# Predicting Students’ Performance in Secondary school Exams

| **Team Name** | Group 10 |
| --- | --- |
| **Team Members** | Prasanna Shanmugakumarasamy  Youssef Shehata  Akhila Sirikonda  Nikhil Reddy Muppidi  Ankita Singh  Rushil Shah  Surya Teja Chintala |
| **Domain Selected** | Education |

## Introduction

In today’s world of education, a student’s academic performance is influenced by various factors such as the economic status, geographical location, social factors (family setup, going out with friends etc.,), cultural influences (gender bias), education background of parents, and demographic conditions. Further, students' performance also varies by the standard of the schools/teaching staff and the academic performance of the students themselves in their primary education.

This data analysis aims at analyzing such a students academic performance dataset to understand the factors that contribute to their secondary school math grade outcomes.

## Research Questions

* What factors contribute to students' performance in the tests ?
* How effective are the extra paid classes and extra educational support in determining the students' success in the final grades?
* Are there specific factors that affect the performance of students that belong to different schools?
* Does the parents' education have an impact on the student’s performance?
* Does family size affect the performance of the student?

## Data Resources

This dataset is sourced from the UCI Machine learning repository. It approaches students' achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. The scope of this study is to only consider the math scores dataset.

### Dataset Link:

<https://archive.ics.uci.edu/ml/datasets/student+performance>

<https://www.kaggle.com/dipam7/student-grade-prediction>

The dataset comprises 395 observations and 33 attributes of student records from 2 schools.

## Data Preprocessing

The students dataset contains the students grades for 3 years viz., G1, G2 and G3. These 3 are highly correlated with each other. For the purposes of the data exploration study, 2 new attributes **avgscore** and **y\_binary** are created.

* Avgscore is the computed average of G1, G2 and G3 for each student rounded upto 2 decimal places.
* Y\_binary is a binary classification target variable with 0 and 1 where 0 is a score less than or equal to 10 of the avgscore variable and 1 is a score greater than 10.

The dataset characteristics and possible values of categorical variables are mentioned in the below table.

For all the data exploration activities, the data is treated as is to understand the proportions, influence of the dependent variables to the Target attribute avgscore or y\_binary.

To build a heatmap of the dependent variables and understand the correlation between the target variable avgscore and other variables, some of the categorical variables were dummy coded using a **feature engineering technique called one-hot encoding**.

| **Column Name** | **Proposed Data Type** | **Possible Values** | **Comments** |
| --- | --- | --- | --- |
| school | Category | GP, MS | Dummy Code to 0 and 1 |
| sex | Category | F, M | Dummy Code to 0 and 1 |
| age | int |  |  |
| address | Category | U, R | Dummy Code to 0 and 1 |
| famsize | Category | GT3, LE3 | Dummy Code to 0 and 1 |
| Pstatus | Category | A, T | Dummy Code to 0 and 1 |
| Medu | Category | 0-4 | Retain Categories |
| Fedu | Category | 0-4 | Retain Categories |
| Mjob | Category | at home, health, other, services, teacher, | Dummy code with One-hot Encoding |
| Fjob | Category | at home, health, other, services, teacher, | Dummy code with One-hot Encoding |
| reason | Category | course, home ,other, reputation | Dummy code with One-hot Encoding |
| guardian | Category | father, mother, other | Dummy code with One-hot Encoding |
| traveltime | Category | 1-4 | Retain Categories |
| studytime | Category | 1-4 | Retain Categories |
| failures | Category | 0-3 | Retain Categories |
| schoolsup | Category | Yes/No | Dummy Code to 0 and 1 |
| famsup | Category | Yes/No | Dummy Code to 0 and 1 |
| paid | Category | Yes/No | Dummy Code to 0 and 1 |
| activities | Category | Yes/No | Dummy Code to 0 and 1 |
| nursery | Category | Yes/No | Dummy Code to 0 and 1 |
| higher | Category | Yes/No | Dummy Code to 0 and 1 |
| internet | Category | Yes/No | Dummy Code to 0 and 1 |
| romantic | Category | Yes/No | Dummy Code to 0 and 1 |
| famrel | Category | 1-5 | Retain Categories |
| freetime | Category | 1-5 | Retain Categories |
| goout | Category | 1-5 | Retain Categories |
| Dalc | Category | 1-5 | Retain Categories |
| Walc | Category | 1-5 | Retain Categories |
| health | Category | 1-5 | Retain Categories |
| absences | int |  |  |
| G1 | int |  |  |
| G2 | int |  |  |
| G3 | int |  |  |
| avgscore | float |  | New Attribute |
| y\_binary | binary |  | New Attribute |

