STAT400: Data Mining and Machine Learning

Contributors: Anthony Shishkin, Chris Rooney, Eric Bixby

Student Alcohol Consumption

**Introduction:**

Our main goal of the project was to find what factors were the best predictors of alcohol consumption in students. To find this, we did different forms of data analysis on our dataset. The four different data analyses we did were: Clustering, Memory Based Reasoning, Neural Networks, and Decision Trees.

**Data:**

The data we used came from kaggle.com. It includes responses from a survey of students in Portugal. We spent time cleaning the data, removing a lot of the variables since the dataset started with 33 unique variables, 21 of which we ended up using.

**Variables:**

* sex - student's sex (binary: 0 - female or 1 - male)
* age - student's age (numeric: from 15 to 22)
* address - student's home address type (binary: 1 - urban or 0 - rural)
* Pstatus - parent's cohabitation status (binary: 0 - living together or 1 - apart)
* Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
* Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
* guardian - student's guardian (nominal: 0 - mother, 1 – father, or 2 - 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)
* activities - extra-curricular activities (binary: 1 - yes or 0 - no)
* internet - Internet access at home (binary: 1 - yes or 0 - no)
* romantic - with a romantic relationship (binary: 1 - yes or 0 - 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)
* G3 - final grade (numeric: from 0 to 20, output target)

The focus here is on “Dalc” and “Walc”, which are the two response variables, and the rest are all predictor variables.

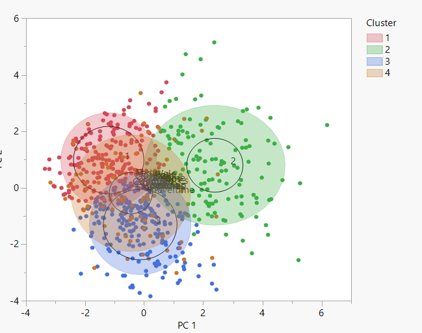
**Exploratory Data Analysis:**

The survey was done on 649 students. The average age of the students was 16.74. The survey was mostly females, 383 compared to 266 males. The measure for alcohol consumption was from 1 to 5, 1 being very low and 5 being very high. The average alcohol consumption was 2.28 on the weekends, and 1.5 during the week.

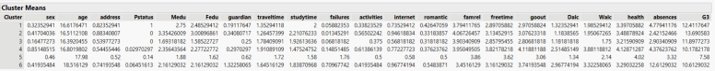
**Clustering:**

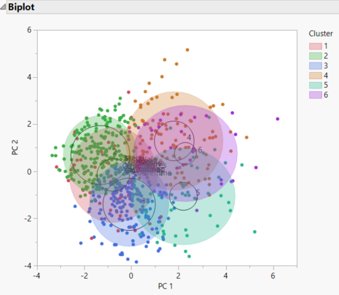
We used jump to do 2 different clustering models of k means = 4 and 6.





These were the results of the k means = 4 clustering. Here, we can see the most telling results from cluster 2. This clustering had the highest alcohol consumption values in both measures, with the lowest grade results, from a mostly male population, the high age, and a low education value of their parents. We can see based on this that older males may be more inclined to drink more. They also may be more likely to have lower grades. Cluster 2 also had the highest value to health, but it is likely this is a result of surveying males, not because they are actually healthier. Cluster 3 had the lowest levels of alcohol consumption. This cluster had a high number of females, the least failed classes, average grades, but the lowest health levels of all 4 clusters. Based on this, we can again see some correlations. Females are more related with lower levels of drinking, and there may be some correlation between health and alcohol, but like mentioned earlier, this could be a correlation between males and their survey answer in the health category and not actual alcohol consumption. Higher alcohol consumption may also be attributed to lower grades, and having parents with less education could also be correlated.





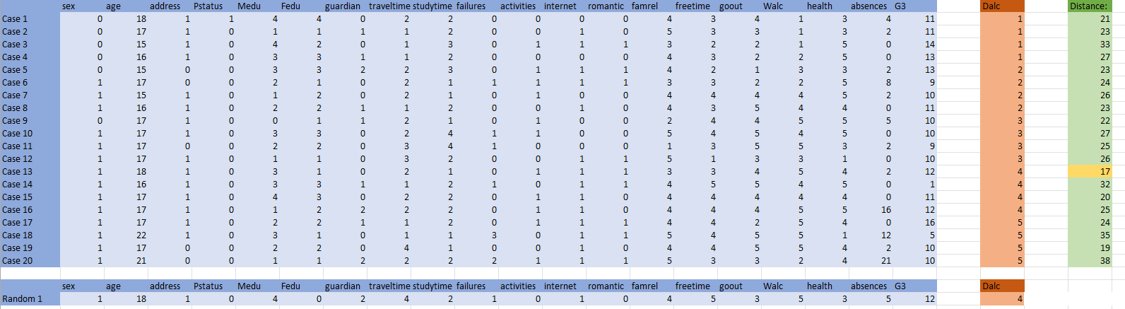
These are the results from the second clustering, with k means = 6 clusters. This clustering was not as good as the last. Much of the data hung closer to the means and seemed to be spread a bit thin. However, cluster 4 provides some of the more interesting results. It had the highest alcohol consumption, the greatest number of males, and the highest health value. This one was very similar to cluster 2 in the previous k means = 4 results. The grades, however, hung around the mean, and the education of the parents wasn’t as bad as the last. Cluster 5 also had some interesting results, with lower alcohol consumption, the lowest grades by far, and below average education level from the parents. It seems like the correlation between parents' education and lower grades may not be as connected to alcohol consumption as it was in the previous clustering results.

