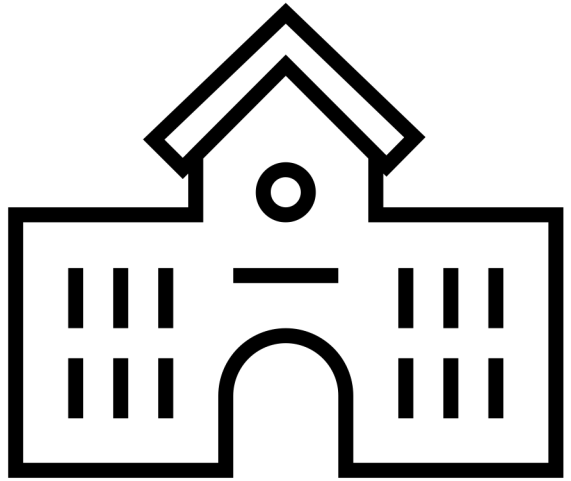
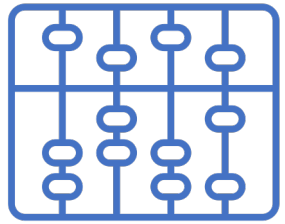
Ebba Cramér

Portuguese students

- ~400 Portuguese students
- Grade 0 – 20, Math
- From a self evaluation the students have gotten a score on a number of features, such as
  - how often they drink alcohol
  - how educated their parents are
  - and whether they want to pursue higher education
- Students divided into three groups based on grade (low, middle, high)



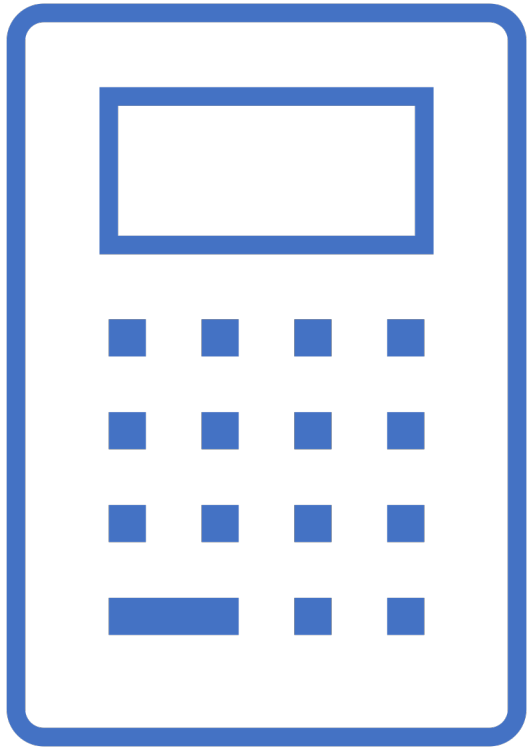
What I wanted to do was try to predict whether a student would get a low grade, a high grade or whether they'd end up in the middle segment based on the students' self evaluation score on the different features



- So, the task is to try to predict which group the students will end up in, and thus if they'll perform well, average or poorly

- If poor performers are caught early measures can be taken to help them further with their studies – this is the main focus

- If high performers are caught early on they can get help to exceed even further – which would be a nice side effect eventhough catching the poor performers is primary focus