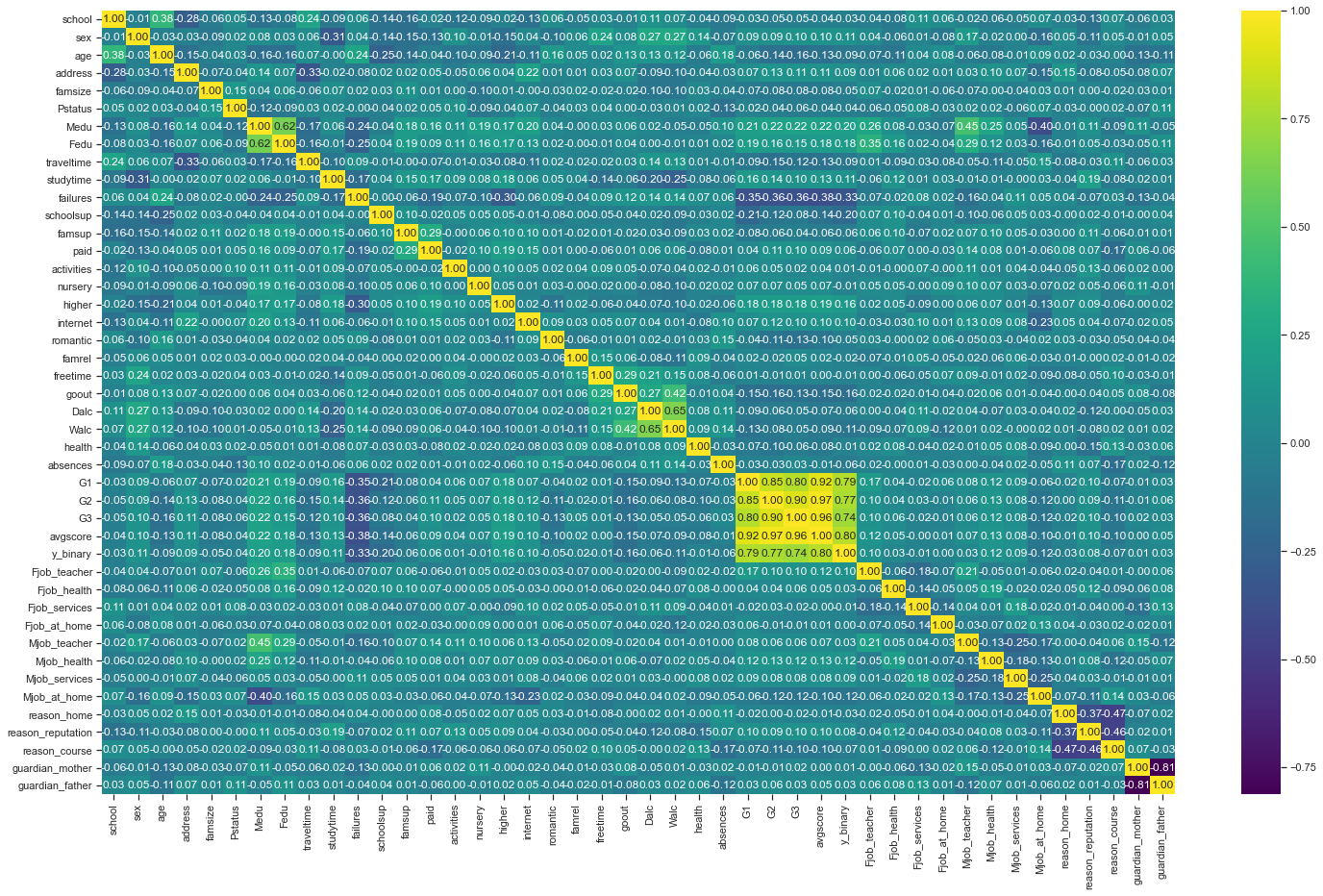
## Data Exploration

### Heat map of all variables - Correlation Analysis

With all categorical variables dummy coded, a heatmap of all the categorical and numerical variables were created to understand the correlation between the avgscore variable and other dependent variables. Fig.1 shows the correlation plot with the correlation meter indicating the nature and level of influence a variable has on the other.

