Ty Moyer
CIS 241
10/18/19
Dr.Hallenbeck

Student Learning Progression Using (xAPI)

Introduction

Different kinds of academic curriculums exist globally, which suggests that student learning progressions may differ. Resources, class sizes, grade level, and other variables can change how a student learns. This project explores students in the Middle East and how certain variables change their learning progression experience as students.

Dataset

The dataset selected from kaggle, consists of data collected by Al-Jarah(2016). Each entry in the dataset is an individual student's academic progress based off their academic behavioral traits which includes raised hands, discussion groups, large, medium or small class size, announcement views, absences under or above 7, parental rating of good or bad, and visited resources. A student is categorized by whether they are in Elementary, Middle, or High school, what subjects they study in a given semester, country, and whether they are Male or Female. Since no grade is included in the dataset, the ultimate decision is based upon parental satisfaction which can be categorized as good or bad. This dataset however contains no units of measurement for the numerics. The first set of questions will be answered using exploratory modeling, and the second part of this paper will answer questions using explanatory modeling. The first questions this dataset conjures includes: Does a student's absences have an effect on how many times they raise their hands? Do students who receive a good level of satisfaction from their parents typically stem from large class sizes? Does class size have an effect on a

student raising their hands, and how often do students who visited resources also view announcements? These will be answered using exploratory analysis.

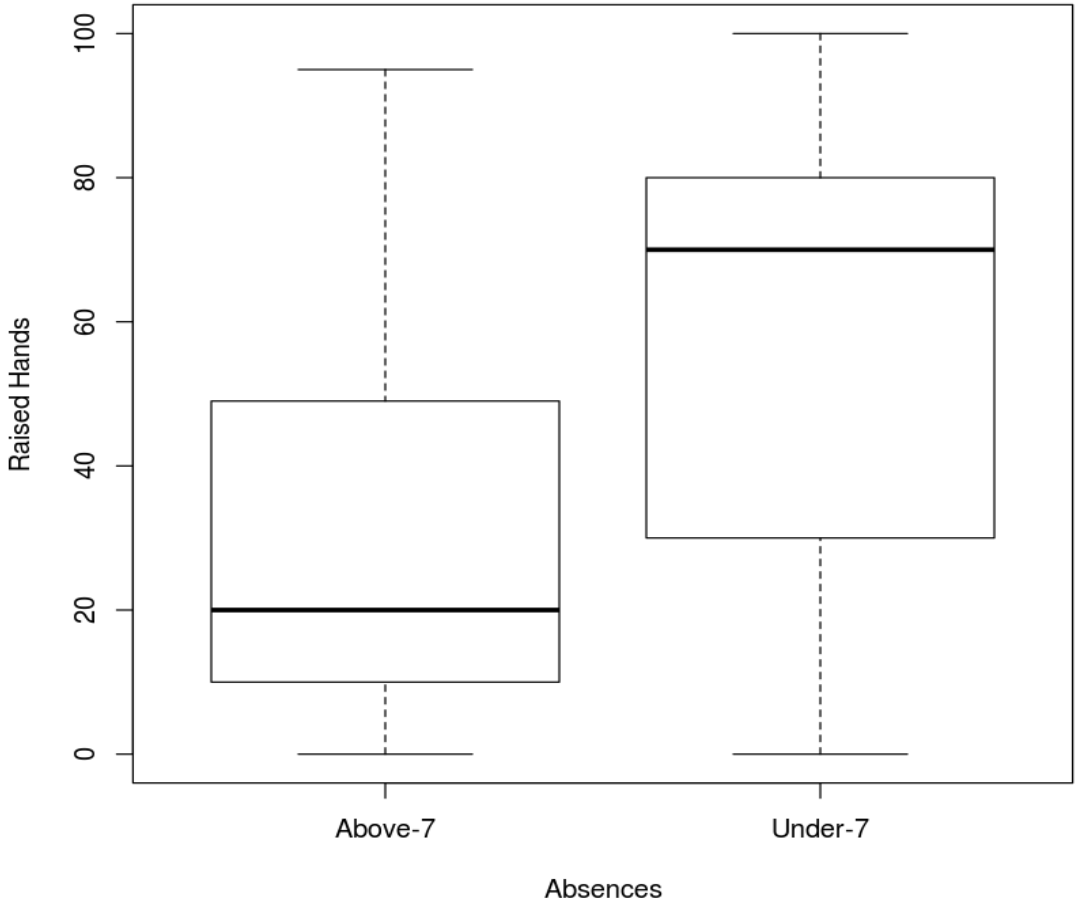
Misc Dataset Info

- Ebuko (2019) was investigating if a student's absences and level of parental intervention in their schooling affects their academic performance? They found that grouping the students into two clusters via k-means clustering, that even students with low absences performed poorly when their parents weren't involved in their schooling.

