Student Academic Performance Project

Analysis of Factors that contribute to Student Academic Performance

Lab Group - Rikki Kendall, Tolulemi Gbile, Geary Stonesifer

Data

Presented below is a view of our dataset.

```
# A tibble: 480 x 17
   gender NationalITy PlaceofBirth StageID GradeID SectionID Topic Semester
   <chr>  <chr>       <chr>        <chr>   <chr>   <chr>     <chr> <chr>
 1 M      KW          KuwaIT       lowerl~ G-04    A         IT    F
 2 M      KW          KuwaIT       lowerl~ G-04    A         IT    F
 3 M      KW          KuwaIT       lowerl~ G-04    A         IT    F
 4 M      KW          KuwaIT       lowerl~ G-04    A         IT    F
 5 M      KW          KuwaIT       lowerl~ G-04    A         IT    F
 6 F      KW          KuwaIT       lowerl~ G-04    A         IT    F
 7 M      KW          KuwaIT       Middle~ G-07    A         Math  F
 8 M      KW          KuwaIT       Middle~ G-07    A         Math  F
 9 F      KW          KuwaIT       Middle~ G-07    A         Math  F
10 F      KW          KuwaIT       Middle~ G-07    B         IT    F
# ... with 470 more rows, and 9 more variables: Relation <chr>,
#   raisedhands <dbl>, VisITedResources <dbl>, AnnouncementsView <dbl>,
#   Discussion <dbl>, ParentAnsweringSurvey <chr>,
#   ParentschoolSatisfaction <chr>, StudentAbsenceDays <chr>, Class <chr>
```

Section 1 - Introduction

Our general research question is as follows: Is it possible to predict a child's class level based off of the information provided in the dataset? If so, which variables are of significance? The information we're going to be using to unearth the answer to this research question is from Kaggle.com. The information was collected from a Learning Management System called Kalboard 360. The data is collected from a learner activity tracker tool which tracks the students behavior, like if they're reading an article, watching an informative video, or doing their classwork. The variables used in this data set and their data types are as follows:

Variables

- Gender - Student's Gender (categorical)
- Nationality - The student's nationality (categorical)
- Place of birth - the student's place of birth (categorical)
- Educational Stage - Elementary, Middle, or High school? (categorical)
- Grade Level - Grade student is in (categorical)
- Section ID - Classroom the student belongs to (categorical)
- Topic - Course topic (categorical)

- Semester - School Semester (discrete numerical)
- Parent Responsible for Student (categorical)
- Raised Hand - How many times the student raises their hand in the classroom (continuous numerical)
- Visited Resources - How many times a student uses the course content (continuous numerical)
- Viewing Announcements - How many times the student checks the new announcements (continuous numerical)
- Discussion Groups - How many times a student participates in discussion groups (continuous numerical)
- Parent Answering Survey - Did the parent answer the surveys provided by the school? (categorical)
- Parent School Satisfaction - Is the parent satisfied with the school? (categorical)
- Student Absence days - How many times the student has missed school (categorical)
- Class - Low, Medium, or High, depending on their grades/marks at the end of the semester (categorical)

Section 2 - Data analysis plan

The aim of our project is to attempt to predict a child's class based on relevant variables provided in the data set. In order to do so we will be using two different variables: the 'outcome' (dependent, response, Y), which is the child's class (L,M,H) and the 'predictor' (independent, explanatory, X) which can be any combination of the aforementioned variables in section 1.

Predictor

Explanatory Variables: Any combination of the variables listed above.

