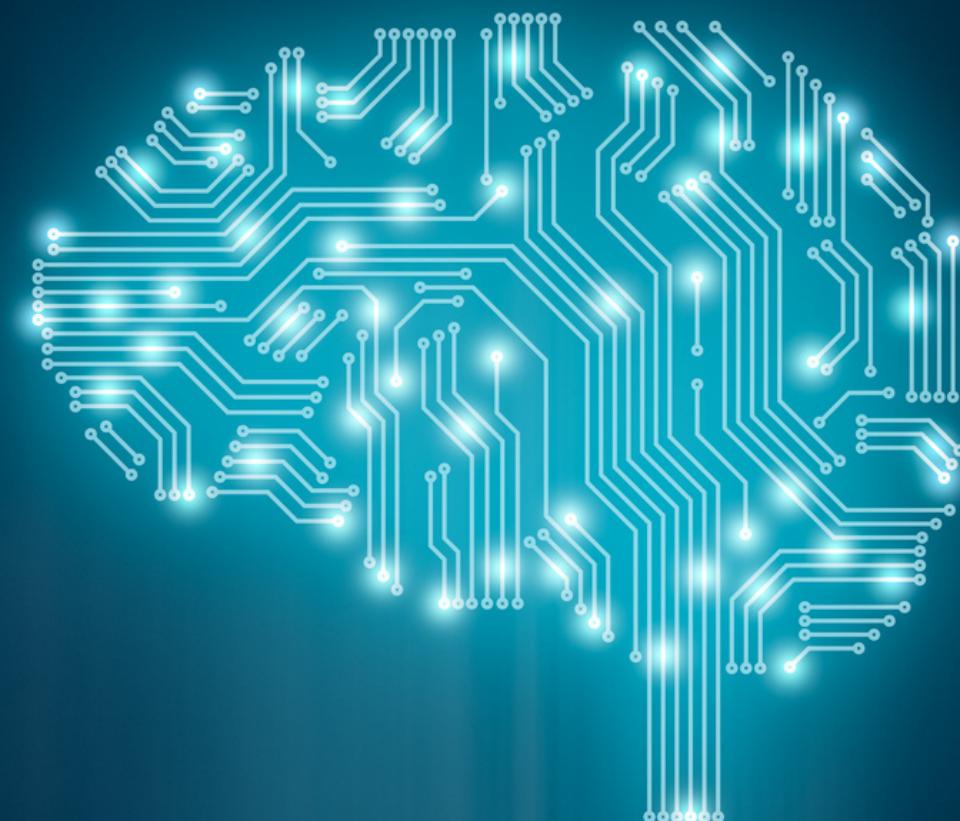


MACHINE LEARNING

19BCE298

MAULIK VIRPARIYA





Data Visulation

Feature Selection

Training and testing the model



DataSet

Student Performance

Characteristic

Type

Attribute
Characteristic

Multivarite

Regression,
Classification

Integer

**There are two Dataset
so first we marge both**

**Added new feature
named subject :**

**Math
Portuguese**

Student-Math

395 x 33

Student-portuguese

649 x 33

Student

1044 x 34

FEATURES

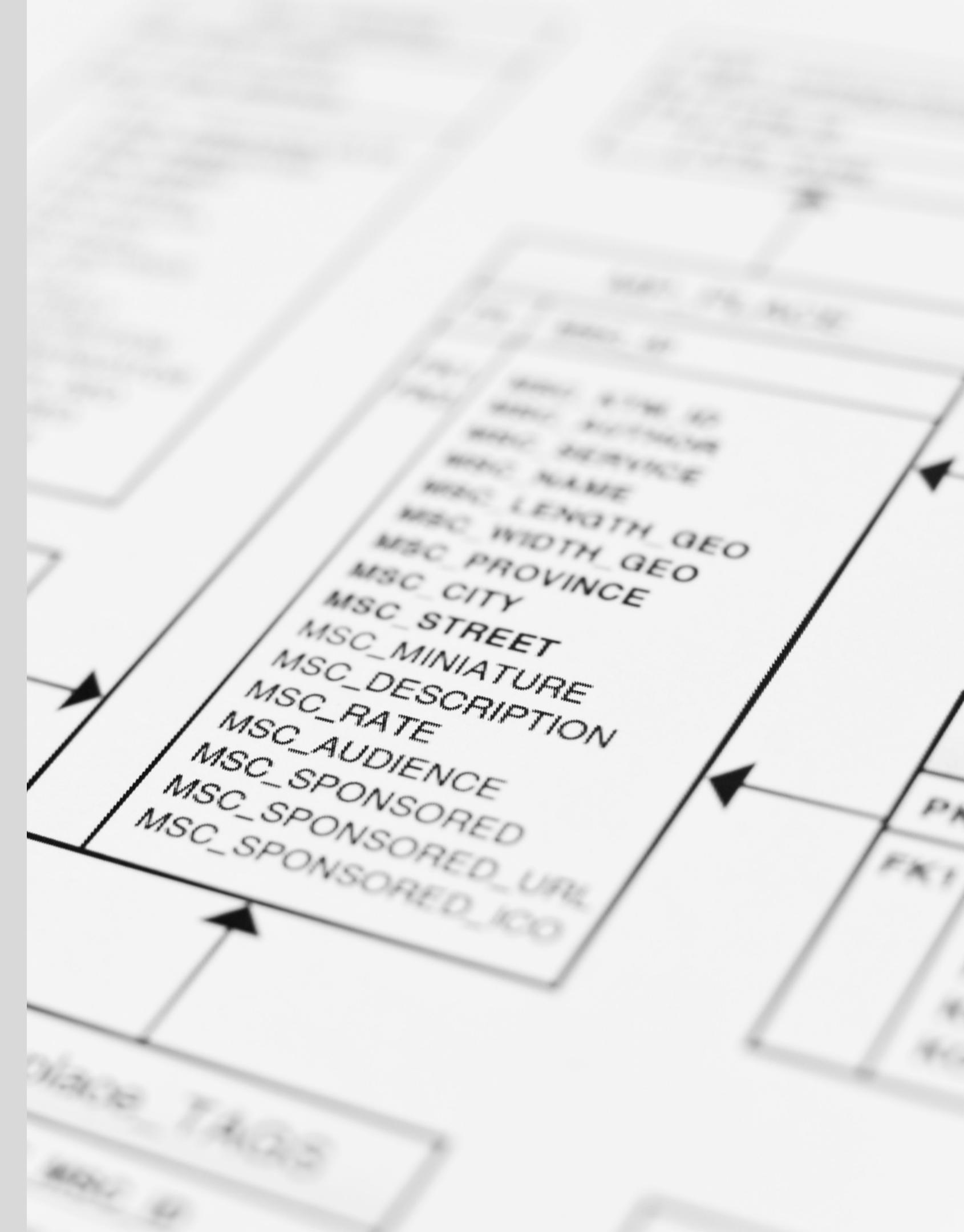
There are two dataset. Both have 29 same features. One has subject Math and one has Portuguese.

Math data set Contains 395 rows (instances)

Portuguese data set Contains 649 rows (instances)

- 1 school - student's school
- 2 sex - student's sex
- 3 age - student's age
- 4 address - student's home address type
- 5 famsize - family size
- 6 Pstatus - parent's cohabitation status
- 7 Medu - mother's education
- 8 Fedu - father's education
- 9 Mjob - mother's job
- 10 Fjob - father's job
- 11 reason - reason to choose this school
- 12 guardian - student's guardian
- 13 traveltime - home to school travel time
- 14 studytime - weekly study time
- 15 failures - number of past class failures
- 16 schoolsup - extra educational support
- 17 famsup - family educational support