In conclusion, we can say from the clustering results that sex seems to be the indicator of higher levels of drinking. Education level from the parents, grade results, and health may also be related, but seem to be a bit all over the place. Many of the other variables are much more self-explanatory, for example, as free time and going out increase, so does drinking.

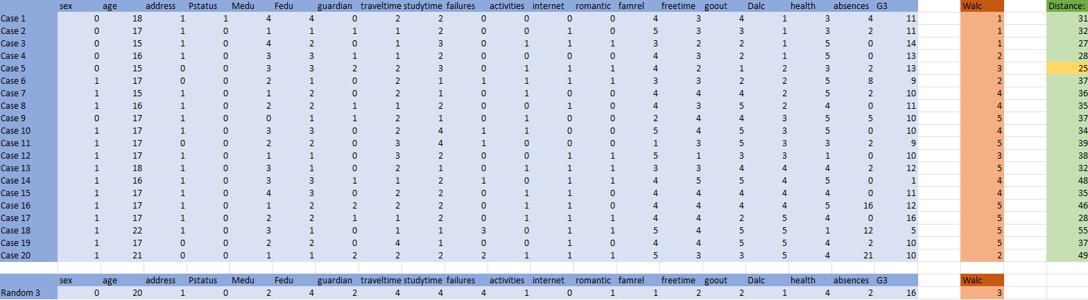
**Memory Based Reasoning:**

We wanted to generate models that would be able to predict student's weekday alcohol consumption and weekend alcohol consumption based on similar cases. We created two samples with 20 cases that allowed us to calculate Euclidean distance and predict randomly generated cases.

For the first model, we set our “diagnosis” as Weekday Alcohol Consumption (DALC) and grabbed 4 separate cases for each of the 5 diagnoses. We made sure that each of the cases were generated at random and were not missing any data to allow us to predict with the best accuracy possible. From there, we looked at each of the associated variables and randomly generated values to fill out a table. The random case was an 18-year-old male with an urban address. Despite his mother having higher education, his father was his primary guardian. He claims to go often and drinks often during the weekends. His final grade was a higher C. Based on the Euclidean Distance, our model predicted that he would classify the amount he drinks during the weekdays as high.



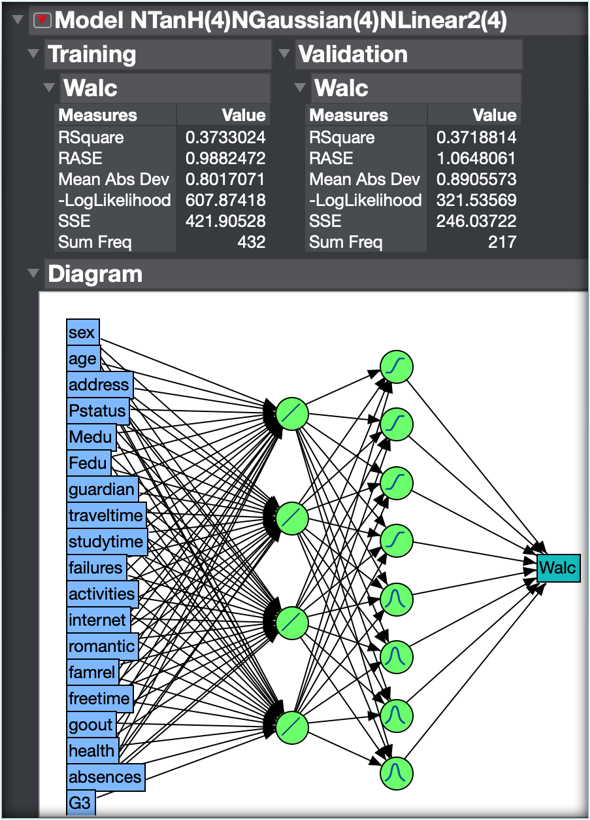
For the second model, we set our “diagnosis” as Weekend Alcohol Consumption (WALC) and reused the same 20 cases *(One thing to note was that there was not an even number of diagnoses, so in the future we think it is best to grab another sample of 20 cases that have an even split)*. We rearranged the table by taking out WALC and putting DALC back into the table as a variable. From there, we looked at each of the associated variables and randomly generated values to fill out a table. The random case was a 20-year-old female with an urban address. Her mother has an education but stopped somewhere in the range of 5th to 9th grade. Her father is her primary guardian, with a higher education. She claims that she does not go out that often and has a high B as her final grade. Based on the Euclidean Distance, our model predicted that she would drink moderately during the weekend.