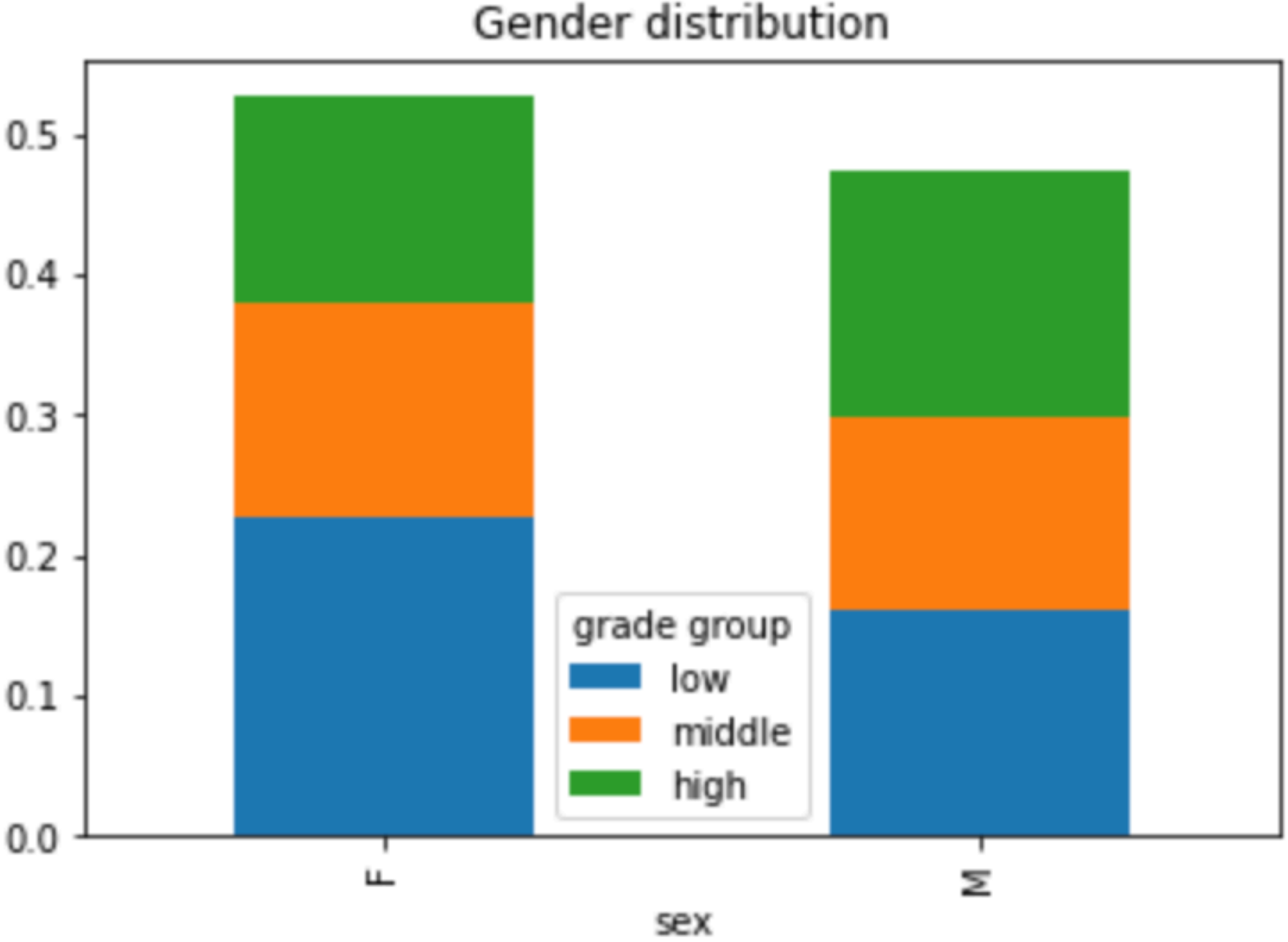


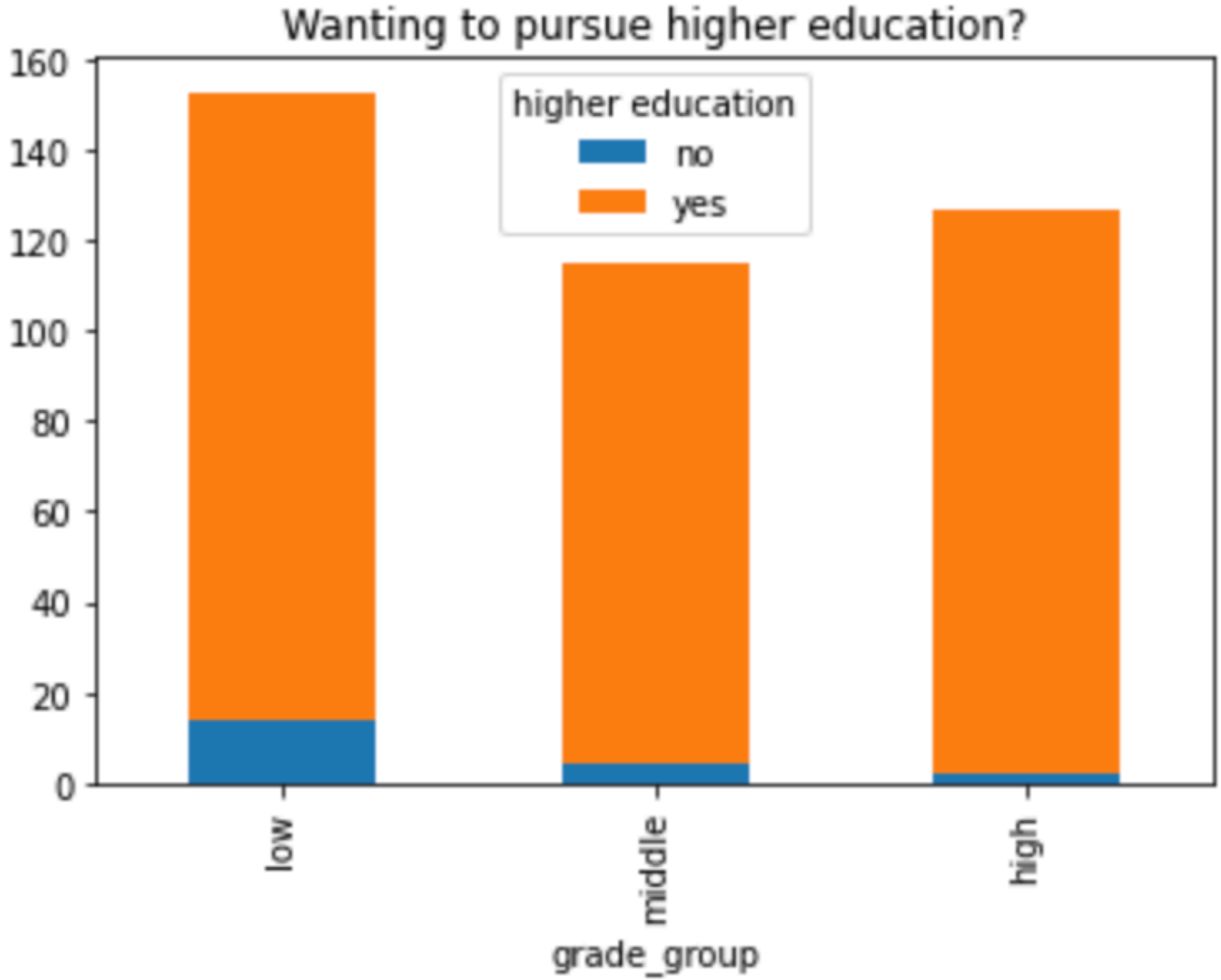
- Since we want to make sure to class as many of the poor performers as possible as poor performers recall, and particularly recall for the low-group, is of particular interest