Outcome

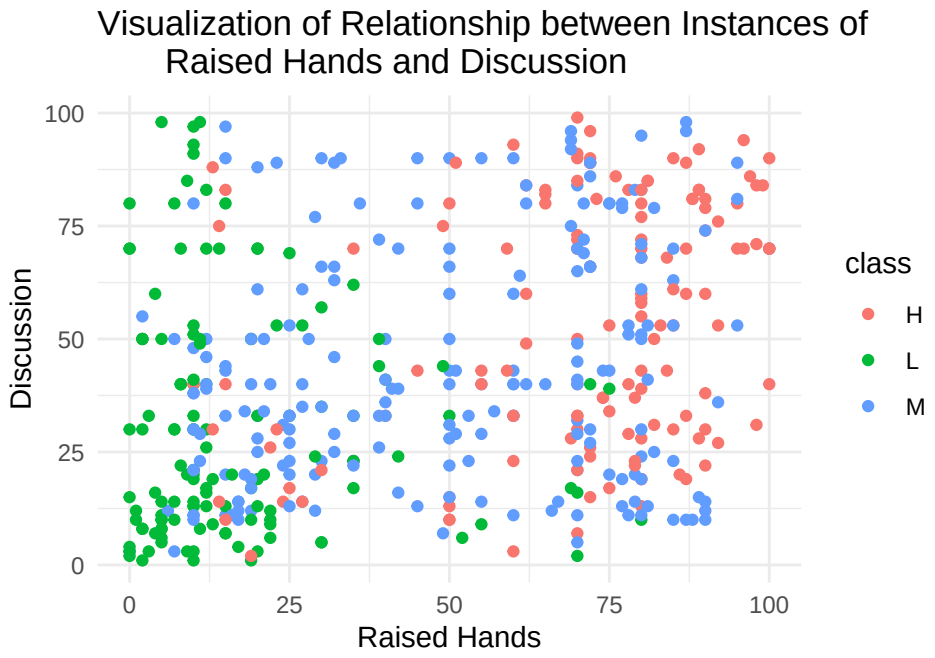
Response Variable: Class (3 Levels: High, Medium, Low)

Summary Statistics/Visualizations

In order to attempt to prove that class is predictable, we will run k-NN on the 3 levels of class as well as build logistic models. With the given correlations expected, we should be able to extrapolate a child's class. However, what is crucial is to pick the most relevant of the variables. Initially, we will screen the variables based on intuition as to which are obviously not useful in determining class.

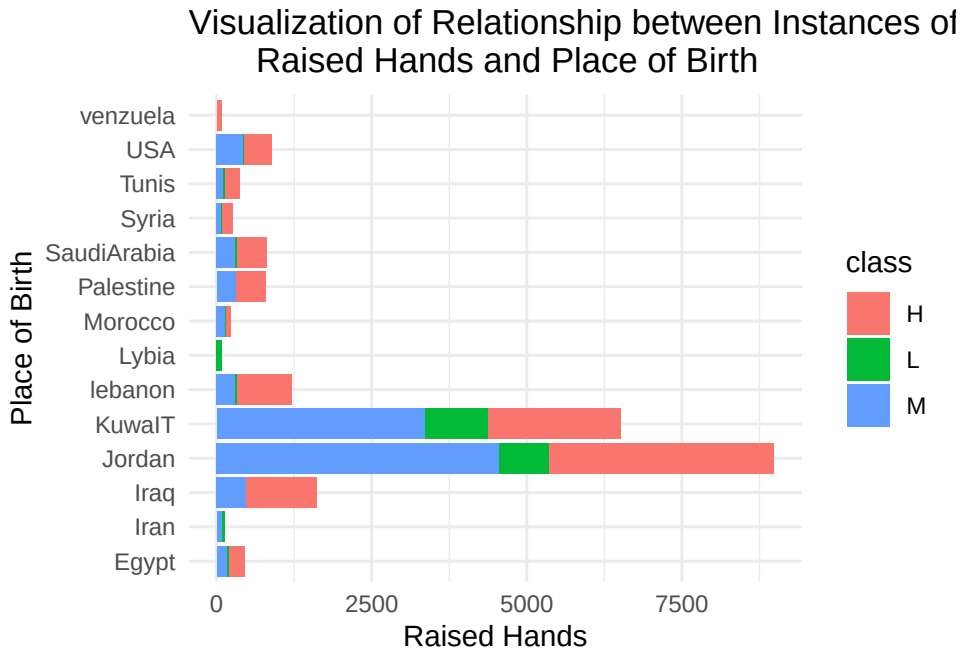
However, at first to portray the relevance of the variables we have provided several visualizations below to learn more about the data:

The first graph explores the relationship between the instances of raised hands and discussion. These two variables, although not necessarily dependent on one another, intuitively go hand in hand since both variables are related to class participation.