Since our target variable is avgscore, the following variables seem to have some influence on the students performance.

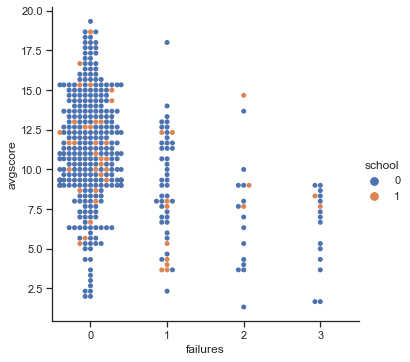
1. Failures (No. of Failures)
2. Higher (The interest expressed to pursue higher education)
3. Medu (Mother’s education level)
4. Fedu (Fathers’ education level)
5. Fjob\_teacher (Result of one-hot encoding of the Fjob variable for the value teacher.



**Fig. 1. Heatmap/Correlation Plot**

### Avgscore Vs. Failures

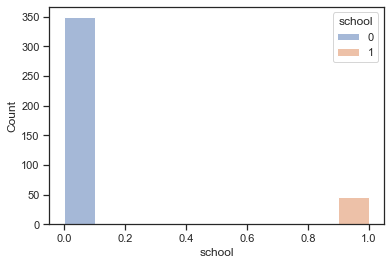
A categorical plot of Failures and avgscore with a crossection by the school indicates that With higher number of failures, the chances of gettin a higher final avgscore diminishes. Students with 3 failures don't seem to cross the 50% threshold



**Fig. 2. Avgscore Vs. Failures**

### Students’ distribution by school

The student population in the data set does not seem to be well distributed. The 2nd school has only about 50 records in the dataset



**Fig. 3. Students Distribution**

### Father’s education Vs. Avgscore

Father's education is differentiated by the present job of Father and their relationship to a student's exam score.

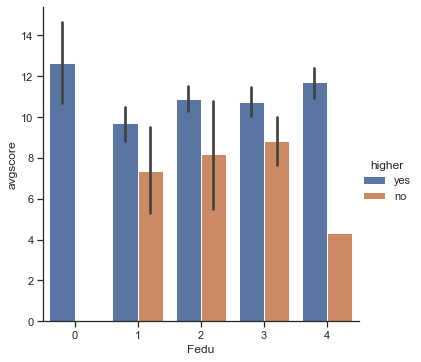
* The FJob = teacher bucket seems to have all students > 10 score that meets our threshold for success.
* All other Fjob buckets have at least one group that has score < 10
* Students whose fathers are in the civil services tend to have the lowest mean and lower confidence range. (2nd plot)

|  |  |
| --- | --- |

**Fig. 4. Fedu Vs Avgscore with cross section of Fjob**

### Father’s education influence towards students desire for Higher Education

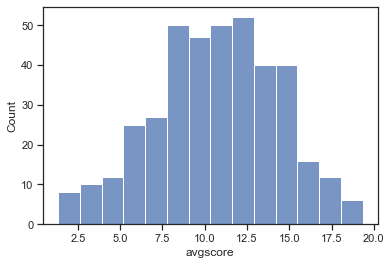
* Although not surprising, it is an important observation that with Fedu=0 (No education), All students in this bucket want to do a higher education.
* Otherwise, the common trend in all buckets is to aspire for higher education.
* Parents with Highest education bucket have fewer cases of not aspiring higher education.
* In other buckets, there is a steady prevalence of not desiring higher education.



**Fig. 5. Fedu Vs Avgscore with cross section of higher education interest**

### Distribution of avgscore

The avgscore attribute derived using G1, G2 and G3 seems to be normally distributed with a high population between 7.5 and 12.5 with very few high scorers and very few low scorers.



