Cognitive Science 106: Data Analysis Project Paper by Happy Han

Data

Data selected came from Kaggle collected by UCI machine learning. Data includes social information, gender, alcohol consumption, and academic performance. Data types are objects and integers. There was no missing data. Binary data with “yes” and “no” objects (i.e. family support, school support, etc.) were encoded into ones and zeros, respectively.

Important Variables for Observation:
Parents’ education, family support, school support, alcohol consumption on workday, and grades.

Exploratory Data Analysis

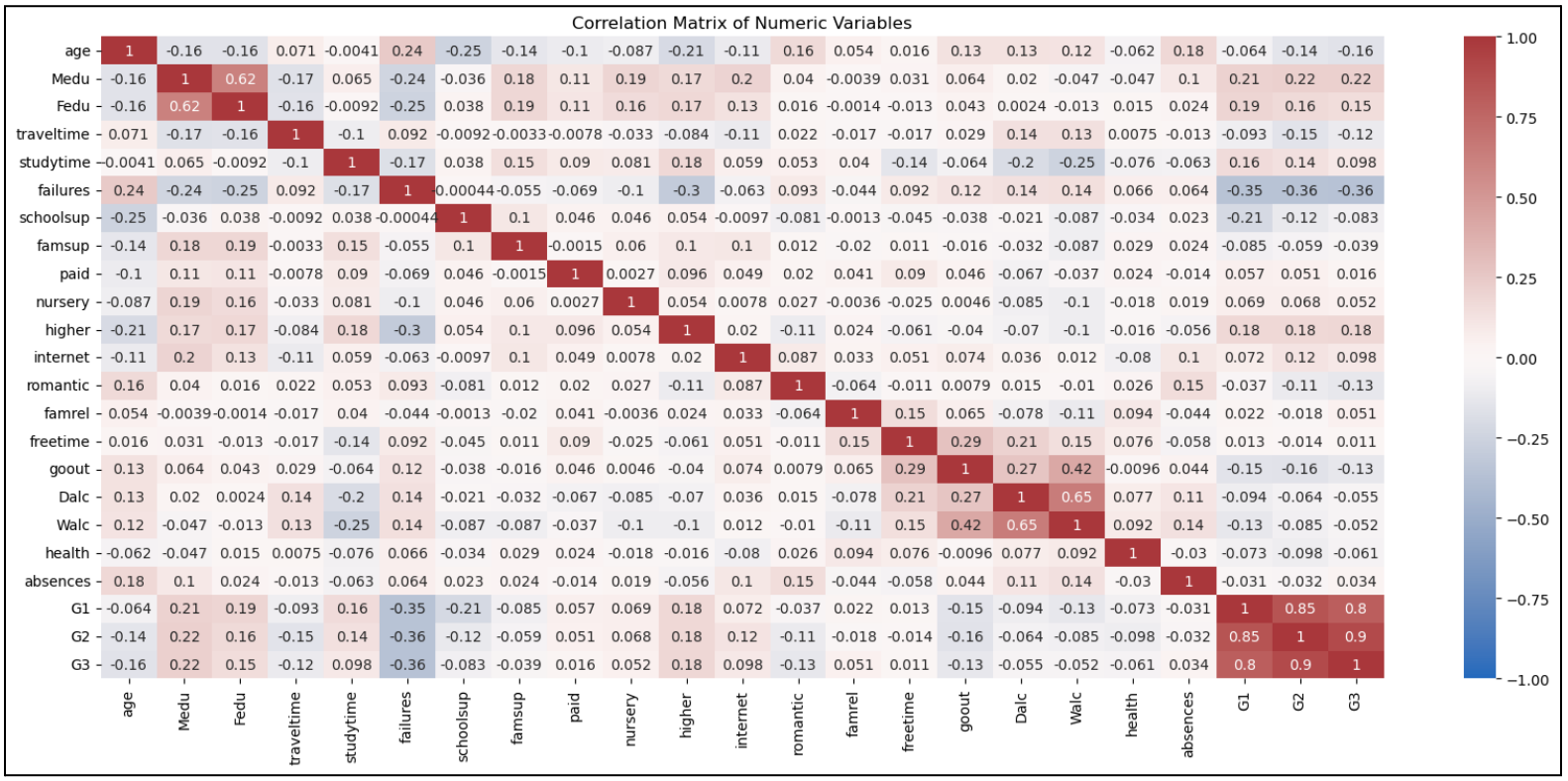

Figure 1: Correlation plot of all the numerical data including the binary yes and no after encoding into 1’s and 0’s.