- Since there is no harm in accidentally calssifying an average or high performer as a poor one precision is of moderate interest

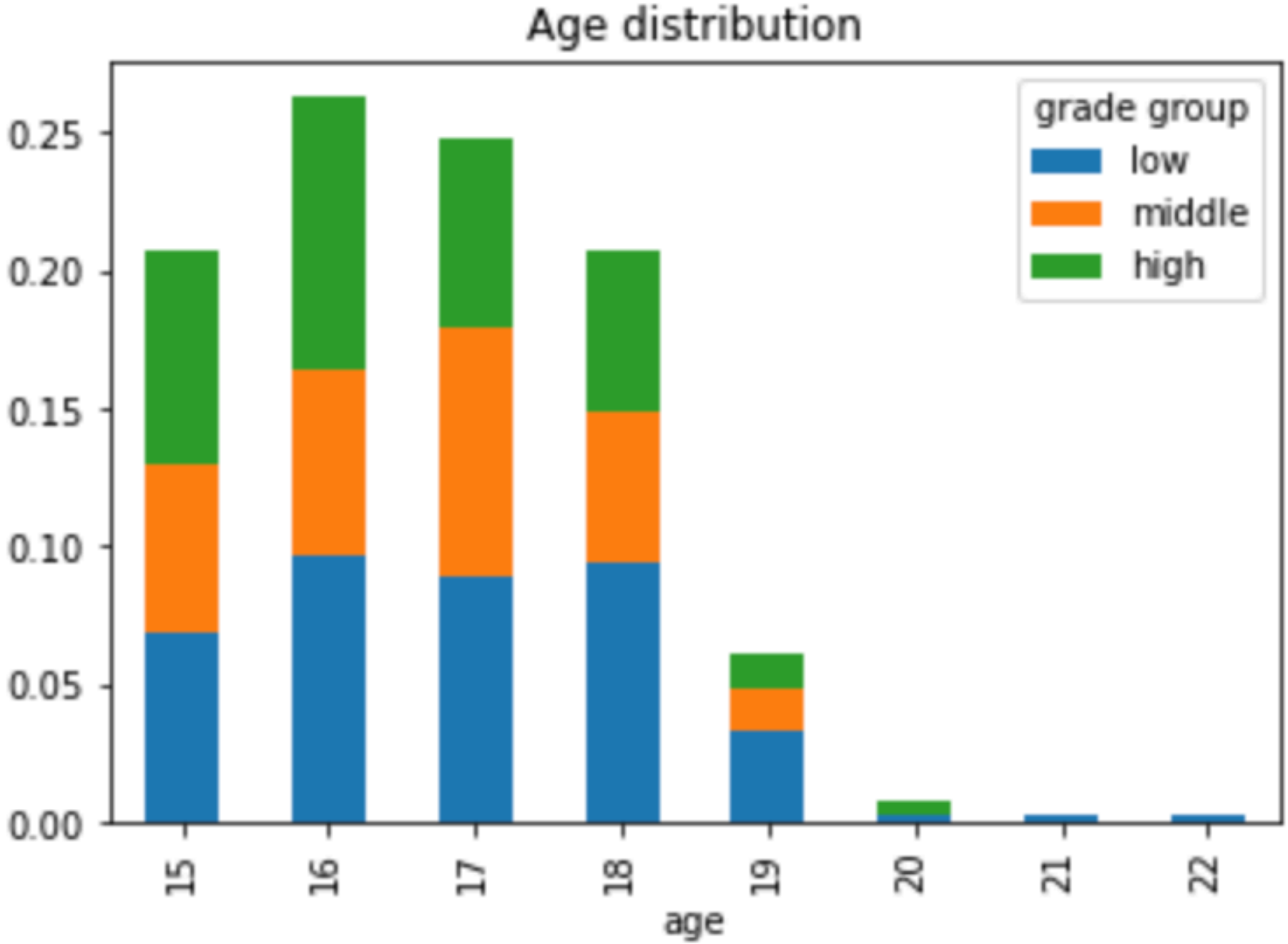
Found a model that could predict the students belonging to the ’low-group’ with a recall of 75% for the validation data and 73% for the test data

