Mackenzie Houser, Blessy Thomas, Nathan Widlake
IST 707 - Applied Machine Learning
Final Project Report
06/20/2023

Introduction

How can we say society is investing in the future if it is not keeping an eye on the next generation? When students finish their High School experience, they are considered, in most cultures, legal adults. The courses they are taught in the long run might not be the most valuable; however, they provide a deeper meaning that goes beyond the four walls of the classroom. Whereas the grade will be forgotten over time, the true value added from these courses is the sense of accomplishment and reward they feel when an effort is given. Students who succeed in these rudimentary topics can take that adulation from teachers and family to the “real world” and look to chase that feeling in more important jobs that add value to society. The challenge in High School isn’t catering to the successful, but nurturing the development of those who struggle to grasp the meaning of their education. No student should be left behind since their struggle is a loss to society as a whole. Identifying these students early provides educators with a chance to assist in a way that the student can learn these lessons and walk out their doors prepared for what lies ahead.

High school students may struggle or fail their classes due to a number of challenges. These factors may include difficulties inside the classroom as well as externally. In order for the staff at these schools to understand how to work with struggling students, they need to have insight into what goes on in their lives. No two people are the same, but if there is a trend of failures following certain personal characteristics, then an assumption can be made based on students with past struggles to support future growth. Identifying these struggles is where schools have difficulties, from a vast number of questions, picking the important ones can quickly become overwhelming. From educational, social, and family events occurring in the life of the student body, discovering the perfect combination of variables is the key that allows teachers to attempt to unlock their students' potential.

These schools can implement various techniques to identify academically struggling students earlier on before it is too late to increase their grades. Techniques may include regular assessments where students are given a quiz on every chapter or an exam on every unit. The grade that a student receives can inform the teacher of how well the student knows and understands the material. Another technique is to mail progress reports where a report card is sent home to a student’s parents or guardians at the end of every grading period describing the student’s strengths and weaknesses in each subject. Progress reports not only allow the parents and guardians but the teachers to track how well a student does throughout the school year. Similarly, parent-teacher conferences are a technique that allows teachers to discuss the academic progress of a student with their parents or guardians face-to-face as well as any behavioral or social concerns that the teacher might have. These techniques can help to bring high school students who are at risk of failing to the attention of teachers and their parents or guardians.

Additional assistance can be provided to failing students by high schools in various ways. Strategies may include early intervention programs such as tutoring, mentoring, study groups, and after-school programs. These programs provide not only failing students but all types of students with academic support. Another form of assistance that is offered by high schools is counseling services. These services can psychologically help students who are dealing with mental health issues, substance abuse issues, and relationship issues. High schools can also work with the parents and guardians of students as well as community organizations to provide and support struggling students with additional resources whether it may be academic, psychological, or medical. By providing struggling students and their families with various assistance resources, high schools can help students to succeed academically and reach their full potential.

Analysis

The Data

Schools cannot seem to identify students who may need additional support before they are unsuccessful in the course. Two Portuguese schools have established a contract to begin analysis focusing on high school math students. Assuming both schools have similar class sizes, the same course syllabus, and the same amount of designated time per week. The goal was to determine what is contributing to the students failing the classes and what these schools can do to identify them early and offer additional assistance.

The data set was obtained from Kaggle and it contains 33 variables and 395 rows that include information about student grades (three periods), demographics, and social and school-related features. The target audience for analysis was students who have a grade of less than 55%. In order to reflect the target audience in the study, a dummy variable was created that states: 1 if grade <= 55%, 0 if grade > 55%.

There was no missing data to perform data cleaning. However, in order to perform all methods of analysis, it was necessary to convert certain forms of variables. These variables included: school, sex, address, famsize, pstatus, medu, fedu, schoolsup, famsup, activities, nursery, higher, internet, romantic, famrel, freetime, goout, dalc, walc, and health. In addition, a second data set was created from the original by using the check feature importance to only select the top 10 variables; this was used quite frequently.

