How Can Schools Improve Students' Math Grades?

Bing-Je Wu and Jason Maloney

IST 707 – DATA ANALYTICS

Table of Contents

Introduction

Education plays a vital role in our life. Knowing what factors can affect student's performance on test scores would be helpful to educators. We want to explore the "Student Grade Prediction" data set from Kaggle to understand the influence of the parent's background, test preparation and other factors on student's performance. There are three questions we are interested in:

1. What are the fundamental factors that will affect students' performance on their final grade?
2. Which factors influence poor performance on the final grade the most?
3. What would be the best way to improve student scores on their final grade?

Dataset

The "Student Grade Prediction" data set[i] contains 395 observations and 33 variables in total of 13,035 values. The data set can be found from Kaggle: https://www.kaggle.com/dipam7/student-grade-prediction

Attribute	Description	Data Type
school	Student's school: 'GP' – Gabriel Pereira or 'MS' – Mousinho da Silveira	Factor
Sex	Student's sex: 'F' – Female or 'M' – Male	Factor
Age	Student's age: '15 – 22'	Numeric
Address	Student's home address: 'U' – urban or 'R' – rural	Factor
Famsize	Family size: 'LE3' – Less than or equal to 3 or 'GT3' – Greater than 3	Factor
Pstatus	Parent's cohabitation status: 'T' – Living together or 'A' – Apart	Factor
Medu	Mother's education: '0' – None, '1' – Primary education (4th grade), '2' – 5th to 9th grade, '3' – Secondary education, or '4' – Higher education	Numeric
Fedu	Father's education: '0' – None, '1' – Primary education (4th grade), '2' – 5th to 9th grade, '3' – Secondary education, or '4' – Higher education	Numeric
Mjob	Mother's job: 'teacher', 'health' care related, civil 'services' (administrative or police), 'at_home', or 'other'	Factor
Fjob	Father's job: 'teacher', 'health' care related, civil 'services' (administrative or police), 'at_home', or 'other'	Factor
Reason	Reason to choose this school: close to 'home', school 'reputation', 'course' preference, or 'other'	Factor
Guardian	Student's guardian: 'mother', 'father', or 'other'	Factor
Traveltime	Home to school travel time: '1' - <15 min, '2' – 15 to 30 min, '3' – 30 min to 1 hr, '4' - >1 hr	Numeric
Studytime	Weekly study time:	Numeric

	‘1’ - <2 hours, ‘2’ – 2 to 5 hours, ‘3’ – 5 to 10 hours, ‘4’ - >10 hours	
Failures	Number of past class failures: ‘n’ if 1 ≤ n <3, else 4	Numeric
Schoolsup	Extra educational support: ‘yes’ or ‘no’	Factor
Famsup	Family educational support: ‘yes’ or ‘no’	Factor
Paid	Extra paid classes within the course subject: ‘yes’ or ‘no’	Factor
Activities	Extra-curricular activities: ‘yes’ or ‘no’	Factor
Nursery	Attend nursery school: ‘yes’ or ‘no’	Factor
Higher	Wants to take higher education: ‘yes’ or ‘no’	Factor
Internet	Internet access at home: ‘yes’ or ‘no’	Factor
Romantic	With a romantic relationship: ‘yes’ or ‘no’	Factor
Famrel	Quality of family relationships: (1 to 5) 1 – very bad 5 – excellent	Numeric
Freetime	Free time after school: (1 to 5) 1 – very low 5 – very high	Numeric
Goout	Going out with friends: (1 to 5) 1 – very low 5 – very high	Numeric
Dalc	Workday alcohol consumption: (1 to 5) 1 – very low 5 – very high	Numeric
Walc	Weekend alcohol consumption: (1 to 5) 1 – very low 5 – very high	Numeric
Health	Current health status: (1 to 5) 1 – very bad 5 – very good	Numeric
Absences	Number of school absences: 0 to 93	Numeric
G1	First period grade: 0 to 20	Numeric
G2	Second period grade: 0 to 20	Numeric
G3	Final grade: 0 to 20 (output target)	Numeric

Methods for Analysis

We start with data exploration to understand the distribution of each variable and perform transformation on some of variables in order to derive more insights. Then, we will implement Association Rule mining technique using the Apriori algorithm to learn the key factors affecting students’ performance from the strong rules. Further, Clustering technique, SVM technique, Naïve Bayes technique and Random Forest are applied for the analysis and prediction. We will split data set into train data and test data. Then, use train data to build prediction models and validate models with test data. Last, we will summarize and conclude the 3 research questions based on the exploratory analysis and data modeling analysis.