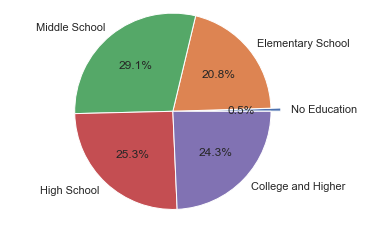
**Fig. 6. Distribution of the avgscore variable**

### Parents’ education Vs. Students Performance

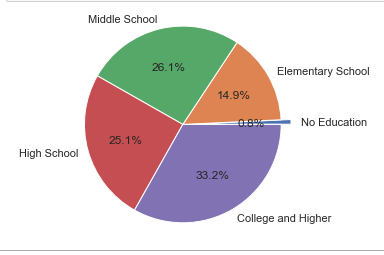
We looked into the parents' education level to determine if it had an affect on the student’s performance.

First, I put the Father’s and Mother’s education level in a pie chart to look at the overall education level of the population of parents.

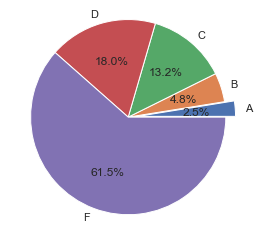
Fathers:



Mothers:



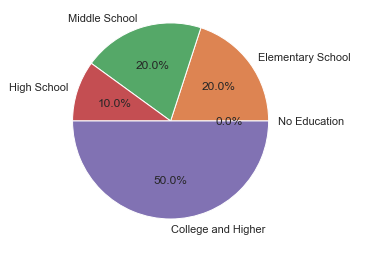
Then I got the total average grades of the students:



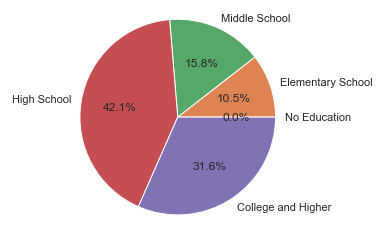
Looks like a tough school!

Then what I did next was, put the fathers’ education in a pie chart, but made 5 pie charts, one for each grade. This gave a nice visual representation of how the parent’s education affects the student's performance in school.

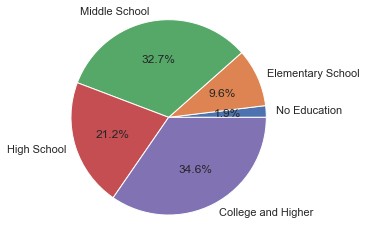
Father’s education for students that got an A average:



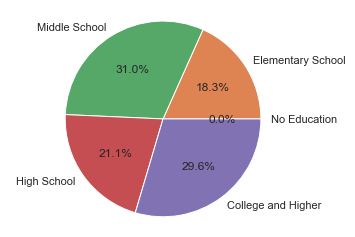
Father’s education for students that got a B average:



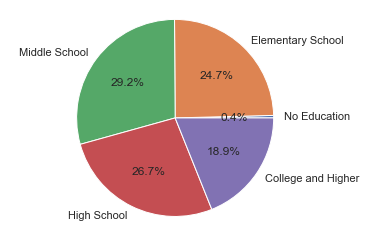
Father’s education for students that got a C average:



Father’s education for students that got a D average:

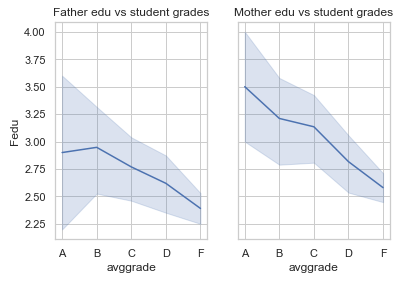


Father’s education for students that got a F average

:

We then did the same thing for the mothers, which had a similar result if not, more telling that the parent has an influence on the student’s performance.

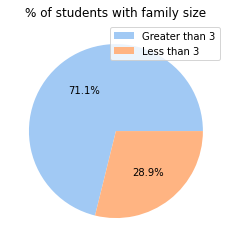
Finally, to put it in perspective, I compiled a final chart that shows the two parents' education level side-by-side with the y-axis a numerical average of the parent’s education level where 0 is No education and 4 is College and Higher and where the x-axis is the student’s average grade. This result looks like this:



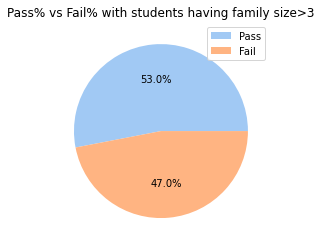
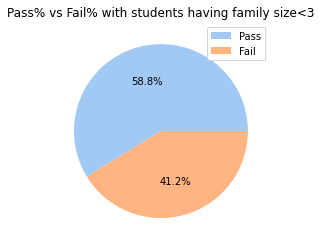
What you are seeing here is a correlation between the Average education level of the parents and the average grade of the students. As you can see, when a student has an A average, the parent generally has an average education level somewhere between middle school and College (and higher). However, as the grades go from an A average to a D or F average, you can see that their parents are all less educated than their higher scoring counterparts having a middle school average education level - indicating that the education level of the parent influences the performance of the student heavily.