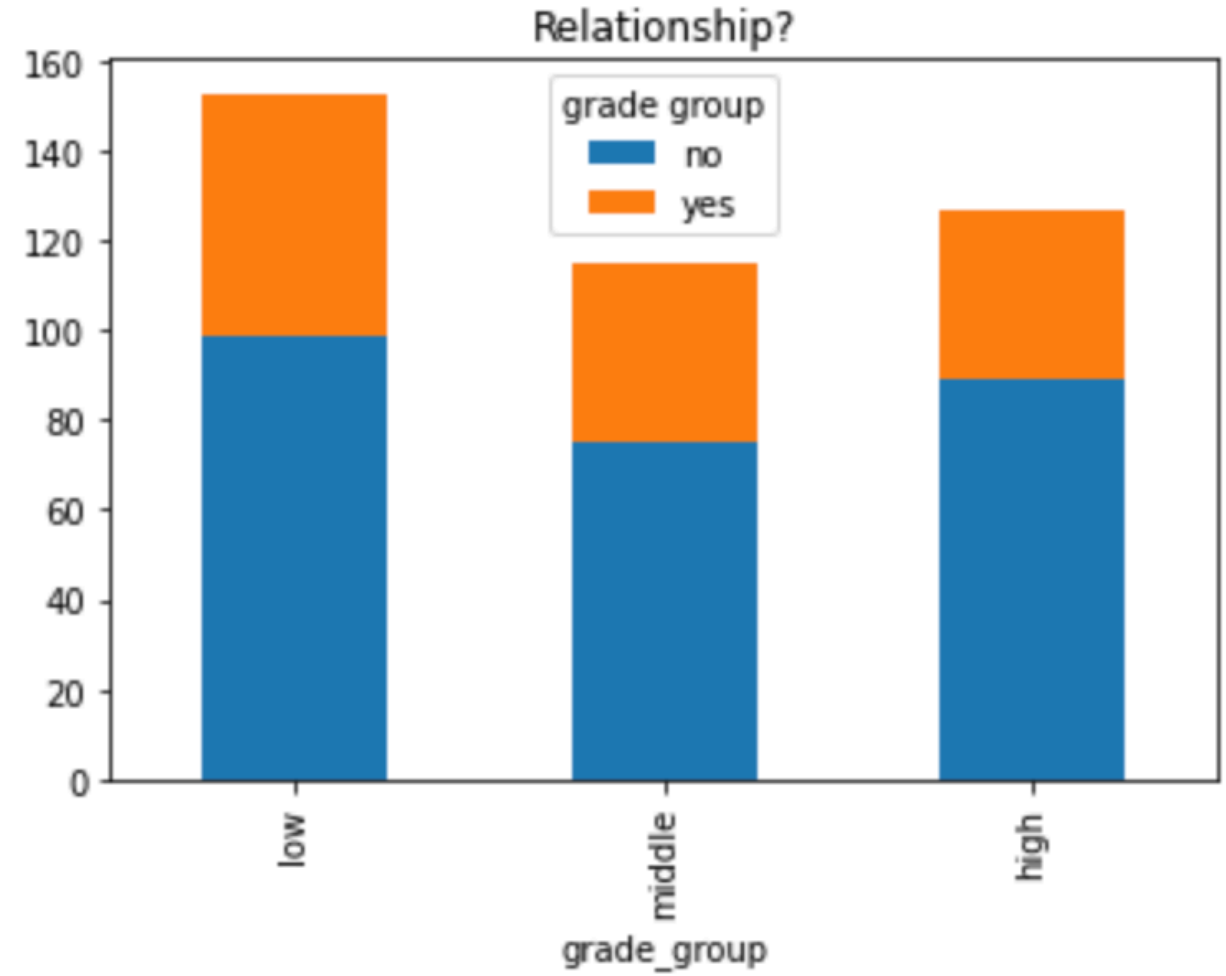
GaussianNB

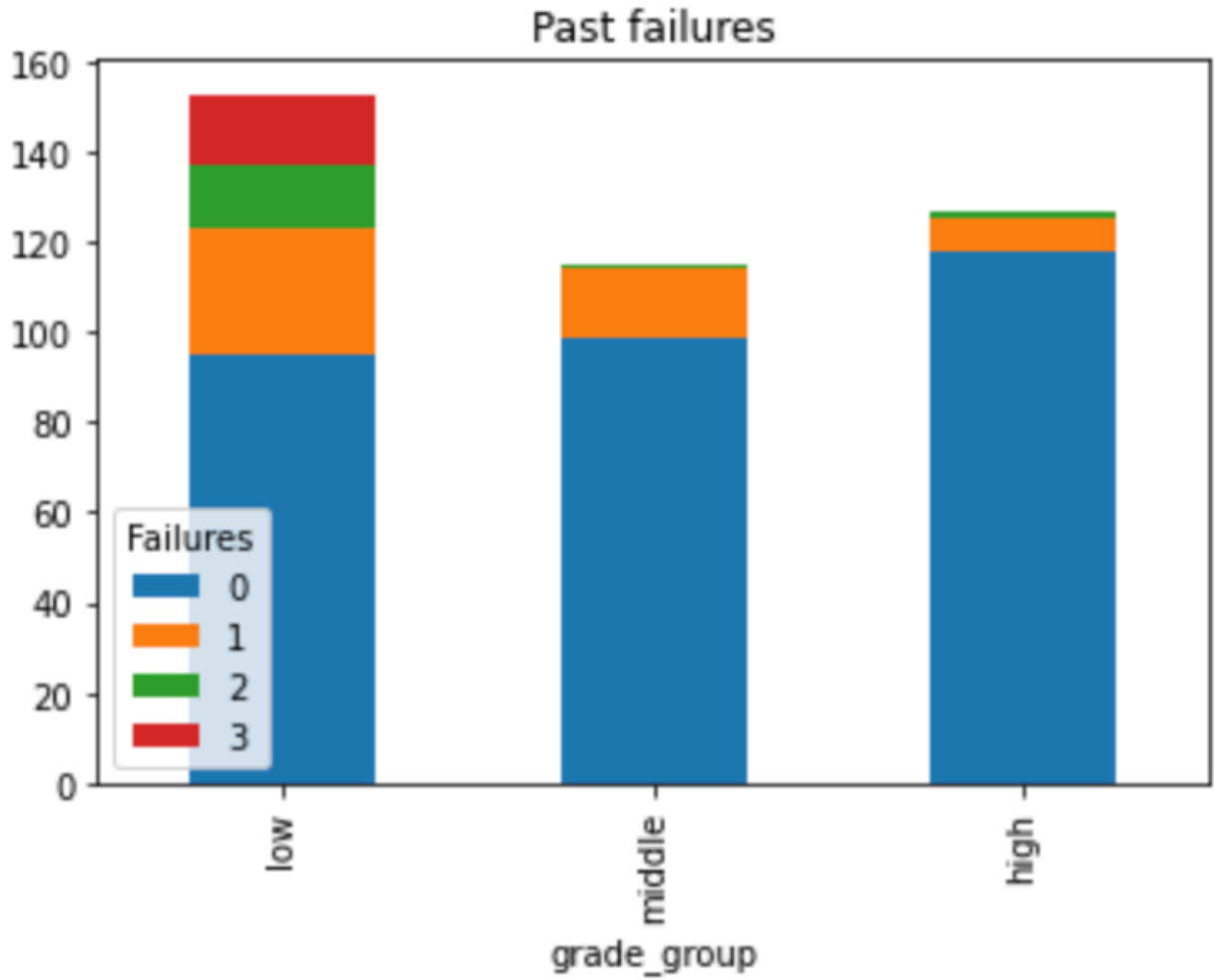
To get a brief idea about characteristics of the groups