Data and Data Processing

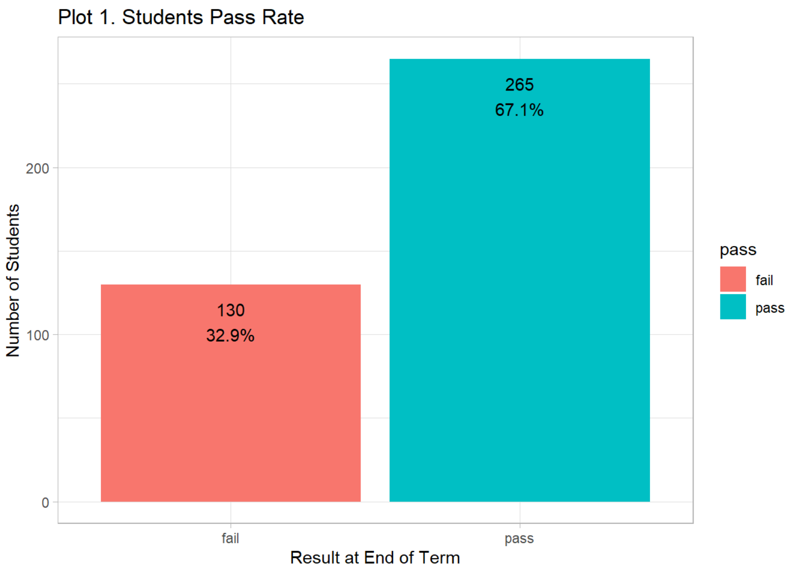
After importing the dataset, first thing to do is check if there is any missing values and the structure of the dataset. The output we got showing that there are no missing values in our dataset. Str() command was used for overviewing the dataset from high level:

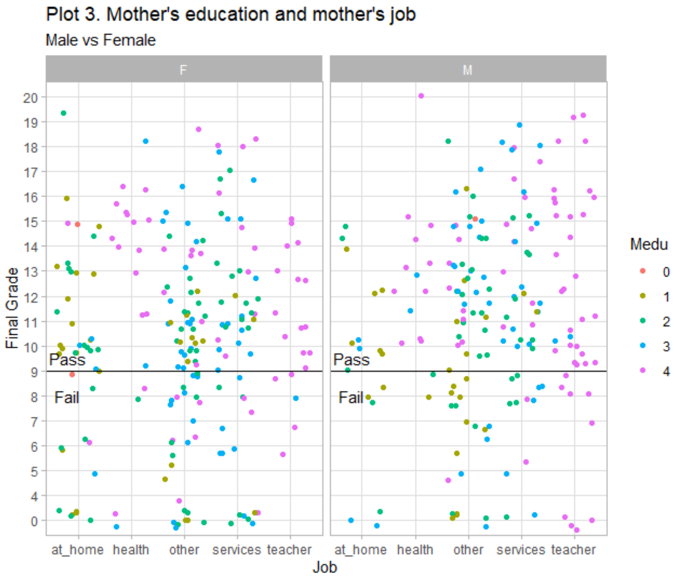
```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 395 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : num  18 17 15 15 16 16 16 17 15 15 ...
## $ address   : chr  "U" "U" "U" "U" ...
## $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
## $ Pstatus   : chr  "A" "T" "T" "T" ...
## $ Medu      : num  4 1 1 4 3 4 2 4 3 3 ...
## $ Fedu      : num  4 1 1 2 3 3 2 4 2 4 ...
## $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
## $ Fjob      : chr  "teacher" "other" "other" "services" ...
## $ reason    : chr  "course" "course" "other" "home" ...
## $ guardian  : chr  "mother" "father" "mother" "mother" ...
## $ traveltime: num  2 1 1 1 1 1 1 2 1 1 ...
## $ studytime : num  2 2 2 3 2 2 2 2 2 2 ...
## $ failures  : num  0 0 3 0 0 0 0 0 0 0 ...
## $ schoolsup : chr  "yes" "no" "yes" "no" ...
## $ famsup    : chr  "no" "yes" "no" "yes" ...
## $ paid      : chr  "no" "no" "yes" "yes" ...
## $ activities: chr  "no" "no" "no" "yes" ...
## $ nursery   : chr  "yes" "no" "yes" "yes" ...
## $ higher    : chr  "yes" "yes" "yes" "yes" ...
## $ internet  : chr  "no" "yes" "yes" "yes" ...
## $ romantic  : chr  "no" "no" "no" "yes" ...
## $ famrel    : num  4 5 4 3 4 5 4 4 4 5 ...
## $ freetime  : num  3 3 3 2 3 4 4 1 2 5 ...
## $ goout     : num  4 3 2 2 2 2 4 4 2 1 ...
## $ Dalc      : num  1 1 2 1 1 1 1 1 1 1 ...
## $ Walc      : num  1 1 3 1 2 2 1 1 1 1 ...
## $ health    : num  3 3 3 5 5 5 3 1 1 5 ...
## $ absences  : num  6 4 10 2 4 10 0 6 0 0 ...
## $ G1        : num  5 5 7 15 6 15 12 6 16 14 ...
## $ G2        : num  6 5 8 14 10 15 12 5 18 15 ...
## $ G3        : num  6 6 10 15 10 15 11 6 19 15 ...
```

