Predicting Whether a Student Will Fail Given Their Alcohol Consumption

Machine Learning

Project 2

Author: Cian Roddis

Student No: R00140685

Table of Contents

Page No. || Contents

Predicting Whether a Student Will Fail Given Their Alcohol Consumption

Abstract

My first thought in seeing this data set was to see if heavy alcohol consumption could be highly correlated to failure with students. This dataset provides categories on weekday alcohol consumption, weekend consumption and study time which seem like the most contributing factors to me. Failures are probably the most disheartening news you can receive as a student (in academic terms). Also we know that socializing is a huge part of the college life. Too many time have I seen people fail or even worse, drop out because they can’t keep up with the work. Granted, alcohol consumption is not the sole factor of a failure but we will be trying to prove the correlation.

Introduction

Machine learning is exploding in usage. If any major IT company has yet to use ML these days, in some aspect, they are well behind the curve. The goal of this study is figuring out whether alcohol consumption and student exam failures, correlate. I will be implementing a couple of machine learning algorithms to see if a machine can predict whether a student will fail or not. I have implemented 3 algorithms that will attempt to correlate these results.

The motivation for this project is that I have a major interest in machine learning and know from experience that too much alcohol consumption can indeed, affect your grade. I, myself have failed modules due to staying out late and allowing my attendance to drop. Granted I was working in a bar and this is the key reason because we would finish at 4 some nights which in turn ruined my sleeping schedule. Regardless, it is an easy correlation to make but here we will prove it. I will now discuss related research such as papers done by other students and professionals, then I will describe my model and give an evaluation of my findings. I will then conclude with a summation of my results and discuss future work.

Related Research

Identifying the cutting-edge research in machine learning is difficult to pin-point as there are a plethora of companies, conferences, studies, etc. invested in Machine Learning and AI. However, I am going to evaluate some related . In order to get an idea of what algorithms gave the best results I turned to Kaggle, where I got my dataset. People have the option to upload their work as “Kernels”. These Kernels are practically research papers.

1. The first Kernel I looked at was the top ranked study. This study was done by Datai and they pose the question “Does Alcohol Affect Success”[1].This study wants to test exactly what I’m trying to test. They use a mixture of python (with libraries such as pandas, numpy, mathploylib) and R programmimg language. This is where I first saw work on this dataset so I got to see graphs and results from a professional entity which is probably as good of a start anyone could get. They use mathplot library to generate graphs showing correlation of features.

