Predicting Alcohol Use of College Students

Marisa Mitchell

Springboard Capstone Project 1

The Problem

- 60% of college students ages 18-22 drank alcohol in the past month
- Nearly 2 out of 3 of those students engaged in binge drinking
- Binge drinking can lead to a variety of harmful consequences such as:
  - Death
  - Assault
  - Sexual assault
  - Academic problems
- Colleges may want to provide targeted interventions to students at risk of binge drinking to prevent these harmful consequences

The Data

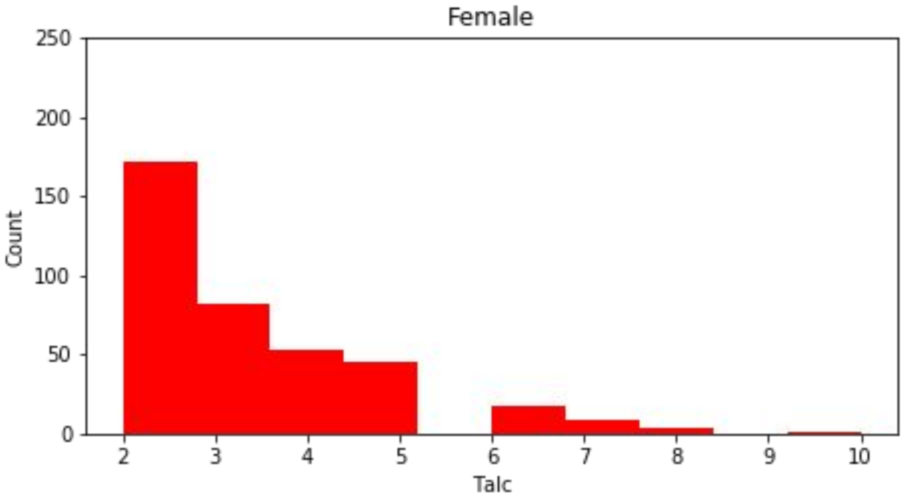
- UCI Machine Learning Student Alcohol Consumption dataset located on Kaggle (https://www.kaggle.com/uciml/student-alcohol-consumption/data)
- 33 variables
- 649 students at one college

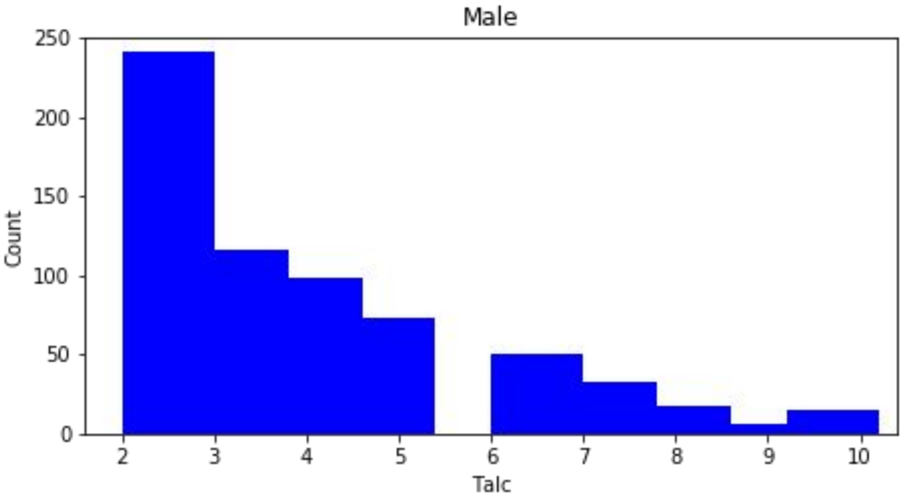
Data Cleaning Steps

Total Alcohol Consumption (TALC)	Creating Dummy Codes	Binning TALC
Creation of TALC variable by totaling the numeric ratings of workday and weekend alcohol consumption. Scale 2(very low)-10(very high)	Dummy variables for several categorical variables such as sex, address (urban or rural), family size, mother’s education, and father’s job	TALC variable was binned into categories of low(≤3), medium(4-6), and high (≥7).

