ITEC 620

BUSINESS INSIGHTS/ANALYTICS PROJECT -

STUDENTS PERFORMANCE ANALYSIS

TAZVITYA AUBREY CHIHOTA

Executive Summary

This study investigates the factors influencing student performance utilising a dataset of 649 students from the UCI Machine Learning Repository in Portuguese language. The study filtered 33 variables, which included variables on demographics, study habits, and the family relationship of the student to identify patterns and relationships that contribute to academic outcomes.

Using a combination of descriptive statistics and predictive modelling, including T-tests, k-Nearest Neighbors (KNN), multiple regression, and regression trees, the analysis identified multiple regression as the most effective method for uncovering key factors impacting performance based on this study. The study found out that Students who express aspirations for higher education tend to have higher final grades and a history of academic failures was associated with lower final grades. Again, there was a significant performance between males and females, where females performed significantly better than males. Finally, the fact that they addressed of as students, whether they lived in rural or urban areas also affected their performance based on the analysis.

Based on these insights, the recommendations were proposed, such as encouraging higher education aspirations through mentorship programs, which can motivate students to achieve better outcomes. Establishing remedial support systems for students with a history of failure can ensure that those who have had more failures receive the assistance they need to improve their grades. Increasing investment in rural schools is also crucial for addressing disparities in educational resources and opportunities. Furthermore, creating gender-responsive learning environments can help cater to the diverse needs of all students, fostering engagement and inclusivity.

By addressing these key factors, this study provides a comprehensive framework for fostering educational equity and improving student outcomes in Portuguese secondary schools.

1.0 Introduction:

In today's education system, improving student performance is a key objective for schools, educators, and policymakers. Understanding the diverse factors that influence student achievement is crucial for creating targeted interventions, optimizing resource allocation, and fostering environments that support learning. However, the path to academic success is multifaceted, shaped by elements such as demographics, family background, study habits, and the broader educational environment. Addressing this complexity requires a data-driven approach to uncover patterns and insights that can guide decision-making.

This project aimed to address a significant challenge faced by educational institutions: identifying and supporting students at risk of underperforming. It sought to bridge the gap by identifying key determinants of success and creating predictive models that can help educators proactively address these challenges. By focusing on final grades (G3) as a primary metric of academic performance, this project sought to uncover actionable insights that can inform interventions and improve student outcomes.

1.2 Problem or Challenge

Educational institutions often face challenges in accurately identifying and supporting students at risk of underperformance. Despite extensive data on demographics, social background, parental involvement, and other factors influencing academic success, many schools lack effective, data-driven methods to predict academic outcomes or design personalized intervention strategies. This leads to generalized approaches to student support, which may fail to address the unique needs of each student, ultimately impacting retention rates, resource allocation, and overall student success.

2.0 Data:

This analysis utilizes a dataset named Student performance from the UCI Machine Learning Repository and we used the dataset for Portuguese language which comprises of data, gathered from two secondary schools in Portugal, contains 649 data points. It includes 33 attributes that provide insight into various factors influencing student performance, such as demographic information, study habits, and parental education levels, however for our analysis we used a total of 16 variables shown in **Table 1** below:

Table 1: List of Variables

Name	Description	Type

Sex	Student's sex (binary: female -'F' = 0 or Male-'M' =1)	Binary
Age	Student's age (numeric: from 15 to 22)	Integer
Address	Student's home address type (urban - 'U' =0 or Rural -'R' = 1)	Categorical
Medu	Mother's education (0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)	Integer
Fedu	Father's education (0 - none, 1 - primary education (4th grade), 2 5ths to 9th grade, 3 secondary education or 4 higher education)	Integer
Guardian	Student's guardian (Mother=1, Father=2, Other=3)	Categorical
Traveltime	home to school travel time (1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)	Integer
Studytime	weekly study time (1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)	Integer
Higher	Wants to take higher education, No=0, Yes=1	Binary
Internet	Internet access at home, No=0, Yes=1	Binary
Famrel	Quality of family relationships (numeric: from 1 - very bad to 5 - excellent)	Integer
Freetime	Free time after school (numeric: from 1 - very low to 5 - very high)	Integer
Goout	Going out with friends (numeric: from 1 - very low to 5 - very high)	Integer
Paid	Extra paid classes within the course subject, No=0, Yes=1	Binary
G3	Final grade (numeric: from 0 to 20, output target)	Integer
Failure	Number of past class failures (numeric: n if 1<=n<3, else 4)	Integer

