Jinwoo Lee, Kunal Jangam, Sean Park

**Abstract**

This project analyzes data about students from two Portuguese schools. The data contains information about the academic and private lives of each student and about the amount of alcohol the student ingests.

Our goal for this project is to determine if there exists a significant impact alcohol consumption has on the performance of each student, and if there is such an impact, we want to determine how significant the impact is. We also want to determine what factors, if any, contribute to high alcohol consumption among students.

**Introduction**

Students around the world often drink alcohol, despite age restrictions. There may be multiple reasons for student alcohol consumption, such as spending a lot of time going out, or for stress relief. Our purpose for this project is to determine the impact alcohol consumption has on students and to determine what causes high alcohol consumption among students.

The data we are using contains the Math scores for students in 2 Portuguese Schools, along with data about the students and their lifestyles. Specifically, the data contains information about:

* Which school the student attends
* Sex of the student
* Age of the student - legal drinking age in Portugal is 18
* Home address of the student
* If the parents of the student live together or apart
* Education level of the student’s mother (0-4)
* Education level of the student’s father (0-4)
* Job category of the student’s mother
* Job category of the student’s father
* Reason the student chose to attend the school
* Legal guardian of the student
* How long it takes for the student to arrive at school
* How many hours the student spends studying weekly (0-4)
* Number of times the student has failed the class in the past
* If the student gets extra educational support from the school
* If the student gets extra educational support from their family
* If the student pays for extra classes
* If the student participates in extracurricular activities
* If the student attended nursery school
* If the student wants to take higher education
* If the student has internet access at home
* If the student is in a romantic relationship
* The quality of the student’s family relationship (1-5)
* How much free time the student has after school (1-5)
* How often the student goes out with friends (1-5)
* The student’s workday alcohol consumption (1-5)
* The student’s weekend alcohol consumption (1-5)
* The student’s health status (1-5)
* The number of times the student has been absent (1-5)
* The grades of the student (0-20) - 10 is considered a passing grade

We also approximated the class year of each student with age - number of failures - 6.

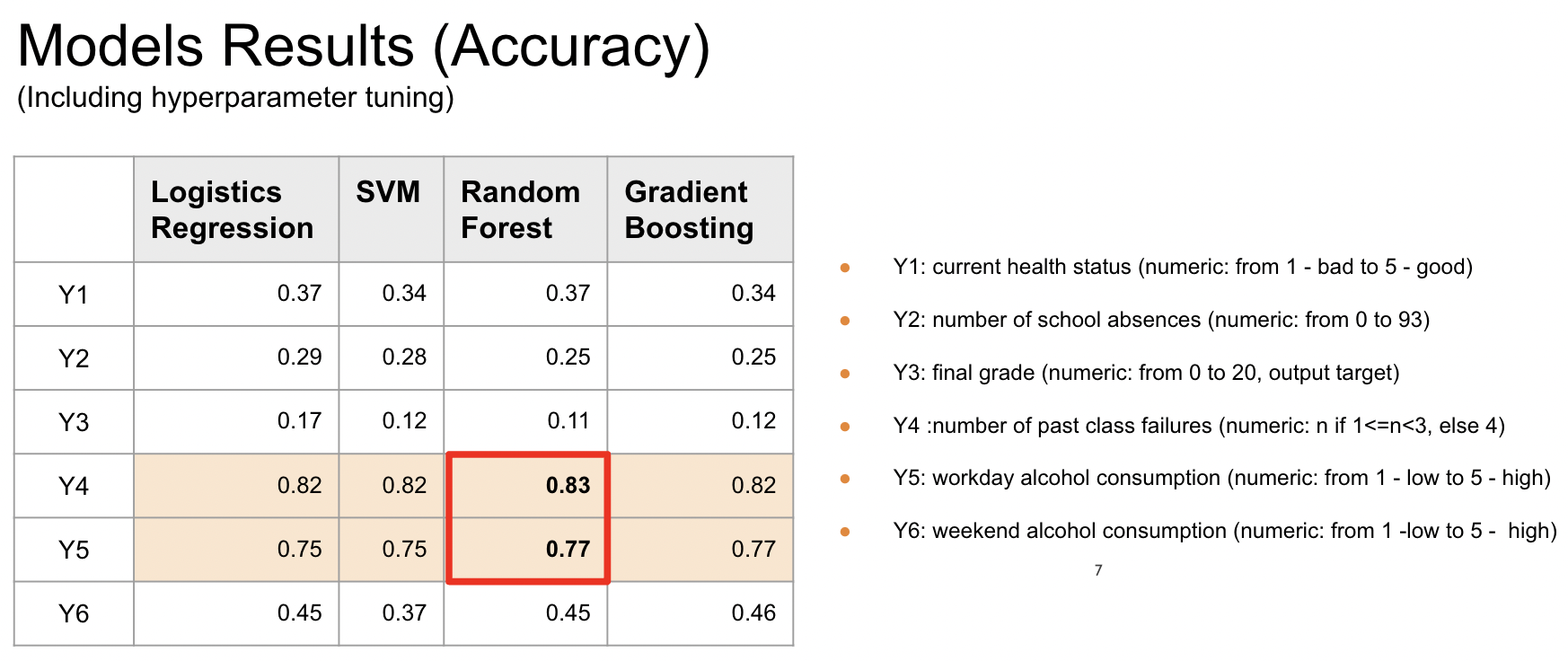
The experiments we will perform fall into two categories: using the data to predict the alcohol consumption of the student, or using the data to predict the performance of the student. For metrics of alcohol consumption, we use workday alcohol consumption and weekend alcohol consumption. For metrics of student performance, we use their health status, number of absences, number of failures, and grade. For predictors of student performance, we use which school they attend, their age, their family size, if their parents live together, their parent’s education levels, how much time they spend studying, if they get extra help from the school, their family, or from a paid source, if they want to pursue higher education, if they have internet at home, how good their family’s relationship is, their workday alcohol consumption, and their weekend alcohol consumption. For predictors of alcohol consumption, we use the student’s sex, age, class year, family size, if their parents live together, if they have internet at home, if they are in a romantic relationship, their family’s relationship, how much free time they have, and how often they go out.