Once the data was cleaned, exploratory data analysis was performed. During this process, basic discoveries were made about the data being investigated. Some outliers were revealed with students who had a final grade of zero. In this case, outliers did not want to be removed, because they were considered the target audience. Another discovery was that both schools showed a decrease in grades and an increase in the number of students receiving zeros as the academic year progressed. Additionally, between the schools, it seemed that Males had a slightly better average than Females. Furthermore, both schools had around the same proportion of students who failed, suggesting that the environment was not coming into play. Lastly, the exploratory data analysis revealed that paying for extra classes did not seem to provide help. Now that basic discoveries have been made, more advanced methods of analysis were used to arrive at a solution for the problem.

Variable	Description
school	student's school (binary: 'GP' : Gabriel Pereira or 'MS' : Mousinho da Silveira)
sex	student's sex (binary: 'F' : female or 'M' : male)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: 'U' : urban or 'R' : rural)
famsize	family size (binary: 'LE3' : less or equal to 3 or 'GT3' : greater than 3)
Pstatus	parent's cohabitation status (binary 'T' : living together or 'A' : apart)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian (nominal: 'mother', 'father' or 'other')
traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first semester math grade (numeric: from 0 to 20)
G2	second semester math grade (numeric: from 0 to 20)
G3	final math grade (numeric: from 0 to 20, output target)
Target	Anyone who finished with a grade at or below 11

Table 1

The table above contains a description of all the variables from the data set. The top ten variables from the check feature importance includes: failures, freetime, fedu, dalc, walc, schoolsup, absences, age, famrel, and famsize.

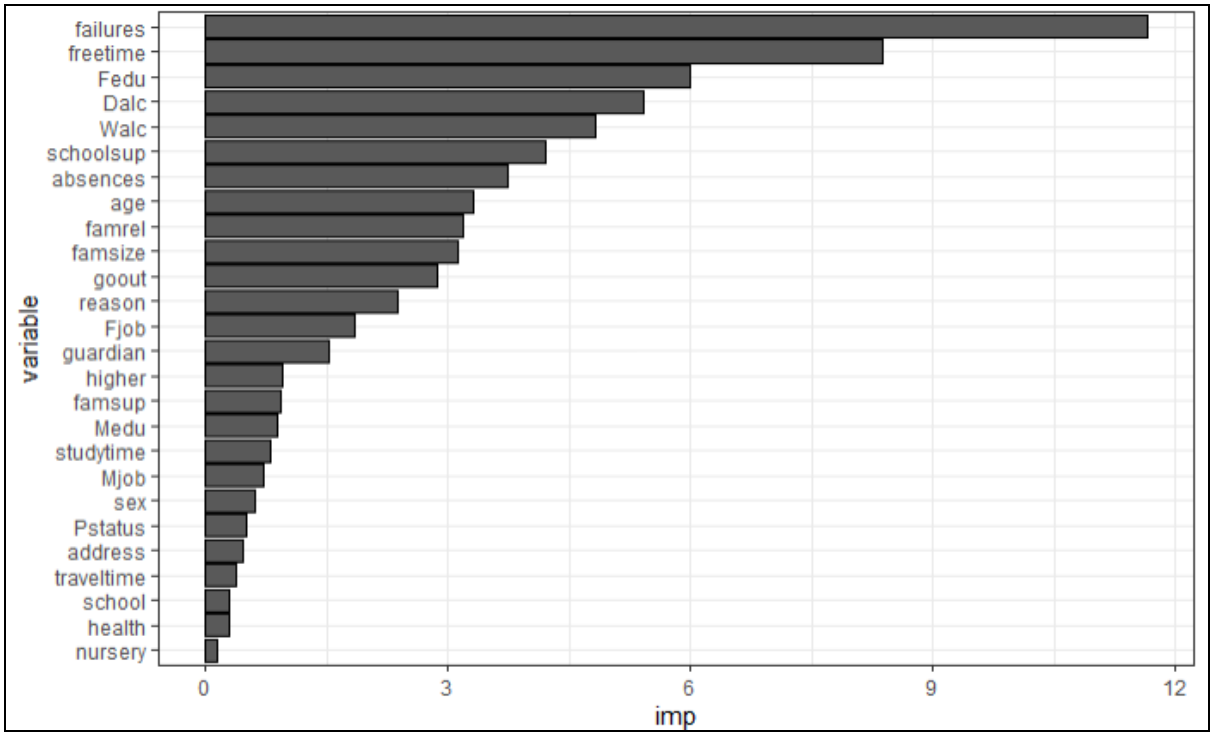

Figure 1

Check feature importance: The top variables' importance from the unpruned tree can be seen above. The model will be returned without the non-important variables with pruning.