2.0.1 Snapshot of Dataset

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	1	0
2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	0	1
3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	0	1	0
4	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	0	1
5	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	0	1
6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	0	1

	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
1	0	0	1	1	0	0	4	3	4	1	1	3	4	0	11	11
2	0	0	0	1	1	0	5	3	3	1	1	3	2	9	11	11
3	0	0	1	1	1	0	4	3	2	2	3	3	6	12	13	12
4	0	1	1	1	1	1	3	2	2	1	1	5	0	14	14	14
5	0	0	1	1	0	0	4	3	2	1	2	5	0	11	13	13
6	0	1	1	1	1	0	5	4	2	1	2	5	6	12	12	13

2.0.2 Data Cleaning

We observed highly correlated variables, such as prior grades (G1 and G2), which were then excluded from the analysis to mitigate overfitting and focus on cumulative performance, represented by the final grade (G3).

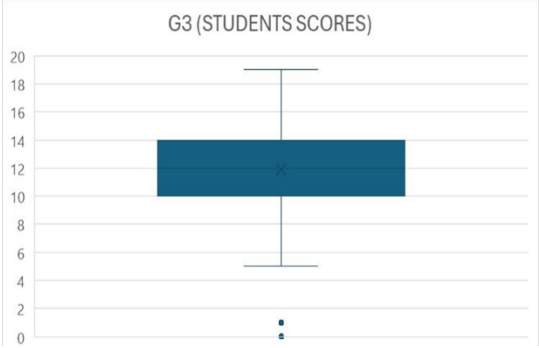
Many of the variables, such as gender, home address, and parental education, were categorical. Therefore, for compatibility with modeling techniques such as KNN, we recoded them into numerical

variables. For instance: Gender was coded as 0 for female and 1 for male, Home Address was coded as 0 for urban and 1 for rural and Parental Education levels were assigned numerical values, with 0 for no education, 1 for primary education, 2 for secondary education, and 3 for higher education.

2.0.3 Outlier Detection:

Figure 1: Box plot showing the distribution of G3

As shown in Fig.1 students clustered around the median of 12 with fewer students achieving extreme values (e.g 0 or19). To avoid overfitting, these outliers were handled by applying robust modeling techniques (such as regression trees) that are less affected by extreme values, ensuring that the predictions were not distorted by these outliers.



3.0 Analysis

3.1 Descriptive Analysis

3.1.1 Summary Statistics on Age, Free time and Grade

Table 2: Summary Statistics on Age, Free time and Grade

	Mean	Std.Dev	Min	Q1	Median	Q3	Max	IQR	Skewness	N
Age	16.74	1.22	15	16	17	18	22	2	0.41	649
Free time	3.18	1.05	1	3	3	4	5	1	-0.18	649
G3	11.91	3.23	0	10	12	14	19	4	-0.91	649

The findings showed the mean age of students in the study was 16.74, with a standard deviation of 1.22 years. The youngest student was 15 years, and the oldest was 22. Also, the average free time a student had was 3.18 hours with a standard deviation of 1.05 hours with the least free time being 1 hour and the maximum free time being 5 hours. The average grade of students, which was measured from 0 to 20 had an average of 11.91 and standard deviations of 3.23. The lowest grade was a 0 and the highest grade was 19. Median grade centered around 14, which means there were students at the extreme ends of the data.

3.1.2 Summary on Gender, Address, Mother and Father Education

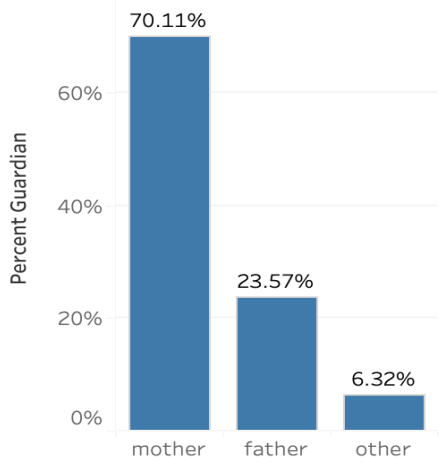
Table 3: Summary on Gender, Address, Mother and Father Education

