Uncovering Insights for Student Success

Amin Feshkari

University of San Diego

ADS-502-01-FA21 Applied Data Mining

Dr. Ebrahim Tarshizi

December 13, 2021

Introduction

The effectiveness of a pupil's education has been tied to multiple factors including the socioeconomic background of the student, the socio-economic background of their family, or other factors outside of the traditionally assumed factors of time and effort (Duncan & Murname, 2011). As the family income gap between poor and non-poor students widens, school performance also has shown to diverge, favoring non-poor students (Duncan & Murname, 2011). As technology advances, teaching methods have evolved for students and teachers with access to newer pedagogical techniques and tools (Chi et al., 2011). In 2017, the United States spent 3.6% of its GDP on elementary and secondary education (U. S. Department of Education, 2021). Despite these efforts of allocating public resources, teachers and administrators, as well as engaged parents, still struggle to consistently improve student performance for considerable swathes of the population regardless of the neighborhood in which they live (Chetty et al., 2020).

Developments in data mining techniques are now available to provide teachers with insights to better assess patterns in student performance. With these tools, teachers and administrators will be able to target their strategies for improving student performance where there is an opportunity for improvement (Ali, 2013). They can use the patterns they discover in student and school performance to assist in crafting new policies aimed at improving the educational experience (Chi et al., 2011).

This proposed data mining assessment is aimed at identifying characteristics of student experience that may not have been obvious in the past, but now through leveraging the large availability of data and advanced data mining techniques can be used to predict student success.

Project Overview

There were two types of data science models that we leveraged to conduct this statistical study on student performance. The first method was to create linear regression models to predict the performance of a student based on various aspects of a student's background, abilities, skills, and

past performance. We leveraged regression models to identify what are the factors that have a significant impact on student performance and how powerful that impact is.

Additionally, this study also leveraged multiple types of classification algorithms including C5 classification, CART, and neural networks, to predict student performance on a pass-fail basis based on factors that may not have been conducive for regression analysis. The study aimed to leverage the power and insight of regression models and classification models to understand what are the important factors and variables that affect student performance and to test preconceived notions on what impacted student performance. Using two types of models allowed us to leverage different features to take a holistic approach for this study. Different insights were learned from the two types of data mining techniques and uncovered that there were some features that we did not expect to significantly impact student performance that indeed had, and conversely showed other factors that had low or no relationship with student performance than we expected to have otherwise.

Data Set Description

The Student Performance Data Set was provided by the UCI Machine Learning Repository (Cortez and Silva, 2008). The data set describes student achievement in two different secondary education Portuguese schools for the specific subjects of Math and Portuguese. The attributes included in the dataset include information regarding student grades, demographic information, social information, school-related attributes, and were collected by using school reporting and surveys. Two datasets were provided with the same set of features and detailed information about students from the same two schools but differentiated by reporting performance in the two separate subjects of Mathematics and Portuguese. Finally, before any manipulation, the dataset contains 33 attributes. Exploratory analysis of the data found that the data set was clean with no missing values, variables of the wrong type, or variables that did not fit what the attribute was attempting to show. A list of the attributes and their description are provided in the appendix.

Data Mining Steps

For this study, we performed 3 main data mining tasks to explore the data set and the relationships between student performance variables. The 3 sections of data mining include Exploratory Data Analysis (EDA), regression modeling, and classification modeling. Performing these tasks provided insights into the data, and allowed us to measure the relationships of the student performance attributes.

Before creating the data mining models, we conducted an Exploratory Data Analysis. The main objective of the Exploratory Data Analysis was to find if there are any concerns in the validity of the data and to explore the distribution of the dataset. Originally, the data as provided by the UCI Machine Learning Repository was provided in two separate files. One file contained student performance data detailing students that were part of a math class and the other data set provided the same attributes but for students from a Portuguese class. As the emphasis of this study was focused on the overall attributes of student performance, and not necessarily keen to that of one specific subject, we decided to concatenate both the data sets into one source while creating a new variable to distinguish the records coming from the math data set vs the Portuguese data set. After loading the data, we searched for missing values or data records not matching the same type as the rest in the attributes. As no such issues were found, we had confidence in the validity of the data and found that the data was recorded without major errors.