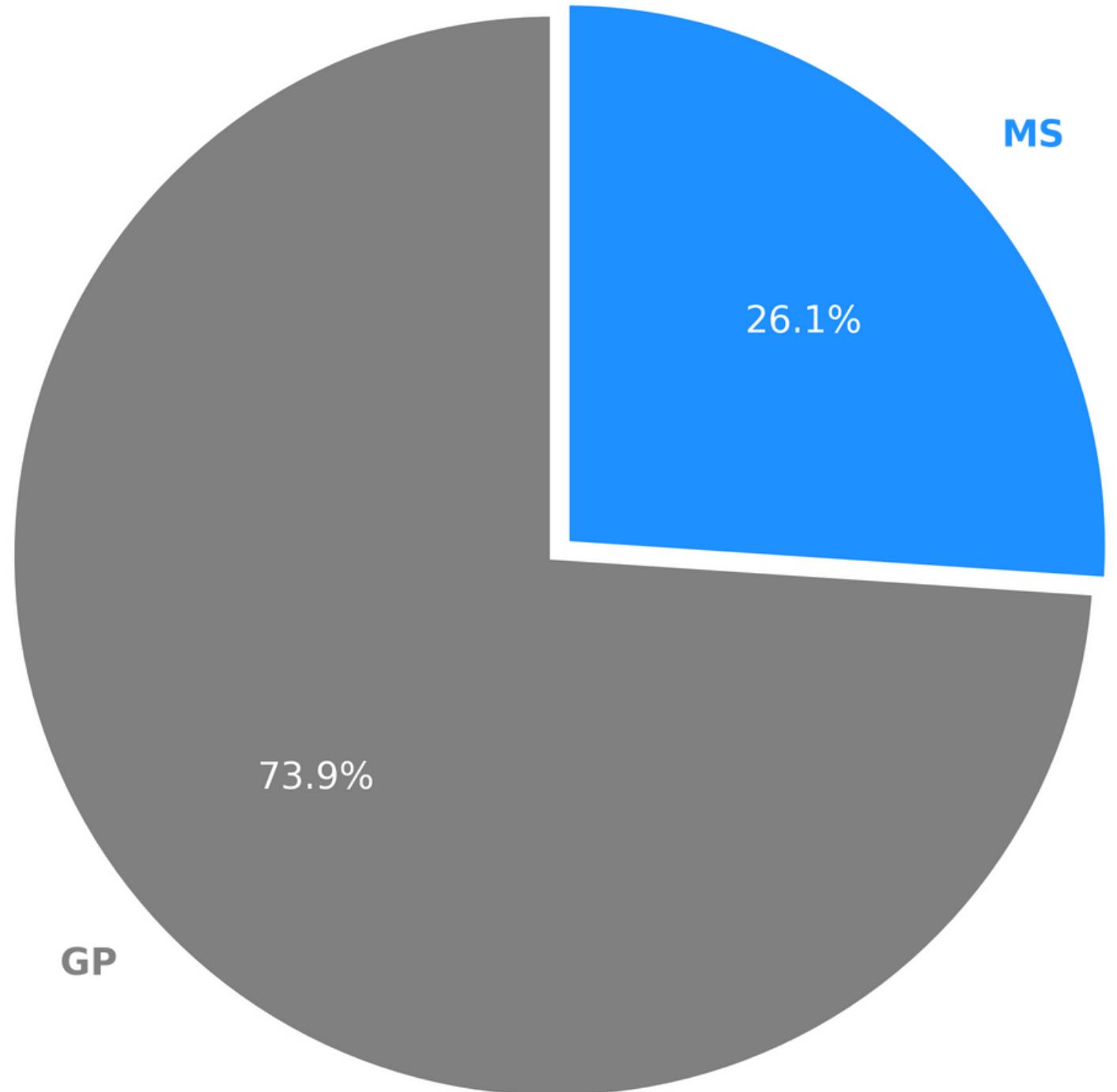
- 18 paid - extra paid classes within the course subject
- 19 activities - extra-curricular activities
- 20 nursery - attended nursery school
- 21 higher - wants to take higher education
- 22 internet - Internet access at home
- 23 romantic - with a romantic relationship
- 24 famrel - quality of family relationships
- 25 freetime - free time after school
- 26 goout - going out with friends
- 27 Dalc - workday alcohol consumption
- 28 Walc - weekend alcohol consumption
- 29 health - current health status
- 30 absences - number of school absences
- 31 G1 - first period grade (numeric: from 0 to 20)
- 31 G2 - second period grade (numeric: from 0 to 20)



