6": "Frequency of Higher Education",
    "9": "12th Year of Schooling - Not Completed",
    "10": "11th Year of Schooling - Not Completed",
    "11": "7th Year (Old)",
    "12": "Other - 11th Year of Schooling",
    "14": "10th Year of Schooling",
    "18": "General commerce course",
    "19": "Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv.",
    "22": "Technical-professional course",
    "26": "7th year of schooling",
    "27": "2nd cycle of the general high school course",
    "29": "9th Year of Schooling - Not Completed",
    "30": "8th year of schooling",
    "34": "Unknown",
    "35": "Can't read or write",
    "36": "Can read without having a 4th year of schooling",
    "37": "Basic education 1st cycle (4th/5th year) or equiv.",
    "38": "Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv.",
    "39": "Technological specialization course",
    "40": "Higher education - degree (1st cycle)",
    "41": "Specialized higher studies course",
    "42": "Professional higher technical course",
    "43": "Higher Education - Master (2nd cycle)",
    "44": "Higher Education - Doctorate (3rd cycle)"
}
```

A.6 Mother's/Father's Occupation

```
data = {
    "0": "Student",
    "1": "Representatives of the Legislative Power and Executive Bodies,
Directors, Directors and Executive Managers",
    "2": "Specialists in Intellectual and Scientific Activities",
    "3": "Intermediate Level Technicians and Professions",
    "4": "Administrative staff",
    "5": "Personal Services, Security and Safety Workers and Sellers",
    "6": "Farmers and Skilled Workers in Agriculture, Fisheries and Forestry",
    "7": "Skilled Workers in Industry, Construction and Craftsmen",
    "8": "Installation and Machine Operators and Assembly Workers",
    "9": "Unskilled Workers",
    "10": "Armed Forces Professions",
    "90": "Other Situation",
    "99": "(blank)",
```

```
    "122": "Health professionals",
    "123": "Teachers",
    "125": "Specialists in information and communication technologies (ICT)",
    "131": "Intermediate level science and engineering technicians and
professions",
    "132": "Technicians and professionals, of intermediate level of health",
    "134": "Intermediate level technicians from legal, social, sports, cultural
and similar services",
    "141": "Office workers, secretaries in general and data processing
operators",
    "143": "Data, accounting, statistical, financial services and
registry-related operators",
    "144": "Other administrative support staff",
    "151": "Personal service workers",
    "152": "Sellers",
    "153": "Personal care workers and the like",
    "171": "Skilled construction workers and the like, except electricians",
    "173": "Skilled workers in printing, precision instrument manufacturing,
jewelers, artisans and the like",
    "175": "Workers in food processing, woodworking, clothing and other
industries and crafts",
    "191": "Cleaning workers",
    "192": "Unskilled workers in agriculture, animal production, fisheries and
forestry",
    "193": "Unskilled workers in extractive industry, construction,
manufacturing and transport",
    "194": "Meal preparation assistants"
}
```

All changes from integer to string are made in appendix B.3

B Appendix B: Preprocessing Code:

The following provides code snippets for certain tasks needed to be accomplished during preprocessing, for sake of replication. Any code in section B.3 onwards should have the first and last few lines of B.3 appended for file reading/writing.

B.1 Semicolons to Commas

Changing data delimiter to comma from semicolon:

```
with open("data.csv", encoding="utf-8-sig") as file:
    f = file.read().replace(';',
',').replace('"','').replace('\t','').replace("'s","")

with open("new_data.csv", "w") as file:
    file.write(f)
```

B.2 Converting Integer to Strings

In reference to section 3.1, taking data from Appendix A