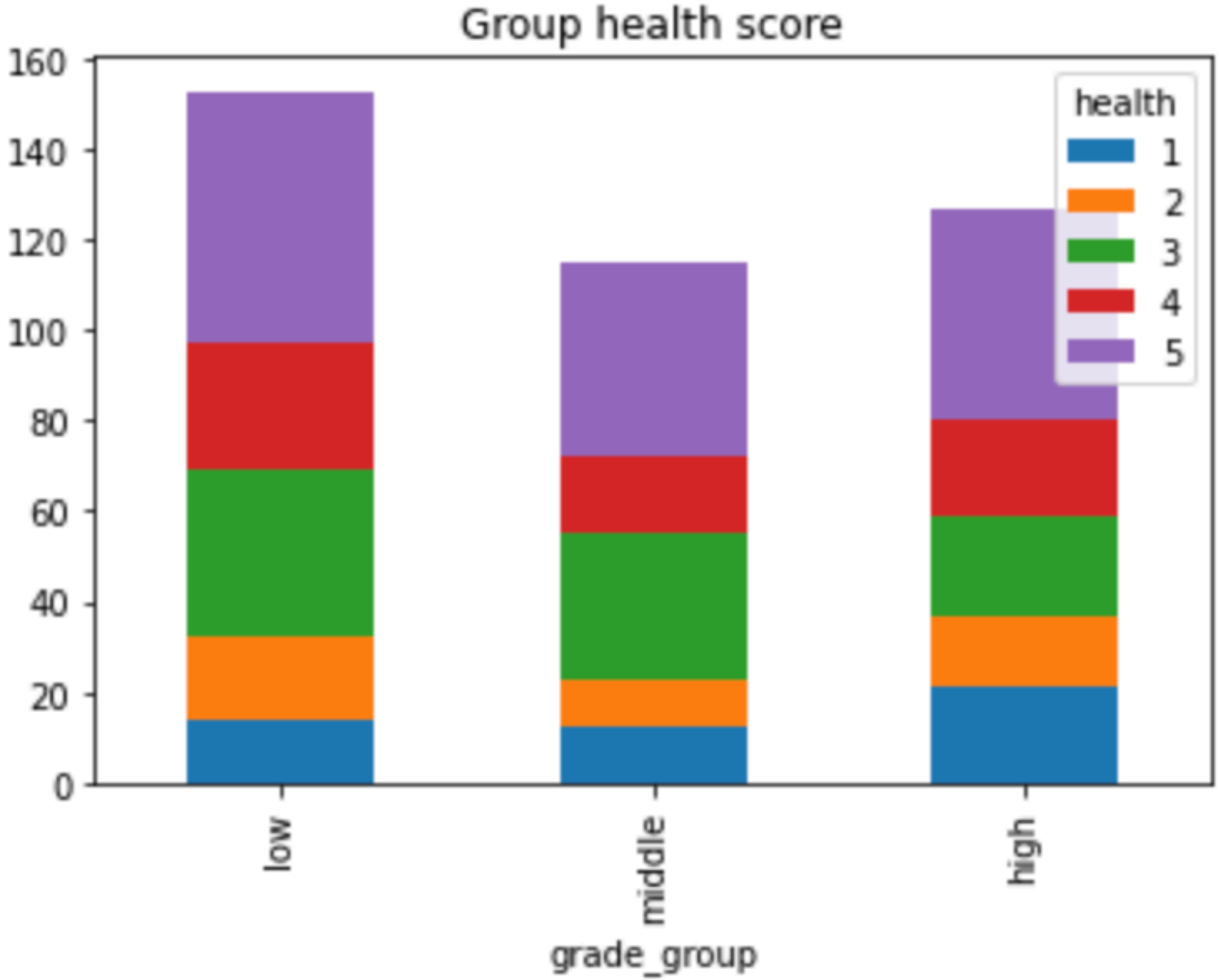

Gender distribution
0.5
0.4
0.3
0.2
0.1
0.0
grade group
low
middle
high
F
M
sex