**Methodologies**

We tried different machine-learning algorithms. We experimented with Support Vector Machine, Logistic Regression, Random Forest, and Gradient Boosting. Some of them worked and others didn't. One of the interesting things that we found out during this experiment is that most of the poor results were due to the overfitting problem. We found out that the training error is significantly higher than the testing error. After the interim report, we worked on the following things. We tried different Machine Learning methods including logistic regression and also conducted hyperparameter tuning as we have done the analysis without tuning previously. Although the tuning did increase the accuracy a bit it didn’t increase it significantly.

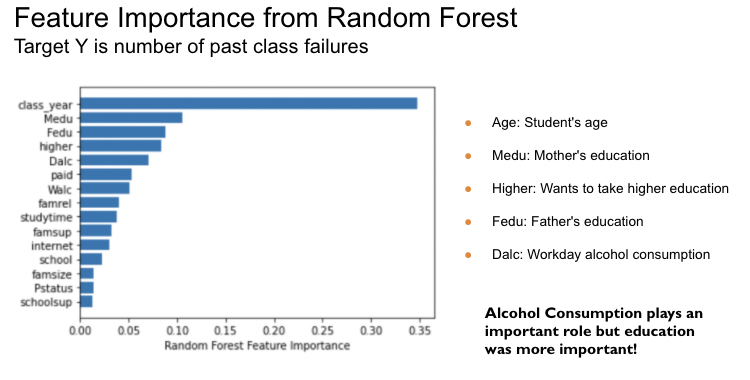
**Results**

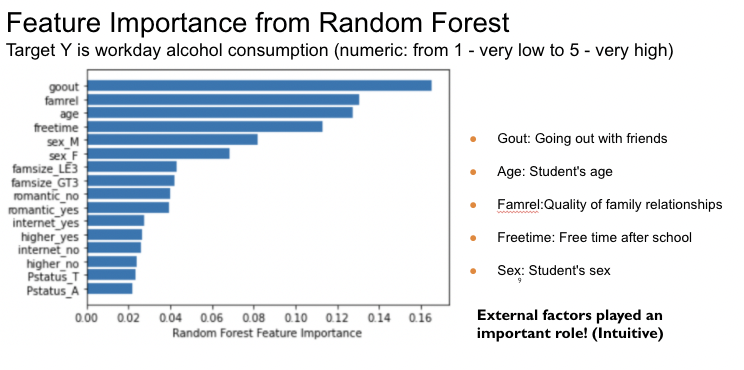
After running the hyperparameters, we got the following results.



Although we couldn’t get all high accuracy for target variables, we did have significant high accuracy results when predicting the number of past class failures and workday alcohol consumption. Also, among the methods, the Random Forest model showed the best performance. Our team thinks this result is because of the overfitting problem. When we were experimenting with the models, we observed low training errors but high testing errors. The Random Forest model solved the overfitting problem.

Furthermore, we derived the feature importance from the Random Forest model when predicting the number of past class failures and workday alcohol consumption.





**Conclusion**

For our first question, we wanted to know how alcohol consumption affects students. After extracting the feature importance based on our highly accurate model, although alcohol consumption was an important factor; its rank was 5th from the feature importance. For our second question, we wanted to know what factors, if any, cause high alcohol consumption for students. From the feature importance from our model, the most important factors that impact workday alcohol consumption are time spent going out with friends, age, and quality of family relationships. If a student goes out often with their friends and if they do not have a good relationship with their family, it’s intuitive that they're more likely to consume alcohol during the workday.