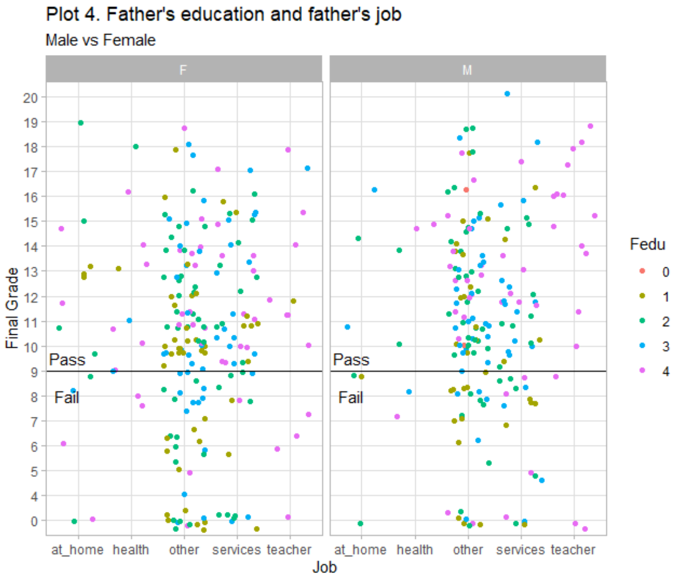
To perform clustering analysis and association rule mining analysis, some preprocessing steps need to be done. The dataset needs to be changed to a certain format for each type of algorithm. For clustering analysis, we converted nominal and ordinal variables into numeric variables. And for association rule mining analysis, we converted all attributes into factor variables.

Exploratory Analysis

When we looked at the data, the first question came to our minds was how many students pass the Math course? Students whose final grade, G3, are greater than 9 are considered pass. The Plot 1 shows that there are 130 students (32.9%) that failed the Math course. The pass rate is 67.1% in combination of the two schools. Clearly, there is room for improvement for educators to find out the reasons why students failed the Math course and help them overcome the exams. We utilized association rule mining technique to identify some of the rules that lead to passing the final grade. We noticed *parents' job*, *parents' education*, *gender* variables are appeared quite often. Thus, we visualized those variables and try to see the patterns.

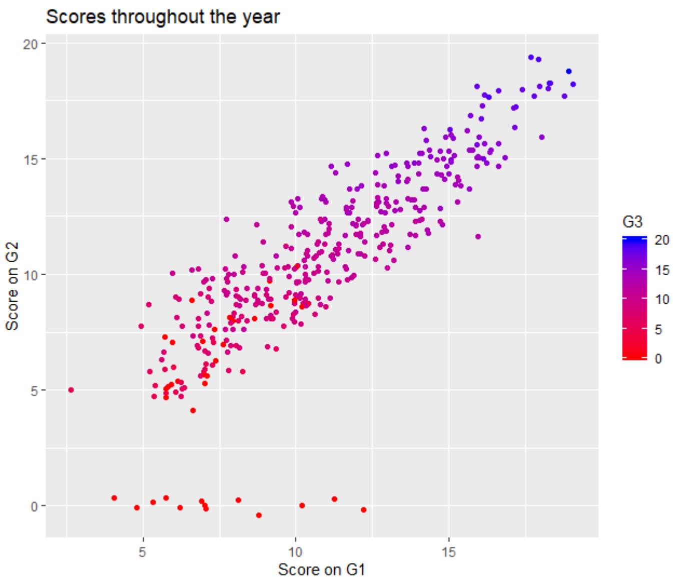






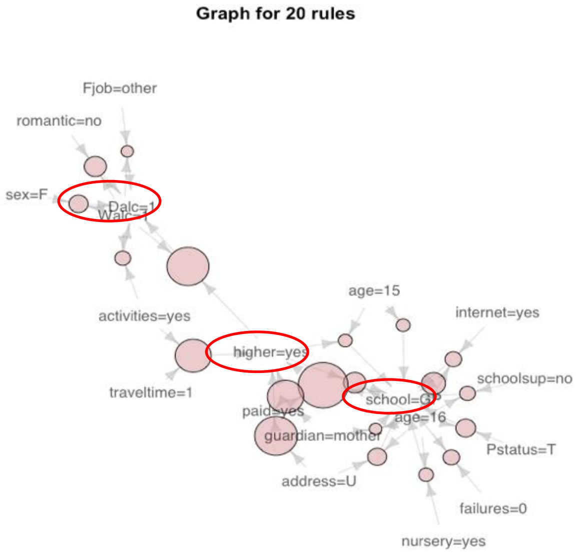
The Plot 3 above shows that males with their mother working as a teacher tend to score high on the final grade and Plot 4 shows that males with father working as a teacher tend to score high on the final grade. Both plots also show that there is large variability in the final grade with students whose parents are in the service industry or have a different occupation. This may discount these variables as high level predictor variables.