	Frequency	Percent
Sex		
Female	383	59.01
Male	266	40.99
Total	**649**	**100.00**
Address		
Rural	197	30.35
Urban	452	69.65
Total	**649**	**100.00**
Father Education		
None	7	1.08
Primary Education (4^{th} grade)	174	26.81
5^{th} to 9^{th} Grade	209	32.20
Secondary education	131	20.18
Higher education	128	19.72
Total	**649**	**100.00**
Mother's Education		
None	6	0.90
Primary Education (4^{th} grade)	143	22.03
5^{th} to 9^{th} Grade	186	28.66
Secondary education	139	21.42
Higher education	175	26.96
Total	**649**	**100.0**

The findings showed that most of the respondents were females (59%) as compared to males (40.99%). Close to 70% of the students lived in urban areas compared to 30% who lived in rural areas. Only seven fathers did not have education (1%), but the majority (32.%) had a 5^{th} to 9^{th}-grade level of education and less than 20% had higher education. Mothers, on the other hand, had majority having 5^{th} to 9^{th}-grade education (28%), and only six mothers (0.9%) did not have any education. Generally, the majority of fathers, as well as mothers, cumulatively had more than 98% having an educational level between primary to higher education.

3.1.3 Distribution of Guardian

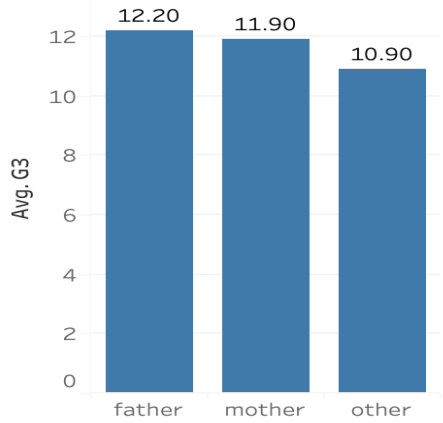
Figure2: Frequency of Guardian



About 70% of the students had a mother as a guardian compared to less than a quarter (23.5%) who had a father as a guardian, and less than 7% had another type of guardian. In this majority of guardians were mothers.

3.1.4 Student Performance based on the type of Guardian.

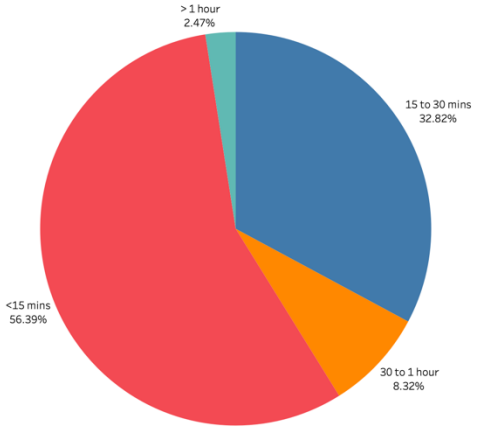
Figure 3: Performance based on the type of Guardian.



The Chart summarizes the average score of students based on the type of Guardian they had at home. It is observed that students with a father as a guardian had the highest average (12.20), followed by those with mothers as the guardian (11.90) with those with other types of guardians (10.9) being the least.

3.1.5 Travel Time

Figure 4: Distribution on Travel Time of Students



More than half of the students (56.39%) travel for less than 15 minutes to school, followed by those who travel between 15 to 30 minutes (32%). Less than 9% of students travelled between 30 to one hour and less than 3% of students travelled more than one hour. Cumulatively, more than 80% of students travelled between 15 to 30 minutes to school.

T-Test

3.2 T-Test Analysis of Performance based on Student's Gender

3.2.1 Comparing the mean Grade of Student based on Gender

Table 4: Comparing the mean Grade of Student based on Gender

Sex	Average Score
Female	12.25
Male	11.40

The table summarizes the average score between males and females. An independent sample t-test was conducted to see if there was a significant difference between male and female students.

Interpretation: To analyse the difference in the mean grade of students by gender, an independent sample t-test was conducted. The results of the t-test indicated that there was a significant difference in the average grade of males and that of females $t(547) = 3.27$, $p = 0.0011$ leading to the rejection of the null hypothesis which assumes that there were not difference between male and female student.. (See screenshot in Appendix A).