- For context, the sample pool of this data set is rather small due to the Middle Eastern / North African children, especially in war torn countries, are less likely to be in school. Girls are 1.5 times more likely to not be in school according to UNICEF (2019). It would be interesting to see the distribution in satisfactory reports between Males and Females because of this. Students in more developed countries in MENA still argue that fair advising is absent, thus leading to less motivation as their schooling progresses through the year Khalifa (2016).

Exploratory Modeling

The first question asks, ‘Does a student’s absences have an effect on how many times they raise their hands’? The variables the question compares are absences which is measured by whether or not a student has missed above, or under seven classes making this a ranked numeric variable. The amount of times a student raises their hand is a numeric variable. Using exploratory analysis to show the distribution between the two variables, I created a box plot.

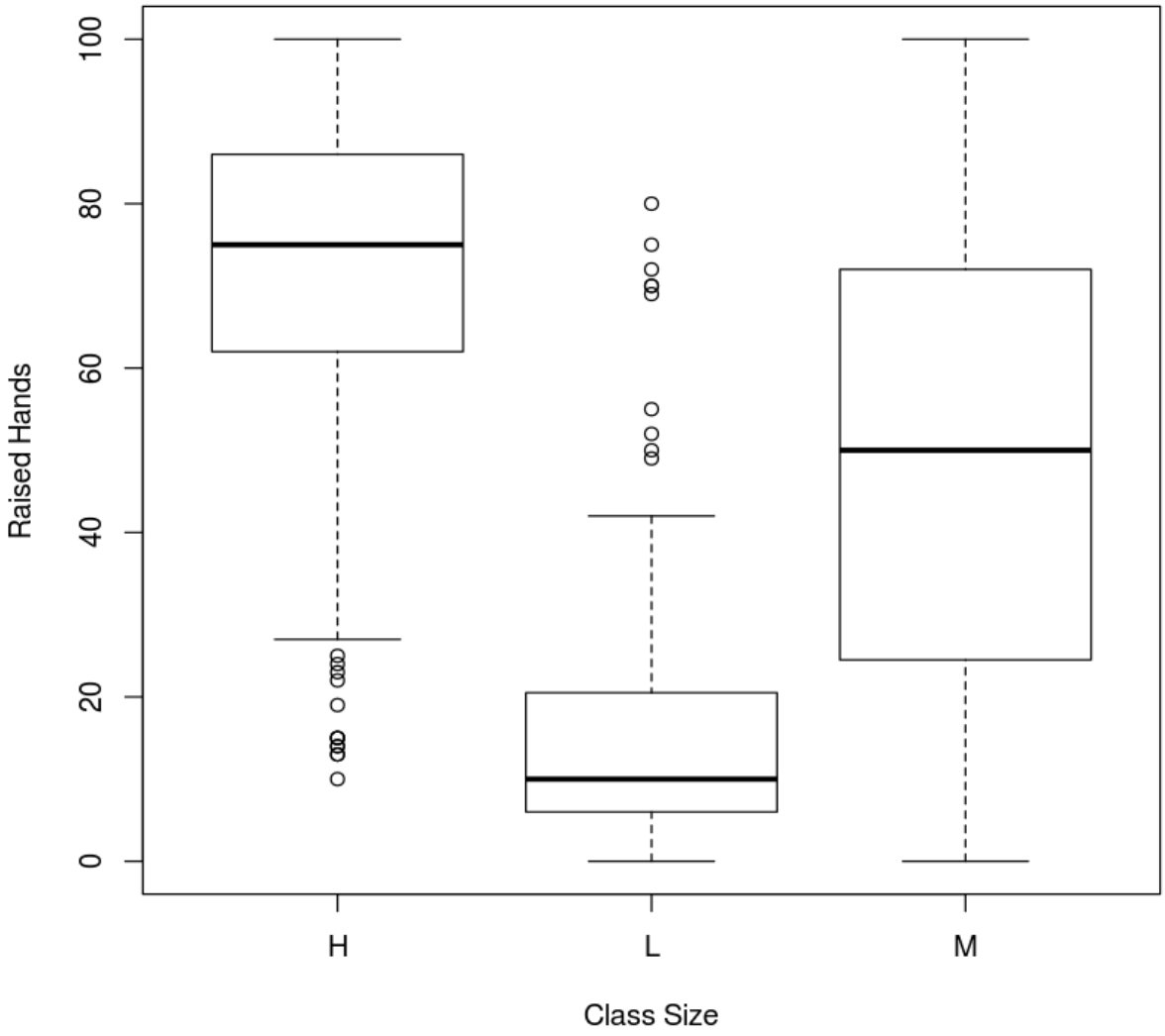


The results of the above box plot suggest that students with above seven absences raise their hands less often, and students who have under seven absences raise their hands more often.