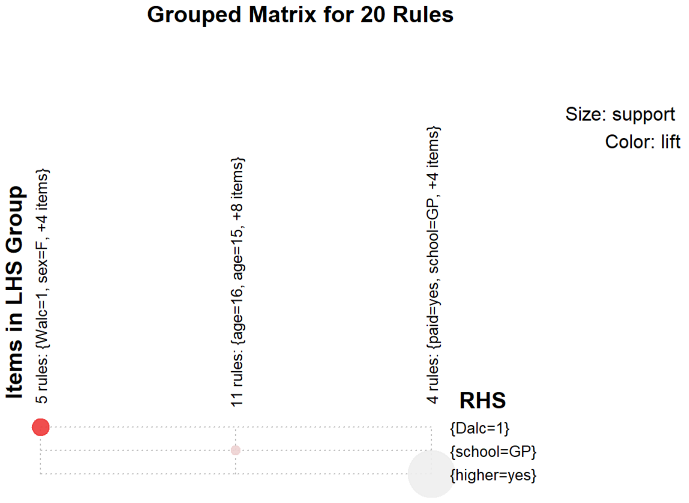
In order to answer the research question 1, "What are the fundamental factors that will affect students' performance on their final grade?", we first looked at the relationship on G1 Score, G2 Score and G3 Score. From the plot below, we can see a pattern that students who do well in the first and second term tend to score high in their final grade. There are several students who perform poorly on the first and second term and do not finish the school year or do not earn a grade at the end of the year. The main factors that influence the performance on the final grade is the performance in the previous two terms, G1 and G2. We will look at the dataset further to find others that are not as obvious.



Data Modeling

Association Rules Mining





To find the rules on passing a Math course, we discretized the final grade into 5 categories, A, B, C, D, and F. Category A, B, C and D are considered Pass. Initially, we ran the apriori algorithm from high level without setting an output constraint. Some of the interesting variables include:

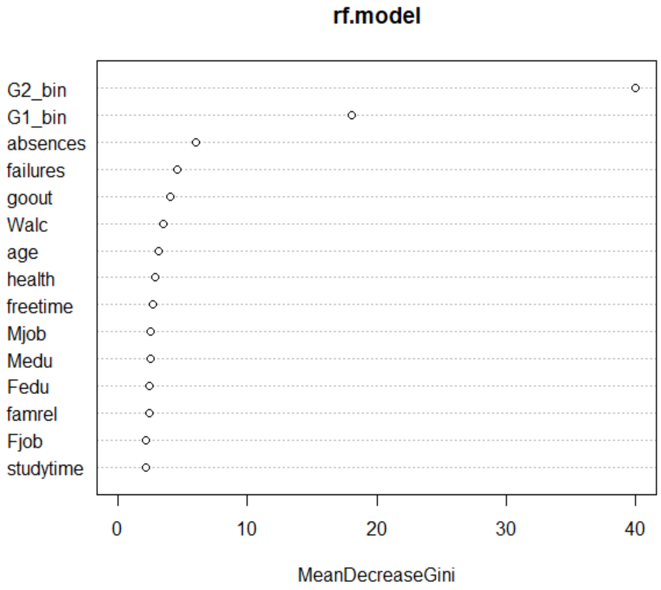
- Female students have low alcohol consumption level
- Students at age 15 want to get higher education
- Students who have less travel time, have taken extra paid classes tend to get higher education as well
- Female students want to get higher education

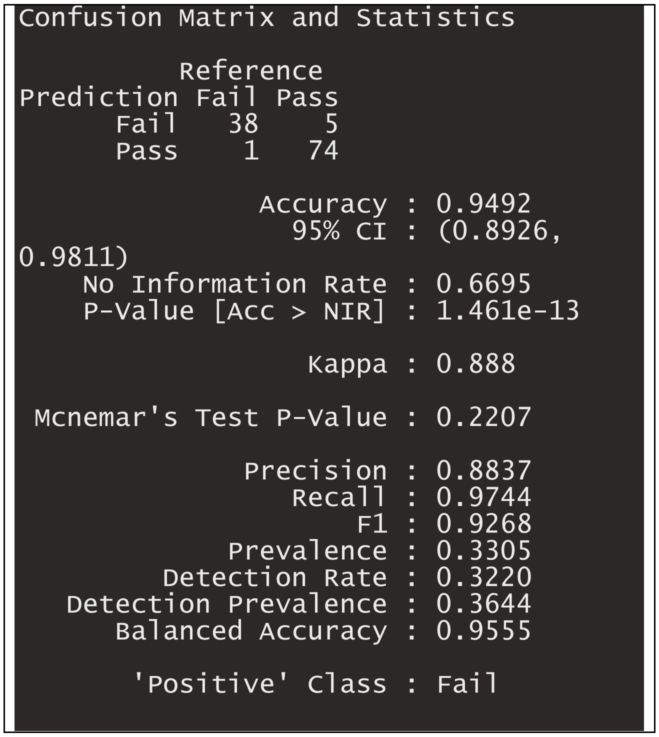
When we run the apriori algorithm to find predictors of passing grades at the end of the term, some common associations are comprised of:

- Mother is a teacher
- Parents have high education level
- Reason choosing school is because of courses offered
- Have free time after school
- Alcohol consumption is high

Random Forest

The random forest model allows us to easily identify which variables have the strongest influence in the predicted variable. By utilizing the feature analysis, we can clearly see that G2 grades in letter binning, G1 grades in letter binning, the number of absences, and the number of previous course failures have effective impacts on predicting students passing the Math course.

