A short analysis on Portuguese maths students

Danila Kurganov

Introduction

Secondary school students from Portugal are considered based on a variety of features in the unsupervised learning setting. Pairwise plotting is used to discern the distribution of students of each type found in a school, whilst multidimensional scaling is an attempt to cluster different types of student based on features they share. Few successful clusters were determined. On the other hand, some features were seen to impose an easy-to-discern labelling onto the clusters such as parental education levels and student's final grade. These were then lightly analysed and several potential explanations about the students were then introduced.

The Data Set

The data set chosen was 'The Student Performance Data Set' recorded by Paulo Cortez (1) (2). This was donated by Paulo Cortez for public use, and it was downloaded from UCI's Machine Learning Repository. The data represents 395 students from two Portuguese secondary schools monitored against their grade, demographic, social, and school related features. A total of 33 attributes for each student was recorded, Table [1] describes each attribute.

The specific subset chosen to study was that of students who studied and completed (by way of examination) Maths. This was chosen due to my interest in student characteristics with regards to grades, and whether a clustering of them could be done given the recorded attributes. As a maths student I was interested to know what the backgrounds of peers might be. Features chosen were due to my interest in factors that correlate with high student marks, and whether student background makes any discernible effects. Due to the number of features to assess, some had to be missed in the plots provided. Specific criteria for choice of predictors is made below.

Pairwise analysis and Philosophy

The first analysis considered is that of pairwise correlations and density. This emphasizes the distribution of students found over any two chosen variables. The general criteria for choice of variable was that of something that may reinforce or negate my current ideas and theories about students.

Parental education status is a commonly assessed question asked in many application forms to educational institutions, so I find it insightful study this variable. I wanted to also see if this factor had any influence on grade of student, so this factor was also considered. As a fun theory, I believe there may be correlation between the amount each student studies and strict parenting or family environment. Strict parenting can also be extended to student failure rate and going out time. Assuming a non-linear relationship may exist, the possibility to consider clustering is interesting. To add explanation to the above queries, a pairwise plot of 'Medu', 'Fedu', 'famrel', 'studytime', 'goout', 'failures', and 'G3' was made,shown in Figure 1. Outside this writing, a pairwise plot was made over all variables, and further questions were posed. This is outside the scope of this paper, however.

Before coming up with conclusion based off observations of pairwise plots, it's important to consider the implicit assumptions this visual tool carries. Mousinho da Silveira (MS) school only had data from 46 students, whilst Gabriel Pereira (GP) had 349 sampled. The small amount of data from MS students indicates that the sampled data is more non-representative of the types of students that generally attend MS, or would have attended MS given the world's conditions at the time. Specifically, with regards to interpreting contours of two feature types, MS may not have had at this sampling a rich range of students, and therefore contours will under-represent and under-granulate some number of students and student type. Finally, the data collection process itself could be faulty, especially with study and going out time observed. Unless students clearly record minute by minute these interactions, the validity of records remains unspecific.

A note on the analysis made, it was revealed in preliminary analysis that each school admits a different age range of student, with MS accepting students aged 17+ compared with 14+ students at GP. Comparing features like study time, going out time, or final end-of-year mark seems unreasonable between the two schools unless only the same-aged students were compared. Therefore, only analysis of students aged 17+ from GP is to be considered, reducing the number of GP students to 163. With pre-processing and assumptions out of the way, inductive claims now follow.

There is a weak correlation between mother's education level and decreased family relations at both schools, however this is not so strongly seen when it comes to the father's education level. I wonder if this is due to societal perceptions for what is considered the 'norm', or if the reasons are different. At the same time, study time and end of year marks are correlated with the education level of both parents. Oddly, there is a correlation between family relations and class failures. Perhaps the more friendly family environment didn't punish failures as much as the less friendly one did - parents seeing them as not so important for the development of their child. An increased study time saw a decreased failure rate, which is good to hear. Not going out saw a decrease in final grade at the lower end, and a minor decrease at the higher end. Perhaps not leading a balanced life was an expense paid in the long-run, even for those who solely studied without going out. Of those that failed, the number of failures was correlated with going out time, seen strongly at the GP school, but of those that didn't no correlation existed. Perhaps there were students knew what they were doing with their time and those that were lost.