DATA VISUALIZATION

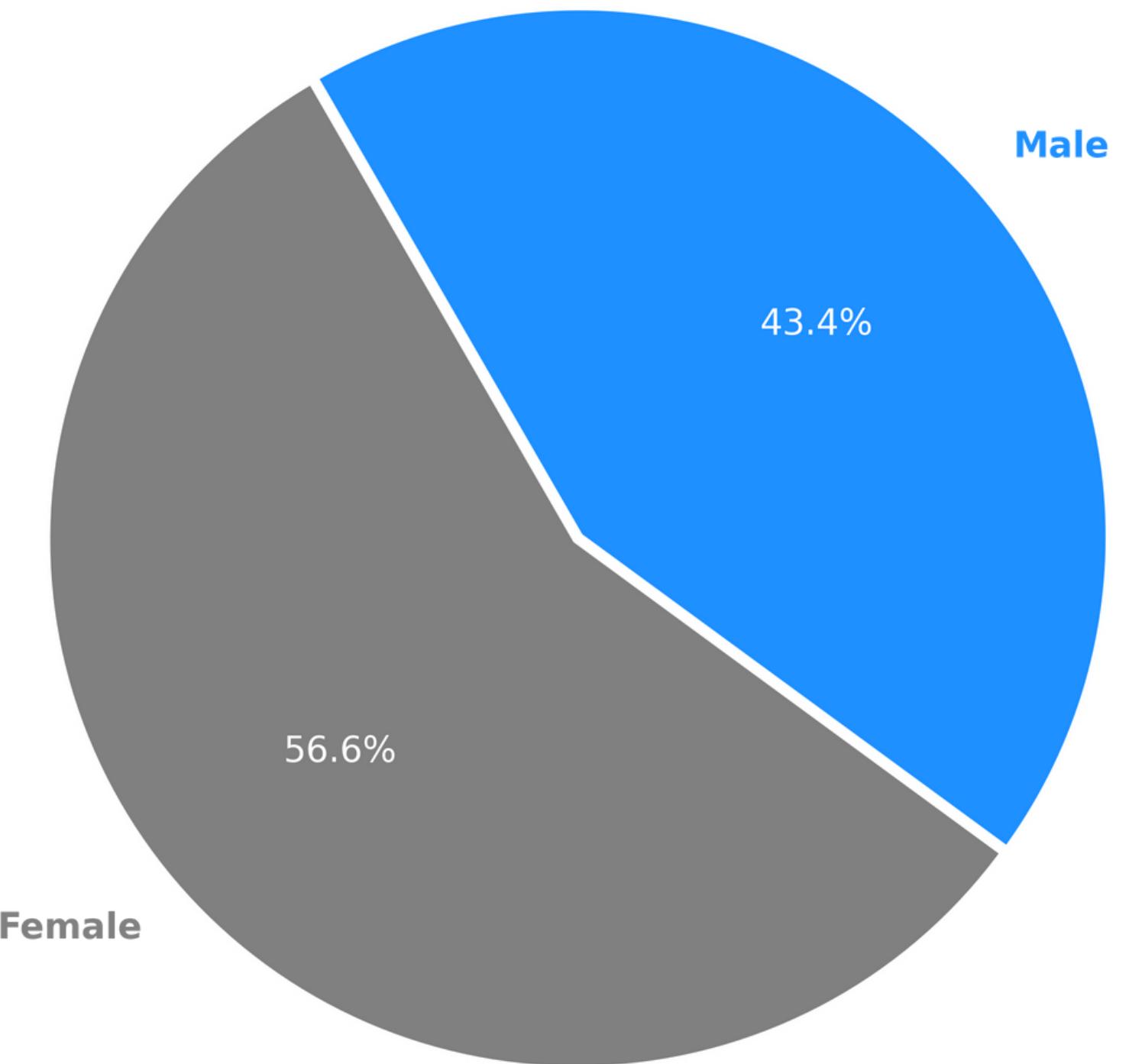


School



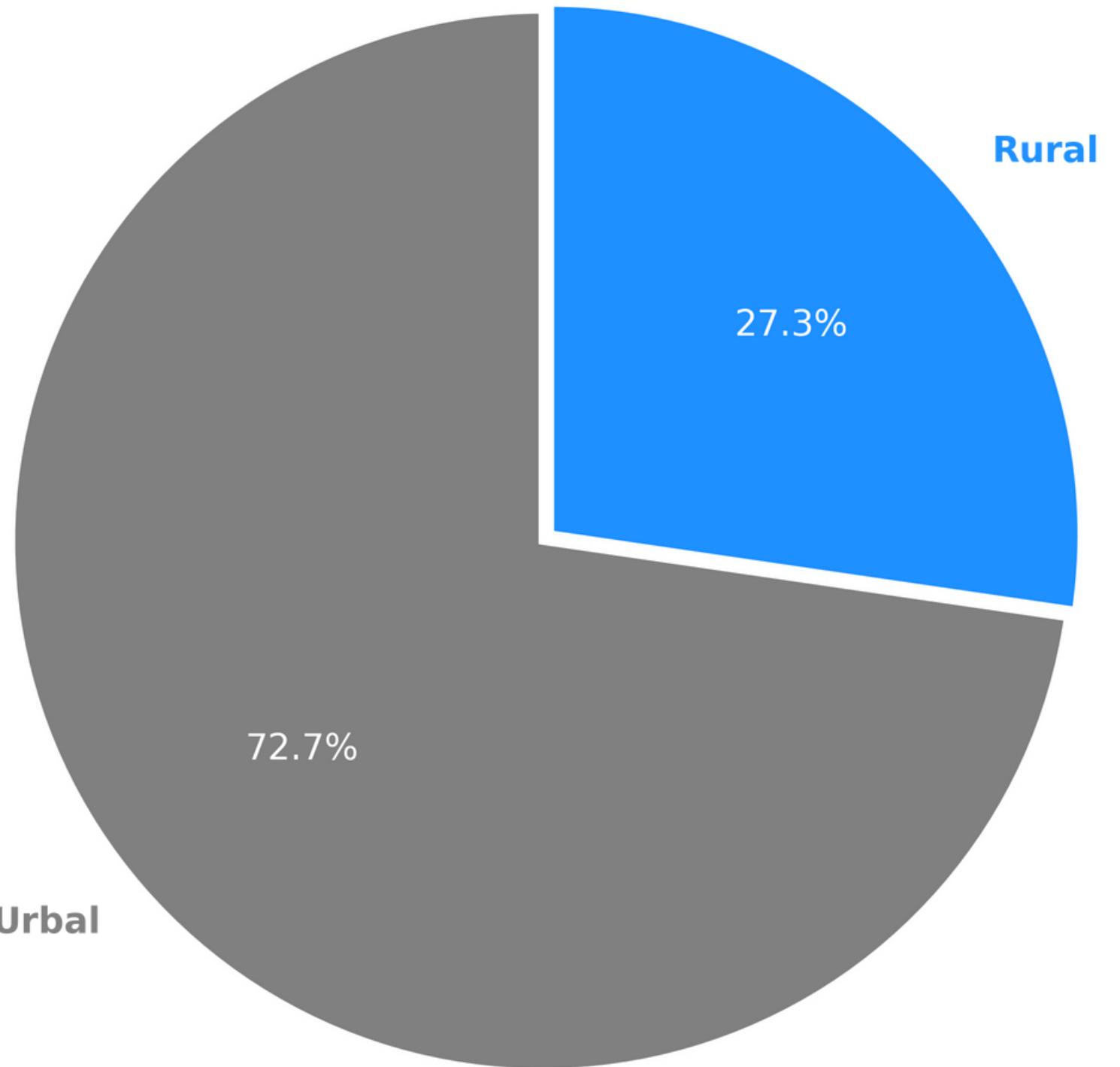
Student count based on
school

Gender



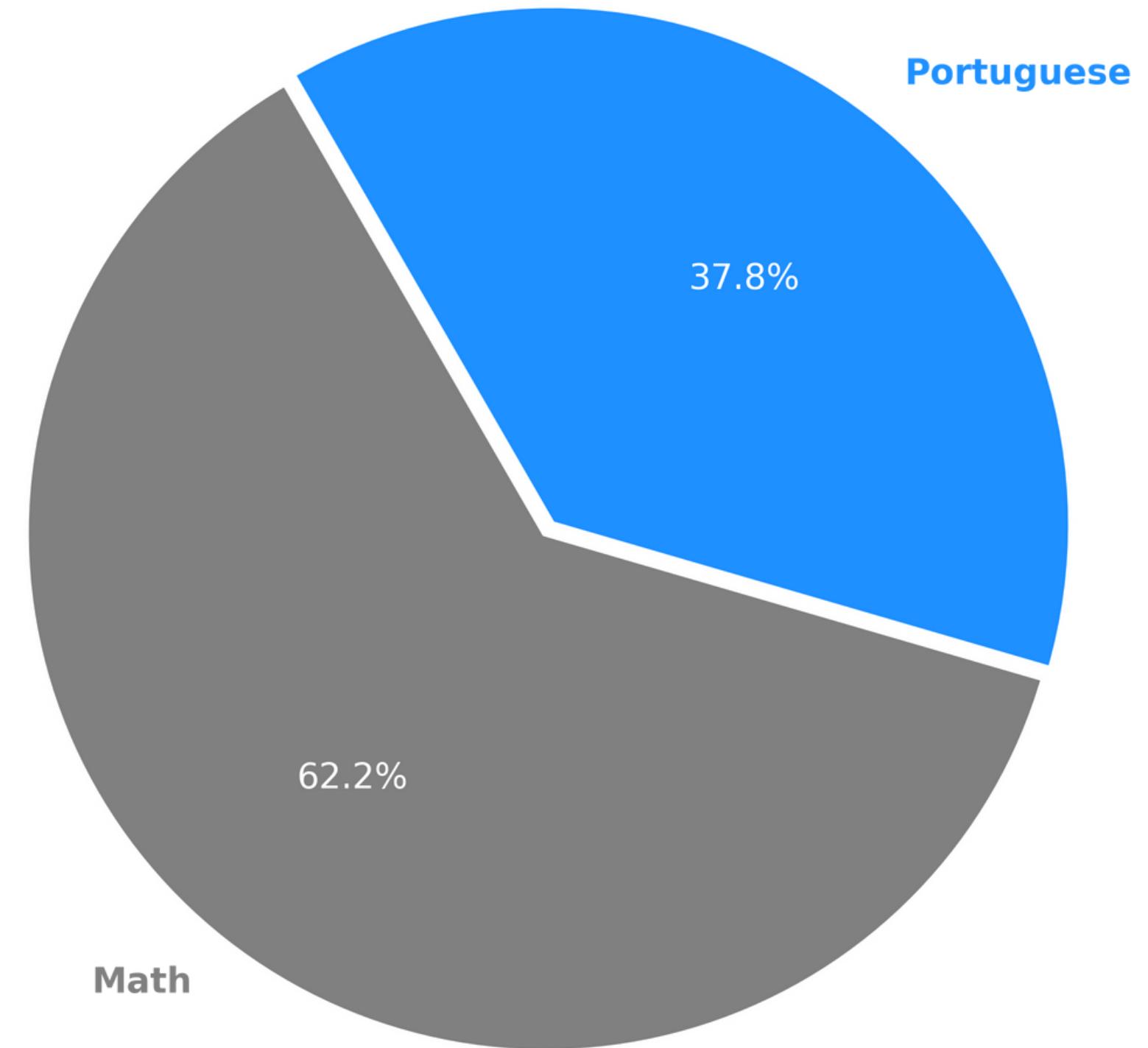
Student count based on
Gender

Area



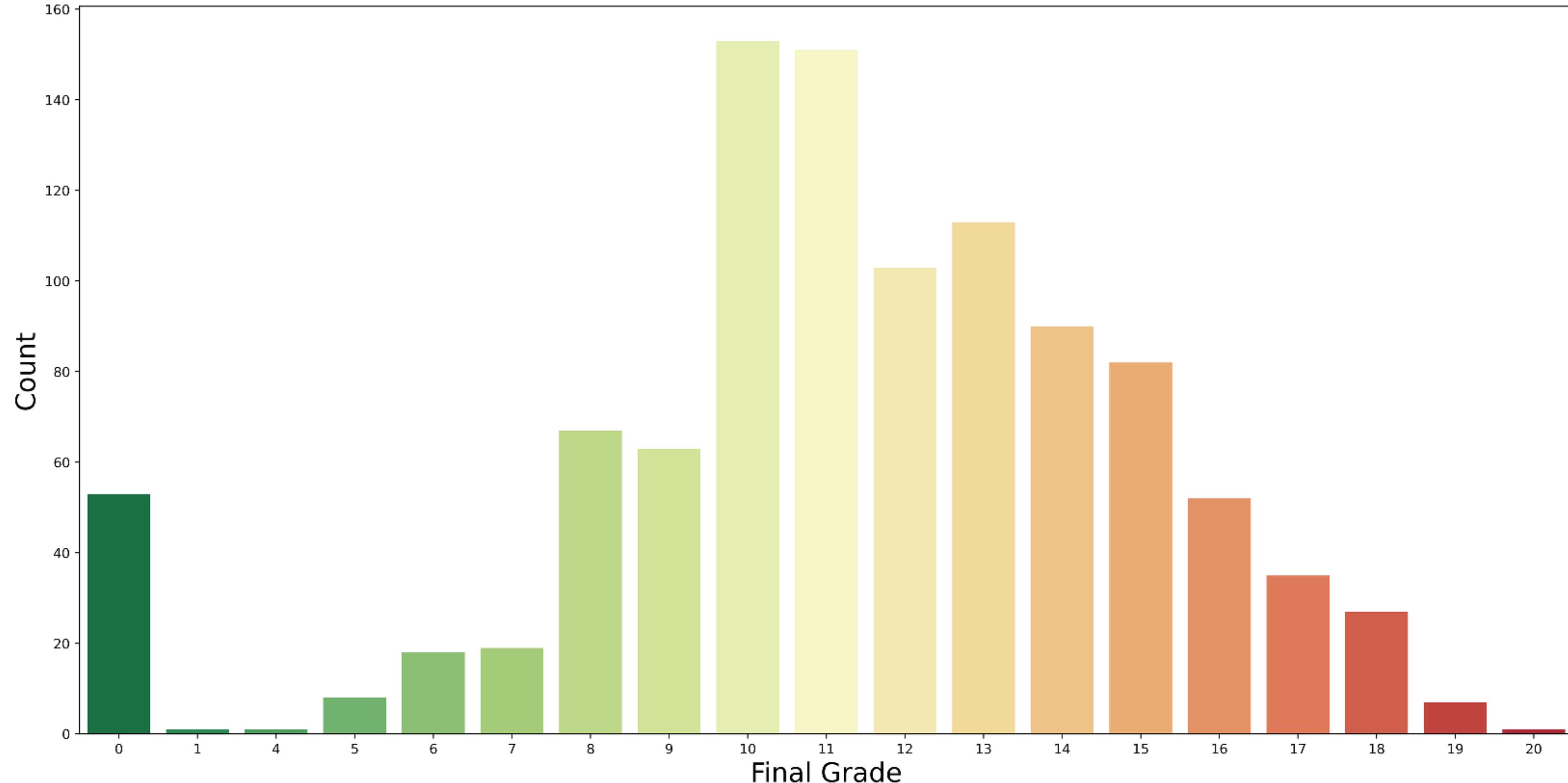