```
mappings = [
    ("Marital status", marital_status),
    ("Application mode", application_mode),
    ("Course", course),
    ("Previous qualification", prev_qualification),
    ("Nacionality", nacionality),
    ("Mother_qualification", mother_qualification),
    ("Father_qualification", father_qualification),
    ("Mother_occupation", mother_occupation),
    ("Father_occupation", father_occupation)
]
def main():
    with open("data.csv") as f:
        data = [l.strip().split(";") for l in f]
    ind_mapper = {}
    print(data[0])
    for attr_name, mapper in mappings:
        ind_mapper[data[0].index(attr_name)] = mapper
    for l in data[1:]:
        for ind, mapper in ind_mapper.items():
            l[ind] = mapper[int
                            (l[ind])]
    with open("out_data.csv", "w") as f:
        for l in data:
            f.write(",".join(l)+"\n")
```

B.3 Converting to Binary Classification

All code of reading the file and writing to a new file is included here once, but removed in future references for sake of brevity; fields will always be given by the list `fields`, rows by list `rows`, and `import csv` is always implied.

```
import csv
rows = []
with open("data.csv", "r") as f:
    csvreader = csv.reader(f)
    fields = next(csvreader)
    for row in csvreader:
        rows.append(row)

data = []
for row in rows:
    if row[-1] in ['Graduate', 'Enrolled']:
        row[-1] = 'Non_Dropout'
    data.append({fields[i]: row[i] for i in range(len(fields))})

with open('data_out.csv', 'w', newline='') as csvfile:
```

```
    writer = csv.DictWriter(csvfile, fieldnames=fields)
    writer.writeheader()
    writer.writerows(data)
```

B.4 Data Normalization

In reference to section 3.3

```
NEWMIN = 0
NEWMAX = 5
fields_to_change = [12, *range(19, 35)] #detailed list found in 3.4
for idx in fields_to_change:
    vals = [float(row[idx]) for row in rows]
    minval = min(vals); maxval = max(vals)
    for row in rows:
        row[idx] = str((float(row[idx])-minval)/(maxval-minval) *
(NEWMAX-NEWMIN) + NEWMIN)
```

C Appendix C: Attribute Selection Results:

The following details the complete result of 4 different attribute selection analyses, pasted from Weka.

C.1 CorrelationAttributeEval

```
=== Attribute Selection on all input data ===

Search Method:
      Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 36 Target):
      Correlation Ranking Filter
Ranked attributes:
 0.57179   31 Curricular units 2nd sem (grade)
 0.5695    30 Curricular units 2nd sem (approved)
 0.48067   25 Curricular units 1st sem (grade)
 0.47911   24 Curricular units 1st sem (approved)
 0.42915   17 Tuition fees up to date
 0.25422   20 Age at enrollment
 0.24535   19 Scholarship holder
 0.22941   16 Debtor
 0.20398   18 Gender
 0.155     29 Curricular units 2nd sem (evaluations)
 0.14151   28 Curricular units 2nd sem (enrolled)
 0.13806    6 Previous qualification
 0.13316    2 Application mode
 0.12463   22 Curricular units 1st sem (enrolled)
 0.11039    1 Marital status
 0.10723   14 Displaced
```

```
 0.09581   13 Admission grade
 0.09012   23 Curricular units 1st sem (evaluations)
 0.0805     5 Daytime/evening attendance
 0.0799    32 Curricular units 2nd sem (without evaluations)
 0.07821    7 Previous qualification (grade)
 0.07049    3 Application order
 0.06487    4 Course
 0.05423   26 Curricular units 1st sem (without evaluations)
 0.05006    9 Mother_qualification
 0.04632   35 GDP
 0.04378   10 Father_qualification
 0.03304   27 Curricular units 2nd sem (credited)
 0.02931   21 Curricular units 1st sem (credited)
 0.02783   34 Inflation rate
 0.02336   11 Mother_occupation
 0.02204   12 Father_occupation
 0.01298   33 Unemployment rate
 0.01037    8 Nacionality
 0.00281   15 Educational special needs

Selected attributes:
31,30,25,24,17,20,19,16,18,29,28,6,2,22,1,14,13,23,5,32,7,3,4,26,9,35,10,27,21,
34,11,12,33,8,15 : 35
```

C.2 InfoGainAttributeEval