Exploratory Analysis

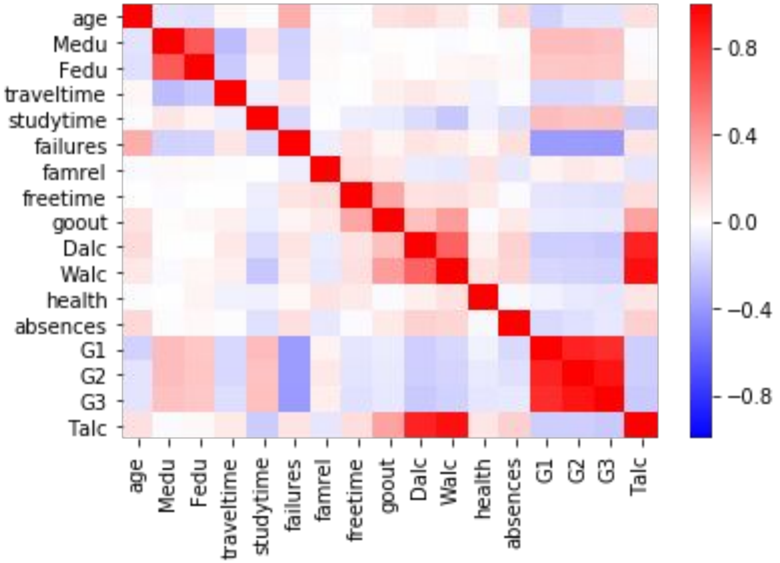
Total alcohol by Sex





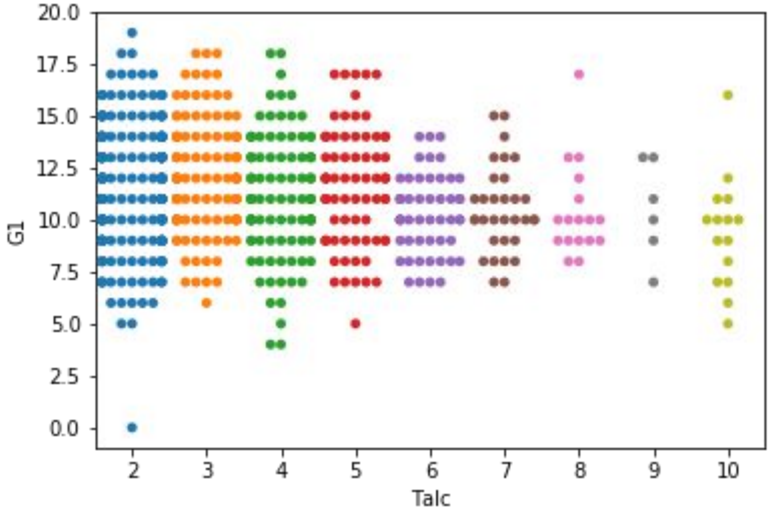
Exploratory Analysis

Heatmap showing correlations of all variables in the dataset



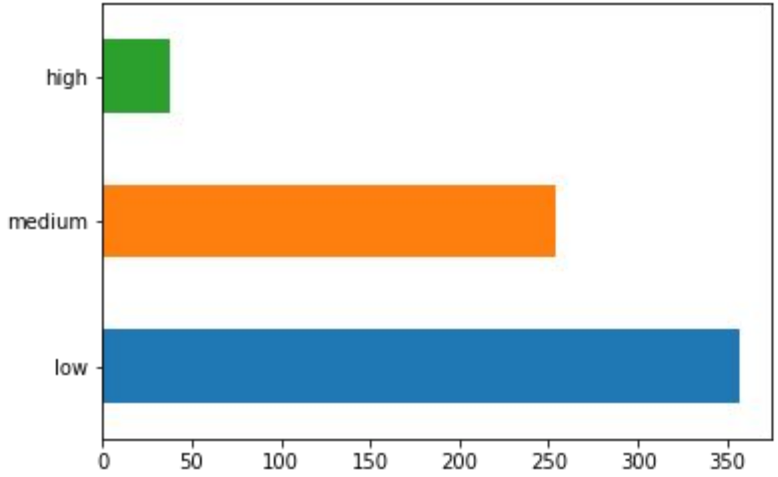
Exploratory Analysis

Swarmplot of total alcohol vs grade 1 before binning

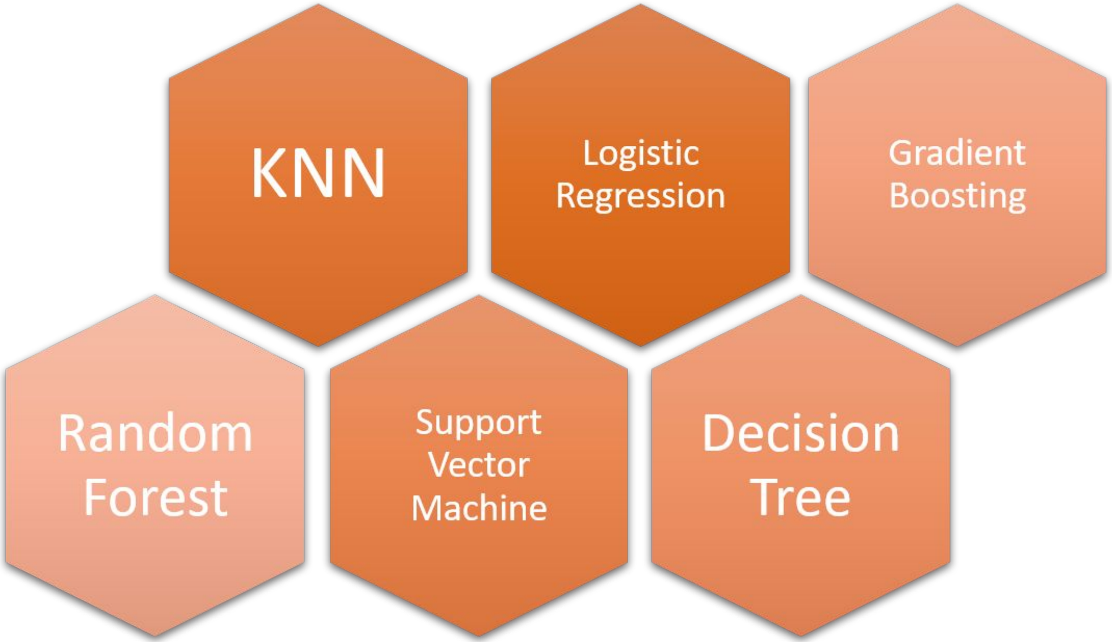


Exploratory Analysis

Bar chart of total student alcohol level after binning



Machine Learning Algorithms Used for Classification



Feature Engineering

Dummy variables were created for the following categorical variables for the KNN, Logistic Regression, and SVM models:

- School
- Sex
- Address
- Family Size
- Parent Status
- Mother's job
- Father's job
- Reason
- Guardian
- School Support
- Family Support
- Paid
- Activities
- Nursery
- Higher
- Internet
- Romantic
- Mother's Education
- Father's Education
- Travel Time
- Study Time
- Family Relationships
- Free Time
- Going Out
- Health

Feature Engineering

Numerical values were created for each of the string values the following categorical variables for the tree-based models:

- School
- Sex
- Address
- Family Size
- Parent Status
- Mother’s job
- Father’s job
- Reason
- Guardian
- School Support
- Family Support
- Paid
- Activities
- Nursery
- Higher
- Internet
- Romantic

Model Evaluation Performance Metrics

Model	Accuracy Score (Before Tuning)	Accuracy Score (After Tuning)