Between schools, there are also some differences. By observing contour peaks between the grades, study times, going out times, and number of failures variables, it seems as if GP school has a cluster of students quite separate from the others. In unseen analysis of pairwise plots that don't consider student with final grade < 5, several contour peaks are still visible, indicating that these students come from different backgrounds and aren't totally disconnected with respect to their features from the rest of cohort. So the separate masses seen in the contours of GP are of different students. GP certainly sees more students with a 0 for their final grade, students that study for longer times than those from MS, some with not as good family relations to those of MS, and slightly higher grades than those from MS. MS looks as if they host more lowly educated parents. Apart from these, both schools look to be similar.

Multidimensional scaling

Another way of understanding data with many variables is that of clustering. Several clusterings can be made, such as comparing students with respect to the different label variables available using multidimensional scaling, and further comparing student clouds between MS and GP using Procrustes analysis. Scaling is a distance-based method, therefore the units of distance that each feature takes plays a significant role in the final embedding. The final grade 'G3' takes values between 0 and 20, whilst family relations 'famrel' takes values between 0 and 4. Scaling will therefore consider 'G3' more significantly, even if this isn't a valid assumption. On the other hand, standardising values so that all with have values between 0 and 1 also brings with it the assumption that all features are equally valuable, and in the case of 'G3', that the difference in family relations from 0 to 4 is in some way equivalent to a difference in final grade from 0 to 20, although intuitively this doesn't sound reasonable. Due to a lack of understanding of the significance that each variable plays, standardisation of values won't be done, and a sacrifice will have to be made by considering each variable in the strength at which it is recorded. I hope the clustering is not significantly uninformative in this way.

'sklearn.manifold.MDS'was used in Python to produce the clusters. The 'MDS' methods applies gradient descent to best product clusters, applying several starting configurations and choosing the scaling with the lowest stress. 10 starting configurations were chosen, along with 1000 iterations of gradient descent. Euclidean distances are used, with a final embedding dimension of 2.

Using the same data as in the pairwise plots, MDS was considered for each school separately, with plots showing the distribution of different clusters in accordance to a specific label type. 'Fedu', 'Medu', and 'G3' were considered labels of interest, and all others didn't produce coherent label clusters or noticeable patterns. Plots can be seen in figures 2-7.

Figures 2 and 3 show classical multidimensional scaling with respect to the father's education level. From both schools it's clear that a distinction of students can be made in some way according to this feature. Different education levels can be seen as layers of education level stacked on top of one another. The order of clustering between MS and GP is also the same, signifying that the father's education level plays a similar role in both schools. In fact, it's visible that students whose Fathers have a low education level are more similar to those that have an intermediary education level vs. high level. The mother's education level told the same story with regards to classifying student and type of student.

Figures 6 and 7 show that the grades of students also plays a role in the clustering of them, as well as a similar one between schools. A similar clustering between both schools can be made. It looks as if the top scoring students from MS are more different than the intermediary scoring students, whilst this isn't quite the case for GP.

Procrustes analysis between both schools would have been interesting, however due to the mismatch in student population this couldn't be done. An analysis on non-maths subjects could also have been done to see whether trends shared in maths-taking students carry over to those of non-maths taking ones, however, it was also discovered that not every student did both, so this also could not be done.