Following the first exploration and manipulation, we continued the EDA process by observing the distribution of the data. Knowing the distribution was important to find where there may be low density areas in the data set specific to certain features, but also to find if we there simply was enough data to explore certain relationships.The team quickly discovered that the data was not evenly balanced as there was a significantly higher index of students from one school than the other school detailed in the set. We also found that this imbalance was evident in other attributes where variables were not completely balanced or followed an usual distribution path. This revealed even before building out any models that it may not be purposeful to explore

relationships or make comparisons between the student populations of the two schools as the imbalance in the data may simply result with less records than needed to draw these insights.

Similarly, we found other features that were not balanced such as the distribution of age, exam scores, time spent studying, and more. Ultimately, we were not concerned about the distribution of these other features but kept it in mind for the interpretation of the results of the model. After our time spent exploring, we partitioned the data into a training and dataset, and we tested to make sure the distribution of the test data would match the distribution of the training data set.

After completing the initial exploratory data analysis, we decided that we had a strong enough understanding of the data to build a linear regression model. The purpose of the linear regression model would be to leverage a subset of attributes to develop a model that could predict the student's performance for their Final Grade. We considered the Final Grade score to be a marker of the student's overall performance. When picking the subset of variables for the model, we were selective in which variables. The reason being that if we included too many variables, we could unintentionally introduce multicollinearity into the model, causing our model to be inaccurate. Lastly, while the direct purpose of the linear regression models would be to predict student performance, we would be evaluating the model to find and understand what features are significant to the regression model and if those features can provide us insight into action that can be taken by schools, families, and policy makers to bolster student performance. After subsetting the data set and selecting our initial collection of attributes, we built the first version of our regression model using our training data set to learn the model. We then leveraged the test data set to find predicted values from the model and calculated the Mean Average Error of the regression by calculating the difference between the predicted values and the actual values of the test data set using R's "mae" function.

The results showed that while our model was more accurate than the Mean Average Error baseline, there were many variables we had initially included which were actually not significant to the model. To find if the model would be more accurate without those models, we created

another regression model but leveraging only the variables that were significant, after testing the initial model with the test data. When testing this 2nd model, we found that the Mean Average Error further decreased, which meant that it was indeed more accurate. We continued this process of refining the model by removing attributes of lower significance levels but found that outside of the initial revision where we removed features from the list of predictors that showed no significant relationships, that removing attributes did not make a difference in the error.

Following the regression model, classification models were created to determine whether a student would pass or fail using various predictors. Classification models used to predict student performance include C50, CART, and neural networks. The purpose of using three different models is to evaluate each model's performance to determine which model is the most efficient for the project's specific dataset. To proceed with the classification models, the dataset was subsetted for only certain attributes where the data was standardized to optimize model performance.

Creating the different classification models showed that each model interpreted the different features in a different way, and gave weights that differed from each type of model. Using the classification technique allowed us to have a different perspective in interpreting the features and showed insights different from the regression analysis.

Results and Discussion

Regression Model

We created the first regression model, model 1, using the features age, traveltime, studytime, failures, famrel, freetime, Dalc, absences, G1, and G2 from the training data set. After leveraging the model to create predictions, we found that multiple features that we expected to be significant were actually insignificant including age, studytime, famrel, freetime, and Dalc. The feature most surprising to see as insignificant was studytime as it showed that the amount of time a student from this data set spent studying did not necessarily have an impact on their future performance. This conflicts with the preconceived notion that many hold that anybody can perform well in school given that they spend a large amount of time studying. The results of this

initial model show otherwise, in that instead what is significant for success is past performance. This is shown by the fact that the features that do show significance include failures, absences, G1, and G2. These 4 features all detail a student's past performance either on exams or assignments, or in performance to be available.