```
=== Attribute Selection on all input data ===

Search Method:
      Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 36 Target):
      Information Gain Ranking Filter

Ranked attributes:
 0.31227   30 Curricular units 2nd sem (approved)
 0.24983   31 Curricular units 2nd sem (grade)
 0.24574   24 Curricular units 1st sem (approved)
 0.17996   25 Curricular units 1st sem (grade)
 0.12686   17 Tuition fees up to date
 0.0927    29 Curricular units 2nd sem (evaluations)
 0.07577   23 Curricular units 1st sem (evaluations)
 0.06836   20 Age at enrollment
 0.06402    2 Application mode
 0.04985    4 Course
 0.04913   19 Scholarship holder
 0.04354    7 Previous qualification (grade)
 0.03519   16 Debtor
 0.03422   11 Mother_occupation
 0.03185   28 Curricular units 2nd sem (enrolled)
```

```
 0.03137    6 Previous qualification
 0.02945   18 Gender
 0.02923   10 Father_qualification
 0.02909   12 Father_occupation
 0.029      9 Mother_qualification
 0.02271   22 Curricular units 1st sem (enrolled)
 0.01107   13 Admission grade
 0.00918    1 Marital status
 0.00828   14 Displaced
 0.0052     3 Application order
 0.00448    5 Daytime/evening attendance
 0.004     35 GDP
 0.00373    8 Nacionality
 0.00357   32 Curricular units 2nd sem (without evaluations)
 0.00257   26 Curricular units 1st sem (without evaluations)
 0         15 Educational special needs
 0         33 Unemployment rate
 0         34 Inflation rate
 0         21 Curricular units 1st sem (credited)
 0         27 Curricular units 2nd sem (credited)

Selected attributes:
30,31,24,25,17,29,23,20,2,4,19,7,16,11,28,6,18,10,12,9,22,13,1,14,3,5,35,8,32,2
6,15,33,34,21,27 : 35
```

C.3 GainRatioAttributeEval

```
=== Attribute Selection on all input data ===

Search Method:
      Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 36 Target):
      Gain Ratio feature evaluator

Ranked attributes:
 0.2405    17 Tuition fees up to date
 0.13978   31 Curricular units 2nd sem (grade)
 0.12913   30 Curricular units 2nd sem (approved)
 0.10939   24 Curricular units 1st sem (approved)
 0.09234   25 Curricular units 1st sem (grade)
 0.06887   16 Debtor
 0.06074   19 Scholarship holder
 0.05893   29 Curricular units 2nd sem (evaluations)
 0.04429   20 Age at enrollment
 0.03889   23 Curricular units 1st sem (evaluations)
 0.03147   18 Gender
 0.02889    6 Previous qualification
 0.02435    2 Application mode
```

```
 0.02414   26 Curricular units 1st sem (without evaluations)
 0.02408   32 Curricular units 2nd sem (without evaluations)
 0.01932   13 Admission grade
 0.01896   28 Curricular units 2nd sem (enrolled)
 0.01876    7 Previous qualification (grade)
 0.01496    8 Nacionality
 0.01461   22 Curricular units 1st sem (enrolled)
 0.01439    1 Marital status
 0.01297    4 Course
 0.01161   11 Mother_occupation
 0.01084   10 Father_qualification
 0.01055    9 Mother_qualification
 0.00901    5 Daytime/evening attendance
 0.00842   12 Father_occupation
 0.00833   14 Displaced
 0.00578    3 Application order
 0.00469   35 GDP
 0         27 Curricular units 2nd sem (credited)
 0         33 Unemployment rate
 0         34 Inflation rate
 0         15 Educational special needs
 0         21 Curricular units 1st sem (credited)

Selected attributes:
17,31,30,24,25,16,19,29,20,23,18,6,2,26,32,13,28,7,8,22,1,4,11,10,9,5,12,14,3,3
5,27,33,34,15,21 : 35
```

C.4 OneRAttributeEval

```
=== Attribute Selection on all input data ===

Search Method:
      Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 36 Target):
      OneR feature evaluator.