3.3 Models

A correlation test was done to measure the strength and direction of the relationship between the independent variables using correlation (See Appendix A). Also, various models were tested and compared to know which one performed well in the case of the student performance data and was found that linear regression was the best among the three modelling techniques used. (Screenshot of various models in Appendix A)

Attribute	Regression	KNN	Regression Tree
RMSE	2.643045	2.71721	2.736133
Interpretability	High	Low	Medium
Handles non-linearity	Low	High	High

3.3.1 Linear Regression Model

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.3602     2.6061   2.441 0.015132 *
sexMale          -0.6310     0.3152  -2.002 0.046030 *
age               0.1447     0.1364   1.061 0.289472
addressRural     -0.7469     0.3487  -2.142 0.032837 *
Medu              0.1301     0.1786   0.729 0.466695
Fedu              0.2553     0.1786   1.429 0.153795
guardian_rFather  0.2553     0.3708   0.689 0.491456
guardian_rOthers  0.5246     0.6359   0.825 0.409917
traveltime        0.2100     0.2157   0.974 0.330852
studytime         0.3678     0.1876   1.960 0.050692 .
higherYes         1.9734     0.5383   3.666 0.000283 ***
internetYes       0.3070     0.3636   0.844 0.399124
famrel            0.1416     0.1567   0.904 0.366758
freetime         -0.1055     0.1575  -0.670 0.503356
goout            -0.1535     0.1366  -1.124 0.261761
paidYes          -0.1064     0.6221  -0.171 0.864267
failures         -1.5659     0.2754  -5.687 2.62e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.889 on 372 degrees of freedom
Multiple R-squared:  0.2481,    Adjusted R-squared:  0.2158
F-statistic: 7.672 on 16 and 372 DF,  p-value: 1.064e-15
```

A multiple linear regression was modelled to predict the grade of student which resulted in a significant model, $F(16, 372) = 7.672$, $p < .000$, $R^2 = .2158$ indicate 21% of the total variation in the student performance (G3) was explained by the independently. Examining the individual predictors further showed that the intercept was significant with ($t = 2.44$, $p < .05$), gender was significant with ($t = -2.002$, $p < .05$), address was significant with ($t = 1.061$, $p < .05$), Higher was significant with ($t = 3.66$, $p < .001$) and failure was significant with ($t = -5.687$, $p < .001$).

Interpretation of variables

Higher: - Estimate = 1.9734, p-value < 0.001, whether a student wanted to take a higher education or not contributed positively towards their performance.

Failures - Estimate = -1.5595, p-value < 0.001 Each additional past failure significantly reduces performance. This indicates that a history of academic failure is a strong predictor of lower student performance.

Address - Estimate = -0.7469, p-value = 0.0328

Where a student stays, whether in the urban or the rural area, negatively affects their performance. This highlights a potential disparity in educational resources or opportunities between rural and urban regions.

SexMale: - Estimate = -0.6310, p-value = 0.0460. The gender of student significantly contributed negatively to their final score.

4.0 Discussion and Conclusion

4.0.1 Discussion

Aspirations and Academic Performance: Students with aspirations for higher education are associated with an average increase of 1.9734 points in their grades ($p < 0.001$). This statistically significant relationship highlights the value of fostering academic ambition. Schools may support this by introducing mentorship programs, goal-setting workshops, and other initiatives that inspire students to aim for higher educational achievements.

The Impact of Past Failures: The analysis revealed that past academic failures adversely impact current performance, with each failure reducing final grades significantly. Each additional failure reduces student grades by an average 1.5595 points ($p < 0.001$). Schools may adopt proactive interventions, including after-school tutoring, counseling, and early identification systems, to support struggling students.

Urban-Rural Disparities: Students from rural areas demonstrated lower performance levels compared to their urban counterparts by 0.7469 points lower than urban students. These findings may give suggestions to the policy makers and educators to come up with strategies that might influence the difference in performance. Without proof or specific evidence, the recommendations should be framed more cautiously, emphasizing general principles rather than definitive claims. Policymakers could focus on improving resource distribution, teacher support, digital access, and parental engagement in rural areas. Using data-driven approaches and regularly evaluating progress may help identify effective strategies to address disparities

Gender Differences: The analysis indicates that gender significantly influences academic performance, with male students scoring 0.6310 points lower than their female counterparts ($p = 0.0460$). This negative estimate highlights the need to address gender-based disparities. Educators may come up with tailor made, gender-sensitive educational strategies that could help to engage male students more effectively and ensure teaching methods cater to diverse learning needs, fostering equitable academic outcomes for all.

Study Time and Effective Habits: additional study time showed a positive association with performance highlighting that an hour increase in study time corresponds to a 0.37-point grade improvement, although only statistically significant (alpha = 0.1).

4.0.2 Conclusion

The findings from this study highlighted the complexity of student performance and the necessity for targeted interventions to address existing challenges and disparities. To establish a more equitable and effective education system, it is crucial for policymakers and educators to collaborate on implementing evidence-based strategies. By fostering students ambitions, supporting those who struggle academically, bridging geographical and gender gaps, and promoting effective study habits, schools can help unlock the full potential of their students. Continued research and innovation in educational practices will be essential to ensure that all students, regardless of their background, can excel academically and achieve their goals. The model built could be used to predict student scores based on variables like gender, aspiration for higher education, failures addressed and study time of the students.

APPENDIX A

T-TEST RESULTS

