James Taylor

Assignment 1

University of Maryland University College

DATA630-9041 Spring 2020

Dr. Gates

james.taylor.ps3@gmail.com

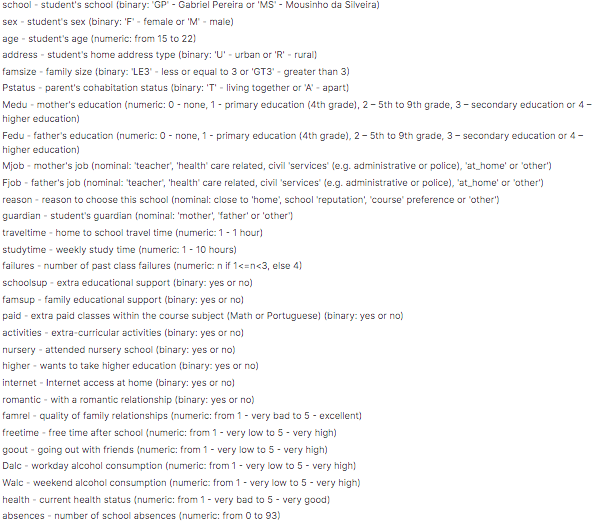
**Introduction**

The purpose of this paper was to find indicators that a high school student may be at risk of excessive alcohol consumption. Not only does excessive drinking pose the risks of violence, inhibition control and motor vehicle deaths, overconsumption of alcohol by teenagers additionally has developmental consequences (Silveri, 2012). Underage drunk drivers are also involved in fatal crashes at twice the rate of age drivers are (Bonnie & O'Connell, 2004). Knowing the correlations and sign of potential alcohol abuse in teens is to help those who don’t fully understand the consequences of their actions yet.

The approach taken was an Association Rules analysis. It is a simple algorithm to discovering correlations in data with multiple metrics to measure the effectiveness of each rule. The goal was to discover correlations found in the Association Rules method that authorities or administrators could use to determine if a teen is at risk of alcohol abuse, in which preventative steps may be taken.

**Analysis and Model Development**

Data Information

The data used for this paper was retrieved from the Kaggle website (url in reference page). It was a survey of secondary school students from two schools. There are 395 observations on 33 variables regarding data on the demographics of the student, how they are in school, their family and alcohol consumption. Below is a figure of all the variables and their values/meanings. ../../Desktop/2.png

*(Source: Kaggle)*

Data Preprocessing

The dataset came in two separate values because of the different classes where the survey was done, so there was a math and Portuguese file. I combined the two files to increase the number of observations. In the description of the data it noted that some student took they survey twice, (once in their math class and again in their Portuguese class). This would bias our results so I decided to remove them. I did this by finding duplicate values on the student demographic variables, like age, mother and father education, family size, place of residence and about 25 other variables. The chances are low that a student has the same values across all those variables as another, so the fear of accidentally removing a unique student was low. There were 85 observations removed.

There were no missing values in the provided dataset. Because the Associated Rules algorithm needs variables as discrete or factors, some were discretized or converted to factors. The only variables that were binned were Age, Absences and Grades (G1, G2 and G3) because their distributions were wide enough that turning them into factors would have too much detail and not smooth. For example, ages were grouped by every two years because the change between them is not dramatic and they very likely may be in the same year of schooling. Absences were also binned so as to not treat one absence differently from two, which is a negligible difference for our purposes. The rest were converted to numerical because they were categorical in nature and at most five unique values or less. After these transformations, the dataset has either factors or discrete bins.

Exploratory Data Analysis

The distribution for weekday alcohol consumption is shown on the right, an ordinal variable. Obviously, lots of students admit to consuming little alcohol during the week, however there is no “none” option. The “very low” level will have student who never drink and drink small amounts. The “high” and “very high” frequencies appear to be similar. As it is a survey, we cannot be certain of the accuracy of this rating system of consumption. Individual students will have different judgments on what these thresholds are.

 Absenteeism is associated is one obvious measure that a student is falling behind their peers. Below is the binned frequency of absences (discrete) for student who admit to drinking ‘high’ or more amounts during the week. It is a small percentage of students who meet this criteria, so the existence of a relationship between a variable and drink ‘high’ or more amounts will be difficult visually, particularly in the left example where drinking a lot doesn’t trend with high absenteeism.

A home in turmoil may be to blame for teen drinking. The parents may be disconnected or inattentive towards their child’s and struggles. Below shows the student’s perceived relationship with their family. Again, these are students who claim to be drink at least a ‘high’ amount of alcohol during the week, and it shows no relationship between bad family life and those teens drinking.

Choosing an individual variable or more and attempting to visualize it to see if a relationship exists between a student drinking heavily and those variables is a long process, especially if one was attempting to view all 30 or so related variables in this dataset. That is why the Associated Rules method is a useful approach for looking at correlations. Not only does it handle normal correlations between two variables, it can also find the relationship between two or more variables to two or more variables.

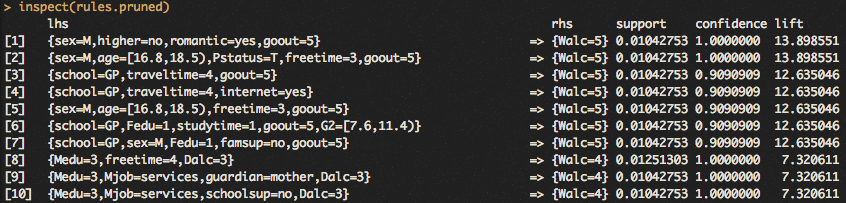
The associated rules method, or apriori algorithm, finds relationships between variables and their values. For example, we may find that good grades are correlated with low absences. It is similar to the statement IF grades are high, THEN absences are low. There are many possible relationships that can be drawn in datasets with many variables.