Exploratory Data Analysis (EDA)

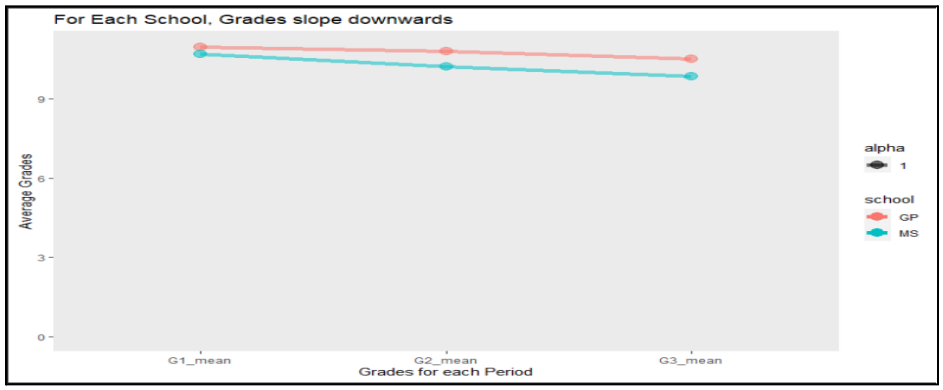

Figure 2

Both schools show a decrease in grades as the academic year progressed.

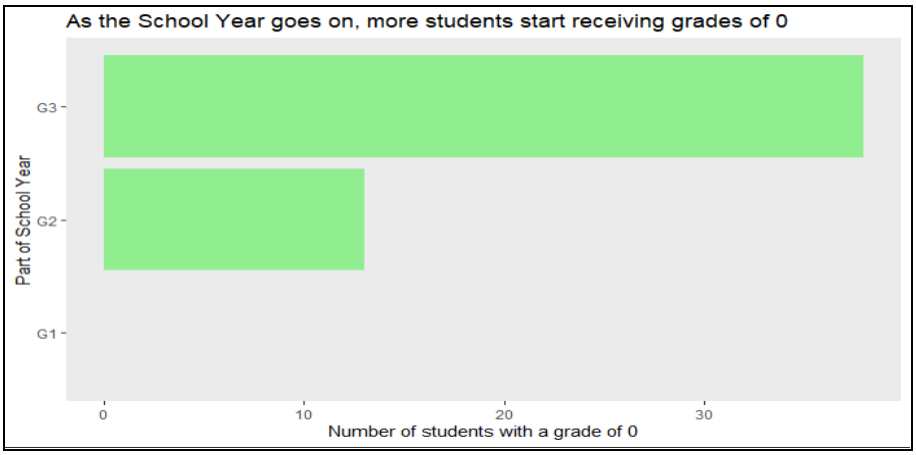

Figure 3

Students who receive grades of zero drastically increase as the year progressed indicating that students grasp the first part of the course well but later topics are the ones causing the challenge.

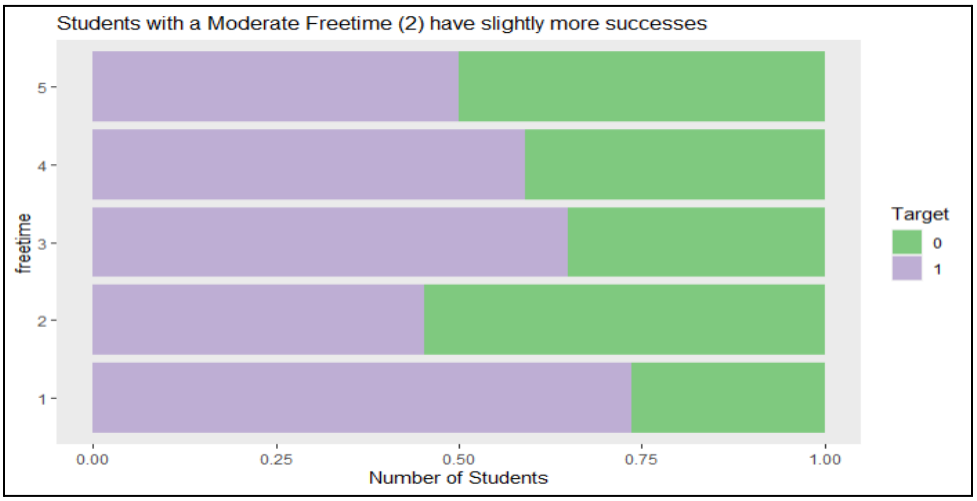

Figure 4

Students with more freetime are most likely the ones that were not taking the class seriously.