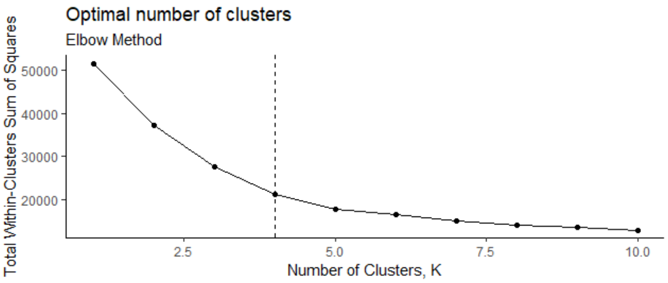


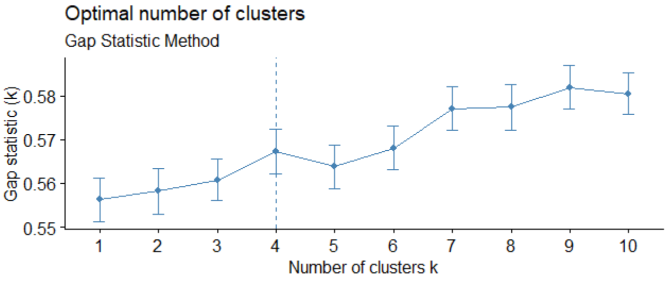


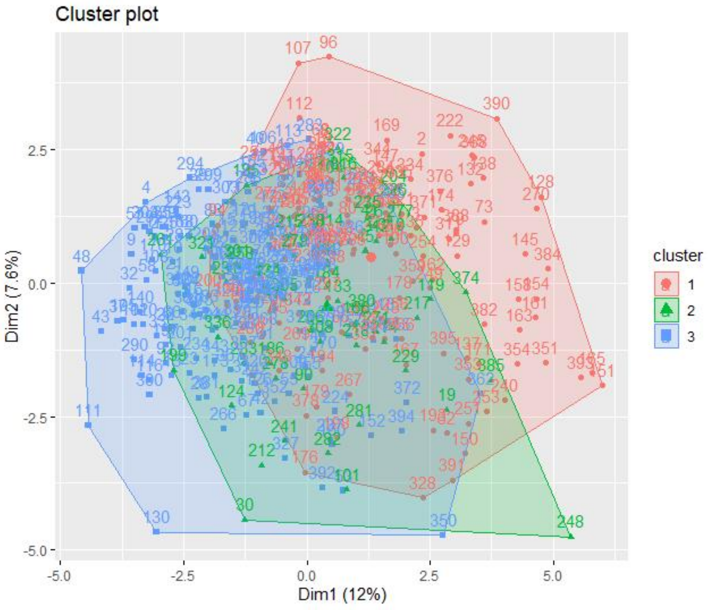
```
Confusion Matrix and Statistics

          Reference
Prediction Fail Pass
      Fail   38    5
      Pass    1   74

               Accuracy : 0.9492
                 95% CI : (0.8926,
0.9811)
    No Information Rate : 0.6695
    P-Value [Acc > NIR] : 1.461e-13

                  Kappa : 0.888

 Mcnemar's Test P-Value : 0.2207

              Precision : 0.8837
                 Recall : 0.9744
                     F1 : 0.9268
             Prevalence : 0.3305
         Detection Rate : 0.3220
   Detection Prevalence : 0.3644
      Balanced Accuracy : 0.9555

       'Positive' Class : Fail
```

K-means Clustering

We utilized clustering technique to find the answer for the research question 2, Which factors influence poor performance on the final grade the most? First, we implemented the Elbow Method and Gap Statistic Method to determine the optimal number of clusters, k = 4. However, when we use 4 or more clusters, there is one cluster has very few observations. Therefore, we decided to set k = 3 as optimal clusters. Using 3 clusters, the algorithm is relatively successful in grouping students who passed into one cluster, students who failed into another, and the third cluster consists of a combination of those students. This suggests that these are the students that are most difficult to classify using this technique.

