Predicting Student Alcohol Consumption using Machine Learning

Ashish Shrestha
College of Computing and Informatics
Drexel University
Philadelphia, PA 19104
http://aasys.io

Abstract—**This paper explores the application of machine learning in effectively predicting if a student is likely to alcohol abuse. Dataset of 1033 high-school students is used for the study. This dataset contains 31 features along with student alcohol consumption habits. Decision tree classification algorithm is used to perform a binary classification which outputs if the student is likely to abuse alcohol or not. The result of predictive model thus created is positive with a maximum observed accuracy of 87.93%. Additionally, features that most heavily determine if a student is likely to alcohol abuse is also discovered.**

I. Introduction

There are many studies that suggest that alcohol doesnt mix well with academics. Although the negative effects might not be apparent with occasional consumption, alcohol abuse can serious short-term and long term implications., as this happens to be the age when the brain is still developing. High alcohol consumption exacts an enormous toll on both the social and intellectual part of a students life. Only in the United States of America, research estimate about 1,825 college students between the ages of 18 and 24 die from alcohol-related unintentional injuries, including motor-vehicle crashes. Furthermore, it is estimated that about 696,000 students between the ages of 18 and 24 are assaulted by another student who has been drinking. Additionally, about 97,000 students between the same age group report experiencing alcohol-related sexual assault or date rape. About 20 percent of college students in US are estimated to fall under alcohol use disorder. These students are not only more likely to perform poor academically, they also have a higher chance of attempting suicide, having other health problems, engaging in unsafe sex, and driving under the influence of alcohol, as well as vandalism, property damage, and involvement with the police. [1][2]

II. Related Work

Research paper titled Using Data Mining To Predict Secondary School Student Alcohol Consumption by Fabio Pagnotta, Hossain Mohammad Amran from Department of Computer Science, University of Camerino also explores the same subject. Using the same dataset, the prementioned researchers used Konstanz Information Miner(KNIME) as their machine learning library. Their algorithm of choice was decision trees and random forest for classification. They could achieve an error rate of 8.018% and accuracy of 92% in their classification. [3]

III. Dataset

This work uses the dataset freely available on UCI Machine Learning Repository. The dataset contains data on 1044 Portuguese secondary school students. That dataset is dated 2005-2006 academic calendar and composed by Paulo Cortez and Alice Silva, University of Minho,Portugal. There are 33 total features in the dataset for each student, separated into math and Portuguese students. The data was collected with questionnaires, along with grade reports. More detailed information of each feature is presented in TABLE 1[4]

TABLE I
About the Dataset

1	school	student's school (binary: 'GP'or 'MS')
2	sex	student's sex (binary: 'F' - female or 'M' - male)
3	age	student's age (numeric: from 15 to 22)
4	address	student's home address type (binary: 'U' - urban?)
5	famsize	family size (binary: 'LE3' - less or equal to 3?)
6	Pstatus	parent's cohabitation status (binary: 'T' - together?)
7	Medu	mother's education (numeric: 0 - 4)
8	Fedu	father's education (numeric: 0 - 4)
9	Mjob	mother's job (nominal: 'teacher', 'health' ...)
10	Fjob	father's job (nominal: 'teacher', 'health' ...)
11	reason	reason to choose this school (nominal: 'reputation' ...)
12	guardian	student's guardian (nominal: 'mother', 'father' 'other')
13	traveltime	home to school travel time (numeric: 1 - 4)
14	studytime	weekly study time (numeric: 1 - 4)
15	failures	number of past class failures (numeric:1-4)
16	schoolsup	extra educational support (binary: yes or no)
17	famsup	family educational support (binary: yes or no)
18	paid	extra paid classes (binary: yes or no)
19	activities	extra-curricular activities (binary: yes or no)
20	nursery	attended nursery school (binary: yes or no)
21	higher	wants to take higher education (binary: yes or no)
22	internet	Internet access at home (binary: yes or no)
23	romantic	in relationship (binary: yes or no)
24	famrel	family relationships (numeric: 1-5)
25	freetime	free time (numeric: 1-5)
26	goout	going out with friends (numeric: 1-5)
27	Dalc	workday alcohol consumption (numeric: 1-5)
28	Walc	weekend alcohol consumption (numeric: 1-5)
29	health	current health status (numeric: 1-5)
30	absences	number of school absences (numeric: from 0 to 93)
31	G1	first period grade (numeric: from 0 to 20)
32	G2	second period grade (numeric: from 0 to 20)
33	G3	final grade (numeric: from 0 to 20, output target)

IV. Learning

The machine learning tool used for this is MATLAB 2016 R2 and its machine learning libraries. First two data set were merged into one single dataset and all the values were adjusted to numeric values. This results into total of 31 features. The other two features, week-end alcohol consumption and week-day alcohol consumption were both merged into one by multiplying the each with amount of days and then dividing by the 7, days in a week. This gives 1-5 value for weekly alcohol consumption.