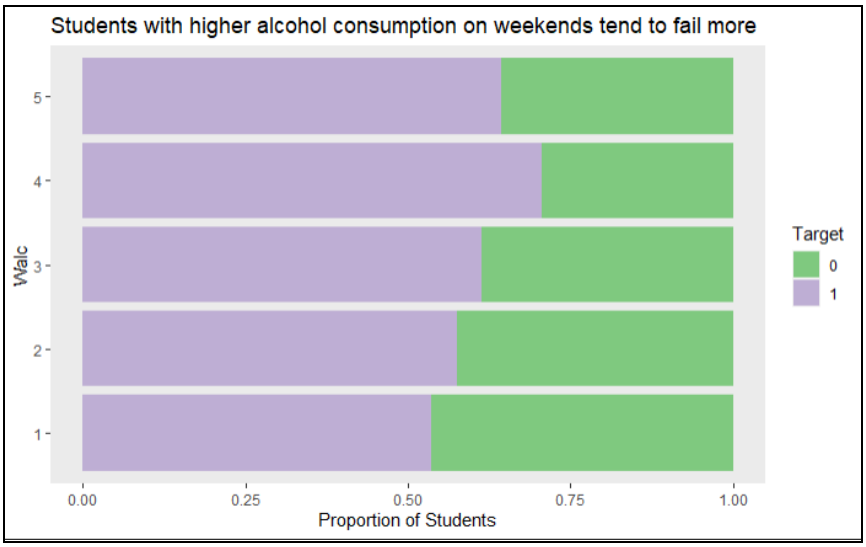

Figure 5

A direct increase was seen in students who failed the class as their weekend alcohol consumption increased.

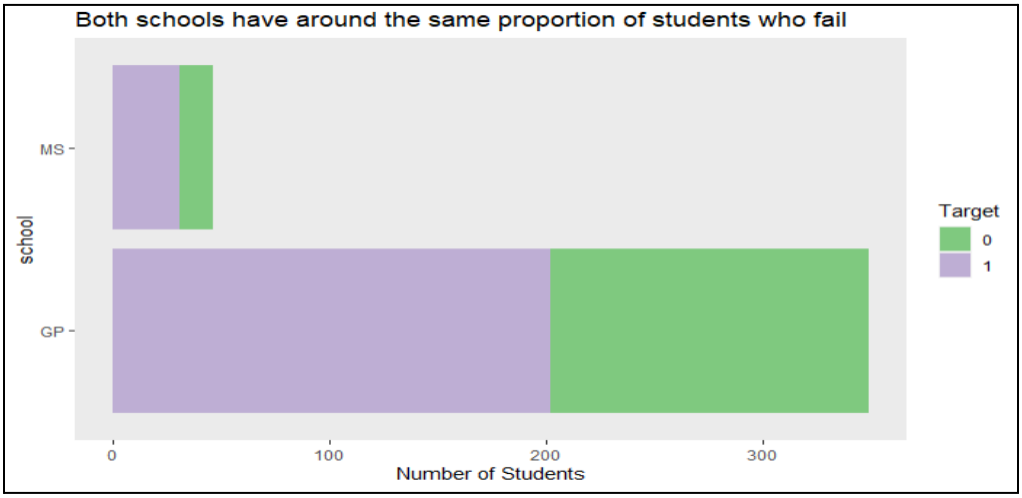

Figure 6

Both schools have around the same proportion of students that failed, suggesting that the environment isn't coming into play, the students are struggling with the course as a whole