After creating the initial regression model, we created a 2nd model, removing the insignificant variables from the first model. The 2nd model contained variables all of which were statistically significant and had a lower mean average error validating that it performed better than the original model. The equation of the improved model and the MAE regression compared to the MAE baseline would be:

$$G3 = (5.86e\text{-}02)failures + (4.47e\text{-}02)absences + (1.02e\text{-}01)G1 + (8.02e\text{-}01)G2$$

MAE Regression: 0.241

MAE Baseline: 0.686

Classification

The C5.0, CART, and neural network models used the same predictors to evaluate individual model performances for this dataset. The predictors used to determine whether a student would pass or fail are age, travel time, study time, failures, famrel (quality of family relationships), free time, school, family size, Psatus (parent's cohabitation status), activities, famsup (family educational support), higher (interest in higher education), and sex. In addition, the models will also be compared against the all negative baseline model with an accuracy of 36.7%. The dataset yields a baseline accuracy of 36.7%. Therefore, any model with an accuracy above 36.7% would be considered useful.

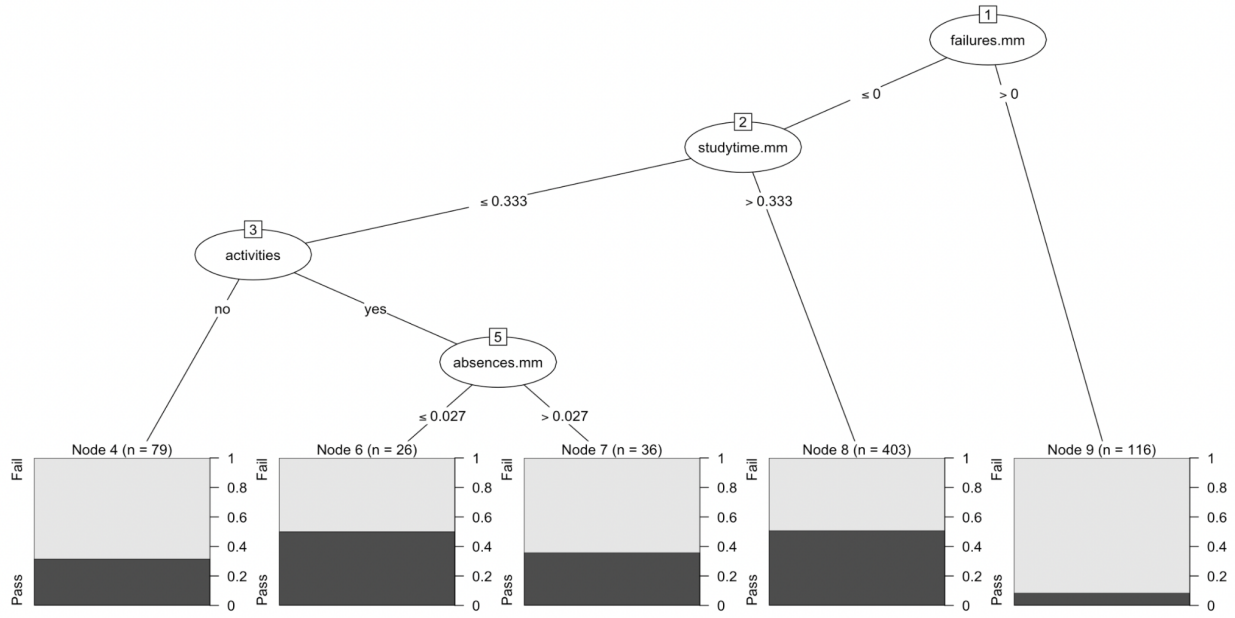
C5.0 Model

A C5.0 model is a supervised machine learning algorithm used to predict a records' class from the values of the other attributes. Executing the C5.0 algorithm resulted in a decision tree with four decision nodes and five leaf nodes as shown in Figure 1. The model begins with a root node for failures where students with more than 1 past failure will immediately terminate in leaf node 9 with a low probability of passing the class. If a failure does not exceed 0, then it will

proceed to Node 2. Node 2's split is on whether or not study time exceeds 0.333. If it does, the branch will terminate in leaf node 8 where there is a roughly close split between passing and failing. If study time does not exceed 0.333, then the tree will split to Node 3. If students respond with no activities, the tree is terminated in node 4 where there is a higher probability of failing than passing. If students respond with yes, then the tree splits into the final node (Node 5). Node 5 terminates the remaining records into Node 6 and Node 7, where having less than the threshold of absences resulted in a higher probability of passing the class.