G1, G2, and G3 has very high correlation so I only used G3 for grades overall. Parents' education has some correlation to grades, and family support. Workday drinking (Dalc) and weekend drinking (Walc) have reasonably high correlation so I only use workday drinking.

Questions

What are the best variables to predict a student's academic performance?
Can we predict family support?
What are the best predictors of alcohol consumption?

Modeling

Model 1: Lasso to Predict Overall Grade

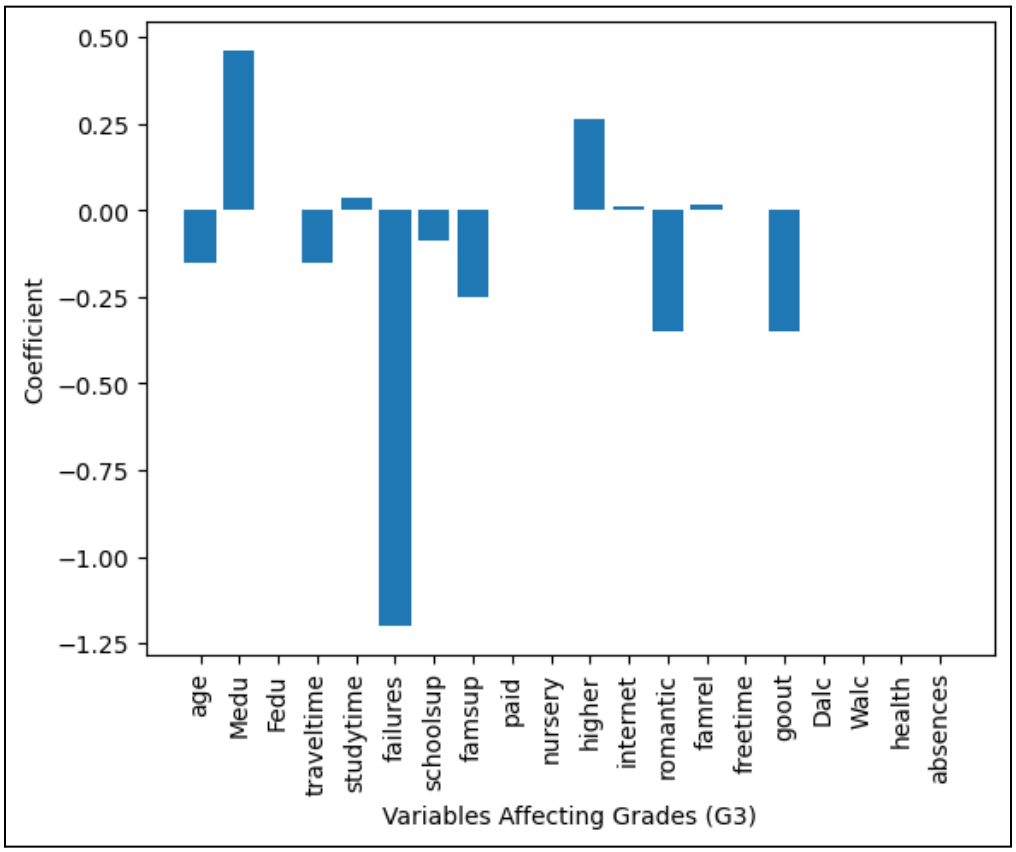

Figure 2: All the coefficients of the variables used to predict overall grade (G3).