Student count based on
Area

Subject



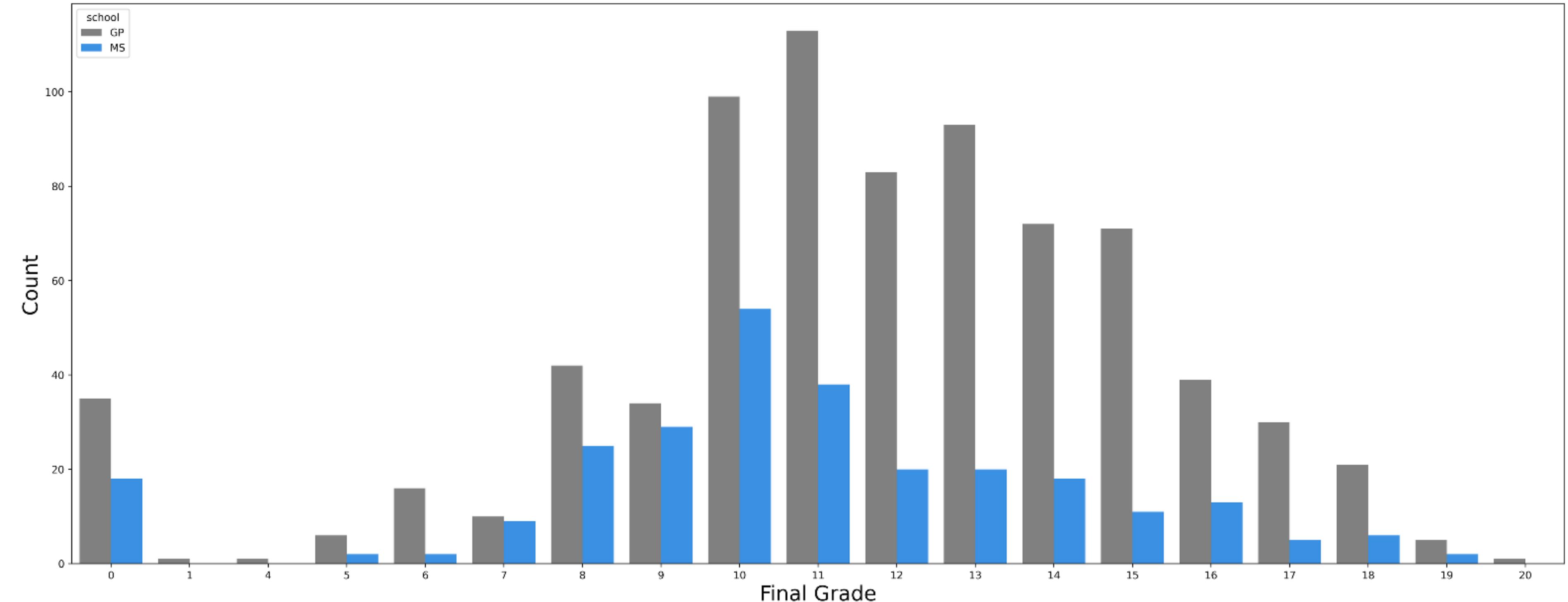
Student count based on
Selected

Distribution of Final Marks of students



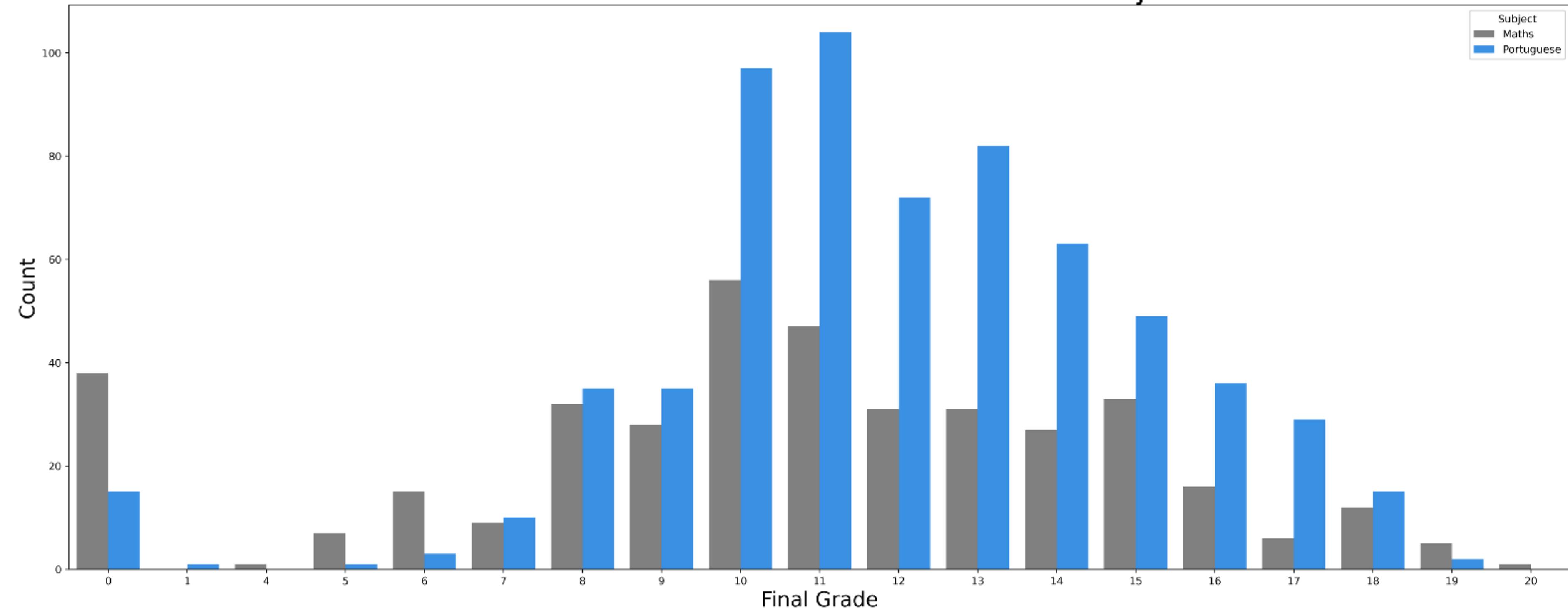
Distributions According to their Marks in G3