### Influence of Family size in determining students performance

We tried to determine if the family size of the students had an influence on their performance. The following pie chart shows the distribution of students among two family sizes in the data set. Most of the students have a family size greater than three.

**Fig 8.1 Pie Chart showing the distribution of students among two family sizes**

Using y\_binary as the target variable, with value 1 being pass and 0 being fail, the following pie charts represent the students pass and fail percentage of two family sizes.



From the above pie charts, we can observe that students with family size of less than three have slightly more pass percentage than the students with family size greater than three. Although there is a slight difference in pass percentage, we can infer that the family size doesn't really have an influence on the student's score.

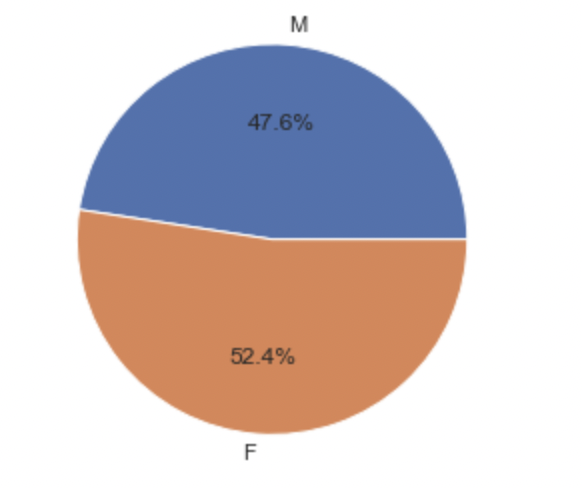
### Are there specific factors that influence students' performance in each school ?

We tried to determine if there are specific factors that influence student's performance in each school.

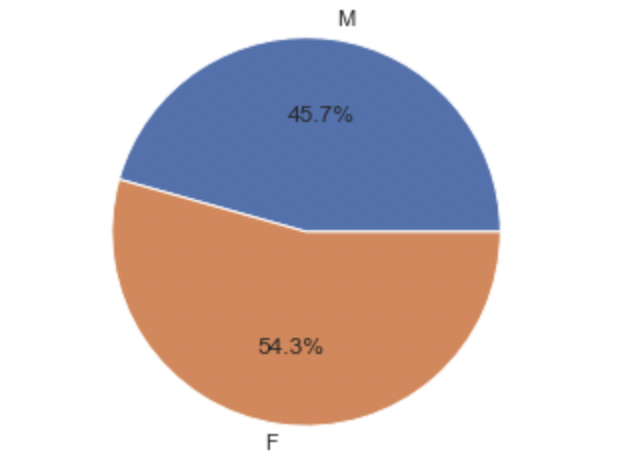
We have two schools in our data set — GP and MS. Furthermore, we tried to take different factors and compare the results(average score) between two other schools.

First, we tried to determine if the sex of the students influenced students' performance in different schools.

The sex ratio of students in GP school is shown below.



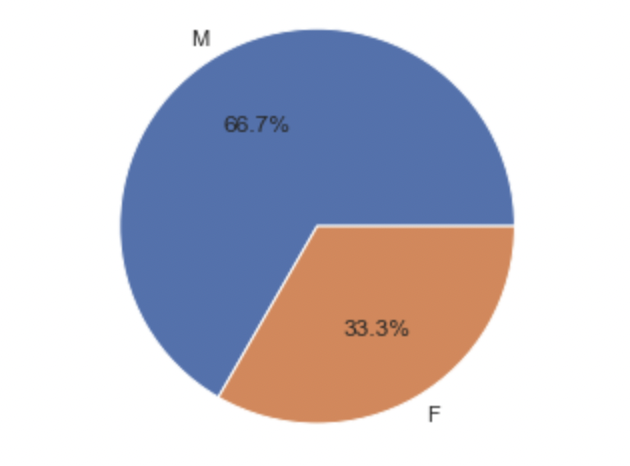
The sex ratio of students in MS school is shown below.