Failures have a correlation of -1.2 being a strong influence of predicting grades. There's a negative correlation so as failures increase, overall grade would decrease. Mother's education has a correlation of 0.46, a positive correlation. The more educated the student's mother is, the better they do academically. One can theorize that mothers would spend their time providing academic support. Oddly enough, family support is negatively correlated so the theory isn't too well supported in this data set.

Model 2: Logistic Regression to Predict Family Support

	precision	recall	f1-score	support
No Support	0.55	0.35	0.43	46
Has Support	0.57	0.75	0.65	53
accuracy			0.57	99
macro avg	0.56	0.55	0.54	99
weighted avg	0.56	0.57	0.55	99

Figure 3: classification report of logistic regression showing the accuracy percentage.

In figure 3, the logistic model is better at predicting if a student has family support (0.57 precision) than predicting if a student doesn't have family support (0.55 precision). But considering the ROC (in figure 4) curve being really close to the reference line, the model isn't a good classification.

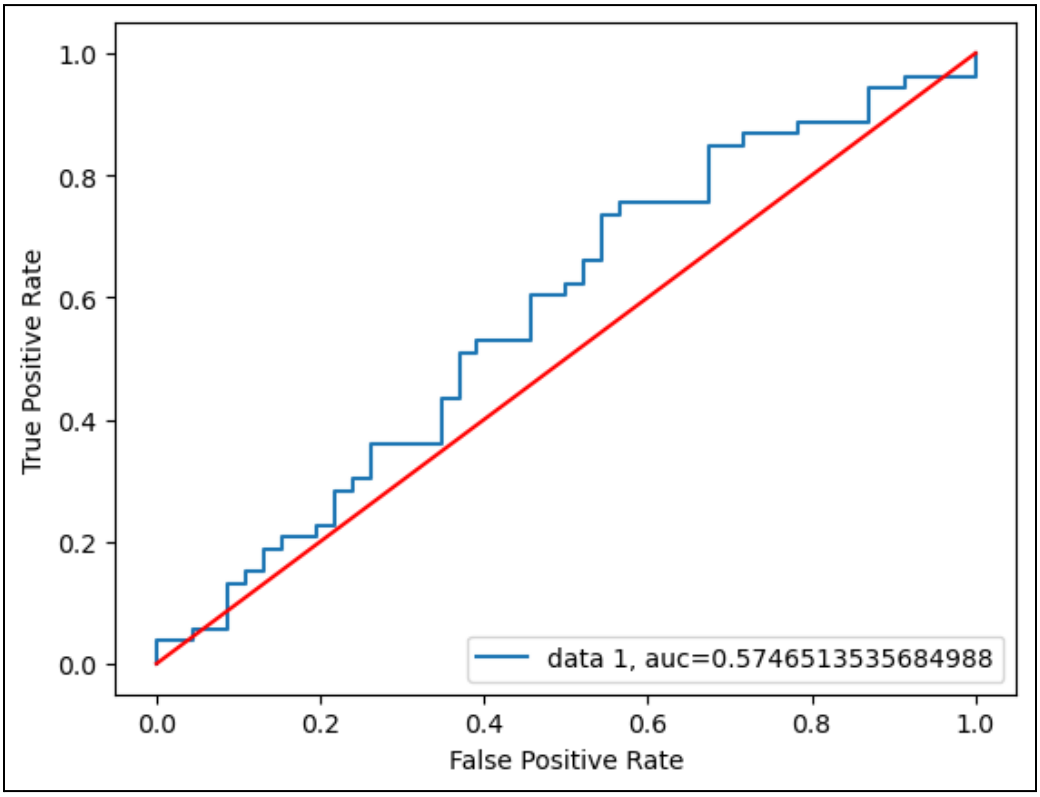

Figure 4: ROC curve of the logistic regression. Logistic regression wasn't too helpful in comparison to just random chance.