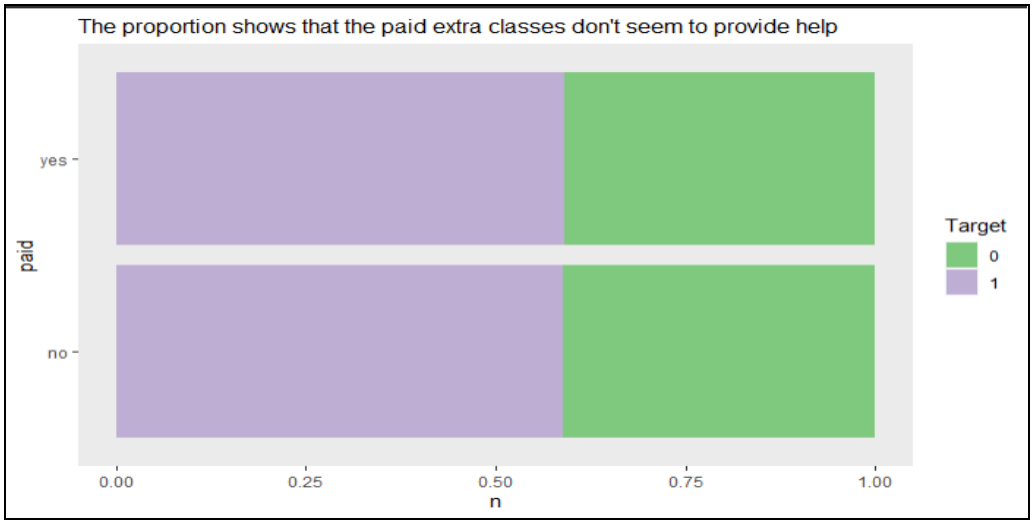

Figure 7

Paying for extra classes did not seem to provide students with help as zero improvement was seen between those who paid and didn't pay for the support.

Models / Results

Decision Tree

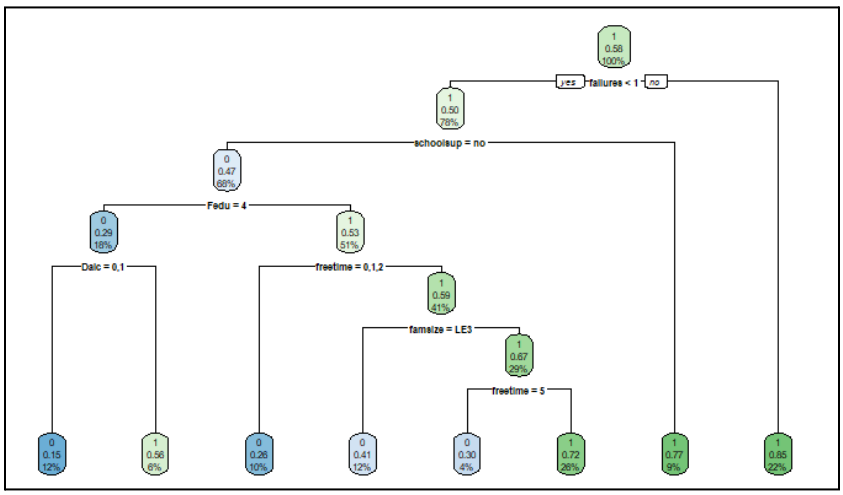

Figure 8

A Decision Tree is a map of possible outcomes of a series of related choices.

Results:
The original unpruned tree was a messy visual and it barely beat the NIR (No information rate - assuming every student failed). Then, after the check feature importance was implemented, the accuracy increased by 12.46%, which is the model shown above.

Logistic Regression

```
Call:
glm(formula = Target ~ ., family = binomial, data = train_student_1)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.18636  176.56193   0.001 0.999158
failures       1.37650    0.36597   3.761 0.000169 ***
freetime.L    -1.17747    0.65343  -1.802 0.071550 .
freetime.Q     0.41363    0.55495   0.745 0.456060
freetime.C    -1.38893    0.44043  -3.154 0.001613 **
freetime^4     0.80258    0.30510   2.631 0.008524 **
Fedu.L        -9.84347  558.29609  -0.018 0.985933
Fedu.Q         7.42671  471.84633   0.016 0.987442
Fedu.C        -4.87438  279.14817  -0.017 0.986068
Fedu^4         1.67696  105.50843   0.016 0.987319
Dalc.L        -0.59442    0.91504  -0.650 0.515943
Dalc.Q        -0.76375    0.70588  -1.082 0.279256
Dalc.C        -0.19206    0.72679  -0.264 0.791580
Dalc^4        -0.12978    0.70220  -0.185 0.853367
Walc.L         0.80760    0.73237   1.103 0.270153
Walc.Q         0.18474    0.55029   0.336 0.737091
Walc.C         0.05596    0.44279   0.126 0.899430
Walc^4        -0.57787    0.37159  -1.555 0.119911
schoolsupyes   1.80990    0.54057   3.348 0.000813 ***
absences       0.03500    0.02817   1.242 0.214121
age            0.14541    0.13151   1.106 0.268865
famrel.L       0.89688    0.69419   1.292 0.196361
famrel.Q      -0.03045    0.60494  -0.050 0.959851
famrel.C      -0.77025    0.59828  -1.287 0.197940
famrel^4       0.61239    0.48418   1.265 0.205942
famsizeLE3    -0.63690    0.33690  -1.890 0.058699 .
```