Distribution of Marks of students acc to school



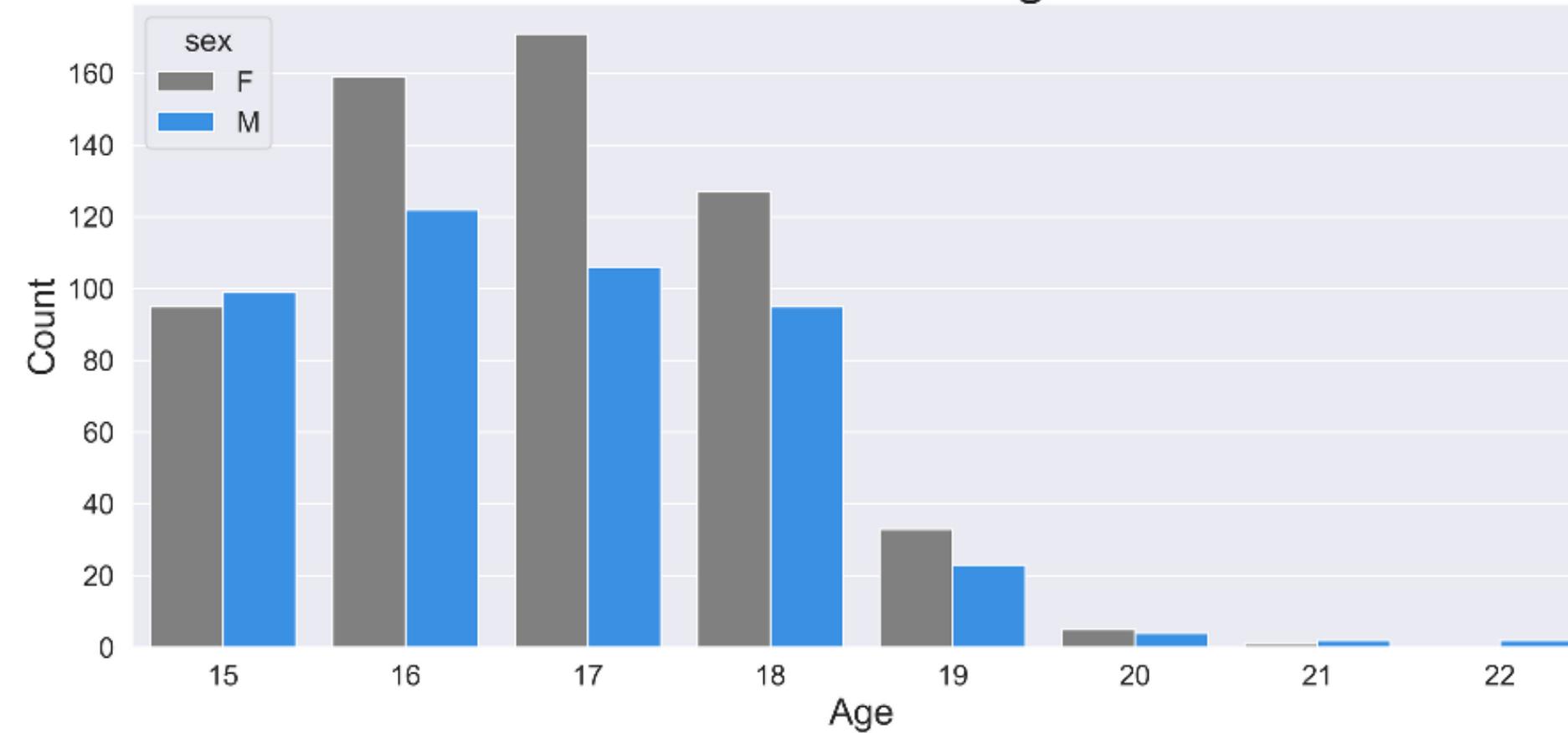
Distributions According to their
School

Distribution of Marks of students acc to Subject

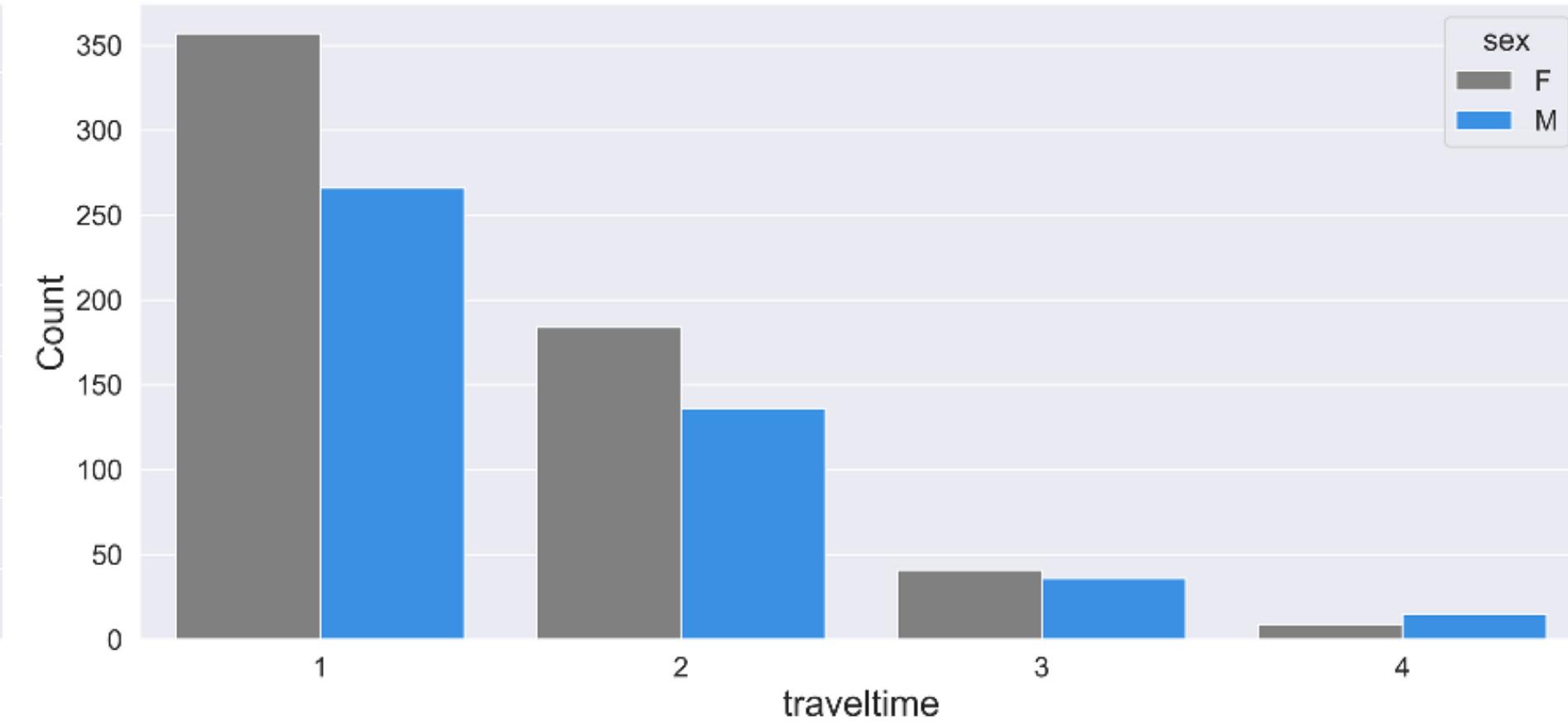


Distributions According to their Subject

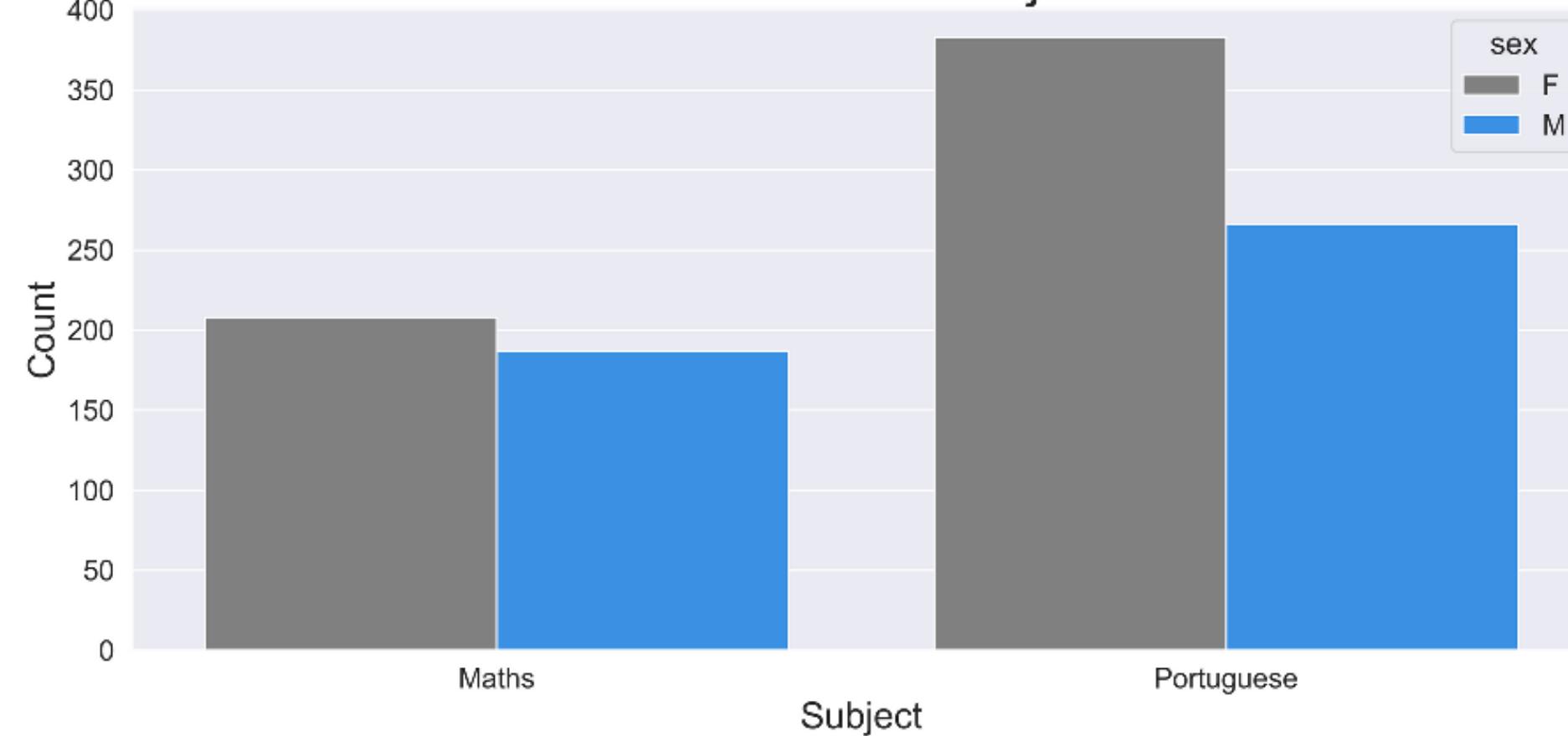
Distribution of Age



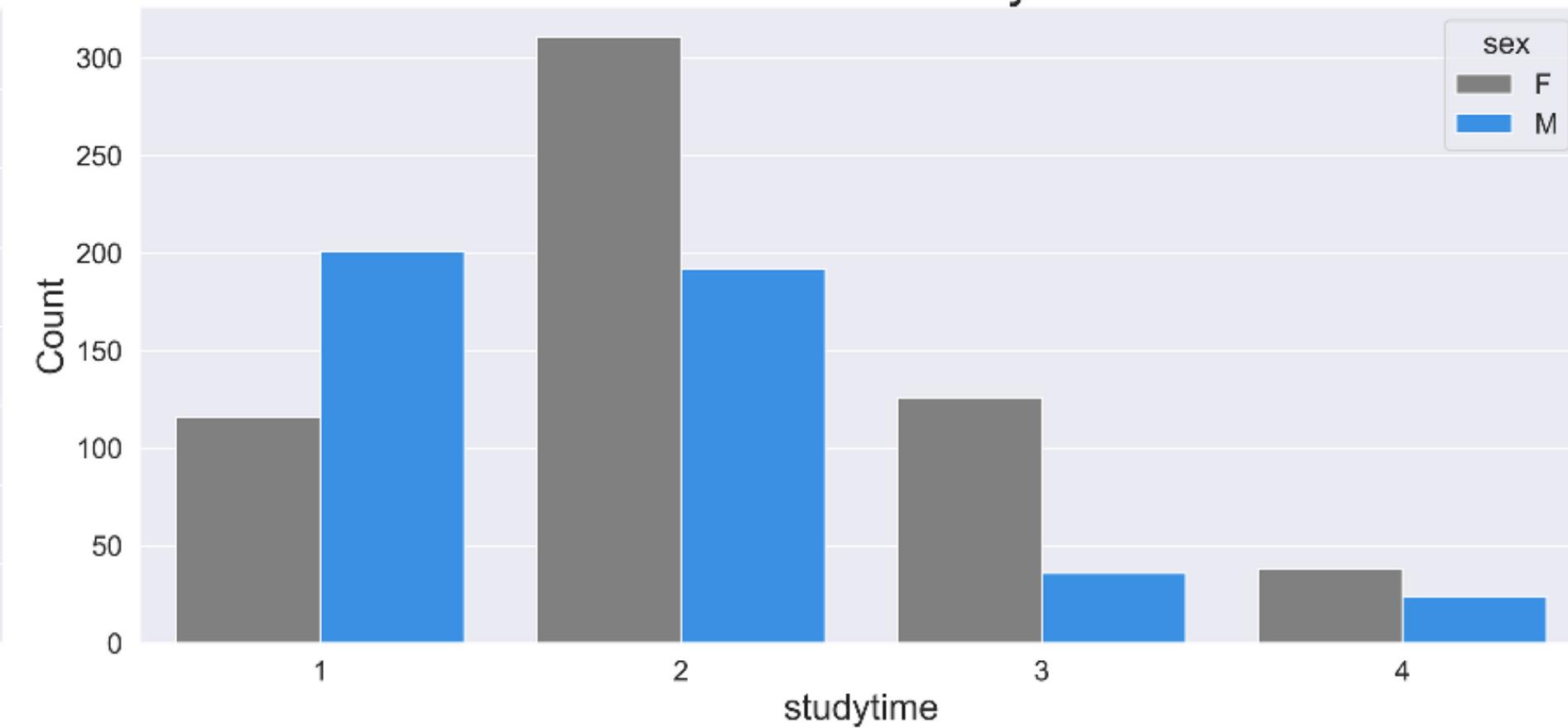
Distribution of traveltimes



Distribution of Subject



Distribution of studytime



Distributions According to Gender



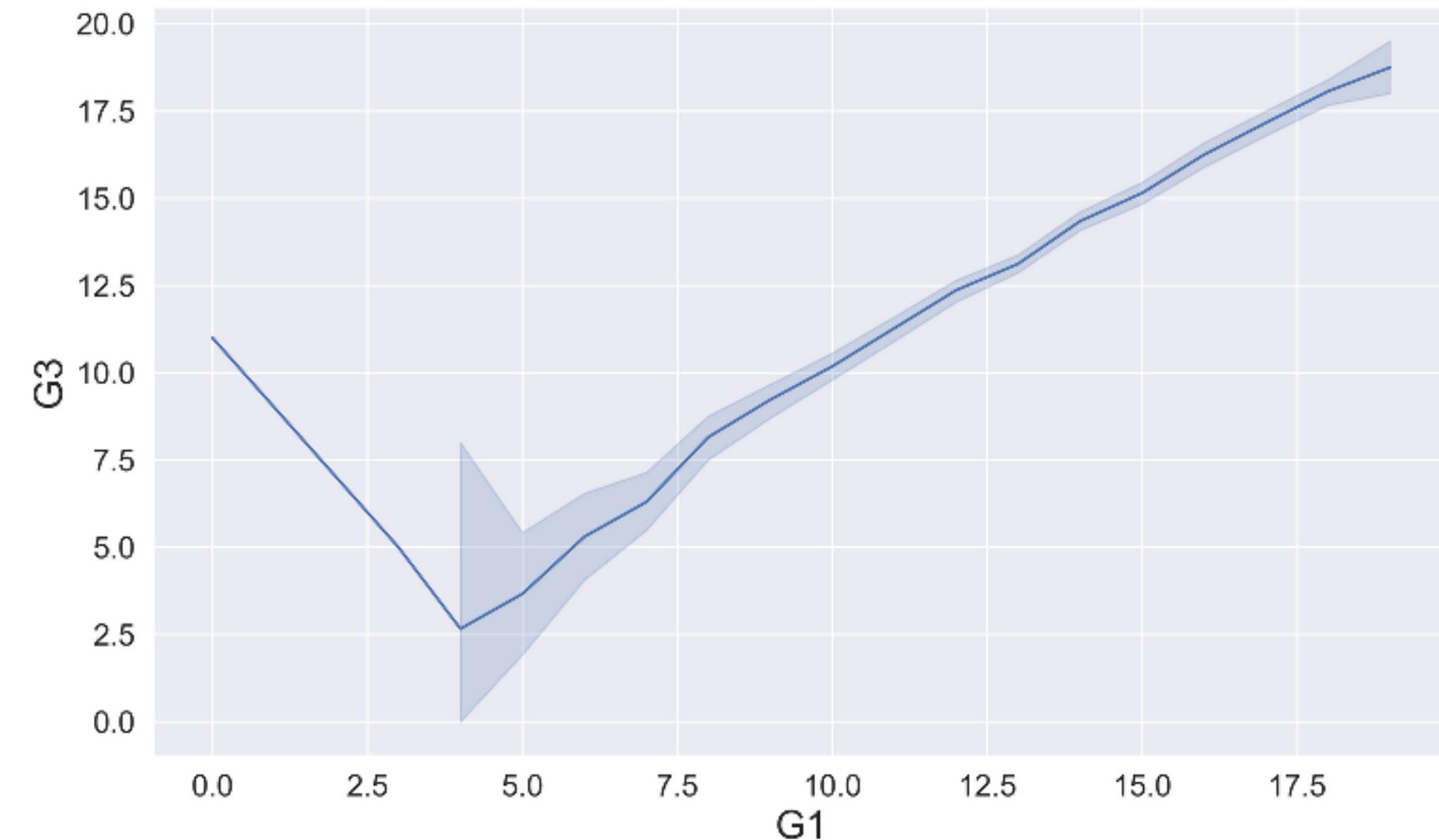
FEATURE SELECTION



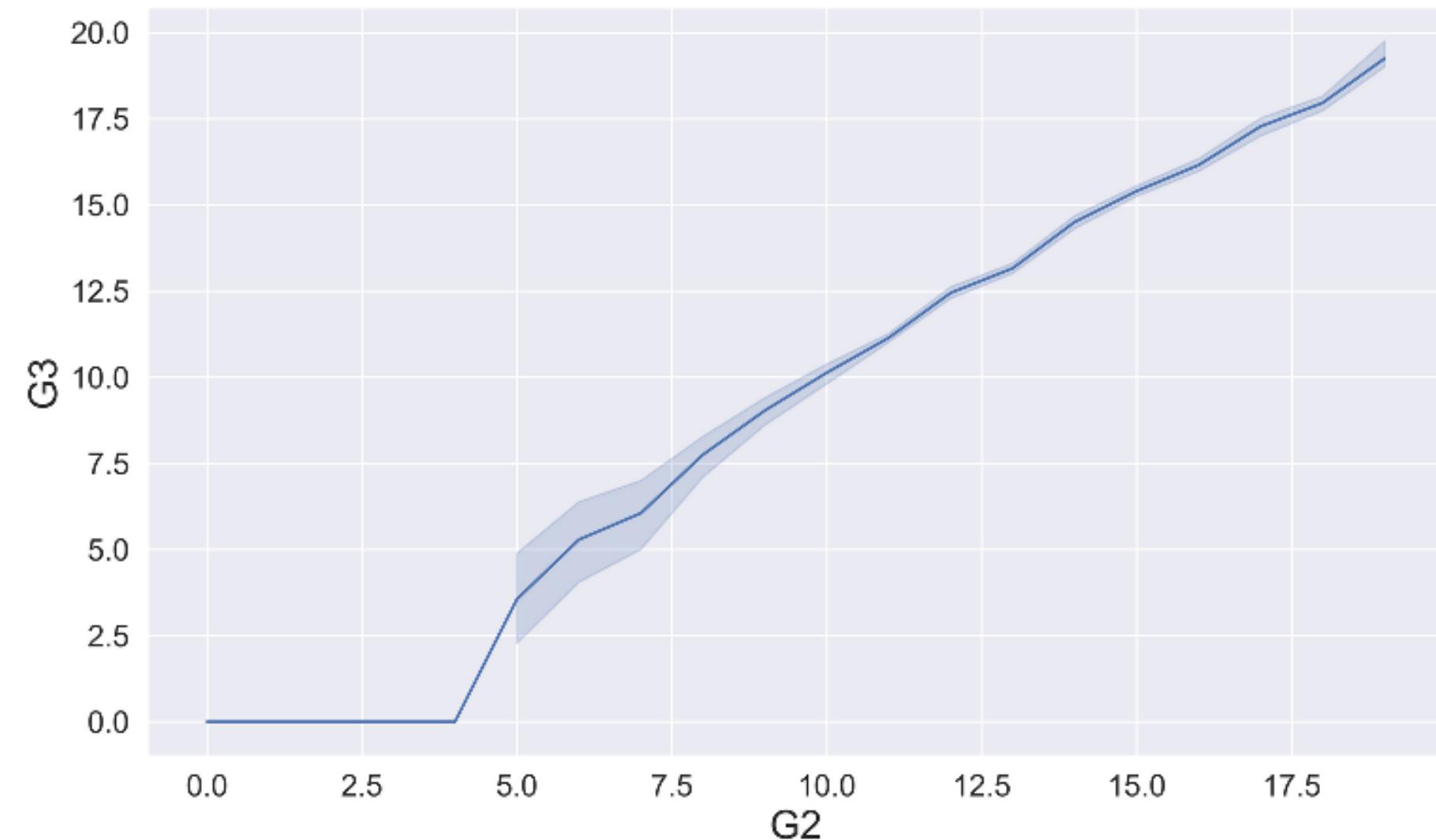
Co-Relation of G3 with Feature G1 and

2

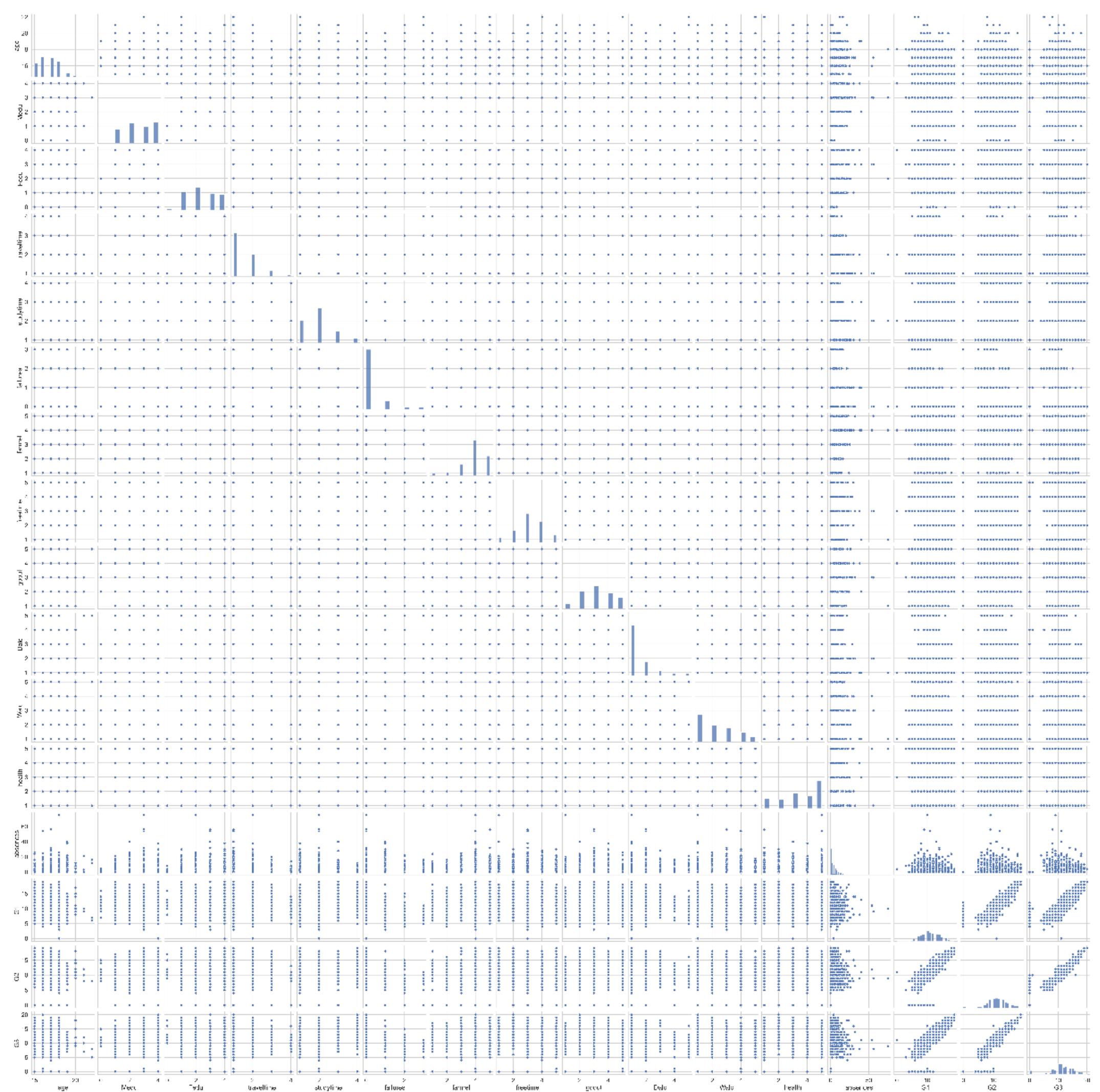
Score G1 vs Score G3



Score G2 vs Score G3



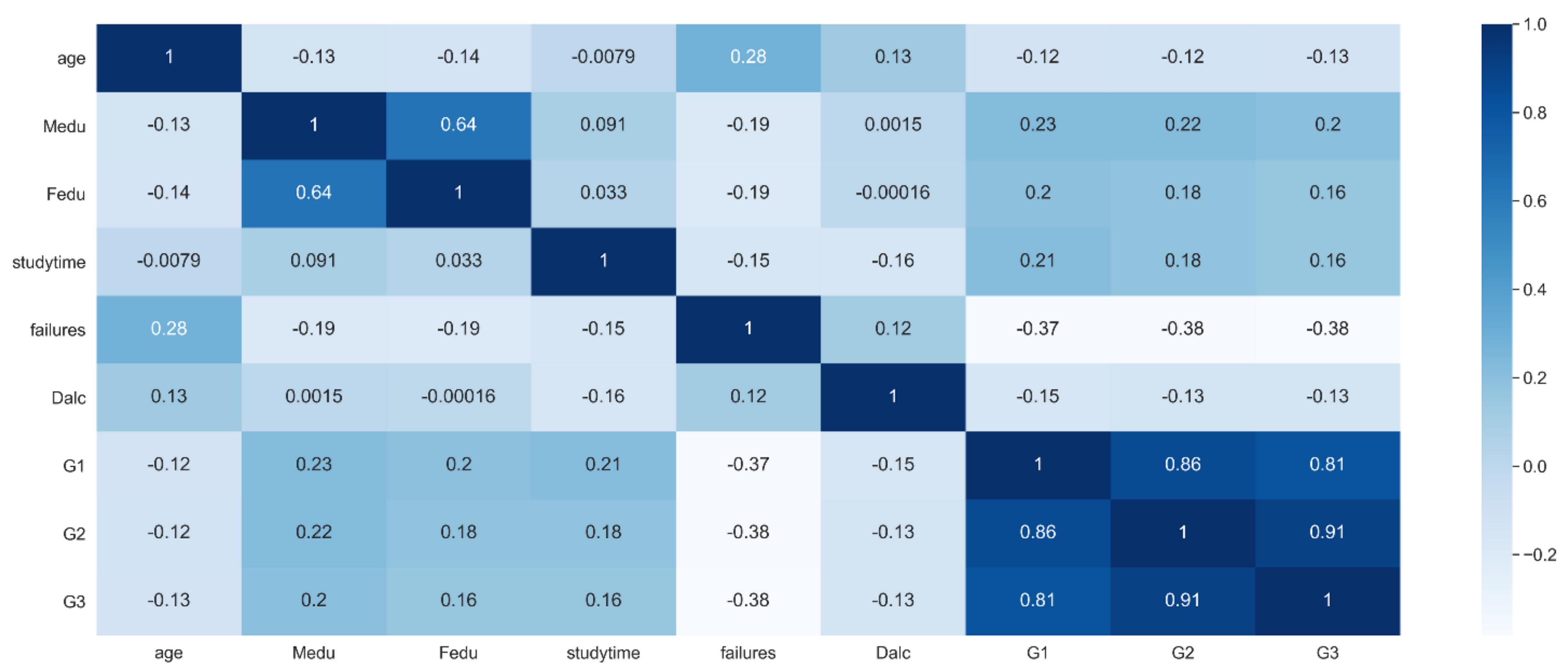
Pair plot with all Feature



A heatmap visualization of a correlation matrix for 15 variables. The variables are listed on both the rows and columns. The color intensity indicates