```
# Variance test
var.test(std_selected$G3 ~ std_selected$sex)

freq(std_selected$sex)

aggregate(std_selected$G3 ~std_selected$sex , data = std_selected, mean)

# T-Test
t.test(std_selected$G3 ~ std_selected$sex)
```

```
data:  std_selected$G3 by std_selected$sex
t = 3.2747, df = 547.44, p-value = 0.001125
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 0.3390334 1.3554639
sample estimates:
mean in group 0 mean in group 1
       12.25326        11.40602
```

Training -Test Partition

```
set.seed(12345)

ycol <- match("G3",colnames(std_selected_r))

training <- sample(1:nrow(std_selected_r), training_size*nrow(std_selected_r))

stds_df.training <- std_selected_r[training,-ycol]
stds_df.training.results <- std_selected_r[training,ycol]

head(stds_df.training)
stds_df.test <- std_selected_r[-training,-ycol]
stds_df.test.results <- std_selected_r[-training,ycol]
```

Linear Regression

```r
# Regression
# Linear regression
stds_df.reg <- lm(G3 ~ ., data=std_selected_r[training,])
stds_df.reg.predictions <- predict(stds_df.reg,std_selected_r)[-training]
(mean((stds_df.test.results-stds_df.reg.predictions)^2))^0.5

summary(stds_df.reg)
summary(stds_df.reg.predictions)
plot(stds_df.reg)
predict(stds_df.reg, new.st)
```

Regression Tree Model

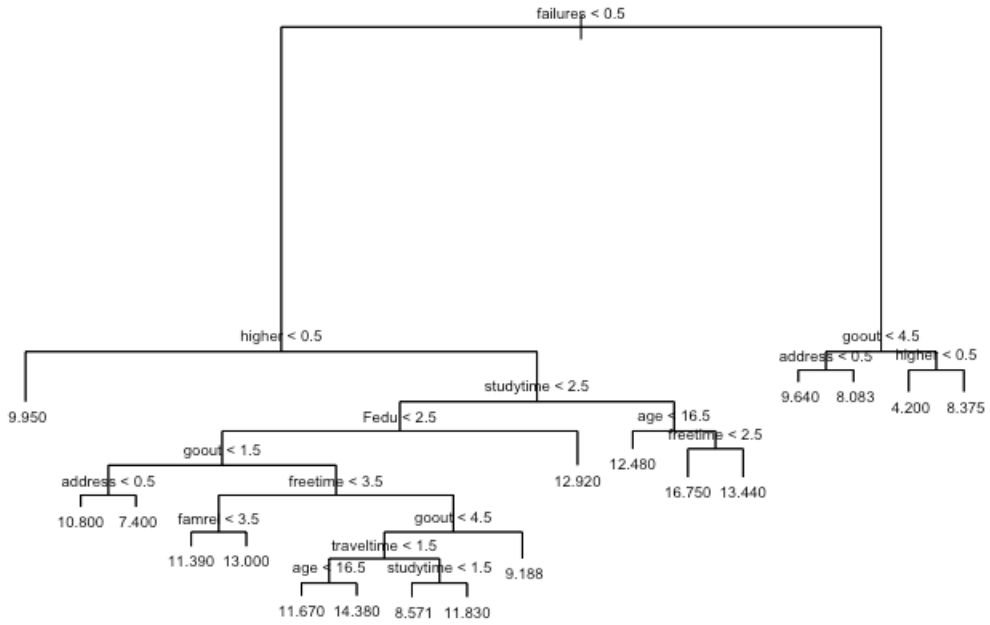
```r
# Regression Tree (using only one value of mindev; use a loop to find the optimal tree)
stds_df.tree <- tree(G3 ~ ., data=std_selected_r[training,],mindev=0.0065)
stds_df.tree.predictions <- predict(stds_df.tree,std_selected_r)[-training]
(mean((stds_df.test.results-stds_df.tree.predictions)^2))^0.5