Figure 1

C5.0 Model to Predict Pass or Fail



Note: Predictors/inputs include age, travel time, study time, failures, famrel, freetime, school, famsize, Pstatus, Activities, Famsup, higher, sex

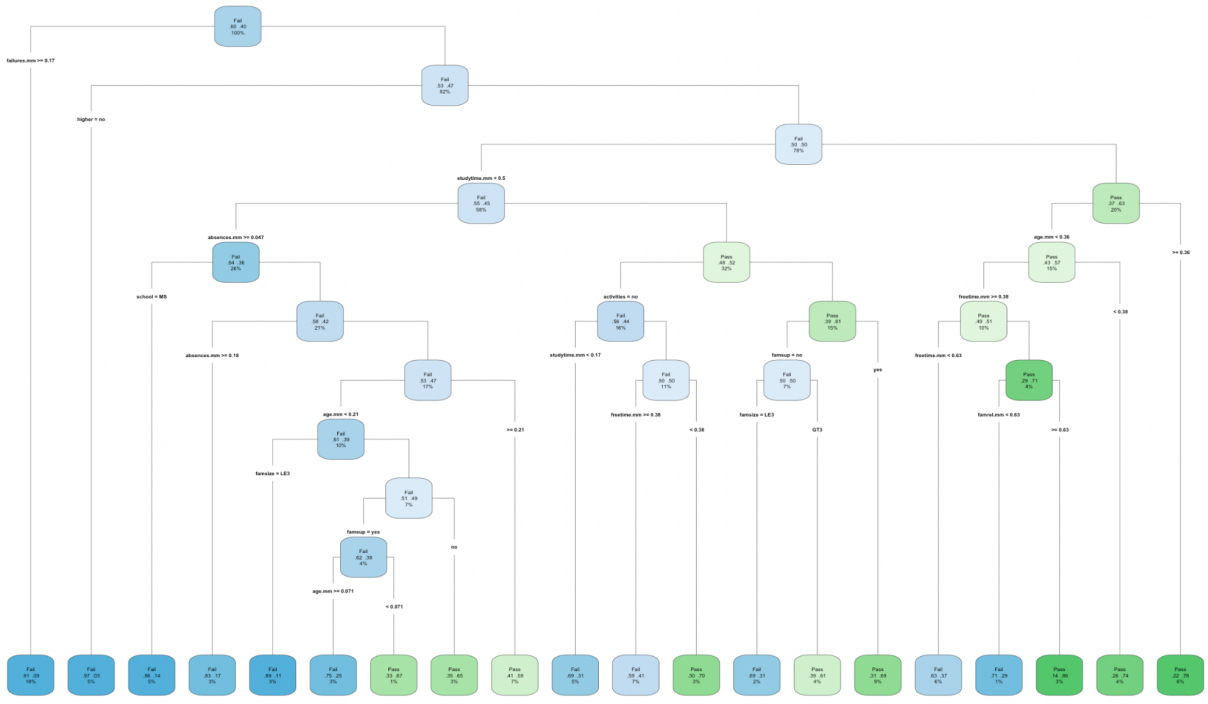
CART Model

The CART model is used to explain how an outcome target's values can be predicted based on other attributes. The method involves selecting an input variable and splitting it where

there are exactly two branches for each decision node. The root node is split into the following decision nodes until it terminates in the leaf nodes. As shown in Figure 2, the CART model has 1 root node, 18 decision nodes, and 20 leaf nodes. Figure 2 illustrates the separate split labels for the left and right directions for all the nodes. In addition, the nodes indicate the probability per class of observations in the node as well as displaying the percentage of observations. Further inspection of Figure 2 will reveal the decision tree pathway indicating the impact of each predictor when predicting whether a student passed or failed.