the strength and sign of the correlation coefficient between pairs of variables. A vertical color bar on the right side provides a scale from -1.0 (dark blue) to 1.0 (light yellow).

	age	Medu	Fedu	travelttime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
age	1	-0.13	-0.14	0.049	-0.0079	0.28	0.0072	0.0026	0.12	0.13	0.098	-0.029	0.15	-0.12	-0.12	-0.13
Medu	-0.13	1	0.64	-0.24	0.091	-0.19	0.015	0.0011	0.026	0.0015	-0.029	-0.013	0.06	0.23	0.22	0.2
Fedu	-0.14	0.64	1	-0.2	0.033	-0.19	0.013	0.0021	0.03	-0.00016	0.02	0.034	0.041	0.2	0.18	0.16
travelttime	0.049	-0.24	-0.2	1	-0.081	0.087	-0.013	-0.0074	0.05	0.11	0.084	-0.029	-0.023	-0.12	-0.14	-0.1
studytime	-0.0079	0.091	0.033	-0.081	1	-0.15	0.012	-0.094	-0.073	-0.16	-0.23	-0.063	-0.076	0.21	0.18	0.16
failures	0.28	-0.19	-0.19	0.087	-0.15	1	-0.054	0.1	0.075	0.12	0.11	0.048	0.1	-0.37	-0.38	-0.38
famrel	0.0072	0.015	0.013	-0.013	0.012	-0.054	1	0.14	0.081	-0.076	-0.1	0.1	-0.062	0.037	0.042	0.054