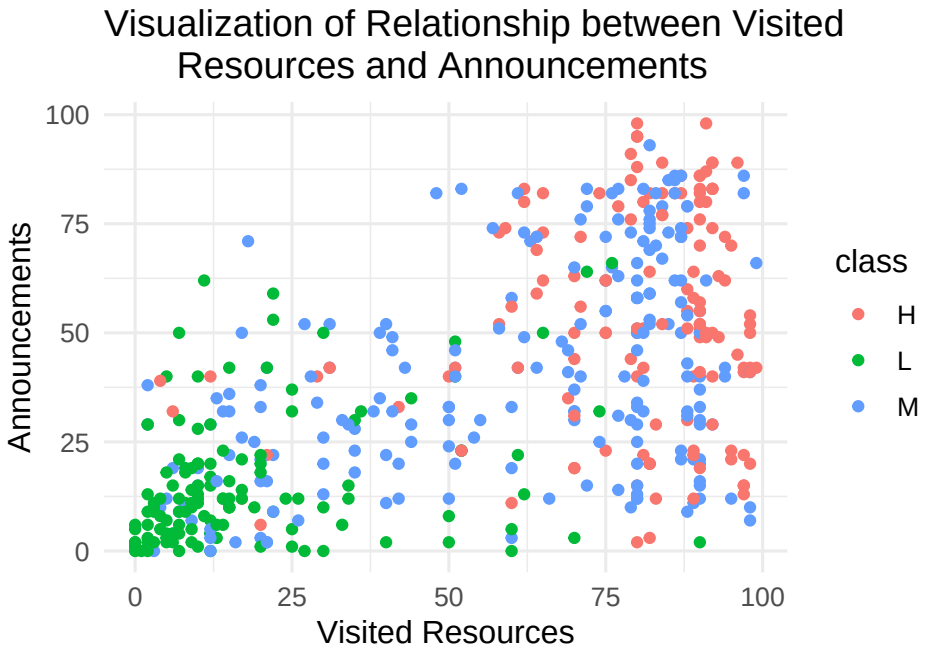


Given the disparity in results, whereby lower class is visibly on the lower spectrum of both Discussion and Raisedhands these variables will be important.

The following visualization corroborates the data in the prior scatter-plot for each country. This is to say that there is a clear trend among class in various countries given the variable, 'raisedhands.'

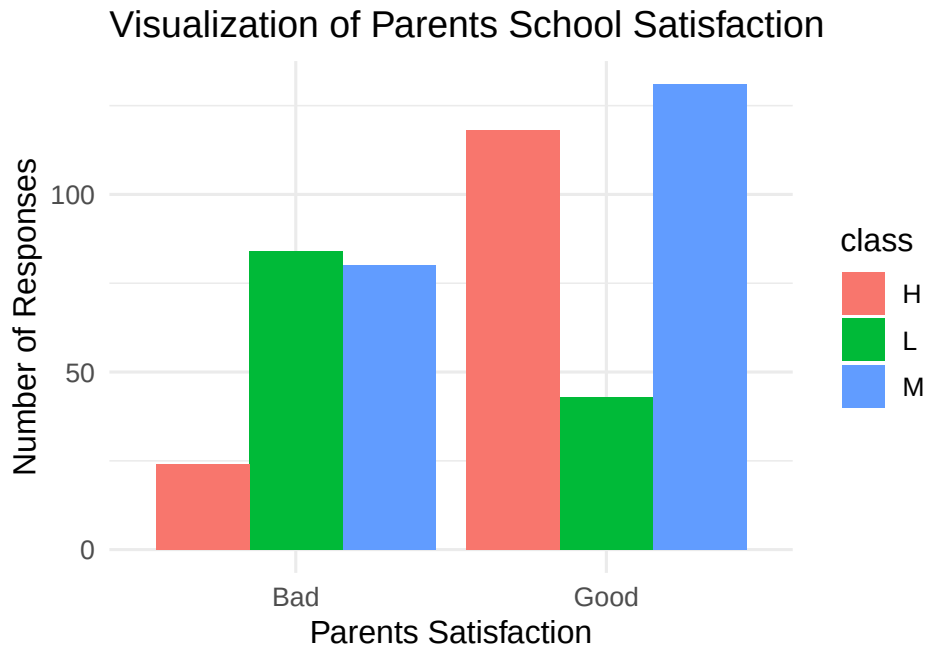


To continue to explore the variables, the next scatterplot takes two more data entries to examine a connection between class and Visited Resources / Announcements.