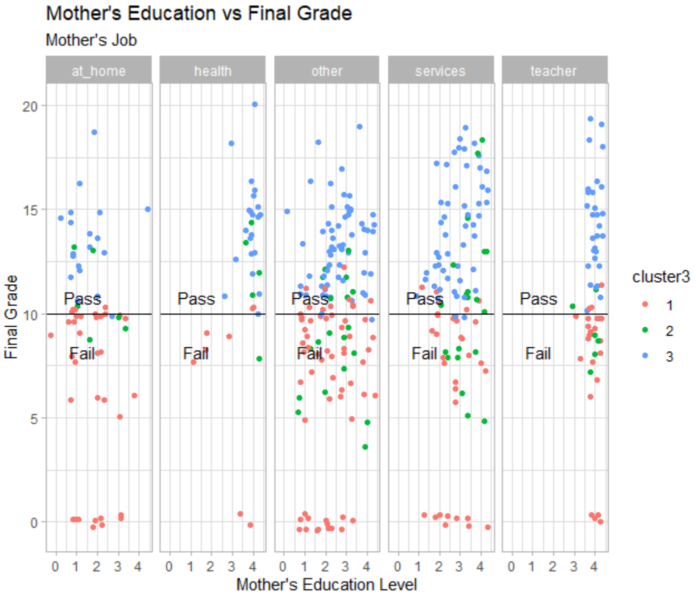


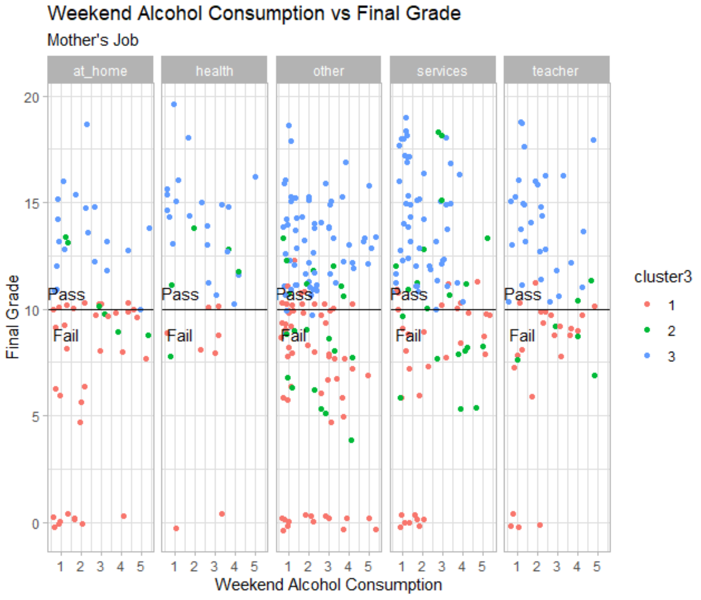


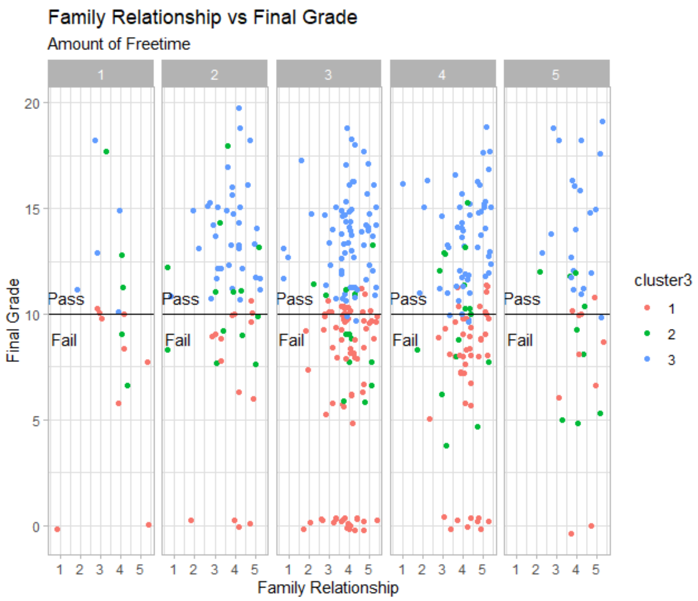


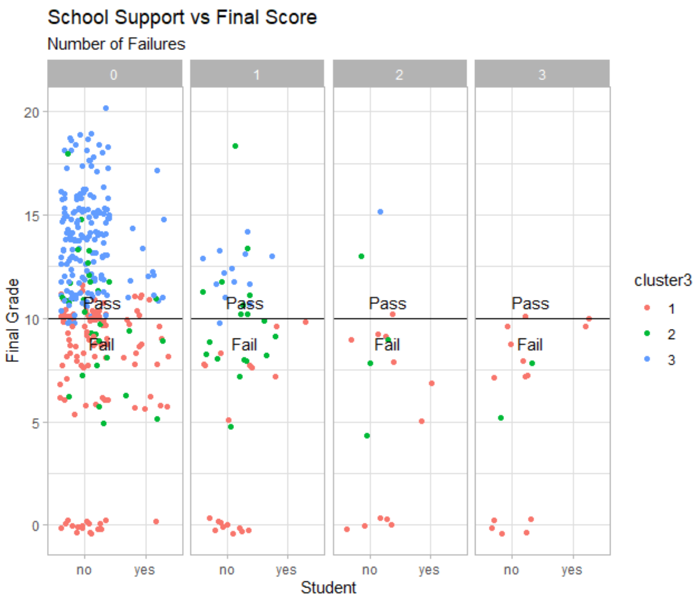
Combining the information from the association rule mining, we investigated the influence the following variables have on students' final grades:

- Mother's education level and job
- Weekend alcohol consumption
- Family relationship with amount of free time
- Number of failures with school support









The plot on the top-left above shows that when a student's mother's education level is at 3 or 4, they are more likely to have a passing grade at the end of the course. Student's mothers who are teachers and in health care have a high pass percentage. Students whose mothers work at home, in services, or in other industries do not have an advantage in passing the course.

