# Predicting Alcohol Consumption of High School Students

Ocean Liu, Derek Modzelewski, George Li, Zenny Chu

## Introduction

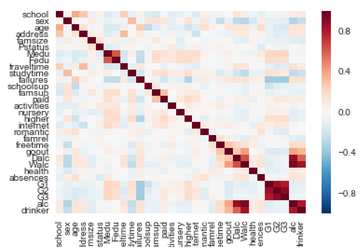
- Alcohol clearly impacts young people's health and education. What causes this behavior? How can we predict the students most likely to become alcoholics? We try to answer these questions by analyzing a survey on the lifestyles of 396 Portuguese high school students.
- The survey asks about a variety of topics: grades, parental status, relationship status. We show that the aggregate of this data is predictive of students' alcohol consumption. Furthermore, we show which features are the most important for this prediction.

## Data Preprocessing

- Nonbinary categorical variables were not used. Binary ones were converted to numerical values
- Weekend and weekday drinking was averaged as the overall drinking level. For classification, overall drinking level was represented as the binary feature "drinker" (being above median drinker or below median drinker)

## Data Visualization

- The heatmap shows that most of the features were not correlated.
- The most significant were weekend and weekday drinking, mother and father occupation, the three grades, and failure and grades



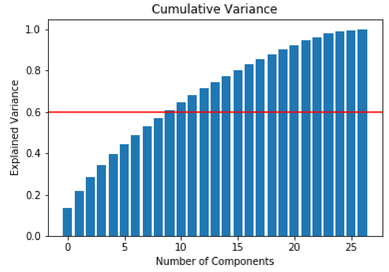

## Linear Regression

- Recall that we threshold a continuous variable to determine who is a drinker. This throws away data. To circumvent this, we use Linear Regression in addition to other classification methods.
- Using Leave-One-Out-Reconstruction (use all but one data sample, predict the label of the missing sample), we found that a linear model explains 25% of the variance in alcohol consumption.
- When we throw away some of the features, the order of importance of the other features varies so we cannot give a strong result of which features are better. As an approximation, we report the average rank (where 0 is the most predictive) of each feature as we remove features:

| Feature | Average Rank |
|---|---|
| goout | 0.00 |
| sex | 1.00 |
| G1 | 2.00 |
| paid | 4.33 |
| studytime | 5.60 |
| **Feature** | **Average Rank** |
| Pstatus | 25.04 |
| medu | 23.39 |
| schoolsup | 22.52 |
| school | 22.50 |
| romantic | 22.30 |

## PCA

- PCA is maximally effective when a small number of components explains a large part of the variance. From the covariance heatmap, we see that the features are largely independent. Furthermore, we see that the variance explained is almost linear w.r.t. the number of components. These facts each indicate that using PCA will not be effective.

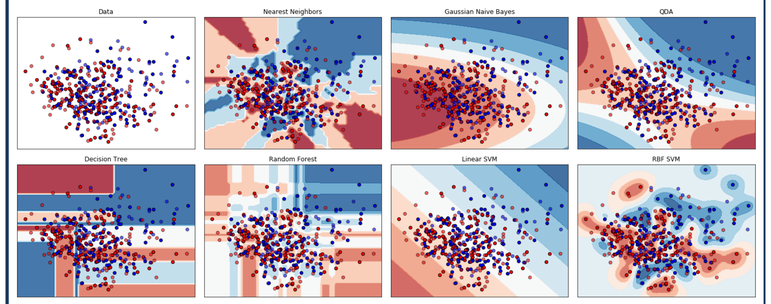



## Classifier Comparison

- Mean scores were computed using 10-Fold Cross Validation
- Random Forest before PCA had parameters trees = 9 and depth = 2
- Random Forest after PCA had parameters trees = 8 and depth =3

| Classifier | Mean Score (Before PCA) | Mean Score (PCA) |
|---|---|---|
| kNN (n=7) | 0.5975 | 0.6279 |
| Gaussian NB | 0.6368 | 0.6148 |
| QDA | 0.5738 | 0.6095 |
| Tree | 0.643 | 0.5542 |
| Random Forest | 0.6731 | 0.6256 |
| SVC Linear | 0.686 | 0.6553 |
| SVC RBF | 0.5468 | 0.5544 |

## Visuals




- The graphs above show the original data plotted against various classification methods

## Conclusions

- We used both linear and nonlinear classifiers on the data, and our tests yielded scores around 60-65%. These unimpressive scores reflect the mediocre reconstruction explained variance of 25%.
- While none of the features are dominant, the three that had the most value in predicting alcohol consumption were "goout", "sex", and "G1".
- The data clouds for "drinkers" and "non-drinkers" heavily overlap, explaining why the decision boundary for kNN is terrible. Linear SVC seems to be the most successful classifier because it has the least model variance and so is least affected by the noisy data.

## Further Improvements

- Use less ambiguous survey questions - ask about ranges of drinks, not an ungrounded 1-5 scale.
- Use other thresholds to demark a "drinker". If we find that only 20% of students are heavy drinkers (whereas we assume 50% are), can re-run analysis.
- Separate features into school, home, and personal life. See how well each feature-set works individually.