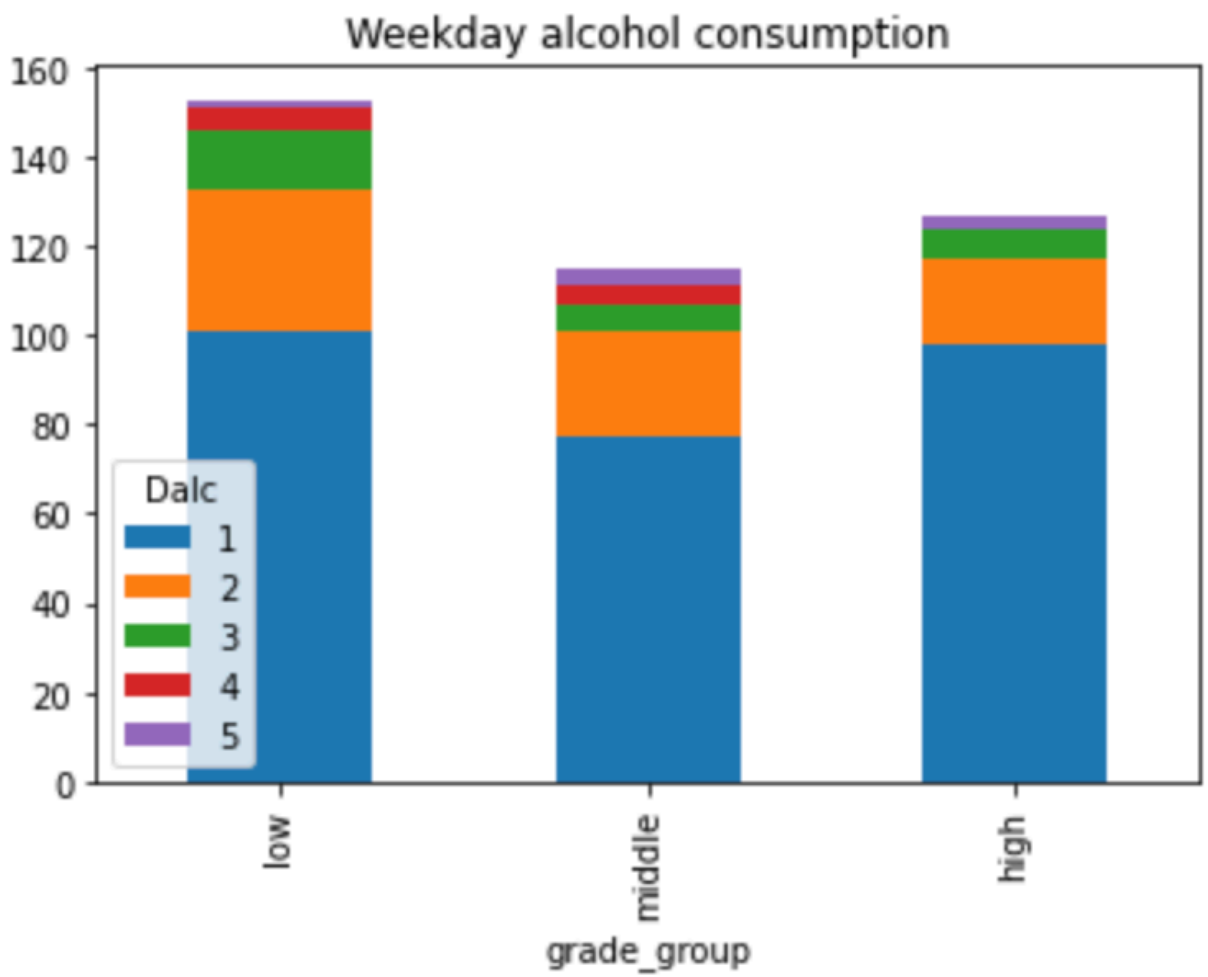

Wanting to pursue higher education?
higher education
no
yes
160
140
120
100
80
60
40
20
0
low
middle
high
grade_group