Again, we begin to see some trends when plotting relevant data side by side. Even more so in this visualization than the first scatter-plot, there is clear distinction between the classes of students.

Lastly, I have attempted to draw a connection to class using Parents satisfaction to examine a potential relationship with class.



This data provides interesting insight to the fact that in the Higher class bracket, there is a significant reduction in instances of negative feedback to any given school. Conversely, it appears the lower class bracket more often has negative responses to their school.

Given the visualizations of the variables above, it appears there are significant trends in the data that we aim to use to answer our research question: Is it possible to predict a child's class level based off of the information provided in the dataset?

Section 3 - Data

```
Observations: 480
Variables: 17
$ gender                  <chr> "M", "M", "M", "M", "M", "F", "M", "M", "F...
$ nationality             <chr> "KW", "KW", "KW", "KW", "KW", "KW", "KW", ...
$ placeofbirth            <chr> "KuwaIT", "KuwaIT", "KuwaIT", "KuwaIT", "K...
$ stageid                 <chr> "lowerlevel", "lowerlevel", "lowerlevel", ...
$ gradeid                 <chr> "G-04", "G-04", "G-04", "G-04", "G-04", "G...
$ sectionid               <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A...
$ topic                   <chr> "IT", "IT", "IT", "IT", "IT", "IT", "Math"...
$ semester                <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F...
$ relation                <chr> "Father", "Father", "Father", "Father", "F...
$ raisedhands             <dbl> 15, 20, 10, 30, 40, 42, 35, 50, 12, 70, 50...
$ visitedresources        <dbl> 16, 20, 7, 25, 50, 30, 12, 10, 21, 80, 88,...
$ announcementsview       <dbl> 2, 3, 0, 5, 12, 13, 0, 15, 16, 25, 30, 19,...
$ discussion              <dbl> 20, 25, 30, 35, 50, 70, 17, 22, 50, 70, 80...
$ parentansweringsurvey   <chr> "Yes", "Yes", "No", "No", "No", "Yes", "No...
$ parentschoolsatisfaction <chr> "Good", "Good", "Bad", "Bad", "Bad", "Bad"...
$ studentabsencedays      <chr> "Under-7", "Under-7", "Above-7", "Above-7"...
$ class                   <chr> "M", "M", "L", "L", "M", "M", "L", "M", "M...
```

Section 4 - Methods and Results

K-NN Approach

```
# A tibble: 10 x 1
     ten
   <dbl>
 1 0.54
 2 0.48
 3 0.54
 4 0.52
 5 0.56
 6 0.580
 7 0.6
 8 0.580
 9 0.54
10 0.5
```

```
# A tibble: 1 x 1
    ten
  <dbl>
1   0.6
```

```
[1] 7
```

In creating the training set, our team took into account the continuous numerical variables of raised hands, visited resources, discussion, and announcements view. The K-NN method on predicting the class produced a 60% prediction accuracy taking 50 observations using binary classification, with 5 being the value of k that gave us the highest accuracy. This could mean that the data was less differentiable with the middle class grades included. Therefore, the next section will conduct a K-NN method taking into account only the high and low class grades. This will be compared to a logistic regression model to compare the prediction accuracies between both methods.

K-NN w/Just High and Low Class Grades

```
# A tibble: 10 x 1
     ten
   <dbl>
 1  0.9
 2  0.92
 3  0.92
 4  0.9
 5  0.94
 6  0.94
 7  0.94
 8  0.94
 9  0.94
10  0.94
```

```
# A tibble: 1 x 1
    ten
  <dbl>
1  0.94
```

```
[1] 5
```

Results:

In creating the training set, our team took into account the same continuous numerical variables of raised hands, visited resources, discussion, and announcements view. The K-NN method on predicting the class produced a 94% prediction accuracy after removing the middle class grades, with 5 being the value of k that gave us the highest accuracy.

1. Motivation to use K-NN: This is a classification problem. In addition to a logistic regression model, K-NN is a good method to for classification and to make comparisons between the two models.
2. Prediction Results: 94% accuracy indicates that the variables that we used are good indicators for classification.