Weekly Alcohol Consumption = (5 x week-day consumption + 2 x week-end consumption) / 7

Taking scale 1 and 2 as a normal non-alcohol abusive drinker and 3, 4 and 5 as heavy drinkers, this problem can be taken as binary classification problem. The resulting 31 feature set and the target binary alcohol abuser value can be now fed into a learning algorithm.

Firstly, lets see how each feature correlates to alcohol abuse. Hanging out often with friends and students gender seem to have highest positive correlation with alcohol abuse, while greater study time seems to have a negative correlation with alcohol abuse.

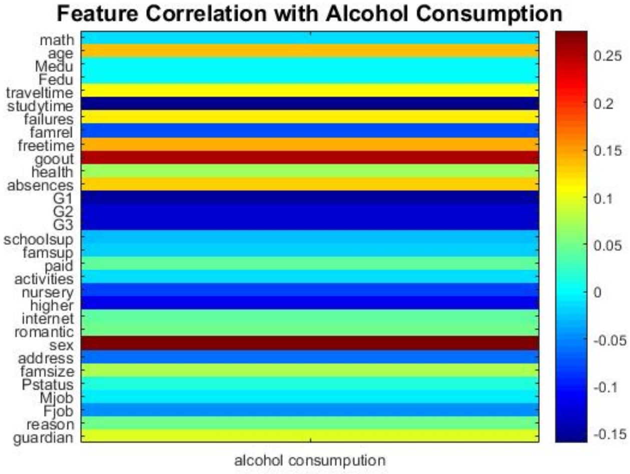

Fig. 1. Correlation of features with alcohol abuse

A. Decision Tree Learning

The learning algorithm of choice here was decision tree learning. This was chosen after preliminary evaluation with multiple classification algorithms which gave lower performance comparatively. Furthermore, decision tree learning also gives more insights to data regarding which features are most likely to affect the output.

The data was randomized and 2/3 of the 1044 students data was separated into as training data and the remaining 1/3 as testing data.

fitctree from MATLAB is used as the decision tree based classification algorithm. This algorithm fits binary classification decision tree for multiclass classification. As in decision tree, is node is a input feature which leads into other nodes with other input features until classification can be done with some amount of confidence.

Using fitctree as it is, gave an accuracy of 81.9% and mean deviation of 0.18. The decision tree resulted from this is show in figure 2. Additionally, figure 3 shows the weight of each predictor in the decision tree which shows peaks in goout and sex which matches that of the correlation chart, with carious other features carrying lesser weights.

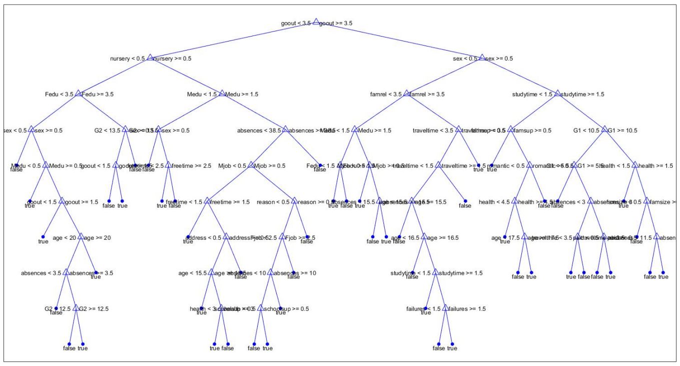

Fig. 2. Decision tree formed by fitctree

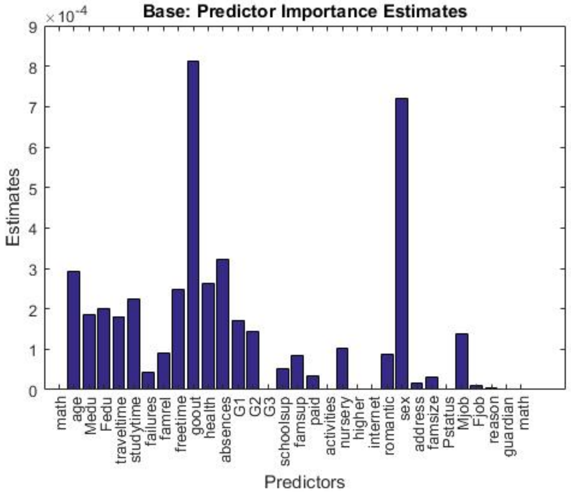

Fig. 3. Predictor Importance Estimate

B. Improvement

Accuracy of 81.9% is acceptable but this can be improved upon. Various methods were explored inorder to improve the predictor model accurracy. First, by pruning the decision tree on various levels, accuracy can be checked against each pruned tree. Figure 4 and 5 shows the effect of on accuracy and class deviation on various levels. Here, level 11 seems to give the highest accuracy and lowest deviation. Figure 6 shows the change in precision and recall in binary classification at each pruning level.