Conclusions

Apart from speculations, the most meaningful insight about each student and school could be found from the pair plots of Figure 1. These show a generally similar distribution of student backgrounds across schools, and highlight the fine-grained differences of GP over MS, first indicating the GP accepted students from a younger age, and when accounted for this, still had a wider range of student type across specific factors considered. Pair plots also highlighted that parental background had a positive correlation to student final mark. This could point towards the abundance of parental education level questions with regards to school and university student applications. Most labels applied to clustering analysis were not easily understood, however the features of student parental education level and final marks achieved were. Labelling scalings by parental education highlighted that students with more similarly educated parents were more similar to each other. Clustering highlighted that final student marks played a role in separating students - with students with a similar final grade being more similar to each other.

Bibliography

1. Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.
2. P. P. Cortez and A. Silva. *Using Data Mining to Predict Secondary School Student Performance*, page 5–12. EUROSIS-ETI, 2008.

Code

Python version 3.8.5 was used for the statistical analysis in this coursework.
Specifically:
‘sns.pairplot‘ and ‘sns.lmplot‘ from the seaborn library (version 0.11.0) were used for the pairwise and multidimensional scaling plots. ‘sklearn.manifold.MDS‘ from sklearn version 0.23.2 was used for the multidimensional scaling. Details for choice of parameters can be found under the ’Multidimensional scaling’ section.

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)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20, output target)