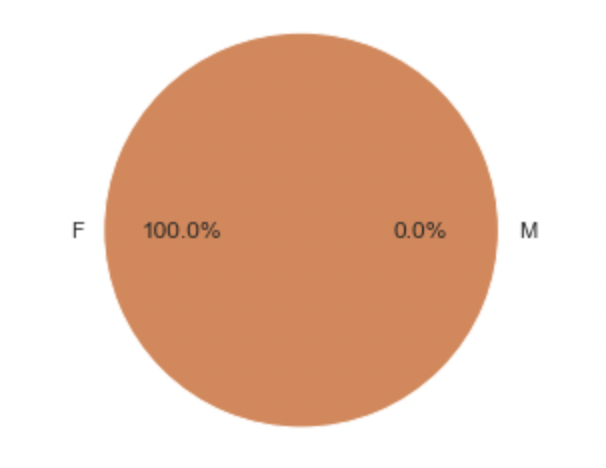
The scores converted to the 5 letter grading scale as shown below: A: >= 90 B: >= 80 C: >= 70 D: >= 60 F: <= 59

We determined the grade distribution for both male and female in GP and MS school, which helps us to compare the results.

**GP School Male/Female for students who got a ‘A’ average**



**MS School Male/Female for students who got a ‘A’ average**

****

Similarly, we determined the data for different factors and compared the result.

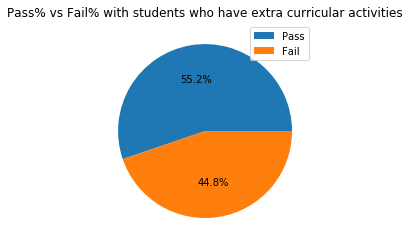
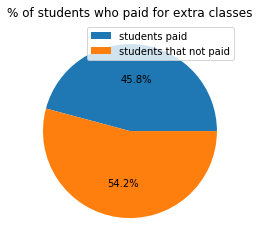
Based on the present data, we did not see any specific factor that influence in different school. Also, the data contains majority of GP school information.

In [57]:

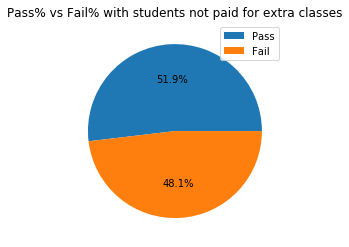
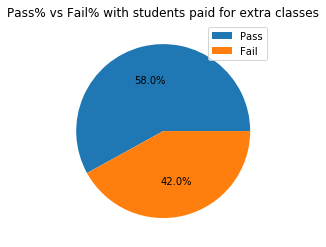
### Are the Extra paid classes and Extra educational support helping students in their final grades ?

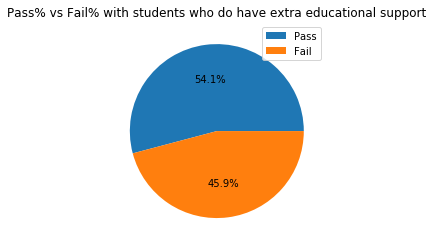
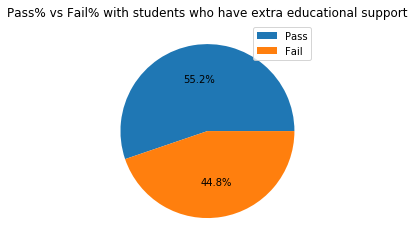
<Surya>

To determine extra paid classes and extra educational support helping students in their final grades we need how many students have paid for extra classes.from below pie chart we observe that 45% of students have paid for classes and 54 % have not taken any extra classes. Respectively 55% of students have educational support and 44% of students don't have educational support



We have to find out the respective pass and fail %, also average score between them for this we are going to use y\_binary and avg score as target variables.