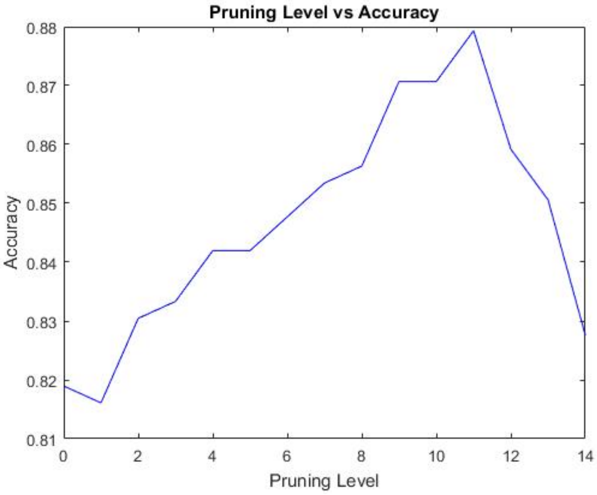

Fig. 4. Pruning Level vs Accuracy

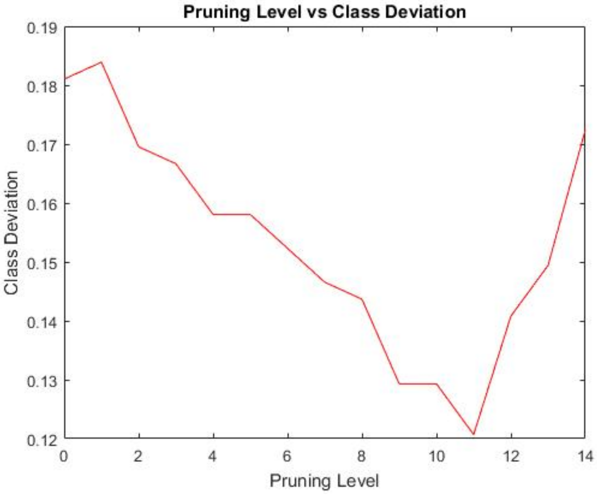

Fig. 5. Pruning Level vs Mean Class Deviation

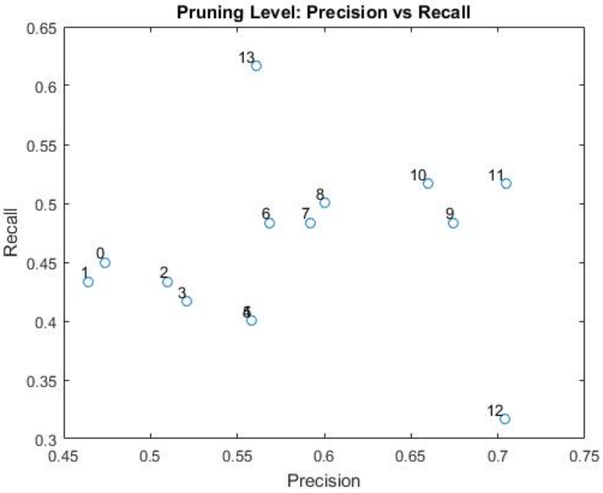

Fig. 6. precision recall graph and Pruning level

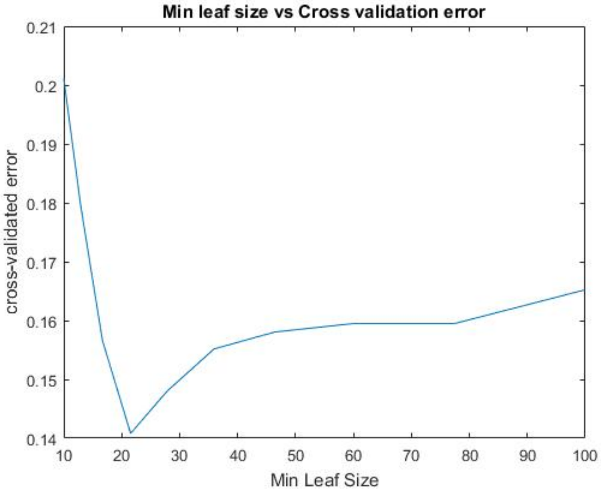

Fig. 7. Min Leaf Size vs cross-validation error

Another method of increasing accuracy is by adjusting the minimum leaf size. Figure 7 shows the cross-validation error when changing each 'MinLeafSize learning parameter. The accuracy of both methods resulted in 87.93%, with a lower class deviation of 0.12 which is considerable improvement. Precision is 0.7045 and Recall is 0.5167 for this new model. Also, both methods result in the same decision tree as shown in figure 8 and 9. Also the predictor importance estimate is updated for the newer models with just a new peaks remaining as shown in figure 10.

V. Conclusion

The final decision tree gives and insight that male students that goes out to often with friends, and spends little time in studying has the highest chance of alcohol abuse. This added with the grade, seems to determine most of the alcohol abusive group in the 1044 student population.

As discussed in the introduction it is established that alcohol consumption and abuse by youngsters have deteriorating effects. Using machine learning in areas like these has huge benefits in prediction of students that are mostly likely or are already alcohol abuse. This