KNN	0.564	-
Logistic Regression	0.636	0.646
SVM	0.641	0.641
Decision Tree	0.503	-
Random Forest	0.615	-
Gradient Boosting	0.621	0.631

Confusion Matrices for top 3 models

Logistic Regression

	high	low	medium
high	1	2	7
low	1	92	23
medium	1	35	33

Gradient Boosting

	high	low	medium
high	3	2	5
low	0	87	29
medium	2	34	33

SVM

	high	low	medium
high	0	2	8
low	0	98	18
medium	0	42	27

Classification Report for top 3 models

Group	Precision	Recall	f1-score	support
Logistic Regression (avg/total)	0.63	0.65	0.63	195
High	0.33	0.10	0.15	10
Low	0.71	0.79	0.75	116
Medium	0.52	0.48	0.50	69
SVM (avg/total)	0.59	0.64	0.61	195
High	0.00	0.00	0.00	10
Low	0.69	0.84	0.76	116
Medium	0.51	0.39	0.44	69
Gradient Boosting (avg/total)	0.63	0.63	0.63	195
High	0.60	0.30	0.40	10
Low	0.71	0.75	0.73	116
Medium	0.49	0.48	0.49	69

Conclusions

- Logistic regression was the most accurate model after tuning with an accuracy of 64.6%
- However, logistic regression was not the best at correctly identifying the high alcohol level group (recall = .10)
- The gradient boosting model may be the best choice due to decent overall accuracy (63.1%) and the best recall (.30) for the high alcohol level group

Next Steps

To deal with small count of students in the high alcohol group some next steps could be:

- Apply resampling techniques such as oversampling or undersampling
- Collect more data
- Generate synthetic samples
- Use a penalized model