freetime	0.0026	0.0011	0.0021	-0.0074	-0.094	0.1	0.14	1	0.32	0.14	0.13	0.082	-0.032	-0.052	-0.069	-0.065
goout	0.12	0.026	0.03	0.05	-0.073	0.075	0.081	0.32	1	0.25	0.4	-0.014	0.056	-0.1	-0.11	-0.098
Dalc	0.13	0.0015	-0.00016	0.11	-0.16	0.12	-0.076	0.14	0.25	1	0.63	0.066	0.13	-0.15	-0.13	-0.13
Walc	0.098	-0.029	0.02	0.084	-0.23	0.11	-0.1	0.13	0.4	0.63	1	0.11	0.14	-0.14	-0.13	-0.12
health	-0.029	-0.013	0.034	-0.029	-0.063	0.048	0.1	0.082	-0.014	0.066	0.11	1	-0.027	-0.06	-0.088	-0.08
absences	0.15	0.06	0.041	-0.023	-0.076	0.1	-0.062	-0.032	0.056	0.13	0.14	-0.027	1	-0.092	-0.089	-0.046
G1	-0.12	0.23	0.2	-0.12	0.21	-0.37	0.037	-0.052	-0.1	-0.15	-0.14	-0.06	-0.092	1	0.86	0.81
G2	-0.12	0.22	0.18	-0.14	0.18	-0.38	0.042	-0.069	-0.11	-0.13	-0.13	-0.088	-0.089	0.86	1	0.91
G3	-0.13	0.2	0.16	-0.1	0.16	-0.38	0.054	-0.065	-0.098	-0.13	-0.12	-0.08	-0.046	0.81	0.91	1
	age	Medu	Fedu	travelttime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3