Table 1. Student Performance Data Set, Labels and Descriptions

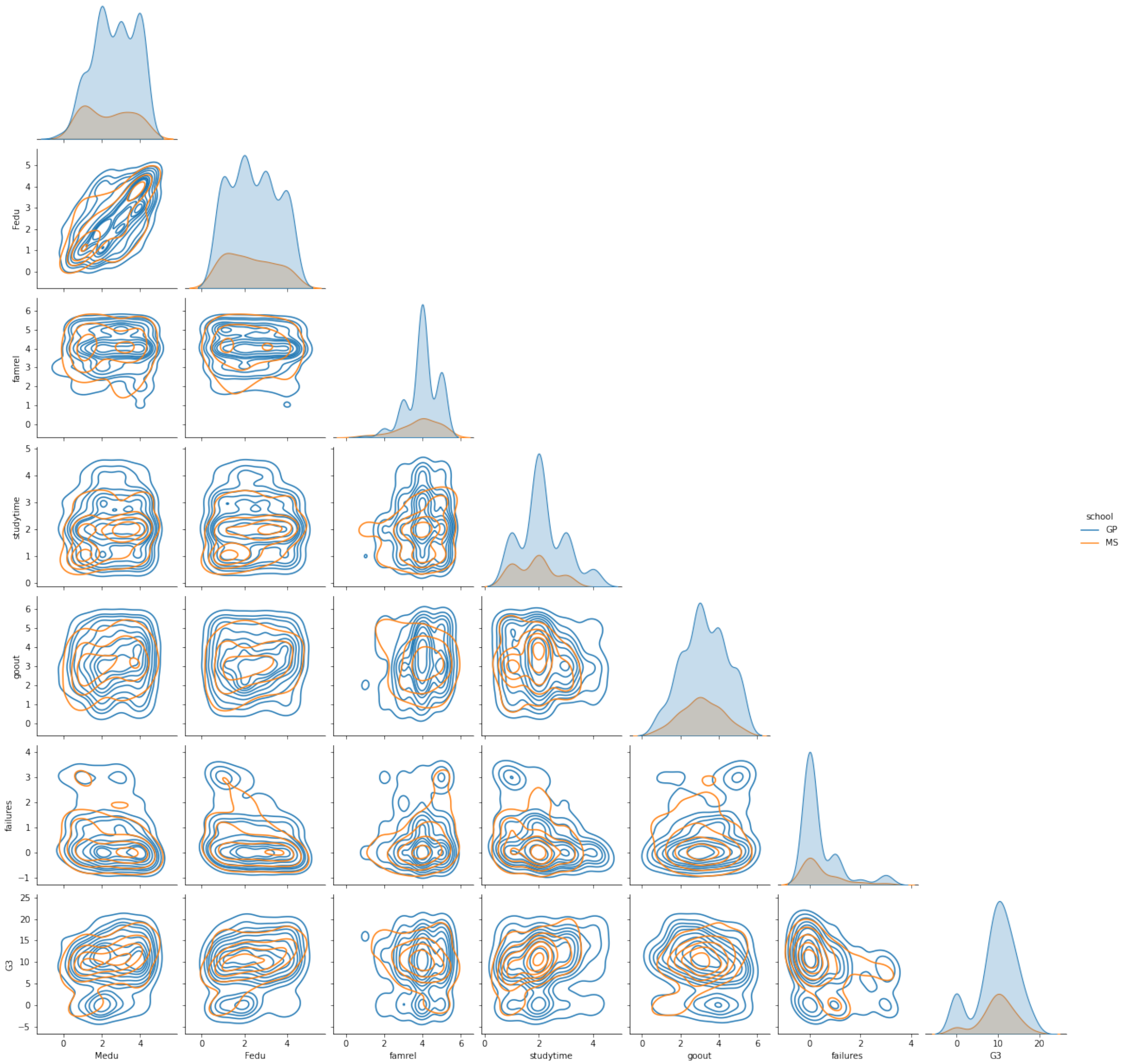

Fig. 1. Pairwise contour and density plots between 'Medu' (Mother's education status), 'Fedu' (Father's education status), 'famrel' (Family relations), 'studytime' (Study time of student), 'goout' (Going out time), 'failures' ('number of failed classes), and 'G3' (final student grade). Blue indicates the statistics of students from the Gabriel Pereira school, whilst Orange from Mousinho da Silveira. For detailed attribute information see 1.

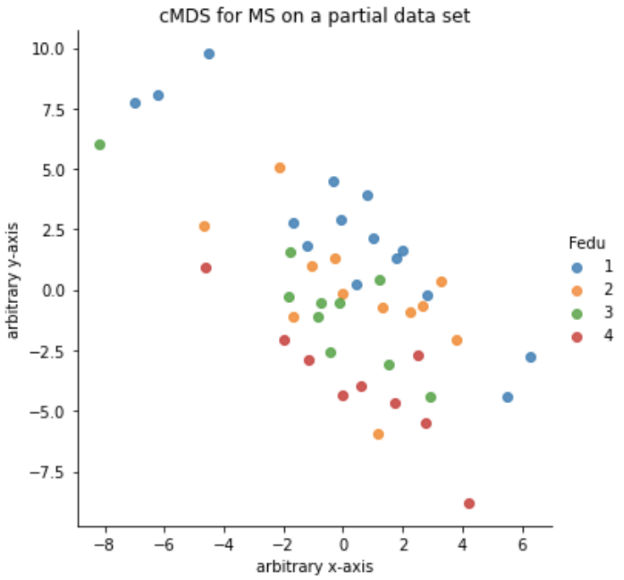

Fig. 2. Classical multidimensional scaling on students from Mousinho da Silveira (MS) using Mother's education level 'Medu', Father's education level 'Fedu', Family relations 'famrel', time spent studying per day 'studytime', going out time 'goout', number of failed modules 'failures', and final end-of-year mark 'G3' as features. Here, a gradient is seen among students according to their father's education level, with similarly father education inferring similar students.