The algorithm has various metrics to determine if a discovered association has mathematical merit. One is the confidence of the association, which is similar to accuracy of the prediction. If an algorithm say X is correlated with Y, the confidence value indicates how many observations with X in it also have Y. The higher this confidence is the stronger the relationship, but it is impossible to get a value greater than one. Another measure of strength is the rule support, which shows the percent of the dataset where the rule is true. Using the previous example, this would be the percentage of the dataset where X and Y are both true. The higher the rule support value, the more evidence the association exists. The last strength metric is lift, which measures the likelihood of X and Y occur together while controlling for the chance of X and the chance of Y occurring on their own. It’s measures how independent or dependent the two conditions are. If the lift is close to one, the two events appearing simultaneously is considered independent. As the lift value becomes greater than one, the more likely the two events occur together. Because there are three separate definitions of strength the algorithm produces for each rule, ones that achieves a balance between them is preferable.

The apriori algorithm also has the ability to take input parameters. Some are thresholds on the different strength categories where rules are omitted if they are not strong enough. The algorithm will also limit itself to finding rules that have a specialized condition. This is very useful for the purposes of this paper since finding correlations with high alcohol consumption is desired. The algorithm used was specified to find rules that have weekday and weekend as ‘high’ or ‘very high’. The thresholds for support was 1% and confidence was 90%.

**Results and Model Evaluation**

The apriori algorithm created 123 rules after redundant rules were removed. Because the sample of students who drink ‘high’ or more amounts is relatively small, the support and counts are just over 1% and around 10 respectively. Below is an output of the top 10 strongest rules.



The most common association with ‘high’ or more weekend drinking was ‘goout=5’, which is the ‘very high’ value of going out with friends. This is a reasonable association to make, since alcohol is normally consumed at higher amounts in social environments. More time spent with friends is also more time spent away from parental supervision, leaving more opportunity for drinking.

Being a male appears to have more predictive value for drinking compared to females. Of the top 10 strongest rules, four have males as an included condition. It’s possible there may be a desire to appear masculine as teenagers place lots of stock in their social status. It may be harder for males to say ‘no’ to an offered drink.

One apparent spurious correlation appears to be in amount of free time. A ‘medium’ amount of free time is in some of the top 10 rules. It’s not as if students with little free time drink more to let off steam, or students with lots of free time drink to pass the time. Travel time appears in some rules as well. This may be evidence of overfitting. Most of the students in this sample don’t drink high amounts of alcohol, so the algorithm is creating associations with only those few students who do. The consequence is it notes correlations that are only true for those few students and not applicable to the rest of the students. This is also evident by the low support values for the rules. This may still be worth it however due to the high costs and damage alcohol consumption can have on their young lives.

The strongest rule (#1 in the above figure) is interesting. A male in a romantic relationship, who goes out very often and has no desire to get higher education. Every student that fits this description in the data drinks heavily on the weekends. This is the only rule in the top 10 that includes higher education. If a prediction was made on what is likely to be correlated with alcohol consumption, no desire for higher education would have been listed. It’s possible those students are quite different from those deciding to got to university. They may feel that school is not for them, or they have less optimism about their future in general. The lift strength metric is large at 13.8. This shows that those the rule is dependent and very unlikely to be random.

The right-hand side of all 123 rules were weekend alcohol consumption and none weekday ones. This means the model could not find associations strong enough for weekday alcohol consumption. A model was created to find associated rules for high weekday consumption, but the correlations were weak and support small.

The levels of lift were significantly high for the models. The issue is the number of observations there were to make rules from. Because the model was specified to create rules with at least a high amount of weekday or weekend consumption, the model could only look at observations that meet that criteria. Looking at the first graph in this paper, not many students admit to drinking those levels of alcohol. Working with these low levels of support gives the level of lift strength less validity. It may be contributing luck and overfitting to the codependence of both sides of the rule occurring together.

**Conclusion**

The Associated Rules approach to finding correlations with student drinking uncovered interesting results. The best evidence is limited to males who spend lots of time with their friends as the best predictors for large consumers of alcohol. There is reasonable evidence for males who are also in relationships and no plan for higher education. The other associations that were discovered either weren’t very strong or it wasn’t clear how it was related to alcohol consumption (possible overfitting). While the support values are not very strong, the lift values are. This is shown above in the visualization of the top 6 six rules. The lowest lift of them still had a 12.6 level, which is far from independent. The issue is the support is very low, which makes an accurate measure of lift difficult.



The results contain no surprises. Students who spend more time with friends are under the watchful eyes of their parents less, leaving more time for them to get and consume alcohol. Those who drink heavily are likely spend time with others who drink significantly as well. Perhaps the lesson here is to learn that traditional signals may not be correct, like number of absences.

References

Silveri, M. (2012). Adolescent Brain Development and Underage Drinking in the United States: Identifying Risks of Alcohol Use in College Populations.National Center for Biotechnology Information. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4669962/>

Bonnie RJ, O'Connell ME. (2004). Reducing Underage Drinking: A Collective Responsibility.National Academies Press. Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK37591/>

Data source: <https://www.kaggle.com/uciml/student-alcohol-consumption>