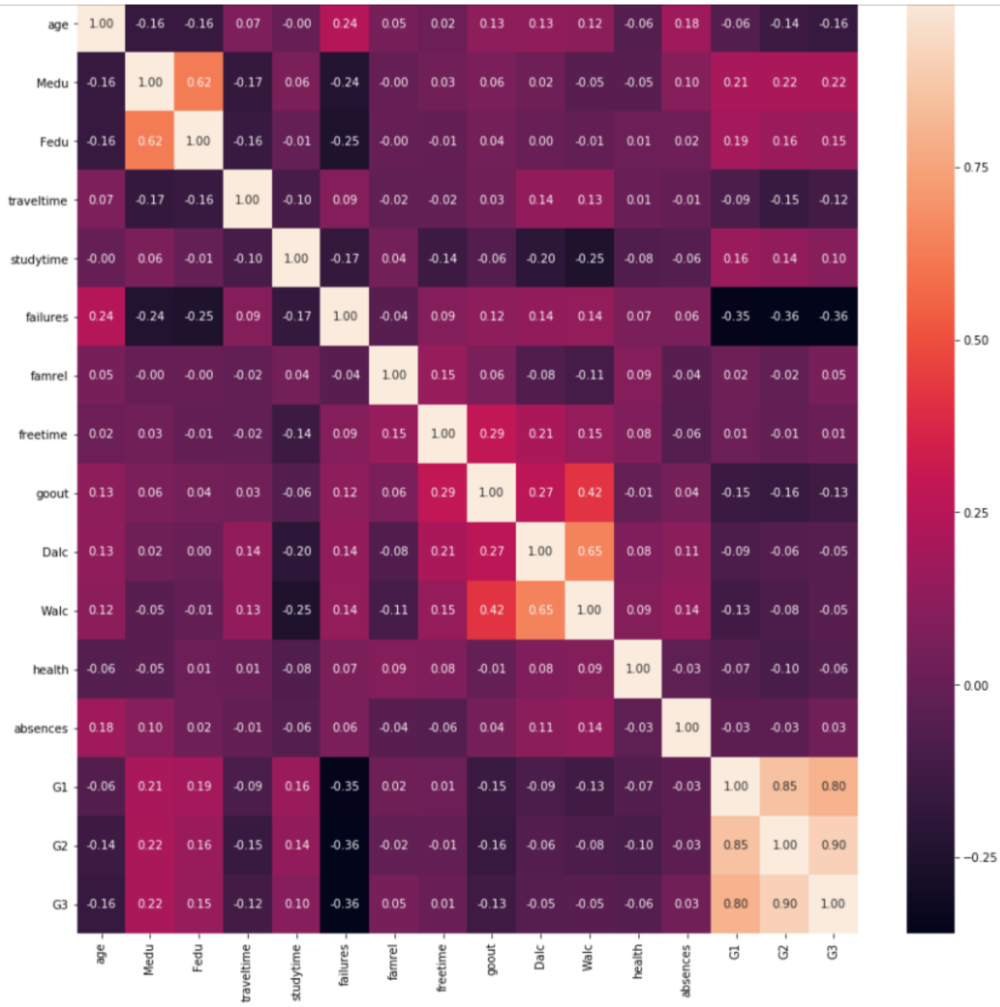


Fig.1 [1]

From this graph we can see if there is a correlation between entries. Obviously the most correlated fields are the Grade fields as most students have a consistent GPA. We can also see the correlation between fails, alcohol intake, etc.

2. The second research paper was very graph heavy and allowed me to visualize the data a lot more clearly. This study did not solely focus on alcohol intake and instead opted for a complete analysis of the whole dataset. Basic EDA and Final Grade Prediction[2] by Dmitriy Batogov is an independent study that shows us a lot of information with plenty of graphs. Here we can see the rate that students drink on weekends and weekdays. We can see percentage of males to females and he cross referenced that with the alcohol consumption graphs.

3. A similar study to this is Effect of Alcohol Use on GPA[3] by Reza Javidi takes the gender seperated graphs to the next level. She focuses on the comparison of data between the genders. This isn't hugely helpful to my study but still informative data. We can see males drink a lot heavier and this in turn is seen in the results. Females have a higher GPA as a whole in comparison to males. Not saying it's the sole factor but it is evident here.

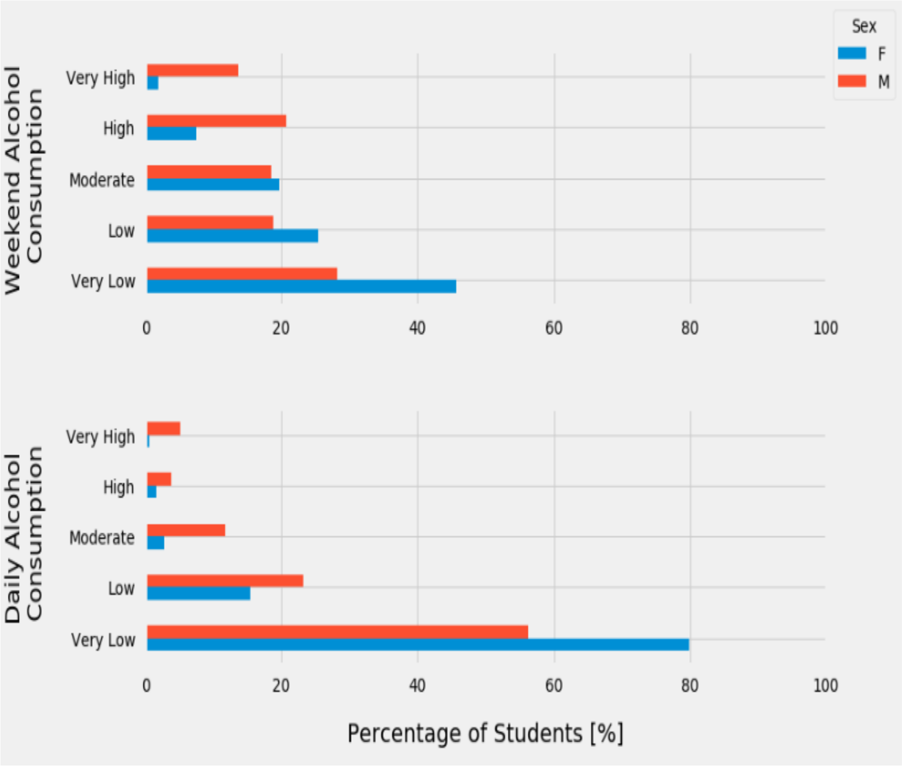


Fig.2 [3]