gives individuals, institution and government the ability to react, to control and prevent alcohol abuse at individual level. Additionally, insight like these can help determine risk groups and thereby give institutions and government the power to be able to reach out for help and awareness, as well as create programs and governing laws at a more specific level to have maximum outreach and efficacy.

VI. Future Work

The result here is less impressive compared to the work done by Fabio Pagnotta, Hossain Mohammad Amran, who could achieve 92% accuracy compared to 87.93% here. This shows that this could be futher improved upon. Furthermore, the dataset belongs to decade old 1044 Portuguese students. Newer and more relevant dataset from more diverse population could

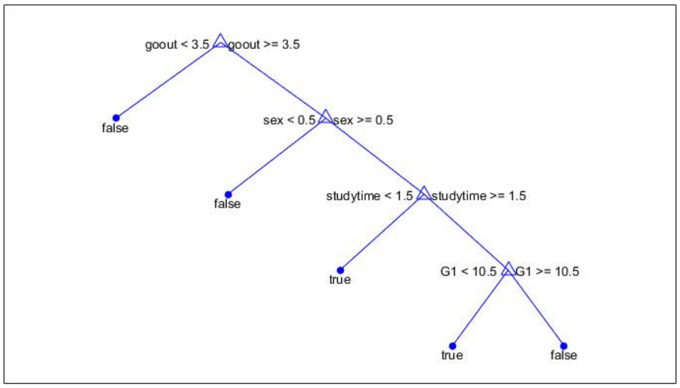

Fig. 8. Decision tree after pruning

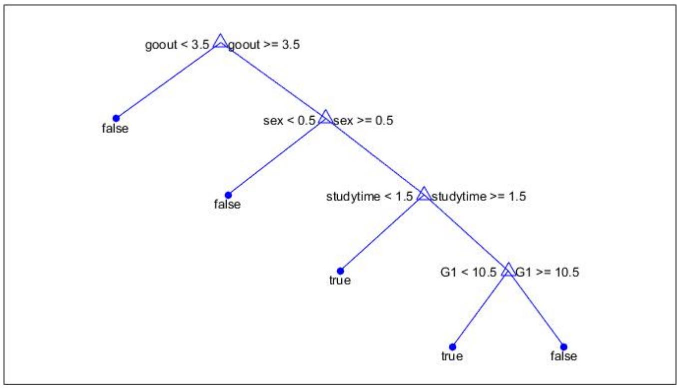

Fig. 9. Decision tree after Min Leaf adjustment

give a more accurate and widely useable model. Additionally, more features like family income, cultural background could also play and important role is determining if someone is likely to alcohol abuse.

References

[1] R. M. . Johnston. (2009) Effects of alcohol on college students. [Online]. Available: http://www.webclearinghouse.net/volume/

[2] C. Drinking. (2007) National institute on alcohol abuse and alcoholism. [Online]. Available: https://niaaa.nih.gov/alcohol-health/special-populations-co-occurring-disorders/college-drinking