I investigated the proportions with a two sample z-test. According to the results of the test, the P-value of 2.22 X $\mathbf{10^{-16,}}$ a significant value, which suggests a high occurrence amongst the two variables. Amongst other tests, the two sample z-test evidently best describes the relationship between these two variables.

The second question asks, do students who receive a good level of satisfaction from their parents typically stem from large class sizes? The class sizes are identified by a large, medium, and small classification. Parents can give a good or bad level of satisfaction. For the sake of comparison, I wanted to investigate those three variables to see if large will come out on top. since the question asks about the probability of these variables occurring and how they affect one another, I used association rule mining for analysis. After turning the variables into transaction, I received 480 rules, however to stick to the focal point of the analysis, I eventually came down to three rules. The results for large classes showed, (confidence = .40(40%, Lift = 1.36) showing a high level of significance in conjunction with good parental satisfaction ratings. What came to my surprise however was that students from medium classes yielded the results, (confidence=.44(44%), lift=1.02) revealing that not only is there a high significance according to the lift, the confidence shows that there is a more frequent occurrence than expected of the statement being true. I questioned the data because one would think that a small class setting, similar to a private school, means more individualized attention. Contrarily small classes showed, (confidence=0.14(14%), lift =0.55). There is a much lower frequency for a good rating and a small class size to occur. The resolution of the data shows that a student has a good chance of receiving a good satisfactory rating if they are in a large or medium sized class but not small.

The third question asks, does class size have an effect on a student raising their hands? A proper way to visualize the data is through a histogram which shows the frequency of raised hands and when the hands raise amongst the different class sizes.



The plot shows a distribution that assumes students in a High(H) raise their hands more frequently then Low(L) and Medium(M) sized classes. To analyze the data further, a spearman correlation test was used to show the strength of the two variables. The results of the test showed

a P-value $2.22X\mathbf{10}^{-16}$, and an R value of 0.64 showing a strong positive correlation amongst raised hands and class sizes.

The fourth and final question I had for this dataset is how often do students who visited resources also view announcements? I took the variable 'Annonucements_View" will use it to determine high or low engagement based off the numeric, and 'Visited_Resources' to determine whether a student is resourceful, or not resourceful based off the number of visitations from the dataset. The test used was association mining to determine the probability to answer how much the students are resourceful and engaged ? After going from 480 rules to 4, I used association rule mining to show the frequencies of how engaged and resourceful students are. After the test, the results showed a high significance when students are resourceful and highly engaged (confidence=.71 (71%) lift= 1.32). What didn't come to my surprise is that students who aren't resourceful are almost definitely not engaged (confidence=.64(64%) lift = 2.56) showing a high significance and frequent occurrence. When I investigated the likelihood of not resourceful students having high engagement, (confidence =.010, lift=.18) and resourceful students having low engagement yielded(confidence=0.099,lift=.39) which evidently shows that the majority of students view announcements and visit resources often, showing high engagement being the most common frequency.