plot(stds_df.tree)
text(stds_df.tree, cex=0.6)

# predict(stds_df.tree, new.st)
```

```r
# Knn regression
stds_df.knn <- knn.reg(stds_df.training, stds_df.test, stds_df.training.results, k=19)
(mean((stds_df.knn$pred - stds_df.test.results)^2))^0.5
```

The Regression tree

```
> # This loop checks values of k from 1 to 100 for the best k.
> best.k <- -1
> RMSE <- -1
> best.RMSE <- 99999999
> for (i in 1:100) {
+   set.seed(12345)
+   std.knn <- knn.reg(stds_df.training, stds_df.test, stds_df.training.results, k=i)
+   RMSE <- (mean((std.knn$pred - stds_df.test.results)^2))^0.5
+   if (RMSE < best.RMSE) {
+     best.k <- i
+     best.RMSE <- RMSE
+   }
+ }
> print(paste("The optimal value of k is",best.k,"with a RMSE of",best.RMSE))
[1] "The optimal value of k is 19 with a RMSE of 2.71721002925135"
>
```

```
> best.mindev <- -1
> RMSE <- -1
> best.RMSE <- 99999999
> for (i in seq(from=0.0005, to=0.05, by=0.0005)) {
+   stddf.tree <- tree(G3 ~ ., data=std_selected_r[training,], mindev=i)
+   std.tree.predictions <- predict(stddf.tree,std_selected_r)[-training]
+   RMSE <- (mean((stds_df.test.results-std.tree.predictions)^2))^0.5
+   if (RMSE < best.RMSE) {
+     best.mindev <- i
+     best.RMSE <- RMSE
+   }
+ }
> print(paste("The optimal value of mindev is",best.mindev,"with a RMSE of",best.RMSE))
[1] "The optimal value of mindev is 0.0065 with a RMSE of 2.73613343648199"
>
```



Correlation between the variables

	sex	age	address	Medu	Fedu	guardian	traveltime	studytime	higher	Internet	famrel	freetime	goout	paid	failures	G3
sex	1	-0.04	-0.03	0.12**	0.08*	0.02	0.04	-0.21***	-0.06	0.07	0.08*	0.15***	0.06	0.08*	0.07*	-0.13***
age		1	0.03	-0.11**	-0.12	0.17	0.03	-0.01	-0.27	0.01	-0.02	0.00	0.11	-0.01	0.32	-0.11
address			1	-0.19	-0.14	0.03	0.34	-0.06	-0.08	-0.18	0.03	0.04	-0.02	0.03	0.06	-0.17
Medu				1	0.65	-0.11	-0.27	0.10	0.21	0.27	0.02	-0.02	0.01	0.11	-0.17	0.24
Fedu					1	0.01	-0.21	0.05	0.19	0.18	0.02	0.01	0.03	0.09	-0.17	0.21
guardian						1	0.09	0.02	-0.14	0.03	-0.04	0.00	-0.03	-0.01	0.14	-0.03
traveltime							1	-0.06	-0.07	-0.19	-0.01	0.00	0.06	-0.04	0.10	-0.13
studytime								1	0.19	0.04	0.00	-0.07	-0.08	0.00	-0.15	0.25
higher									1	0.07	0.05	-0.10	-0.07	0.02	-0.31	0.33
internet										1	0.08	0.06	0.09	0.03	-0.10	0.15
famrel											1	0.13	0.09	0.03	-0.06	0.06
freetime												1	0.35	-0.05	0.11	-0.12
goout													1	-0.01	0.05	-0.09
paid														1	0.07	-0.05
failures															1	-0.39
G3																1