Model 3: RandomForest to find important variables to predict alcohol consumption

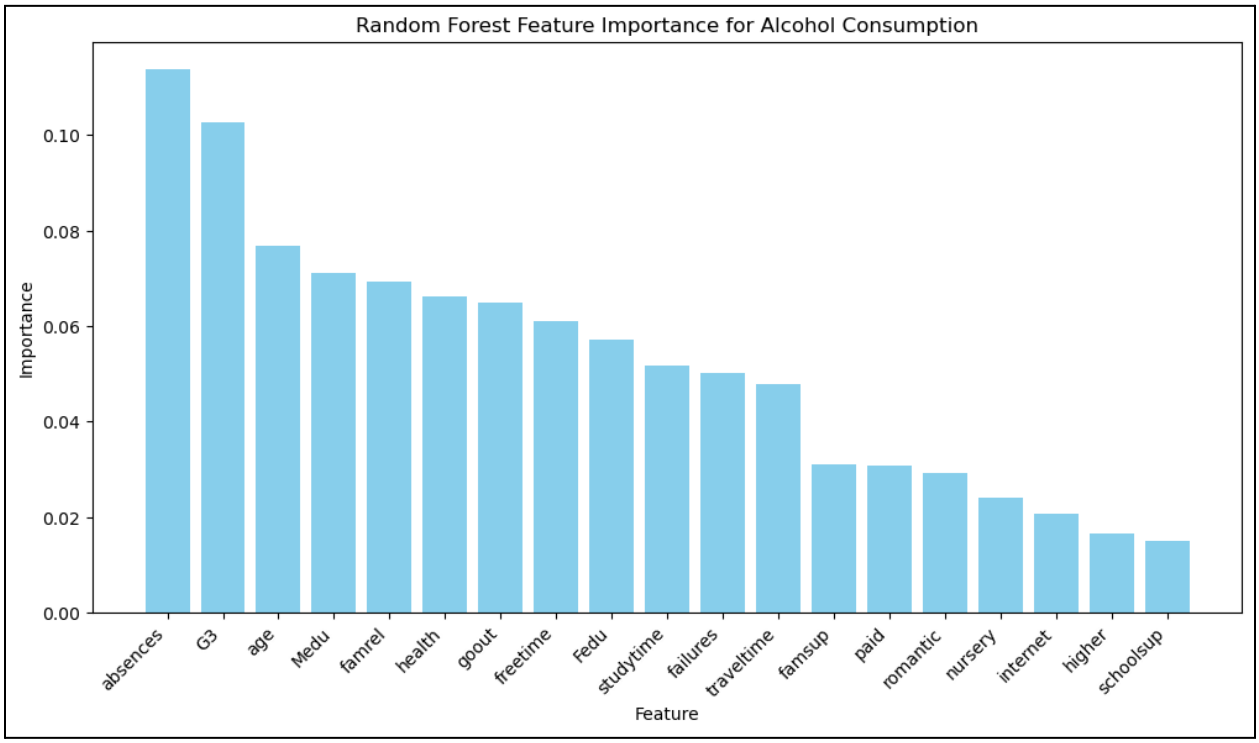

Figure 5: Randomforest to find important variables to predict alcohol consumption.

I just wanted to explore using randomforest. I got the proportion of important variables to predict alcohol consumption. It's oddly surprising to find absences to have the highest percent importance, even if it's only about 0.11.

Conclusion

Using Lasso and Randomforest were helpful in deciding on the variables that are helpful to predicting grades and alcohol, respectively. Besides finding variable correlations and connections, the data analysis doesn't conclude anything.

One bothersome issue is the ranges in the data. An example would be weekly study time: 1 if less than 2 hours, 2 if between 2-5 hours, 3 if between 5-10 hours, or 4 if more than 10 hours. Even worse, the going out and alcohol consumption variable is a subject scale of "very low" for 1 to "very high" for 5. Scaling the data doesn't seem enough.

I've tried to apply logistic regression to predict alcohol consumption. I tried to learn using one vs rest but I couldn't understand how to use it. Thus changing logistic regression to family support which is binary.