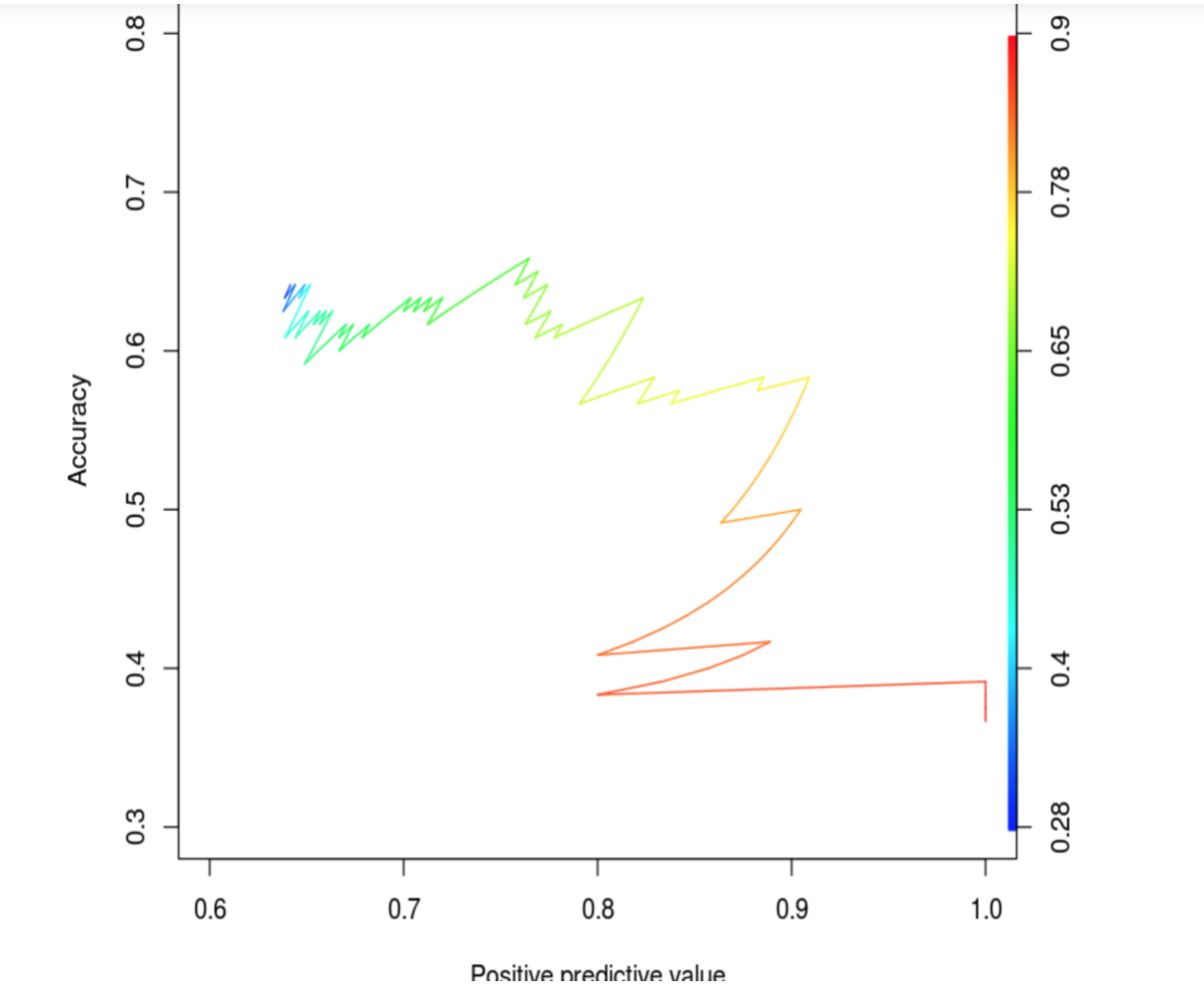
Conclusions

The overall results of my analysis of learning progression of students in the MENA region, is that students in large and medium sized classes are most likely to have a good parental satisfactory rating , students within the under-7 absences category participate more in class via raising their hands, students from larger classes also raise their hands more often, and students who view announcements and visiting resources are more engaged.

Explanatory Modeling

A different way of analyzing the data to answer more questions can be done through explanatory modeling. I will answer four questions utilizing this method in an attempt to develop a prediction. The first question asks, can I create a model that predicts the gender-based off raised hands, visited resources, announcements viewed, number of discussions, parental satisfaction, below or under seven absences, and class size? Within the dataset, gender is defined as a binary of Male, ‘M” or Female, ‘F,” so since this identifies gender as the y variable, the proper type of analysis to use is logistic regression to determine two possible outcomes. After training the dataset, the model yielded significant values from the variables: announcements viewed, number of discussions, large and medium class sizes. Since the question asks about probability, a prediction was created based on the model and will be estimated through an accuracy and positive predictive value calculation. The accuracy, which shows how often the test is correct, is 61%. The positive predictive value, which shows if my test says ‘true,’ and how often is it really true, had a percentage of 87% based on the model with no cut off made. These are good

