Gabriel Solomon Holland

Data Science – Conclusions

**Initial Question**: Do students who drink more and miss classes tend to fail more classes or get lower grades?

**Immediate correlations:**

This project has shocked me in many ways because the reality I’m seeing is that it is incredibly difficult to predict success because of the sheer number of factors. Simply looking at correlations – there are almost none. Alcohol consumption, which is a survey based scale 1-5, had almost no correlation with how many absences there were.

Chart, bar chart

Description automatically generated

I see a trend across dozens of charts showing that there are no correlations amongst two features which at initial glance completely disproved my hypothesis. This lack of correlations was supported by my correlation matrix, nearly every feature had a sub 20% correlation.

**Advanced Conclusions**

Since two features didn’t give any results, I went to machine learning. Machine learning started showing some significantly better answers. My target was number of failures since the target was only a range of 0-15 (roughly) as compared to the trimester grades which was 0-100. Using that target I made several sets of features to test with.

* Set x = Weekday alcohol consumption and number of absences
* Set w = age, mother’s education, father’s education, studytime, and number of times they went out
* Set v = all the features above plus weekend alcohol consumption and how much freetime they had.

Set x and set v were the ones that I was really focused on. The most interesting result was that across all types of learning, across both x and v, machines could only get 85.38% exactly. There was frequently no difference in accuracy between x and v, it seemed almost like the machine was ignoring all of the extra features that v offered. Based on two features predicting one y rather than one feature one y, we were able to jump from almost nothing to an 85% accurate prediction. 85.38 was the exact number for:

* Support vector machine, set v and x, for linear, RBF, and sigmoid kernels.
* Forest ensemble, set v and x, using soft voting, logreg, svm, and random forest machines.
* Mulitilayer perceptron, set v, set x, AND set w.

Other close to 85% results were:

* Decision tree classifier
  + Set x 83.08% accuracy
  + Set w 84.62% accuracy
  + Set v 82.31% accuracy
* Forest Ensemble
  + Set x,w,v with 82.31% accuracy
  + Used:
    - Soft voting
    - Logreg
    - Svm
    - Random forest with 15 predictors
  + Yes the same forest ensemble got 85.38% on other runs

So we see two numbers appear multiple times: 85.38% and 82.31% which both appear multiple times across multiple runs which likely means that they are important and might be the actual correlations between these features, the problem is it is hard to see how they’re deciding this. So I made a decision tree visualizations. In this process I edited the decision tree classifier depths to get the best accuracy. The above accuracies are within 2% of the edited ones but have a uniform 6 depth. The updated ones have varying depths but the accuracies are comparable. These are HUGE pictures so you can zoom in and if Windows permits, it will come into focus. I will also attach some of these to the assignment itself. The color is based on the gini impurity, red is highest, then orange, yellow/beige, green, blue, purple

Set V: A screenshot of a computer

Description automatically generated with medium confidence

Set W: A screenshot of a computer screen

Description automatically generated with medium confidence

Set X:

A screenshot of a computer

Description automatically generated with medium confidence