Figure 9

Logistic Regression is used to obtain the odds ratio in the presence of more than one explanatory variable.

Results:
Three different models were tested using logistic regression on the data set that only contained the top ten variables. The first model, which is shown above, included all of the variables in the dataset. This occurred to be the most accurate model, with an accuracy of 71.74%. The second model reduced the variables to just the ones that were significant in the first model. The accuracy slightly decreased to 69.93%. Lastly, the third model only included the variables that were significant from the second model. This model had the lowest accuracy out of all three, which is 62.68%.

k-NN

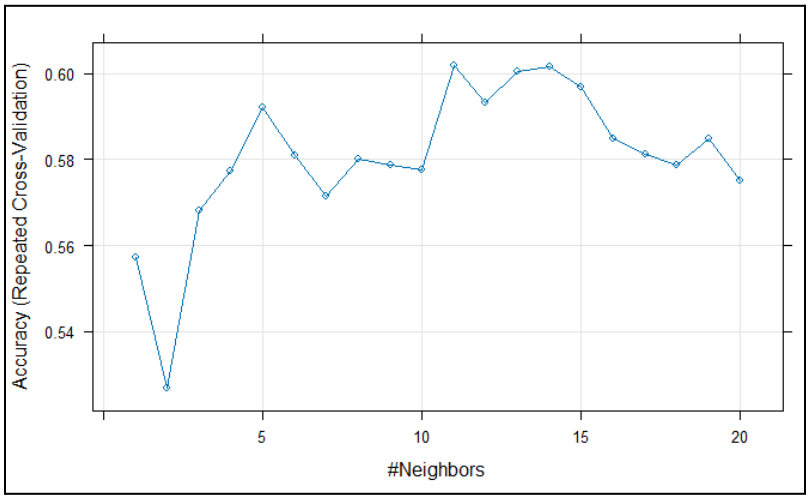

Figure 10a

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  69  39
         1  47 121

               Accuracy : 0.6884
                 95% CI : (0.6301, 0.7426)
    No Information Rate : 0.5797
    P-Value [Acc > NIR] : 0.0001303

                  Kappa : 0.3544

 Mcnemar's Test P-Value : 0.4503513

            Sensitivity : 0.5948
            Specificity : 0.7562
         Pos Pred Value : 0.6389
         Neg Pred Value : 0.7202
             Prevalence : 0.4203
         Detection Rate : 0.2500
   Detection Prevalence : 0.3913
      Balanced Accuracy : 0.6755

       'Positive' Class : 0
```

Figure 10b

K-Nearest Neighbors is used for classification. It classifies the data point on how its neighbor is classified.

Results:
This model was also performed using the data set with the top ten variables. Accuracy was used to select the optimal model using the largest value. Hence, the final value that was used for the model was k = 11. In the plot shown above, a peak can be seen at point k = 11, so that is what the model used. As a result, the accuracy of the model was 68.84%.

Association Rule Mining

	lhs <chr>		rhs <chr>
[1]	{age=[15,16), studytime=[2,4], schoolsup=yes}	=>	{Target=1}
[2]	{Medu=3, studytime=[2,4], schoolsup=yes}	=>	{Target=1}
[3]	{reason=home, nursery=yes, Walc=4}	=>	{Target=1}
[4]	{guardian=mother, goout=5, absences=[0,2)}	=>	{Target=1}
[5]	{schoolsup=no, goout=5, absences=[0,2)}	=>	{Target=1}
[6]	{Medu=1, Fedu=1, Mjob=other}	=>	{Target=1}
[7]	{sex=F, internet=no, absences=[2,6)}	=>	{Target=1}
[8]	{address=R, famsize=GT3, Walc=3}	=>	{Target=1}
[9]	{Fedu=1, romantic=yes, freetime=3}	=>	{Target=1}
[10]	{address=R, famsize=GT3, goout=4}	=>	{Target=1}

Figure 11

Association Rule Mining finds interesting connections within data sets. Specifically, how frequently a specific trend appears.

Results:
This method used the entire data set with all of the variables to produce common trends that lead to students having grades less than 55%. The image shown above were the top ten rules outputted by association rule mining. From top to bottom, the rules were sorted by lift value. All of the displayed rules had a lift of about 1.7, confidence of 1, and support of > 0.03.

SVM