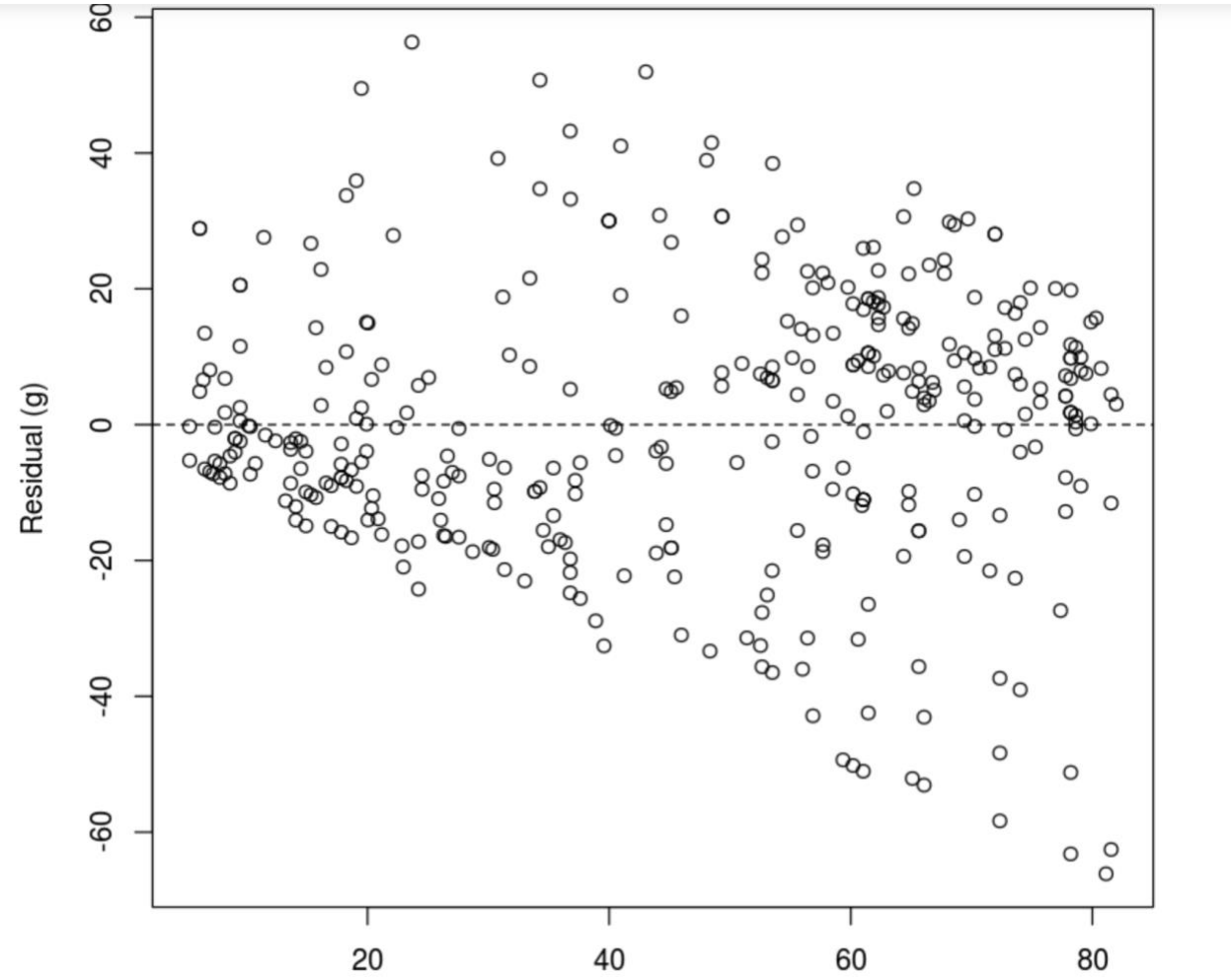
numbers; however, a better prediction could be made with an implemented cutoff of 0.65.