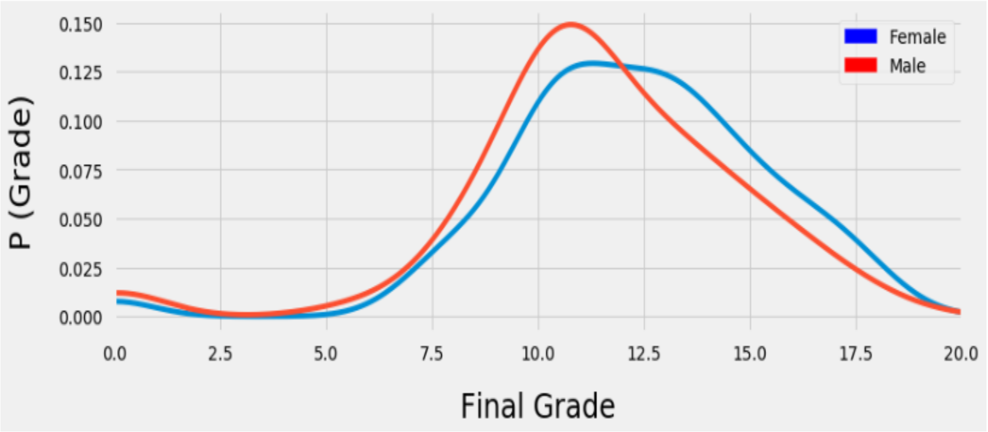


Fig.3 [3]

These graphs were done with the help of mathplot library and show us a significant correlation in alcohol intake and pass/fail rate in males and females.

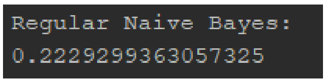
Algorithm/Model Detail

My dataset was quite simple to work with. It was structured in a way that was already normalized. Each category I was working with was a ranking of 1-5 with failures also having a 0 if the student didn't fail. There were a couple of additions I had to make. I added a total alcohol consumption to combine the Weekday and Weekend alcohol consumption categories. This was used to create a binary category too. If the student was over a certain threshold they were considered a heavy drinker so their result for this would be 1, otherwise it would be a 0. I incorporated multiple algorithms into my program to find the most accurate method possible for analyzing my dataset. I started with a primitive method of doing naive bayes. I tried to use it for comparison to sklearn's methods of machine learning.

Naive Bayes (Basic) - I took my knowledge of Naive bayes algorithm and applied it in the beginning. As I was going to be using Bernoulli's Algorithm with SciKit, I didn't see reason to apply that in a primitive way. Using this algorithm I got out a result.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

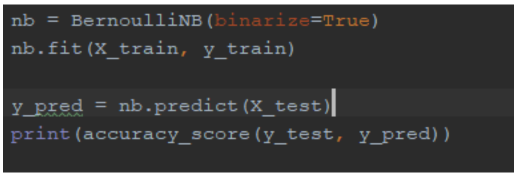
The result was poor and showed a low sense of correlation. With a result of 22.29% accuracy, I didn't try to spend any more time on it than needed.



```
Regular Naive Bayes:
0.2229299363057325
```

This is what I got out and started to doubt if this project was even possible to achieve a high achieving accuracy due to a lack of correlation.

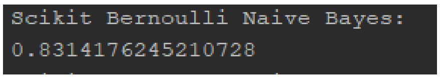
Naive Bayes (Bernoulli) - For Bernoulli's Naive Bayes Algorithm I implemented SciKit Learn's methodology for scoring accuracy. This shortens the code by my estimates, over 80%. The whole thing can fit on about 4 lines.



```
nb = BernoulliNB(binarize=True)
nb.fit(X_train, y_train)

y_pred = nb.predict(X_test)
print(accuracy_score(y_test, y_pred))
```

After implementing this algorithm I saw instant improvement, leaps and bounds above the previous implementation of naive bayes. First result I got back from this test was 83%. I was pleasantly surprised with this result. I did not think it would come back so accurate. The result came back so highly accurate because it isn't just looking at alcohol intake but also looking at study time. I added this feature because I knew that it would be highly correlated with failures if it was a low result. Thanks Bernoulli, very cool.



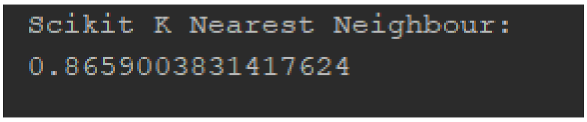
```
Scikit Bernoulli Naive Bayes:
0.8314176245210728
```

K Nearest Neighbour(KNN) - With the previous algorithm being so efficient, I was expecting a similar outcome from KNN. K nearest neighbour uses algorithms that finds K number of nodes located closest to your testing node and gives a prediction on your testing node depending on the nodes surrounding it. In my case, using 5 seemed to yield the best results. Because sklearn does most of the heavy lifting, this algorithm can also be implemented in just a few lines. I did see the best result from KNN at 86% but it could also drop as low as 77%.