Figure 2

CART Model to Predict Pass or Fail



Note: Predictors include age, travel time, study time, failures, famrel, freetime, school, famsize, Pstatus, Activities, Famsup, higher, sex

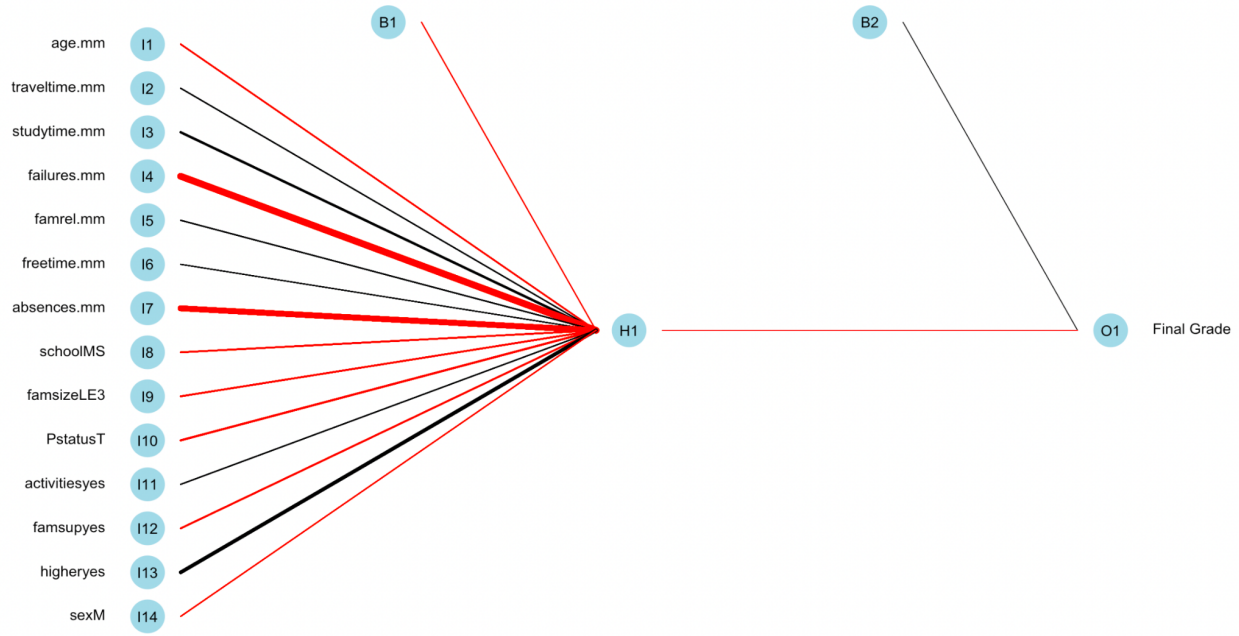
Neural Network Model

Neural network algorithms excel at machine learning models due to their ability to recognize hidden patterns in datasets. Figure 3 illustrates the neural network model used to predict the final grade (i.e, passing grade). Within the figure, there are two constants (B1 and

B2), one hidden layer (H1), fourteen inputs(I1-14), and one output (O1). Moreover, weights in a neural network can give insights into each predictor's significance to the model. With regards to Figure 3, black lines indicate positive connections and red lines indicate negative connections. In addition, the line thickness is proportional to the absolute magnitude of each weight. Therefore, according to the figure below, an interest in higher education and study time have the largest positive weight while failures and absences had the largest negative weights. In other words, the positive weights indicate that having an interest in higher education protects against the probability of failing. On the other hand, the negative weight indicates that an increased number of failures results in an increased probability of failing. This description can also be applied to all other positive and negative weights.