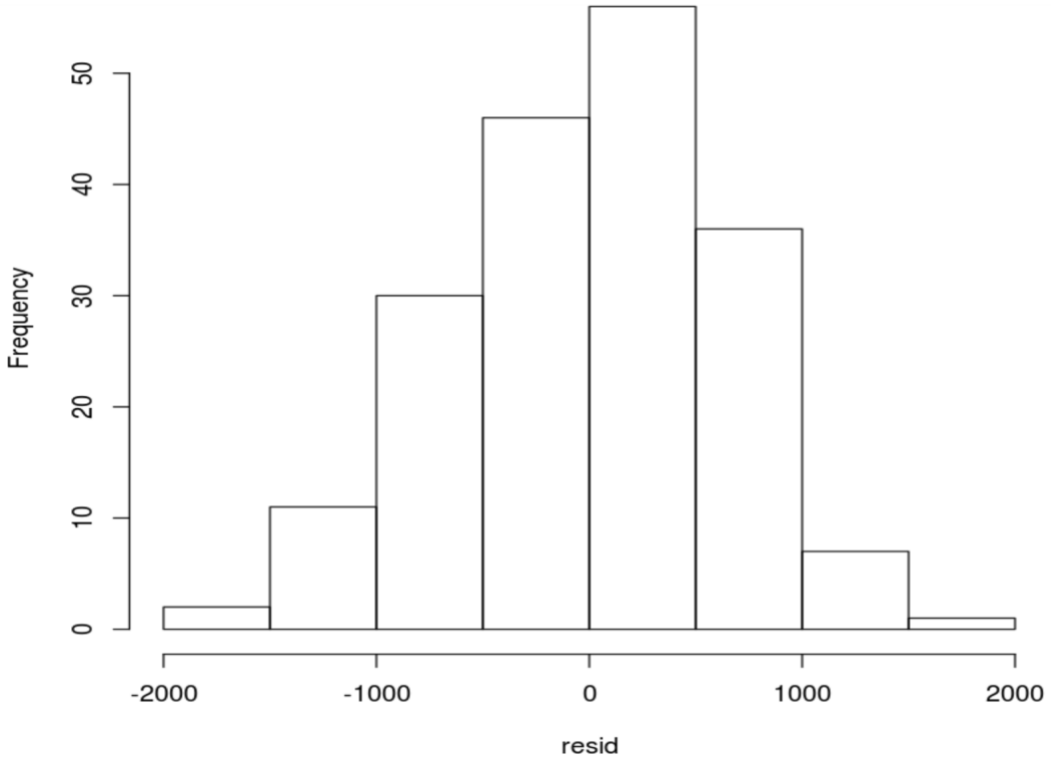
The above graph suggested using a cut off according to the point noted by the yellow and orange tinge pointing towards ‘0.65’. The accuracy increases to 64.1%, and the precision or positive predictive value rises to 76.5%, improving the model.

The second question asks, can I create a model that will predict the number of raised hands based on a student's year, absences, visited resources, and class size? Since the Y variable is numeric, and the x variable can be either, multilinear regression is the appropriate method of analysis in order to create a prediction.

The model yielded that middle schoolers, students who visited resources in large or medium class seizes weighed in as the most significant factors in determining the prediction outcome.