However, K-NN does not look at each coefficient specifically so our team decided to fit a logistic regression model and compare the prediction accuracies using the same variables.

Logistic Regression Model

```
# A tibble: 5 x 2
  term              estimate
  <chr>                <dbl>
1 (Intercept)         -6.56
2 raisedhands          0.035
3 visitedresources     0.064
4 discussion           0.03
5 announcementsview    0.022
```

```
fitting null model for pseudo-r2
```

```
# A tibble: 1 x 2
  names        x
  <chr>    <dbl>
1 McFadden 0.759
```

Analysis of Coefficients:

Raised Hands: If there is a unit increase in raised hands, then the log-odds ratio is expected to increase by 0.035.

Visited Resources: If there is a unit increase in visited resources, then the log-odds ratio is expected to increase by 0.064.

Discussion: If there is a unit increase in discussion, then the log-odds ratio is expected to increase by 0.030.

AnnouncementsView: If there is a unit increase in announcements view, then the log-odds ratio is expected to increase by 0.022.

Hypotheses:

Null hypothesis: The coefficient is 0.

Alternative Hypothesis: The coefficient is not 0.

```
[1] 0.94
```

```
# A tibble: 5 x 4
  term              conf.low conf.high       p.value
  <chr>                <dbl>     <dbl>         <dbl>
1 (Intercept)       -9.08      -4.73   0.00000000172
2 raisedhands        0.0105     0.0611 0.00622
3 visitedresources   0.0419     0.0895 0.000000111
4 discussion         0.00572    0.0565 0.0205
5 announcementsview -0.00916    0.0570 0.182
```

Analysis:

Looking at the confidence interval for raised hands, this CI does not cover 0 with (0.011, 0.061). Its p-value is less than the 0.05 significance level with 0.006 and therefore we can reject the null hypothesis with that p-value. The results from the p-value and the confidence interval are consistent with each other. The next two coefficients of visited resources and discussion are similar to raised hands in that the CI does not cover 0 and the p-values are less than the significance level. However, the last coefficient of announcements view is different. We cannot reject the null hypothesis because the p-value is greater than the significance level and because the CI covers 0. Therefore, our team can conclude that the first three coefficients are most important and statistically significant.

Step-wise Regression

Now that we've run through KNN with Low, Medium, and High class, then KNN with Low and High class, and fitting a logistic regression model using the same predictors to see which model could provide us with the most accurate predictions, we figured that a step-wise selection model could potentially provide us with another way to answer our research question.

```
# A tibble: 1 x 1
    AIC
  <dbl>
1  61.6
```

```
# A tibble: 7 x 2
  term                      estimate
  <chr>                        <dbl>
1 (Intercept)                 -7.19
2 genderM                     -4.38
3 semesterS                    1.75
4 visitedresources             0.062
5 announcementsview            0.06
6 parentansweringsurveyYes     2.23
7 studentabsencedaysUnder-7    5.78
```

```
# A tibble: 1 x 1
    AIC
```

```
  <dbl>
1  43.9
```

```
fitting null model for pseudo-r2
```

```
# A tibble: 1 x 2
  names        x
  <chr>    <dbl>
1 McFadden 0.901
```

Here, we're conducting step-wise regression to find the independent variables that actually do have a significant effect on the dependent variable. The step-wise regression worked fairly well, evidenced by a significant decrease in the AIC values between the full model (61.558) and the step model (43.914). We eliminated factors such as place of birth, nationality, and class topic because we did not think that they would be relevant in the context of our research question, and would make analysis of the data more convoluted.

Coefficient Analysis

Gender: If the student is a male, the log-odds ratio decreases by 4.379.

Visited Resources: For each increase in the number of times a student uses/accesses the course content, the log-odds ratio increases by .062

Semester: If it's Spring semester, then the log-odds ratio increases by 1.752.

Announcements Viewed: Every time a student checks their announcements, there is a resulting increase in the log-odds ratio by .060.