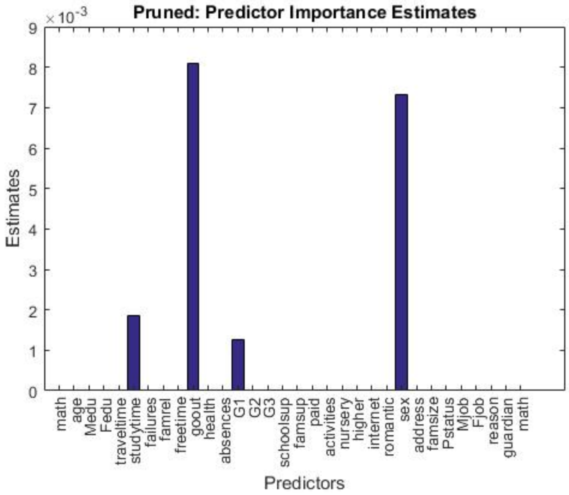

Fig. 10. Predictor Importance Estimate after pruning

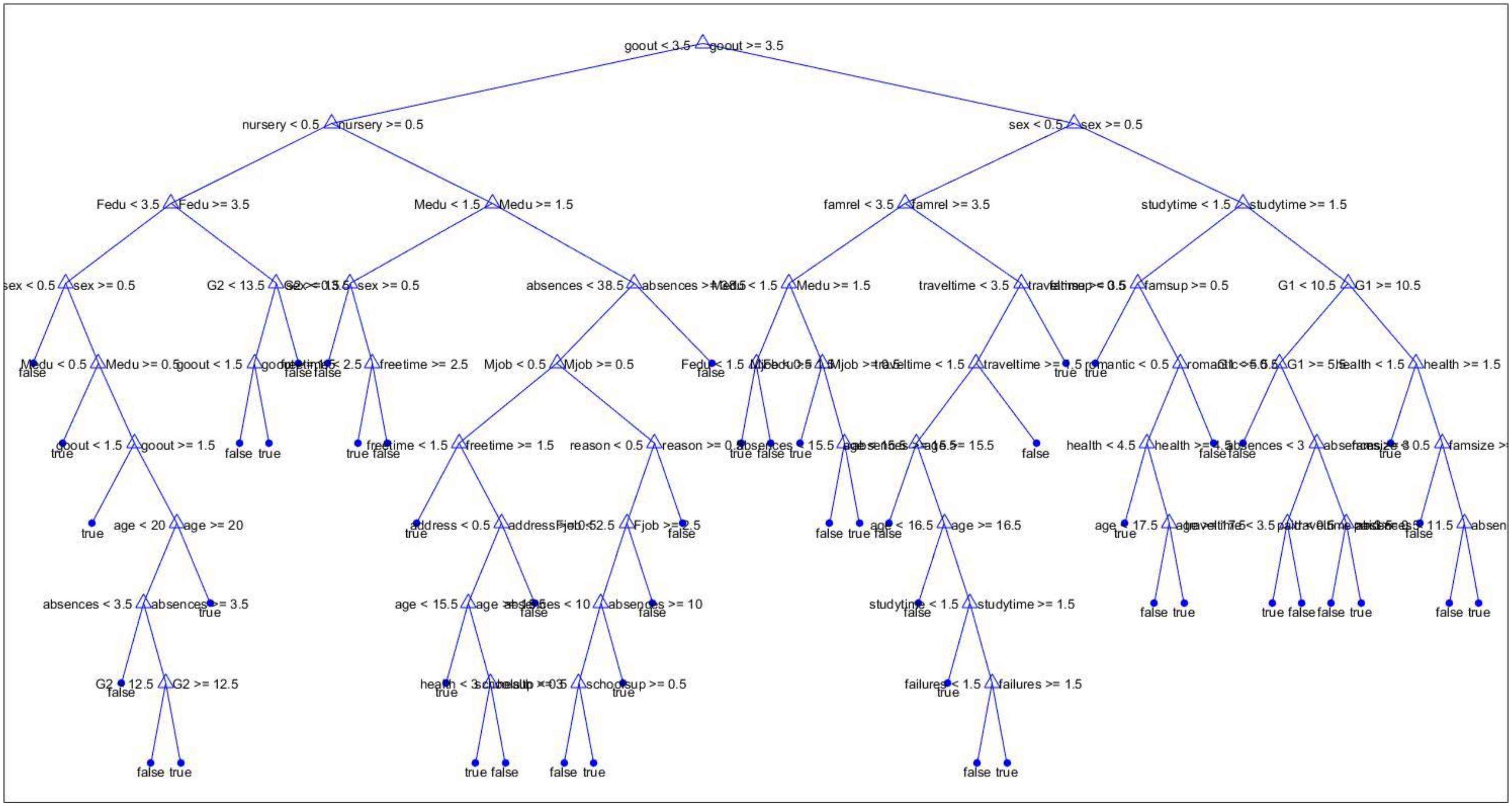

Fig 2. (ENLARGED) Base Decision tree