      Using 10 fold cross validation for evaluating attributes.
      Minimum bucket size for OneR: 6

Ranked attributes:
83.16     30 Curricular units 2nd sem (approved)
80.8092   31 Curricular units 2nd sem (grade)
80.0633   24 Curricular units 1st sem (approved)
78.2098   25 Curricular units 1st sem (grade)
76.6049   17 Tuition fees up to date
73.1465   29 Curricular units 2nd sem (evaluations)
71.3834   23 Curricular units 1st sem (evaluations)
70.6148   16 Debtor
```

```
70.5018    2 Application mode
70.0497   20 Age at enrollment
69.9367   11 Mother_occupation
69.8463    6 Previous qualification
69.6203    7 Previous qualification (grade)
69.3942   10 Father_qualification
69.2812    9 Mother_qualification
69.0778   12 Father_occupation
68.6483   22 Curricular units 1st sem (enrolled)
68.3318   28 Curricular units 2nd sem (enrolled)
68.1736    4 Course
67.9928   32 Curricular units 2nd sem (without evaluations)
67.925    26 Curricular units 1st sem (without evaluations)
67.8797    5 Daytime/evening attendance
67.8797    3 Application order
67.8797   35 GDP
67.8797   18 Gender
67.8797   19 Scholarship holder
67.8797   34 Inflation rate
67.8797   15 Educational special needs
67.8797   14 Displaced
67.8797   33 Unemployment rate
67.8119   27 Curricular units 2nd sem (credited)
67.8119    1 Marital status
67.7667    8 Nacionality
67.6763   21 Curricular units 1st sem (credited)
65.8002   13 Admission grade