Parents Answering Survey: If the parent answers yes to the school satisfaction survey (as in they are satisfied with the school), the log-odds ratio increases by 2.231

Student Absence Days: If the student has under 7 absences, then the log-odds ratio increases by 5.780.

```
[1] 0.96
```

The prediction accuracy and pseudo-R^2 value of this model are our highest yet, being 96% and .90, respectively. Upon our extensive testing, this is sufficient evidence that the following factors are sufficient in predicting a student's class between High and Low:

Gender
Semester
Visited Resources
Announcements View
Parent Answering Survey Student Absence Days

Section 5 - Discussion

Our original research question was asking if it was possible to predict a child's class level within this dataset based off of the information provided in the dataset, and if so, which variables were of significance. To be clear, class level refers to the student's grades at the end of the semester: Low, Medium, or High. To answer our question: yes, it is very much so possible to predict a child's class level using the information provided in the dataset. The second part of our question was very open ended, and upon working extensively with this dataset, we feel as though a better way to phrase our question would be which variables could allow us to accurately predict a student's class, as which variables are "significant" is somewhat of a vague question to be asking.

Thanks to our analysis, we were able to reach a solid conclusion as to which variables could allow us to accurately predict a student's class: Gender, Semester, Visited Resources, Parent Answering Survey, Viewed Announcements, Student Absence Days were the variables that resulted in the highest prediction accuracy

within the various models we created. This result was obtained from a step-wise selection model, which provided us with a prediction accuracy of 96%. We originally planned to use a linear regression model to answer our question, but we were quickly redirected towards using a k-nearest neighbors algorithm or multinomial regression, both of which make much more sense and are actually applicable to answering our research question, as linear regression is used when the dependent variable is continuous. All in all, our updated methods for statistical analysis were appropriate, and our (updated) research question was properly addressed by our methods.

Our original KNN model gave us a prediction accuracy of 60% when we were testing our continuous variables and using all three levels of the class variable. At this point, we decided that it would be a better course of action to instead modify our dataset to only include the high and low classes, as medium seemed to follow no particular trends within our preliminary visualizations. Additionally, we would be able to generate a logistic regression model (as they're used for binary dependent variables) to compare accuracies. After running KNN using the same predictor variables and instead changing our response variable to a binary, we achieved a 94% prediction accuracy. Intuition aside, we figured that it would also be a good idea to utilize a backwards step-wise algorithm, allowing us to remove insignificant predictor variables efficiently. This allowed us to achieve a prediction accuracy of 96%, as stated above. We utilized pseudo r-squared values between our logistic models, focusing specifically on McFadden's version, and received a value of 0.75 and 0.9 for our original logistic model and step-wise logistic models, respectively. This showcases that our predictors within the latter model are the strongest out of our other models.

In retrospect, our methodology is far from perfect, but it's definitely in the right direction. In using KNN with Euclidean distance, we were limited by the fact that we were only allowed to use continuous variables as our predictor variables, which eliminated us being able to test a majority of the independent variables of the dataset using KNN. Additionally, our main determination for the usefulness of a model was the prediction accuracy, which may suffice for the k-nearest neighbors models, but for the logistic regression models, we believe more robust testing should have been used to determine the model fit, such as a chi-squared test or other goodness of fit tests. We tried utilizing the pR2 function from the PSCL (Political Science Computational Laboratory) package, which calculates pseudo-R^2 values, but the multiple values provided by the function are difficult to interpret, and as such, provide little information as to whether or not the model is a good fit. We do feel as though we went into sufficient depth, but we do think our analysis was somewhat hindered by methods we used, which could lead to a weakening of our conclusion. One thing that stood out to me was the rejection of the variable announcementsview in our first logistic model, but the re-inclusion of it within the step-wise model. Yes, we have a higher prediction accuracy with this model, but sufficient testing was not conducted to see if the model truly had good fit. Perhaps there was some co-linearity between some of our independent variables that we were unaware of, or some underlying assumptions that the algorithm made.

If we were to start over on this project, we would definitely attempt multinomial regression, as we would be able to use all three levels of the class variable (High, Medium, Low), and we would utilize the Hosmer–Lemeshow goodness-of-fit test, which is a popular test for goodness of fit. Further, if we were to continue working on this project, we would definitely find more robust measures in finding good model fit for our logistic models, preferably methods with a large amount of documentation, and/or that are still considered relevant. Also, we could use a variant of KNN which uses Manhattan distance, which would allow us to use categorical variables coded in numerical form, allowing us to delve even deeper into our question and reach a more concrete conclusion.