**Final Project Proposal-**

Run Linear Progression models on the student data set we downloaded from Kaggle. The data set breaks down students in 2 classes, a math and a Portuguese language class. Each student’s family demographics, such as family size, parents’ education levels & jobs, home location (urban or rural), as well as other outside inputs such as support from family and school, internet access and whether or not the student attended nursery school. We will train and test the models to determine its accuracy as well as remove a selected data category from the dataset (i.e. internet access, family support, extra-curricular activities) to see if accuracy results change with each new run. The students were binned into categories of likelihood to drink on the weekends (not at all, slightly, moderately, very and extremely), which is what will be modeled.

**Possible Machine Learning Models**

* Logistical Regression
* Support Vector Machines
* Neural Network w/ Keras

**Data sets to adjust (check for accuracy differences)**

* Internet access
* School support
* Extra-curricular activities
* Multiple variations

**Run Predictions**

Create a mock-up of fictional child to run through model to check for probability of child’s likelihood to drink.

**Show data**

We will show test prediction accuracy results, as well as if there were any significant differences in the accuracy levels on the datasets that remove certain categories. These visualizations will be shown in Tableau.

Final presentation will be compiled into either a Power Point presentation or a web page for easier viewing.

<https://prabalb.github.io/prabalb.github.io-alcohol.prediction/>

**DATASET DESCRIPTION**

Secondary school student alcohol consumption data with social, gender and study information.

**SUMMARY**

The data were obtained in a survey of students’ math and portuguese language courses in secondary school. It contains a lot of interesting social, gender and study information about students. You can use it for some EDA or try to predict students’ final grade.

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

* school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
* sex - student's sex (binary: 'F' - female or 'M' - male)
* age - student's age (numeric: from 15 to 22)
* address - student's home address type (binary: 'U' - urban or 'R' - rural)
* famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
* Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
* Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
* Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
* Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')
* Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')
* reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
* guardian - student's guardian (nominal: 'mother', 'father' or 'other')
* traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
* studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
* failures - number of past class failures (numeric: n if 1<=n<3, else 4)
* schoolsup - extra educational support (binary: yes or no)
* famsup - family educational support (binary: yes or no)
* paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
* activities - extra-curricular activities (binary: yes or no)
* nursery - attended nursery school (binary: yes or no)
* higher - wants to take higher education (binary: yes or no)
* internet - Internet access at home (binary: yes or no)
* romantic - with a romantic relationship (binary: yes or no)
* famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
* freetime - free time after school (numeric: from 1 - very low to 5 - very high)
* goout - going out with friends (numeric: from 1 - very low to 5 - very high)
* Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
* Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
* health - current health status (numeric: from 1 - very bad to 5 - very good)
* absences - number of school absences (numeric: from 0 to 93)

Additional note: there are several (382) students that belong to both datasets . These students can be identified by searching for identical attributes that characterize each student, as shown in the annexed R file.

Source: Kaggle

<https://www.kaggle.com/uciml/student-alcohol-consumption>

Show less