Selected attributes:
30,31,24,25,17,29,23,16,2,20,11,6,7,10,9,12,22,28,4,32,26,5,3,35,18,19,34,15,14
,33,27,1,8,21,13 : 35
```

D Appendix D: Model Performance:

The following is a neater formatting of the results listed in section 6.

D.1 Overall Performance

Model Name	Accuracy	TP Dropout	FP Dropout	ROC	TP Total	FP Total
CorrJ48	0.863698	0.712	0.065	0.848	0.864	0.216
CorrNaive	0.836573	0.695	0.097	0.872	0.837	0.238
CorrLogistic	0.869801	0.709	0.054	0.909	0.87	0.215
CorrRF	0.865506	0.713	0.062	0.906	0.866	0.215
InfoJ48	0.861438	0.692	0.058	0.865	0.861	0.228
InfoNaive	0.846745	0.719	0.093	0.884	0.847	0.221

InfoLogistic	0.869575	0.6891	0.062	0.9	0.87	0.207
InfoRF	0.866184	0.701	0.056	0.908	0.866	0.221
GainJ48	0.861212	0.706	0.065	0.849	0.869	0.221
GainNaive	0.839738	0.694	0.091	0.873	0.84	0.237
GainLogistic	0.868897	0.705	0.054	0.903	0.869	0.217
GainRF	0.865958	0.709	0.06	0.904	0.866	0.217
OneRJ48	0.859629	0.685	0.058	0.862	0.86	0.232
OneRNaive	0.845841	0.712	0.091	0.878	0.846	0.225
OneRLogistic	0.869123	0.719	0.06	0.899	0.869	0.21
OneRRF	0.865958	0.695	0.053	0.906	0.866	0.224
IntuitJ48	0.853978	0.715	0.08	0.845	0.854	0.219
IntuitNaive	0.839738	0.685	0.087	0.86	0.84	0.242
IntuitLogistic	0.875226	0.725	0.06	0.914	0.875	0.204
IntuitRF	0.869349	0.72	0.06	0.911	0.869	0.209

D.2 By Accuracy

Model name	Accuracy
IntuitLogistic	0.875226
CorrLogistic	0.869801
InfoLogistic	0.869575
IntuitRF	0.869349
OneRLogistic	0.869123
GainLogistic	0.868897
InfoRF	0.866184
GainRF	0.865958
OneRRF	0.865958
CorrRF	0.865506
CorrJ48	0.863698
InfoJ48	0.861438
GainJ48	0.861212
OneRJ48	0.859629
IntuitJ48	0.853978
InfoNaive	0.846745
OneRNaive	0.845841
GainNaive	0.839738

IntuitNaive	0.839738
CorrNaive	0.836573

D.3 By Dropout TP Rate

Model name	TP Dropout
IntuitLogistic	0.725
IntuitRF	0.72
OneRLogistic	0.719
InfoNaive	0.719
IntuitJ48	0.715
CorrRF	0.713
CorrJ48	0.712
OneRNaive	0.712
CorrLogistic	0.709
GainRF	0.709
GainJ48	0.706
GainLogistic	0.705
InfoRF	0.701
OneRRF	0.695
CorrNaive	0.695
GainNaive	0.694
InfoJ48	0.692
InfoLogistic	0.6891
OneRJ48	0.685
IntuitNaive	0.685

D.3 By Area Under ROC Curve

Model name	ROC
IntuitLogistic	0.914
IntuitRF	0.911
CorrLogistic	0.909
InfoRF	0.908
CorrRF	0.906
OneRRF	0.906
GainRF	0.904

GainLogistic	0.903
InfoLogistic	0.9
OneRLogistic	0.899
InfoNaive	0.884
OneRNaive	0.878
GainNaive	0.873
CorrNaive	0.872
InfoJ48	0.865
OneRJ48	0.862
IntuitNaive	0.86
GainJ48	0.849
CorrJ48	0.848
IntuitJ48	0.845

E Replication:

The following details the steps taken to recreate our process. The steps detailed are specific to a certain attribute selection algorithm/model, but are general and can be applied to any algorithm we have used.

E.1 Attribute Selection

1. Download the data_final.csv file in our google drive.
2. Open Weka Explorer and choose the file to view its contents.
3. Convert the csv to an arff file by clicking save, then save as arff. This is done under the Preprocess tab.
4. Reopen the arff file you just saved, and click on the Select Attributes tab.
5. Click Choose under Attribute Evaluator, and select the attribute selection algorithm you would like to use.
6. Agree to use Ranker search method if prompted and make sure that “Use full training set” is selected under Attribute Selection Mode.
7. Click the class dropdown below the Attribute Selection Mode box to select the class attribute, Target.
8. Click the start button to run algorithm.

E.2 Classifier Algorithms

1. After following the steps in E.1, note down the attributes over the chosen threshold
2. Return to the Preprocess tab and select any attributes that were not chosen

3. Click the remove button
4. Click save to save the intermediate dataset, for future reference
5. Click on the Classify tab
6. Click Choose, then the model you would like to use
7. Make sure Cross-validation is selected with 10 folds (it is the default)
8. Make sure the Target attribute is selected in the class dropdown
9. Click start

9 References:

Breiman, L., & Cutler, A. (2019). *Random forests - classification description*. Berkeley.edu. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Khanna, N. (2021, August 18). *J48 Classification (C4.5 Algorithm) in a Nutshell*. Medium. https://medium.com/@nilimakhanna1/j48-classification-c4-5-algorithm-in-a-nutshell-24c50d20658e

Tung.M.Phung. (n.d.). *Information Gain, Gain Ratio and Gini Index*. Tung M Phung's Blog. https://tungmphung.com/information-gain-gain-ratio-and-gini-index/

Turing. (n.d.). *Naive Bayes Algorithm in ML: Simplifying Classification Problems*. Www.turing.com. https://www.turing.com/kb/an-introduction-to-naive-bayes-algorithm-for-beginners

Witten, I. (n.d.). *Data Mining with Weka Class 2 -Lesson 1 Be a classifier!* https://user.eng.umd.edu/~austin/ence688p.d/handouts/DM-Weka-Class02.pdf