Figure 3

Network Model to Predict Pass or Fail



Note: Predictors/inputs include age, travel time, study time, failures, famrel, freetime, school, famsize, Pstatus, Activities, Famsup, higher, sex

Classification Model Evaluation

Looking at Table 1, all models outperformed the baseline model performance of 36.7% indicating that all classification models were useful in predicting student performance. Furthermore, accuracy represents the overall proportion of correct classifications being made by the model. The CART, C5.0, and neural network model had accuracies of 39.4%, 60.2%, and 64.0% respectively. Comparing the models, the neural network yielded the highest accuracy and lowest error rate.

Additionally, sensitivity measures the ability of the model to classify a record positively, while specificity measures the ability to classify a record negatively. The C5.0 model had the highest sensitivity of 69.1% making it the best model if the study placed higher importance on avoiding false negatives than false positives. On the other hand, the neural network model had the highest specificity of 76.4% by a large margin, making it the best model if false positives were more crucial to avoid.

Table 1

Model Evaluation Metrics

Evaluation Measure	CART	C5.0	Neural Network
Accuracy	39.5%	60.2%	64.0%
Error Rate	60.4%	39.8%	36.0%
Sensitivity	30.8%	75.3%	58.0%
Specificity	54.4%	34.0%	74.6%

Note: All models outperformed the baseline model accuracy of 36.7%.

Conclusion

The results of this initial model show otherwise, that instead what is significant for success is past performance. This is shown by the fact that the features that do show significance include failures, absences, G1, and G2. These 4 features all detail a student's past performance either on exams, assignments, or in performance to be available.

The results of our linear regression model showed that the best predictor of student performance is simply the student's past performance. Features including a student's family background, or the time they spent studying, did not show a significant relationship with the grade of the student on their final exam. These results may suggest that for a student to do well in a class, they cannot rely on their performance on just the final exam or certain assignments but that they should focus on consistent excellence all the way through.

Moreover, A good classification model should have acceptable levels of accuracy, sensitivity, and specificity. Therefore, the neural network model was considered the best model to use when compared to the C5.0 and CART models. In addition, several considerations should be explored to increase model performance. Models with different predictor combinations should be executed to potentially increase model accuracy. Furthermore, a correlation and multicollinearity analysis can be conducted to determine which predictors are most significant when predicting student performance. Additionally, identifying and removing outliers can also potentially increase model performance. Lastly, different classification models can be incorporated to determine if there is a better classification model than the neural network to predict whether a student would pass or fail based on a set of characteristics.

The analysis of student performance for this group of Portuguese students proved that many of the preconceived notions that we had in regards to student performance were not relevant for this group of students. It was found that differences in the family relationships and socioeconomic backgrounds of these students failed to show a significant relationship with their performance and that their performance was directly impacted by only a subset of features related to their past performance in their classes. Nevertheless, we realize that the results of this

study show insight into possible areas for more research, and we would suggest that more research will need to be done with a wider group of students to see if the insights from this study are relevant for all students and can be applied elsewhere.

References

Ali, M. (2013). Role of Data Mining in Education Sector. *International Journal of Computer Science and Mobile Computing, 2*(4), 374 – 383.

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R., & Porter, S. R. (2020). *The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility.* National Bureau of Economic Research. Retrieved December 7, 2021, from https://www.nber.org/system/files/working_papers/w25147/revisions/w25147.rev0.pdf

Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011). Empirically evaluating the application of reinforcement learning to the induction of effective and adoptive pedagogical strategies. *User Modeling and User-Adapted Interaction, 21*(1-2), 137-180. https://doi.org/10.1007/s11257-010-9093-1

Duncan, G., & Murname, R. (Eds.). (2011). *Wither Opportunity? Rising Inequality, Schools, and Children's Life Chances*. Russell Sage Foundation.

Cortez, P., & Silva, A., Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

U. S. Department of Education. (2021, May). Education Expenditures by Country. National Center for Education Statistics. Retrieved December 10, 2021, from https://nces.ed.gov/programs/coe/indicator/cmd

Appendix

Attribute Information as provided by UCI (Cortez & Silva, 2008)

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 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)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)

31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)