Based on the above results from piechart and graphs,If a student either has extra paid classes or extra educational support will help in getting good final grades compared to students who don't have either of them.With respective to their Average scores students with extra classes and extra educational support have better Average score

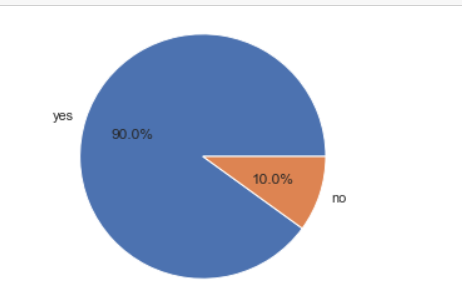
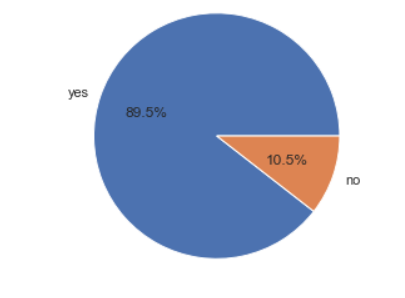
### Influence of Internet availability in Student’s performance

We tried to determine the influence of internet availability in a student's performance.

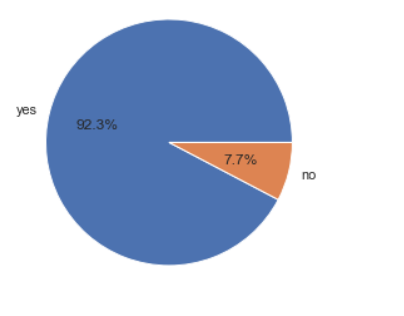
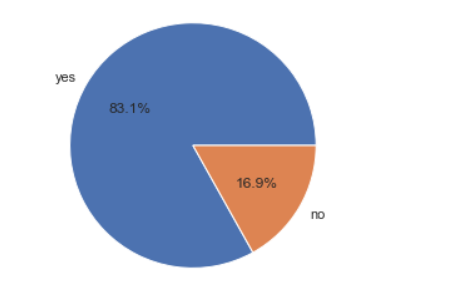
We have internet availability(yes/no) and students average grade in our data set. Furthermore, we tried to take different factors and compare the results(average score) for all students and present whether the actual grading is affected or not due to internet availability.

The scores converted to the 5 letter grading scale as shown below: A: >= 90 B: >= 80 C: >= 70 D: >= 60 F: <= 59

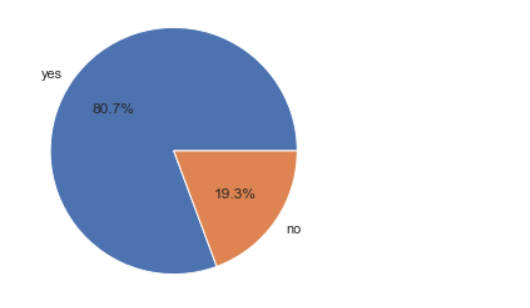
Students who got an A grade and B grade

Students who got C grade and D grade

Students who got F grade



Similarly, we determined the data for different factors and compared the result.

Based on the present data, we did not see Influence of Internet availability in Student’s performance Also, the data contains the majority of GP school information.

## Data Preparation for Modeling

For the research question discussed above, ***what factors influence a students performance in the final exams***, it requires a binary classification target variable that classifies success and failure based on the students exam score. For this purpose, the y\_binary variable introduced has a threshold of avgscore 10 above which a student is considered successful and below which a student is classified as not successful. This y\_binary target variable can then be used with many classification models such as the logistic regression, decision trees and support vector machine.

### Dropping Columns

* **Age:** This is a nominal variable in this dataset and from correlation analysis it shows no significance to the avgscore or y\_binary.
* **G1, G2 and G3 and avgscore:** G1, G2 and G3 were combined to derive avgscore and y\_binary originally. Therefore, they are redundant and will introduce multicollinearity in the data. Also, while the avgscore had been an important attribute for the data exploration process, it cannot be used for the classification modeling.

Therefore the above attributes will be dropped for the modeling.

### DataType Transformation

The one hot encoding introduces new attributes for each value under Fjob, Mjob, reason, guardian attributes.

### Training and Test Sets:

The dataset of 395 observations will be split into **70% training set (276 records) and 30% test set (119 records)** with a random seed so that data from both schools are sampled.

## 

## Predictive Modeling

The classification models that were used for this project are logistic regression, Naive Bayes and decision trees. These techniques allowed us to evaluate multiple explanatory variables to evidently predict the probability of students' success in final exams. The target variable y\_binary is a newly introduced attribute based on the average score from the grades G1, G2 and G3. A score of >10 is considered as success (1) and any score < = 10 is considered as failure (0) for the purposes of classification modeling.