```
Confusion Matrix and Statistics


polymodel1Pred   0   1
             0  89  23
             1  27 137

               Accuracy : 0.8188
                 95% CI : (0.7682, 0.8624)
    No Information Rate : 0.5797
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6265

 Mcnemar's Test P-Value : 0.6714

            Sensitivity : 0.7672
            Specificity : 0.8562
         Pos Pred Value : 0.7946
         Neg Pred Value : 0.8354
             Prevalence : 0.4203
         Detection Rate : 0.3225
   Detection Prevalence : 0.4058
      Balanced Accuracy : 0.8117
```

Figure 12

Support Vector Machine is a linear model used for classification and regression.

Results:

The data set that contained the top ten variables was used again. This model was tuned with a radial kernel. The cost parameter was equal to 0.95 and as a result, the accuracy was equal to 81.88%. This was a black box method that doesn't display the feature importance from the model, but given the accuracy, it was assumed that the top ten variables do play a large role in predicting the target variable.

Random Forest

```
Confusion Matrix and Statistics


rfmodel3Pred   0   1
           0 115   1
           1   1 159

               Accuracy : 0.9928
                 95% CI : (0.9741, 0.9991)
    No Information Rate : 0.5797
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9851

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9914
            Specificity : 0.9938
         Pos Pred Value : 0.9914
         Neg Pred Value : 0.9938
             Prevalence : 0.4203
         Detection Rate : 0.4167
   Detection Prevalence : 0.4203
      Balanced Accuracy : 0.9926
```

Figure 13

Random Forest is used to predict the things that help industries operate efficiently. In this case, which variables lead to student success and which lead to failure.

Results:
Once again, only the top ten variables were used in this method. This model had a ntree parameter equal to 500, which is the default, and a mtry parameter equal to four. The accuracy of the model was 99.28%. Overfitting could be present as a result of the high accuracy.

Results / Model Comparison

Model	Accuracy
No Information Rate	58.97%
Decision Tree	74.64%
Logistic Regression	71.74%
kNN	68.84%
SVM	81.88%
Random Forest	99.28%

Results:
Every model resulted in an accuracy percentage based on the analysis of the top ten variables from the check feature importance. When testing the whole data set versus the top ten variables, the accuracy of each model increased when using the narrowed-down version. Based on that, it can be concluded that those variables had the most weight in predicting the student's grades. From the results of all of the analysis models, it was clear that Random Forest had the highest accuracy. However, it may be overfitting. The model that had the second-highest accuracy was SVM, which is also a good predictor.

Conclusion

The problem trying to be resolved is that schools cannot seem to identify students who may need additional support in math before they are unsuccessful in the course. The goal was to determine what is contributing to the students failing the classes and what the schools can do to identify them early and offer additional assistance. Looking at details such as if the student feels they have educational support going into the year or their weekly alcohol consumption seems to be the key success factors for schools to help students pass these challenging classes.

Based on the findings, there are a few recommendations to offer to the two Portuguese schools in order to support the target students that are struggling. First, given the high accuracy levels of the Random Forest model, it is recommended to apply that method, after assuring that no overfitting has taken place by training and testing on more new data. Next, with the top ten identified variables, the school's survey to students can be drastically reduced which will promote better completion rates and accuracy. Lastly, distribute results to the schools for their upcoming classes to predict who might need assistance in math. Essentially, the main focus of the results is the top ten variables that have the most impact on predicting the students' final grades:

- Failures

- Freetime
- Fedu
- Dalc
- Walc
- Schoolsup
- Absences
- Age
- Famrel
- famsize