Empirical Evaluation

I was experimenting with a decision tree algorithm by sklearn but decided against it as there were “value conflicts”. Even after restructuring the data I was only able to get out a score and not an accuracy score for some reason. Decision trees are better suited to all discrete attributes anyway so I stopped my research after a day of tirelessly trying to get an accuracy score.

K Nearest Neighbour uses a number specified by the programmer so I experimented with numbers 3, 5, 7 and 9 as to have no split decisions for the algorithm, I used only odd numbers. This resulted in a bunch of runs to see what had a consistently higher average accuracy score between the numbers. 5 came out on top with an average of about 83%. 3 had an average of about 82 but it was hard to call between 3 and 5. I opted for 5 as to allow the algorithm a wider base of training nodes. 7 and 9 both had a consistent 80% accuracy. 5 also had the highest overall result with 86.59



```
Scikit K Nearest Neighbour:
0.8659003831417624
```

Notable graphs include one that shows total alcohol intake and the amount of students that drink at those levels.

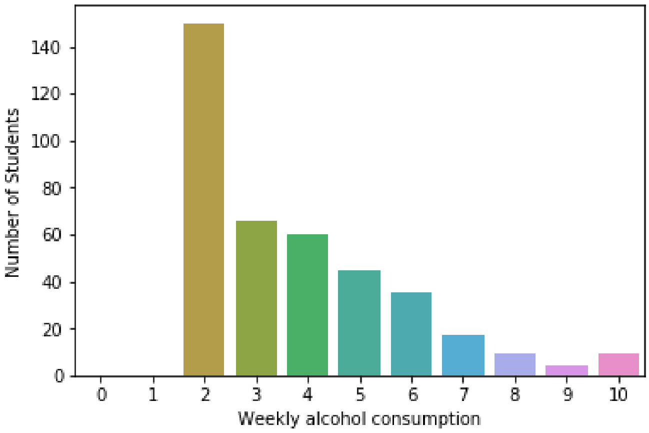


Fig. 4 [1]

This graph by Datai show a compliment to this graph and show student’s grade distribution according to their alcohol intake.

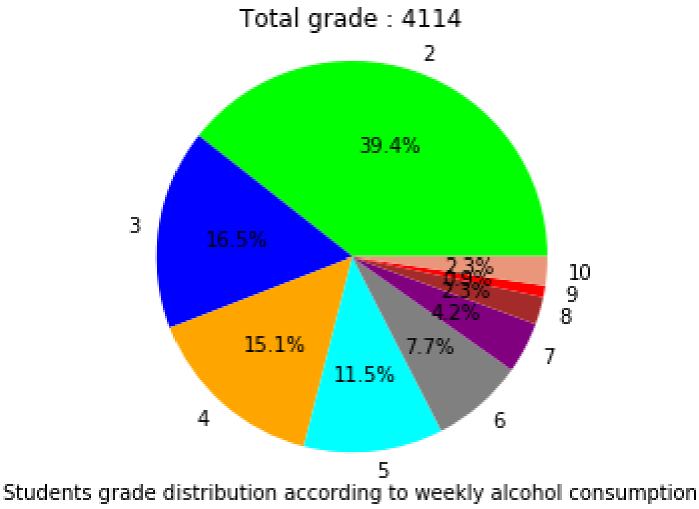


Fig 5 [1]

Conclusion and Future Work

In conclusion I feel like even though K Nearest Neighbour had the highest instance of accuracy score, Bernoulli Naive Bayes was a lot more consistent with 83% so I have to declare it as the best algorithm used in this study. K-NN is definitely more basic in the fact that almost anyone can understand what it is doing immediately. Even if you just told them without any graphs or example I believe the average person could easily wrap their head around the concept of K-NN. Bernoulli's Naive Bayes is also simple. However, both algorithms are very effective in their own right.

I will use the knowledge I gained in this study and apply it to my Final Year project. My FYP heavily incorporates machine learning. I now have tool at my disposal I did not know existed such as the SciKit Learn library and MathPlot. Both incredibly useful libraries that I will definitely be using.

References

1. https://www.kaggle.com/kanncaa1/does-alcohol-affect-success
2. Basic EDA and Final Grade Prediction
3. Effect of Alcohol Use on GPA