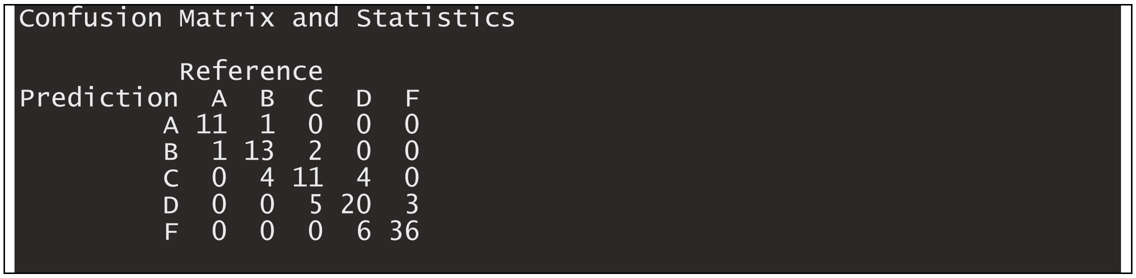
The plot on the top-right above shows that, as noted from Association Rules, students who have high rates of *weekend alcohol consumption* surprisingly do not have significantly lower scores at the end of the course. Students who have a low to moderate level of *weekend alcohol consumption* do have higher rates of passing.

The plot on the bottom-left above shows that students with low to moderate amounts of free time tend to have better grades on G3. In addition, the better the *family relationship*, 1 = very bad and 5 = great, the more likely a student is to perform well on G3 and pass the course.

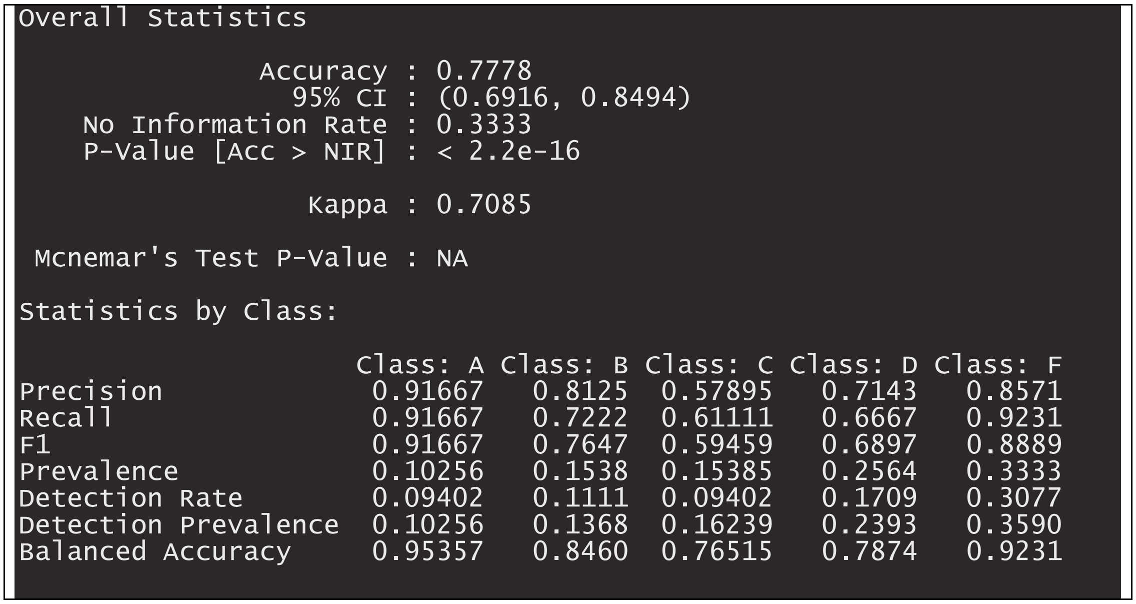
The plot on the bottom-right above shows that most of students do not have school supports and students who have failures in the past without school support tend to fail again in the final. Clearly, school support is a factor that can have a large impact on each grade period.

Naïve Bayes

We used the top 15 important variables that we got from Random Forest model to improve the Naïve Bayes model. By using the primary 15 predictor variables, the Naïve Bayes classifier improves to 77.78% accurate at predicting a student's performance at the end of the course. It is correctly able to identify most students who fail a course. It incorrectly predicted 6 students earning an F and 3 earning a D when they actually earned a D and a F, respectively.