In Conclusion, we believe that our MBR model could use a little bit of improvements, but it seemed like it predicted quite well. We noticed that the first randomly generated case was quite identical to its nearest neighbor (Case 13). The second model had a bit more variation, but that was since there was not an even split in the number of each diagnosis, which created a bias.

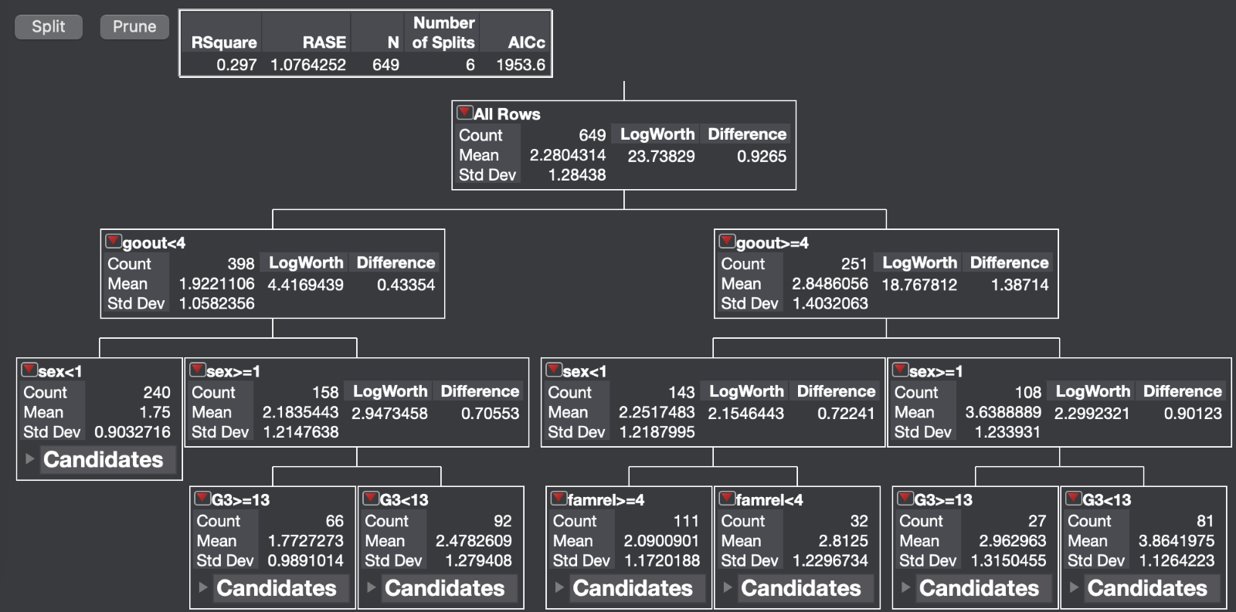
**Neural Networks:**

We used jump to create a neural network. Our neural network ended up using 19 of our 21 variables as factors, with weekend alcohol consumption as the response variable. We adjusted the learning rate to .01 and used all three of the activation functions provided by jump, TanH, Linear, and Gaussian in two layers. There were 4 tanH neurons and 4 gaussian neurons in the first layer with 4 linear neurons in the second layer. The best r-squared value we could get was about 37%.