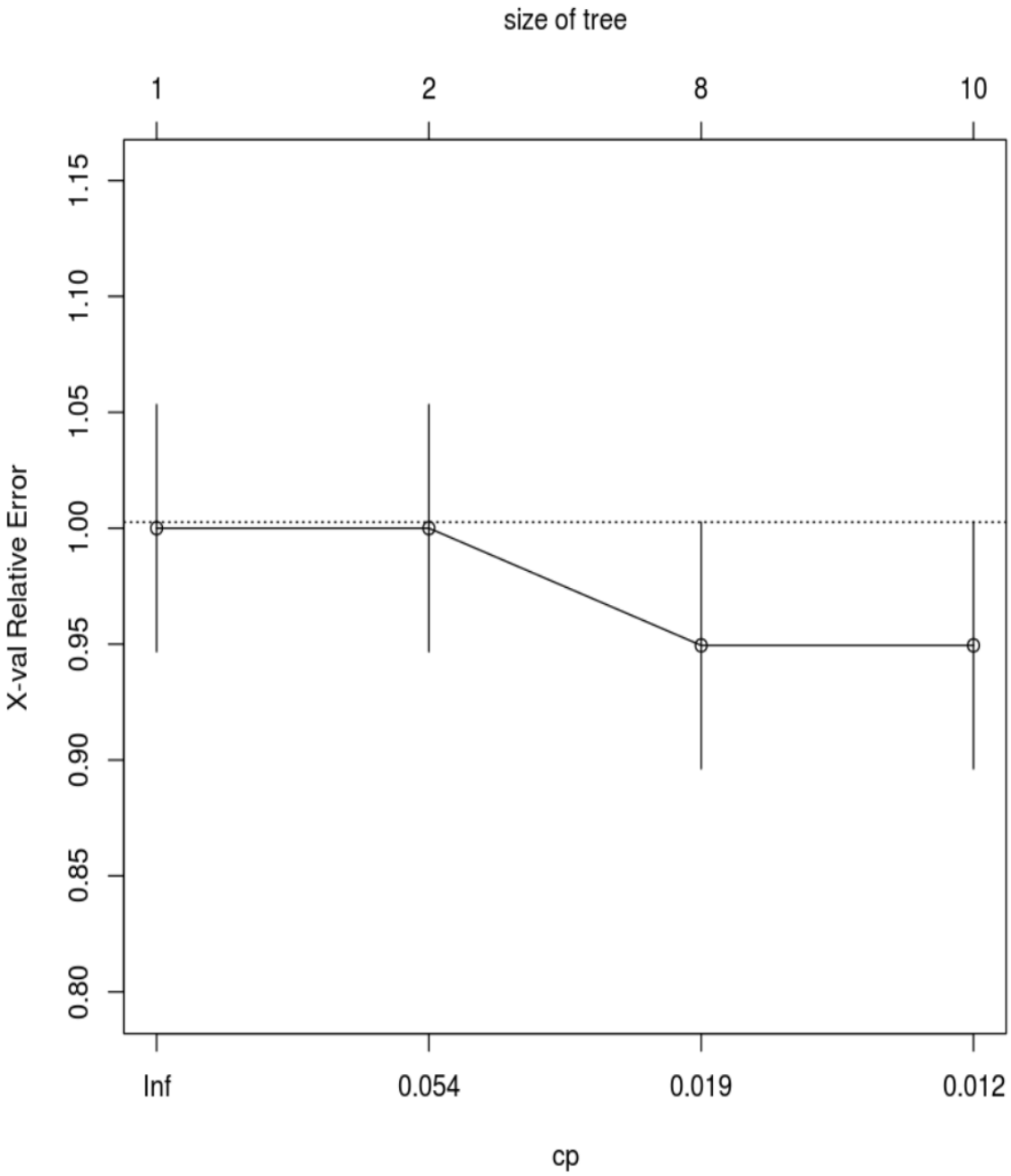


The plot above shows how there is an excellent distribution. However, there is a rather peculiar linear trend from the bottom left to right. After multiple attempts, I could not develop legitimate reasoning as to why it is plotted this way. Fortunately, the argument that this is a good model is backed up by the histogram below, which displays a bell curve.

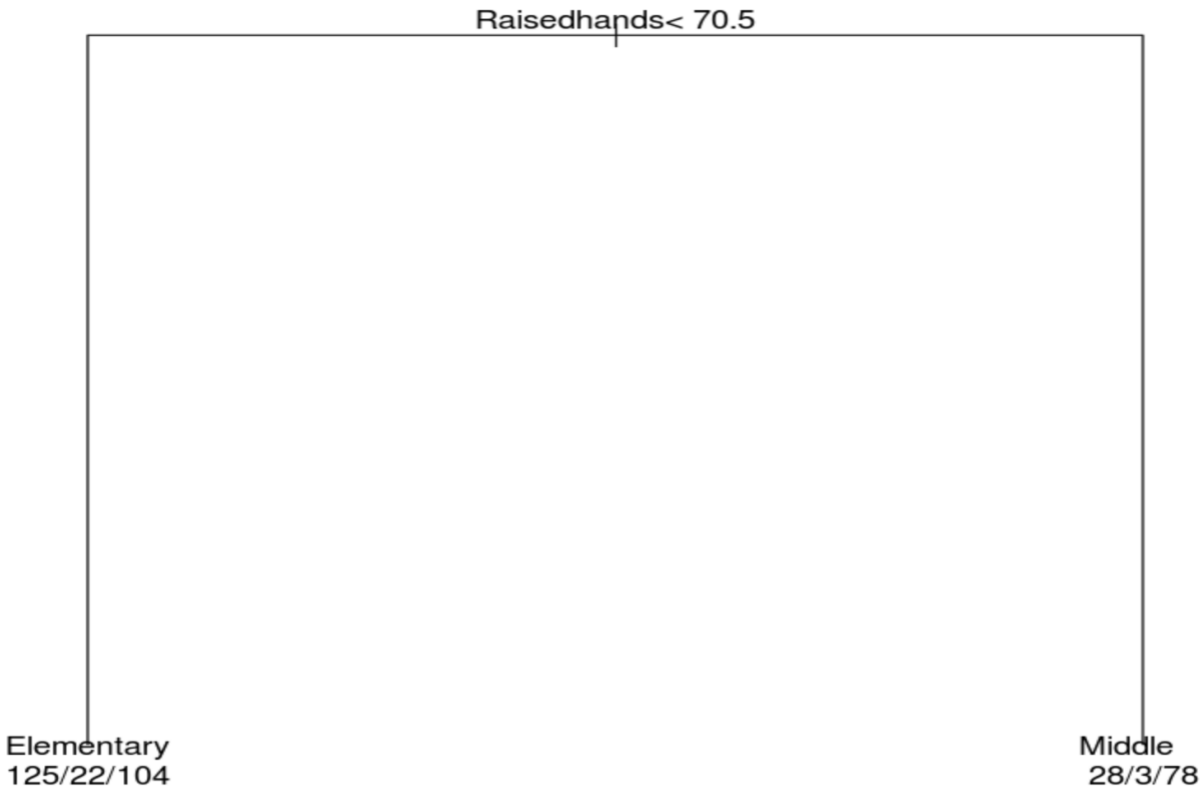


The cor test also was not needed due to the high level of differentiation amongst the variables. Based off the two plots, this model would be good for the prediction.