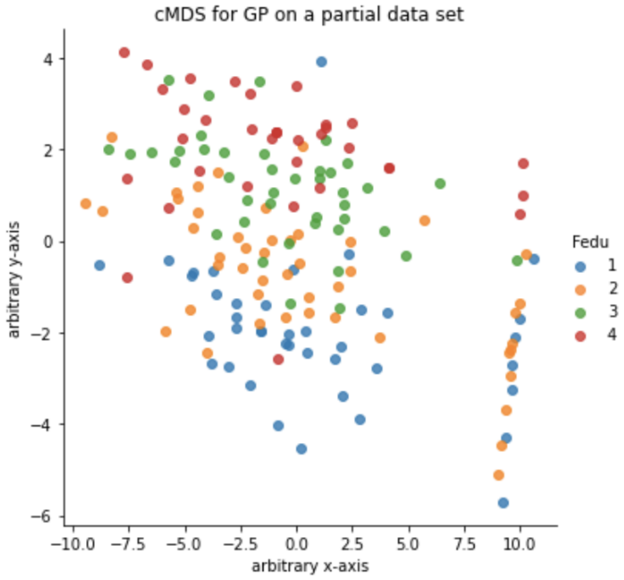

Fig. 3. Classical multidimensional scaling on students from Gabriel Pereira (GP) using Mother's education level 'Medu', Father's education level 'Fedu', Family relations 'famrel', time spent studying per day 'studytime', going out time 'goout', number of failed modules 'failures', and final end-of-year mark 'G3' as features. Here, a gradient is seen among students according to their father's education level, with similarly father education inferring similar students.

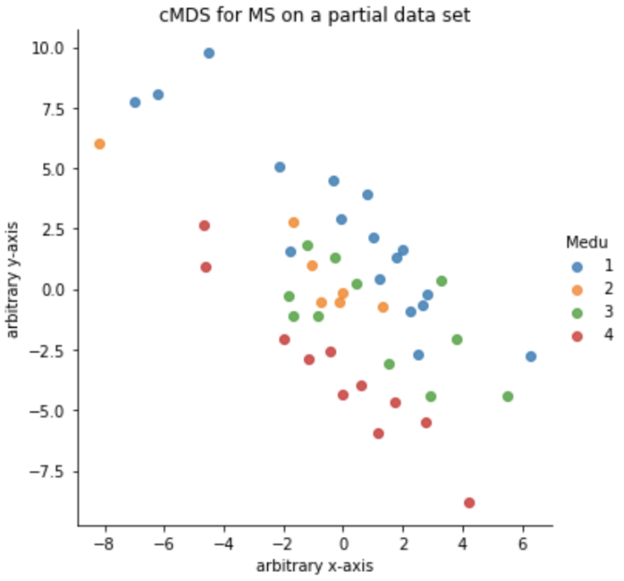

Fig. 4. Classical multidimensional scaling on students from Mousinho da Silveira (MS) using Mother's education level 'Medu', Father's education level 'Fedu', Family relations 'famrel', time spent studying per day 'studytime', going out time 'goout', number of failed modules 'failures', and final end-of-year mark 'G3' as features. Here, a gradient is seen among students according to their mother's education level, with similarly mother educations inferring similar students. The gradient is quite mixed for mother education levels 2 and 3.

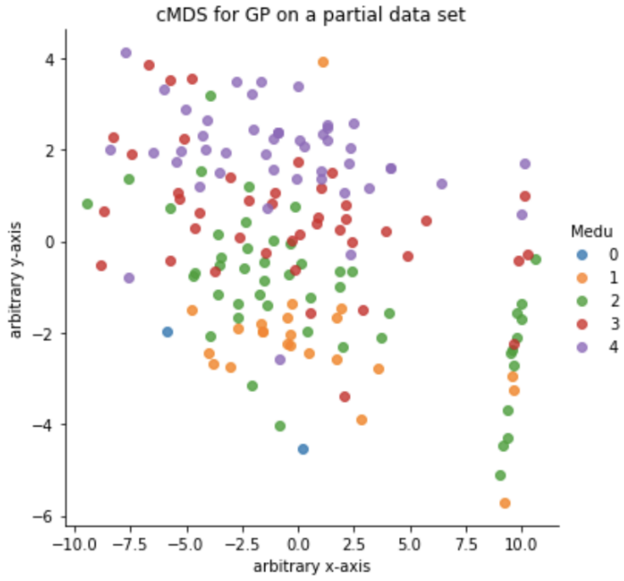

Fig. 5. Classical multidimensional scaling on students from Gabriel Pereira (GP) using Mother's education level 'Medu', Father's education level 'Fedu', Family relations 'famrel', time spent studying per day 'studytime', going out time 'goout', number of failed modules 'failures', and final end-of-year mark 'G3' as features. Here, a gradient is seen among students according to their mother's education level, with similarly mother educations inferring similar students. The gradient is not as clearly defined as in Fig 3.