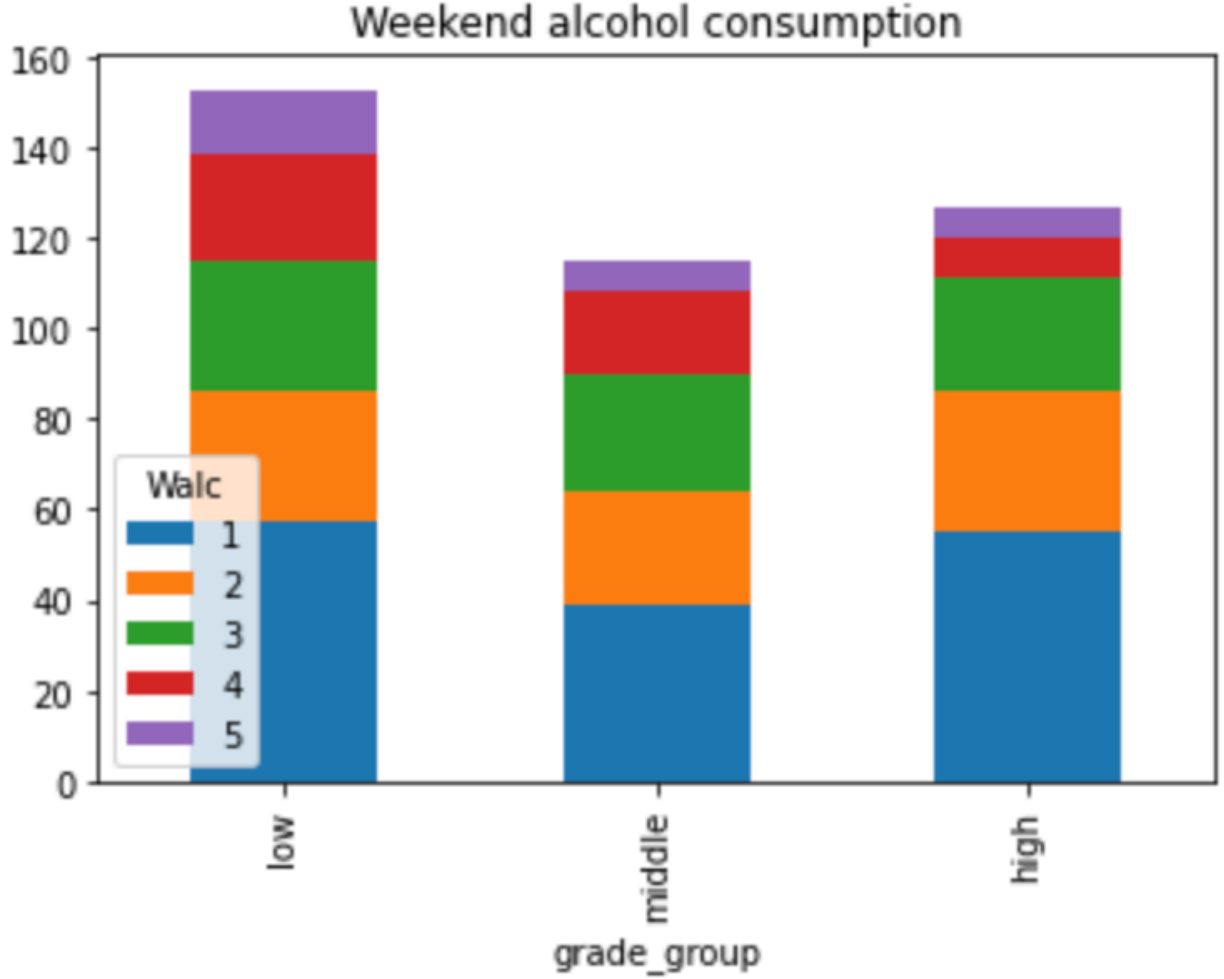

Age distribution
grade group
low
middle
high
0.25
0.20
0.15
0.10
0.05
0.00
15
16
17
18
19
20
21
22
age


Relationship?
grade group
no
yes
160
140
120
100
80
60
40
20
0
low
middle
high
grade_group


Past failures
Failures
0
1
2
3
160
140
120
100
80
60
40
20
0
low
middle
high
grade_group


Group health score
160
140
120
100
80
60
40
20
0
health
1
2
3
4
5
low
middle
high
grade_group


Weekday alcohol consumption
160
140
120
100
80
60
40
20
0
Dalc
1
2
3
4
5
low
middle
high
grade_group


Weekend alcohol consumption
160
140
120
100
80
60
40
20
0
Walc
1
2
3
4
5
low
middle
high
grade_group

- Model selection
  - With and without OneVsRest multiclass classifier
- Tried out various of the models
- Performed gridsearch and tuning om the best performing ones
- Three top contestants
- OneVsRest GaussianNB, GaussianNB and RandomForest

Narrowed down to two contenders

OneVsRest GaussianNB

- Validation data: recall 80% , precision 50%
- Test data: recall: 70%, precision 52%

GaussianNB

- Validation data: recall 75%, precision 47%
- Test data: recall: 73%, precision 30%

- Tight battle between the two models
- For the 3 percentage points recall we gain with GaussianNB we lose 22 percentage points on precision (compared to onevsrest gaussiannb).
- Since we have stated that recall on the low-group is the most important score the tradeoff might be worth it and hence, the best performing model is

GAUSSIANNB