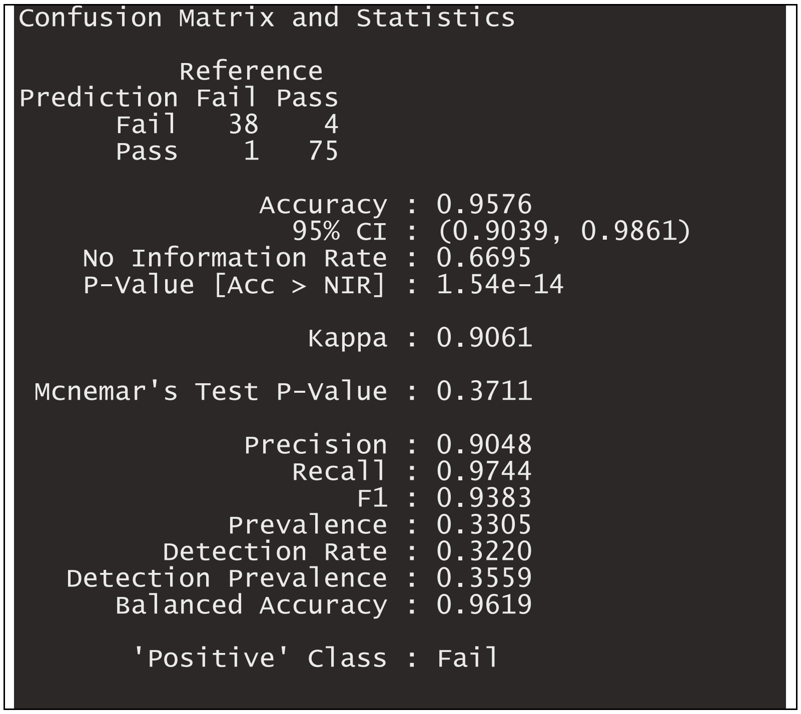
```
Confusion Matrix and Statistics

          Reference
Prediction  A  B  C  D  F
         A 11  1  0  0  0
         B  1 13  2  0  0
         C  0  4 11  4  0
         D  0  0  5 20  3
         F  0  0  0  6 36
```



```
Overall Statistics

               Accuracy : 0.7778
                 95% CI : (0.6916, 0.8494)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7085

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: A Class: B Class: C Class: D Class: F
Precision             0.91667   0.8125  0.57895   0.7143   0.8571
Recall                0.91667   0.7222  0.61111   0.6667   0.9231
F1                    0.91667   0.7647  0.59459   0.6897   0.8889
Prevalence            0.10256   0.1538  0.15385   0.2564   0.3333
Detection Rate        0.09402   0.1111  0.09402   0.1709   0.3077
Detection Prevalence  0.10256   0.1368  0.16239   0.2393   0.3590
Balanced Accuracy     0.95357   0.8460  0.76515   0.7874   0.9231
```

Support Vector Machine

We applied the SVM model with the Linear kernel. It yielded the optimal results of predicting at 95.76% accurate. By letting the objective variable only take on two classes, Pass and Fail, the model is able to predict the final result of a student. Here we cannot accept the prediction on False Negative, type II error, which means students actually Fail but are predicted to Pass. Therefore, we aim to find a model with higher Recall. And the model has the highest Precision and Recall among others. In addition, the SVM model with Linear kernel shows that the Student Grade Prediction dataset is linear separable.



```
Confusion Matrix and Statistics

          Reference
Prediction Fail Pass
      Fail   38    4
      Pass    1   75

               Accuracy : 0.9576
                 95% CI : (0.9039, 0.9861)
    No Information Rate : 0.6695
    P-Value [Acc > NIR] : 1.54e-14

                  Kappa : 0.9061

 Mcnemar's Test P-Value : 0.3711

              Precision : 0.9048
                 Recall : 0.9744
                     F1 : 0.9383
             Prevalence : 0.3305
         Detection Rate : 0.3220
   Detection Prevalence : 0.3559
      Balanced Accuracy : 0.9619

       'Positive' Class : Fail
```

Conclusion

The largest indicators for a student's performance at the end of the course are his or her performance in the previous two terms, G1 and G2. Other factors that have large impacts on these scores are the number of absences, the number of previously failed courses, the mother's and/or father's education level, and the job held by the mother.

If a student fails a course and does not have support from the school via tutoring or afterschool help, the student is likely to fail again. In order to help the students, who have poor performance in the past, a mentor or a tutor providing educational support can make a huge difference on their performance.

Based on our analysis, we recommend the schools provide support to students throughout the duration of a Math course. If students are successful in G1 and G2, it is highly probable they will be successful for the end of the course, G3. Students who had no access to school support systems have higher failure rates. This in turn will help the students to pass a course the first time they take it. Thus, the schools can implement an early warning policy where instructors report to a tutoring center or academic advisor those students who are in danger of failing. This will provide the schools with more information so that they can reach out to struggling students and offer help sooner.

Reference

[i] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. Web Link