**Logistic regression** is a widely implemented modeling approach for classification problems. Given that we might consider feature encoding in tandem with logistic regression, this approach allows us to predict outcome for the binary variable using the response variables of which can be classified into categorical or continuous patterns.

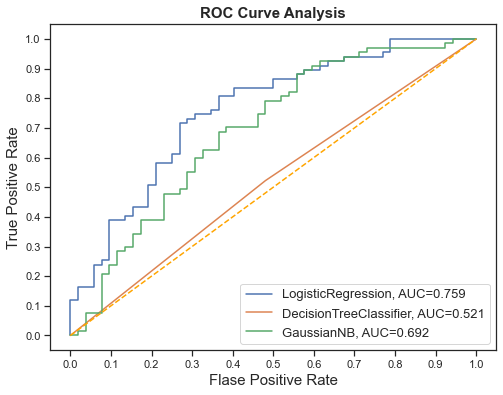
**Decision trees** analyse a data set in order to construct a set of rules, or questions, which are used to predict a class. It evaluates variables to determine which are the most important in partitioning the data, and then creates a tree of decisions (a set of rules) which best partitions the data.

**Naive Bayes Classifier** is another modeling technique that uses a family of probabilistic algorithms and applies the probability theory and Bayes’ theorem. The fundamental Naive Bayes assumption is that each feature makes an independent and equal contribution to the outcome.

## Results Evaluation

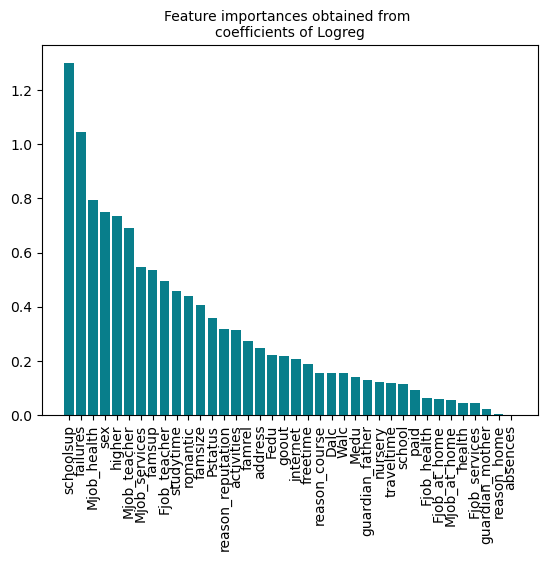
Here we are examining our results from the three models we used on our data set. Generally, for a classification problem we value the AUC more so than the accuracy and precision. This is why we decided to go with the Logistic Regression model as our model of choice as seen below.

|  | Decision Trees | Logistic Regression | NaiveBayes |
| --- | --- | --- | --- |
| Accuracy | 0.5949367088607594 | 0.6430379746835443 | 0.660759493670886 |
| F1 | 0.6307738019178956 | 0.6854021204036026 | 0.7085532439584428 |
| Precision | 0.6341560027274313 | 0.6685131013005577 | 0.6719294724164888 |
| Recall | 0.6298097251585624 | 0.7084566596194503 | 0.7645877378435518 |
| AUC | 0.529276693455798 | 0.7214408725602757 | 0.6382032146957519 |



The feature importance is displayed below to visualize which feature had a greater impact on the performance of the student ordered from greatest to least. As you can see, the top 5 features that influenced the student’s performance were:

* Schoolsup - Extra educational support (yes or no)
* Failures - Former classes the student has failed
* Mjob\_health - Mother’s job is in the health field
* Sex - Student’s gender (Male or Female)
* Higher - If the student wants to achieve higher education (yes or no)



## Future Work

GIven more time, we would look to other schools in different countries to understand if the factors discovered in this report apply globally, or if this is just a regional occurrence. Ideally, we would identify highschools or the highschool-equivalent in that region across the globe and use the same features to discover whether the feature importance applies to the globe as a whole or, if not, where and why there is a difference in one region versus another region. This would help us understand how students around the world are affected by outside factors, further allowing institutions and governments to improve in education with these new findings.

## GitHub Repository

<https://github.com/AkhilaSirikonda/group10kdd>

## References:

P. Cortez and A. Silva. 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.

[Web Link](http://www3.dsi.uminho.pt/pcortez/student.pdf)