Question three asks, can I predict the year based off of how many times a student participates in the discussions, views announcements, and raised hands? The y variable will be the year, which represents the level of schooling a student is in, making it categorical. The x variables are a mix of both numerical and categorical fitting the criteria for a decision tree to do the analysis. The first step is to plot the model to detect any reasons to prune it before making the actual tree. The action of pruning is when we remove the sub-nodes of a decision node. Based on the plot on the next page, it is obvious that there is a need to prune the data at cp=0.0025.


size of tree
1
2
8
10
X-val Relative Error
0.80
0.85
0.90
0.95
1.00
1.05
1.10
1.15
Inf
0.054
0.019
0.012
cp

It would be safe to assume after pruning, the tree itself would be a lot more coherent. Shockingly, this turned out to not be the case as seen below.



On the left-hand side, it suggests that Elementary students raise their hands the most under 70.5 times, High school, in the middle, was predicted to do so 22 times, and High school with 104. Middle Schoolers were predicted to most likely to have 70.5 raised hands or more. High schoolers were predicted to have 70.5 raised hands three times, and Elementary 28, which causes ambiguity. In order to figure out why, the accuracy of the model needed to be calculated. The accuracy value is 56%, and there was also an issue with the table not calculating high school values, as shown below.

	Elementary	High	Middle
Elementary	38	0	8
High	7	0	1
Middle	36	0	30

The above result would possibly also explain why there are only two branches, as multiple factors were used in this prediction. The ending result reveals that this is not a good model to analyze the data.

The final question asks, can I create a model that can predict what topic students will study based on whether or not they have under or above 7 absences, Parental Satisfaction, and Class Size? The y variable is binary categorical as it determines the number of absences a student has accumulated, whether it is under or above 7. The Naive Bayes Classifier would be the right test to use here. After creating the model and formatting it into a prediction, fascinating results were received, meaning I wasn't sure what to make of them.

predict	Arabic	Biology	Chemistry	English	French	Geology	History	IT	Math
Arabic	2	1	1	3	4	0	0	0	0
Biology	0	2	0	1	0	0	0	0	0
Chemistry	0	1	2	0	0	2	0	0	0
English	1	0	0	2	1	0	0	2	1
French	3	1	0	1	11	0	0	2	0
Geology	0	2	2	2	0	7	3	0	0
History	1	0	0	0	0	0	0	0	0
IT	2	0	1	3	2	0	0	12	4
Math	0	0	0	0	0	0	0	1	0
Quran	2	0	0	0	0	0	0	0	0
Science	2	0	0	0	2	0	0	0	0
Spanish	4	0	0	0	0	0	0	0	0

predict	Quran	Science	Spanish
Arabic	0	2	1
Biology	1	0	0
Chemistry	0	1	0
English	0	1	0
French	0	1	0
Geology	2	0	0
History	0	5	0
IT	0	0	0
Math	0	1	0
Quran	1	1	2
Science	0	2	0
Spanish	1	1	1

The Naive model presented to not be a good test as the tables did not accurately account for all the possibilities for the prediction. Additionally, the accuracy was rather low, only being 42.8%. This is a model that seems to be unable to help anyone reach conclusions; even the decision tree was more competent in completing the prediction. However, after further analyzing the data, the fact that I have 12 variables and not two could misconstrued the results from the test.

Conclusions

Logistic regression seemed to have the easiest time in making predictions within explanatory modeling. The outlier in this type of modeling seemed to be the Naive Bayes Classifier, mostly because it was a bad model where it was impossible to articulate a conclusion of any kind.

Works Cited

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.

"Education." *UNICEF Middle East and North Africa*, https://www.unicef.org/mena/education.

Ekubo, Ebiemi Allen. "Attributes of Low Performing Students In E-Learning System Using Clustering Technique." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2019, pp. 480–485., doi:10.32628/cseit1953158.

Khalifa, Batoul, et al. "A Qualitative Study of Student Perceptions, Beliefs, Outlook and Context in Qatar: Persistence in Higher Education." *Qatar Foundation Annual Research Conference Proceedings Volume 2016 Issue 1*, 2016, doi:10.5339/qfarc.2016.sshapp1552.