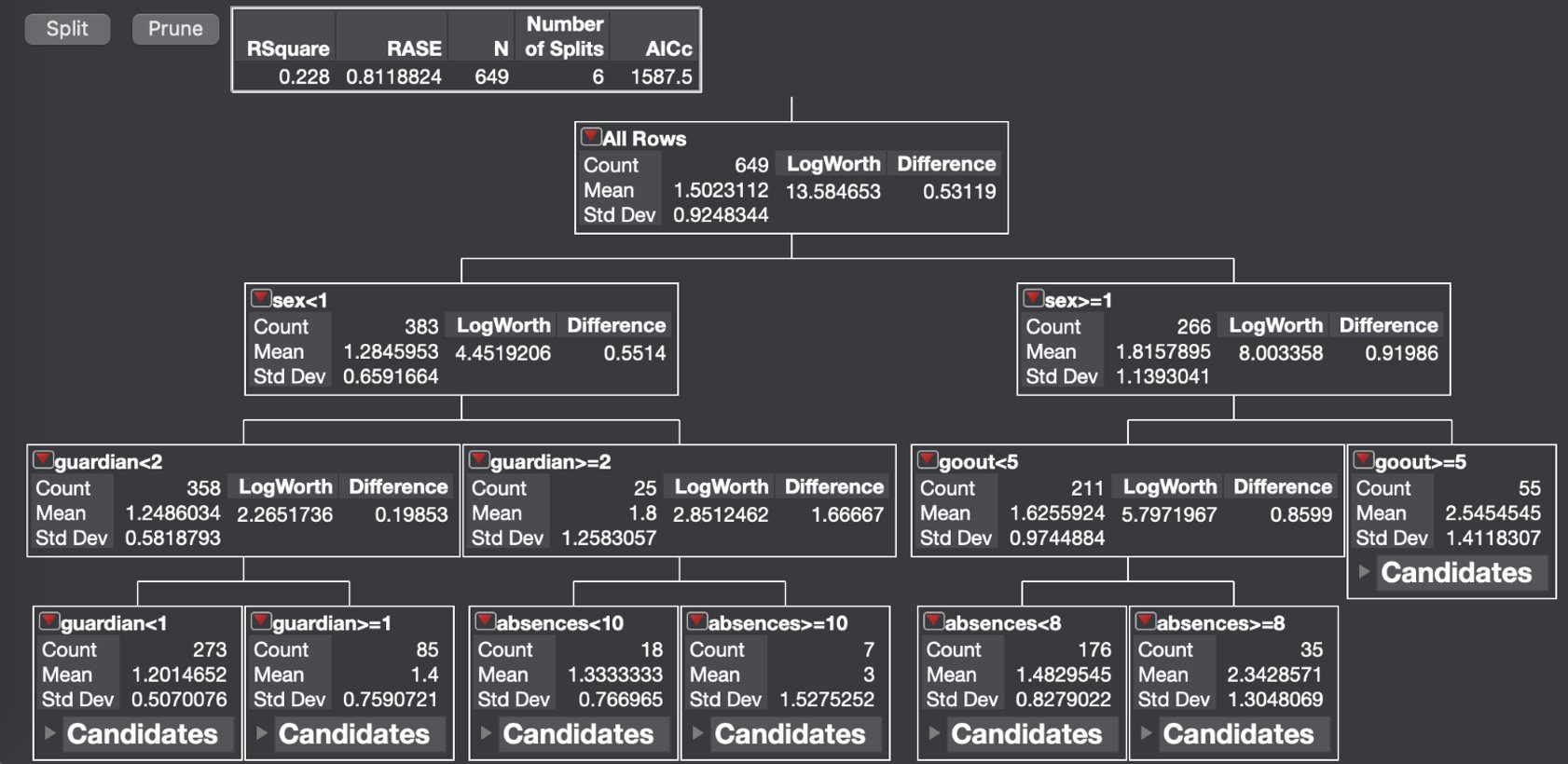


**Decision Trees:**

We created a decision tree for calculating weekend alcohol consumption. The first split was based on how frequently people go out with those who go out more frequently drinking more on the weekends. For both those who go out frequently and those who do not, their level of weekend alcohol consumption can be best predicted by their sex with females drinking during the weekends less. An interesting difference between males and females is that for males who go out frequently, their alcohol consumption is best predicted by their final grade in school while for females, their alcohol consumption is best predicted by the quality of their family relationships. Weaker family relationships and lower grades are associated with more drinking.



We also created a decision tree for calculating weekday alcohol consumption. For this tree, the first split was based on sex rather than frequency of going out. Males appear to be more likely than females to drink during the week. For females, their amount of drinking is more dependent on whether their guardian is either their mother or father or if their guardian is someone else. Females with their dad as their guardian are more likely to drink than if their mom is their guardian.



**Conclusion:**

Our initial goal was to find what factors were the best predictors of alcohol consumption in students. We gained most of our insight from clustering and decision trees. Both models showed sex to be a highly influential factor. Final grade was also a factor we found to be useful in both. From clustering specifically, we also saw that parent education and health were related to alcohol consumption. In our decision trees we found that for predicting alcohol consumption, the strength of family relationships was more important for females that go out while final grade was more important for the males. Our neural network accounted for 37% of the variability in weekend alcohol consumption and our MBR model did a decent job at predicting new cases. Overall, we were not very surprised with the results from our analysis since the predictors were quite intuitive.