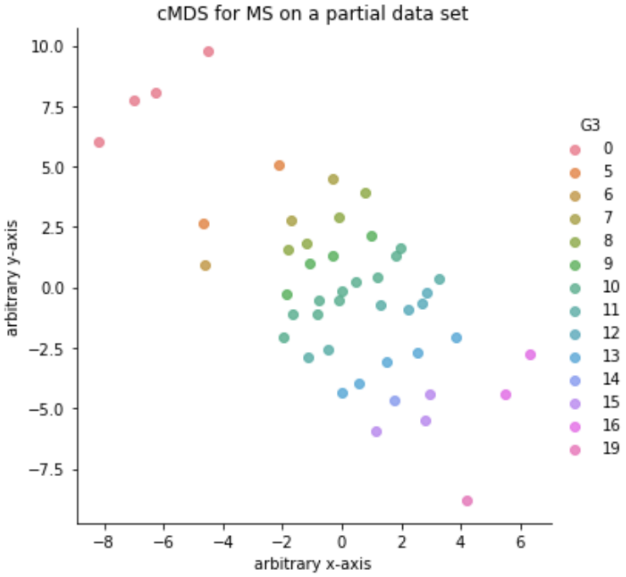

Fig. 6. Classical multidimensional scaling on students from Mousinho da Silveira (MS) using Mother's education level 'Medu', Father's education level 'Fedu', Family relations 'famrel', time spent studying per day 'studytime', going out time 'goout', number of failed modules 'failures', and final end-of-year mark 'G3' as features. Here, a gradient is seen among students according to the student's final end-of-year mark, with similarly performing students being similar in other ways to each other. This distinction is clear and with little error. The failed students and highest marked students are different from the main mass of students in a more significant manner, shown by their distance from this mass. Failed students are similar to each other in background, and successful students are more unique.

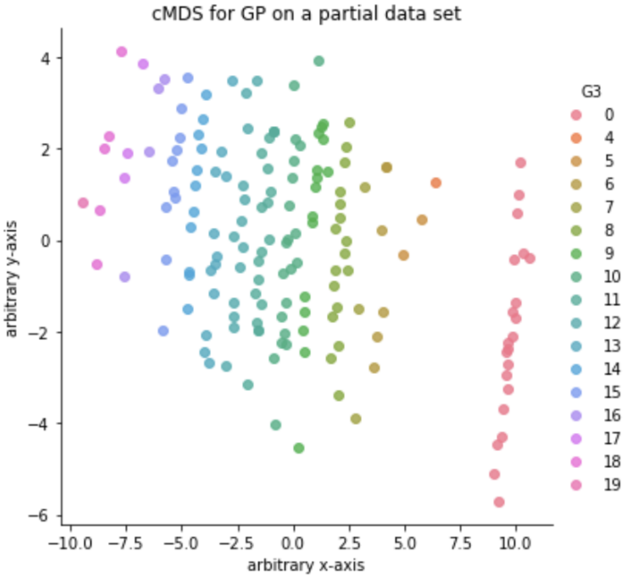

Fig. 7. Classical multidimensional scaling on students from Gabriel Pereira (GP) using Mother's education level 'Medu', Father's education level 'Fedu', Family relations 'famrel', time spent studying per day 'studytime', going out time 'goout', number of failed modules 'failures', and final end-of-year mark 'G3' as features. Here, a gradient is seen among students according to the student's final end-of-year mark, with similarly performing students being similar in other ways to each other. This distinction is clear and with little error. Top performing students are more closely related to to slightly less than top performing ones than to other top students. Perhaps their are friend groups where only one or two students dominate more evidently in their group.