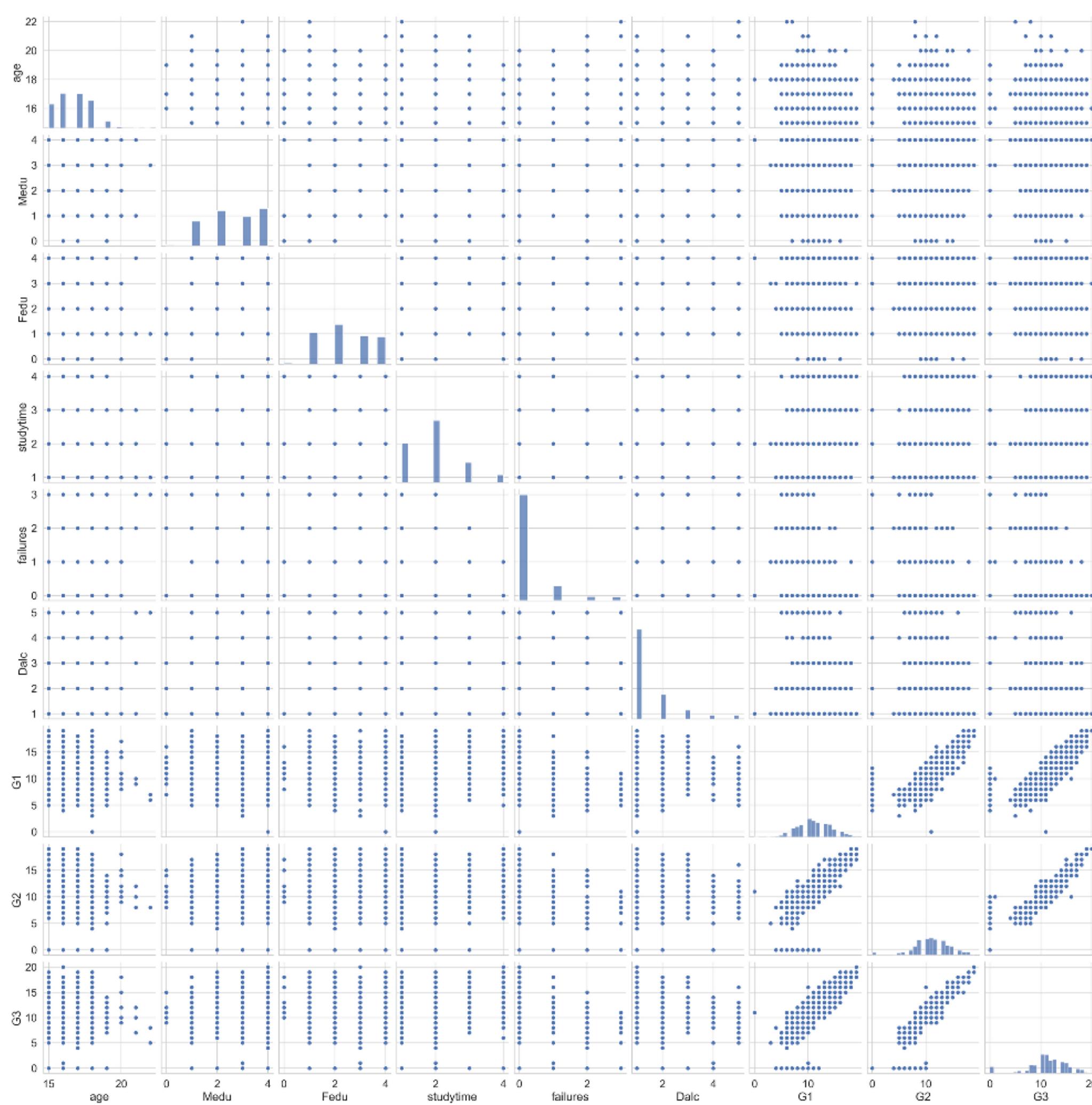
Heat Map For every
feature



Heat map of highly
co related features with
G3

```
fig, ax = plt.subplots(figsize=(26, 10))
mat = student.corr()
top_corr_features = mat.index[abs(mat["G3"])>0.12]
ax = sns.heatmap(student[top_corr_features].corr(), cmap='Blues', annot = True)
plt.show()
```

Pairplot of highly co related features with G3



	age	Medu	Fedu	studytime	failures	Dalc	G1	G2	G3
0	18	4	4		2	0	1	5	6
1	17	1	1		2	0	1	5	6
2	15	1	1		2	3	2	7	10
3	15	4	2		3	0	1	15	14
4	16	3	3		2	0	1	6	10
...
644	19	2	3		3	1	1	10	11
645	18	3	1		2	0	1	15	15
646	18	1	1		2	0	1	11	12
647	17	3	1		1	0	3	10	10
648	18	3	2		1	0	3	10	11



Training and testing
the model

LINEAR REGRESSION

Regression With All features

MAE : 0.829

MSE : 1.501

Regression score 89.031%

Best Regression

MAE : 0.699

MSE : 1.063

Regression score 91.660%

Regression With Most related features

MAE : 0.801

MSE : 1.5

Regression score 89.043%



school Medu Fedu reason failures schoolsup
Dalc Walc absences G1 G2 Subject

DECISION TREE

DATA MANIPULATION

As Decision tree can not predict continuous value so we add new column grade which calculates grade accordings their pervious score.

GRADE

A- Best Performance	>0.9 of max
B- Good Performance	>0.7 of max
C- Keep Going	>0.4 of max
D- Bad Performance	<0.4 of max

Subject	G1	G2	G3	GAvg	grades
0	0	5	6	5.666667	D
1	0	5	5	5.333333	D
2	0	7	8	8.333333	C
3	0	15	14	14.666667	B
4	0	6	10	8.666667	C
...
644	1	10	11	10	10.333333
645	1	15	15	16	15.333333
646	1	11	12	9	10.666667
647	1	10	10	10	10.000000
648	1	10	11	11	10.666667

DECISION TREE

Accuracy: 89.80891719745223

confusion_matrix:

$$\begin{bmatrix} 7 & 1 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 3 & 65 & 3 & 0 \end{bmatrix}$$

[0 7180 8]

[0 0 10 30]

Classification report :

precision recall f1-score support

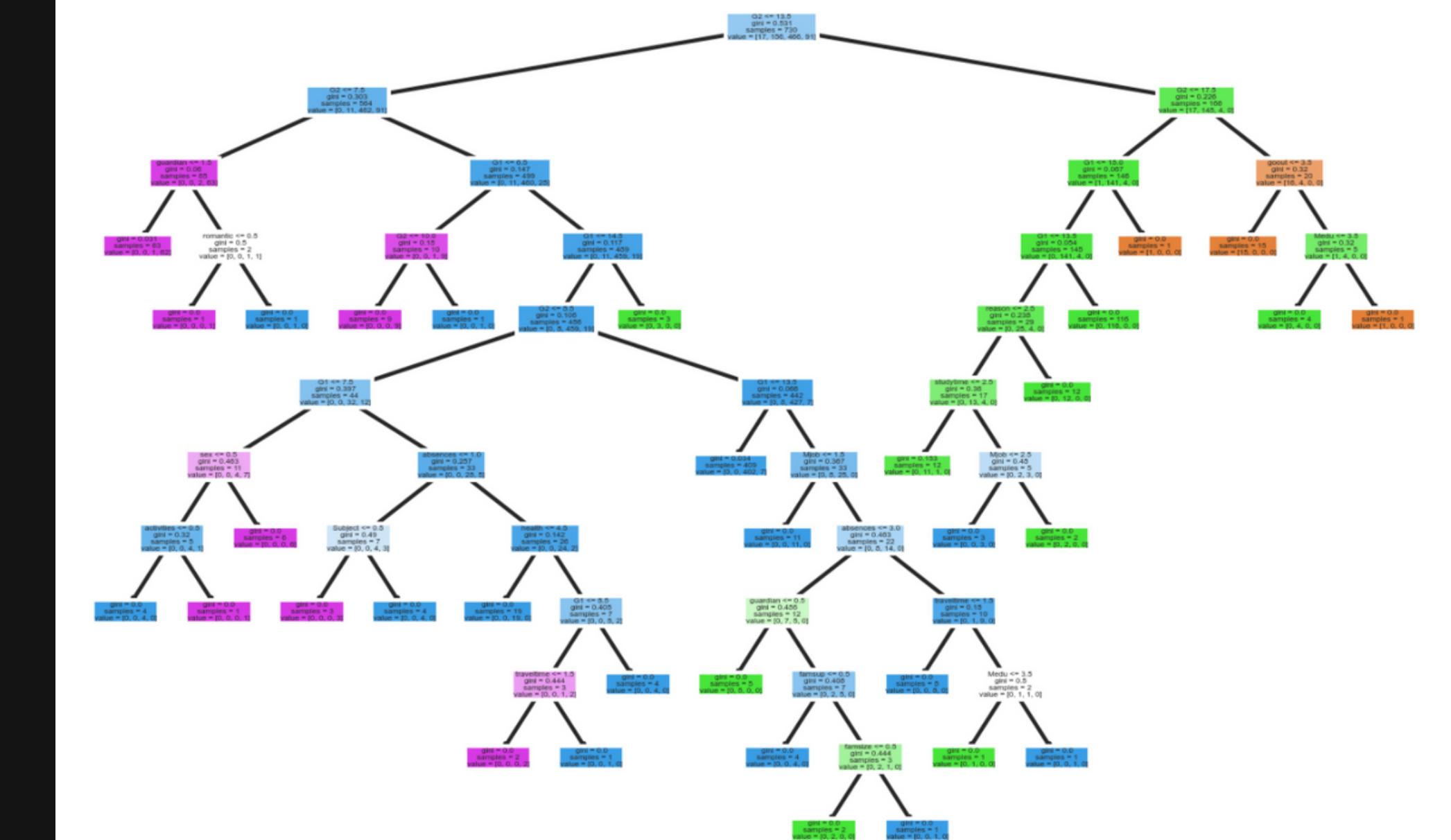
A 0.70 0.88 0.78

B 0.89 0.92 0.90

C 0.93 0.92 0.93 19

D 0.79 0.75 0.77 /

WITH ALL FEATURES



DECISION TREE

Accuracy: 63.05732484076433

confusion_matrix:

$$\begin{bmatrix} 0 & 4 & 4 & 0 \end{bmatrix}$$

[0 22 49 0]

[1 24 165 5]

[0 3 26 11]]

Classification report :

precision recall f1-score support

A 0.00 0.00 0.00 8

B 0.42 0.31 0.35 7

C 0.68 0.85 0.75 195

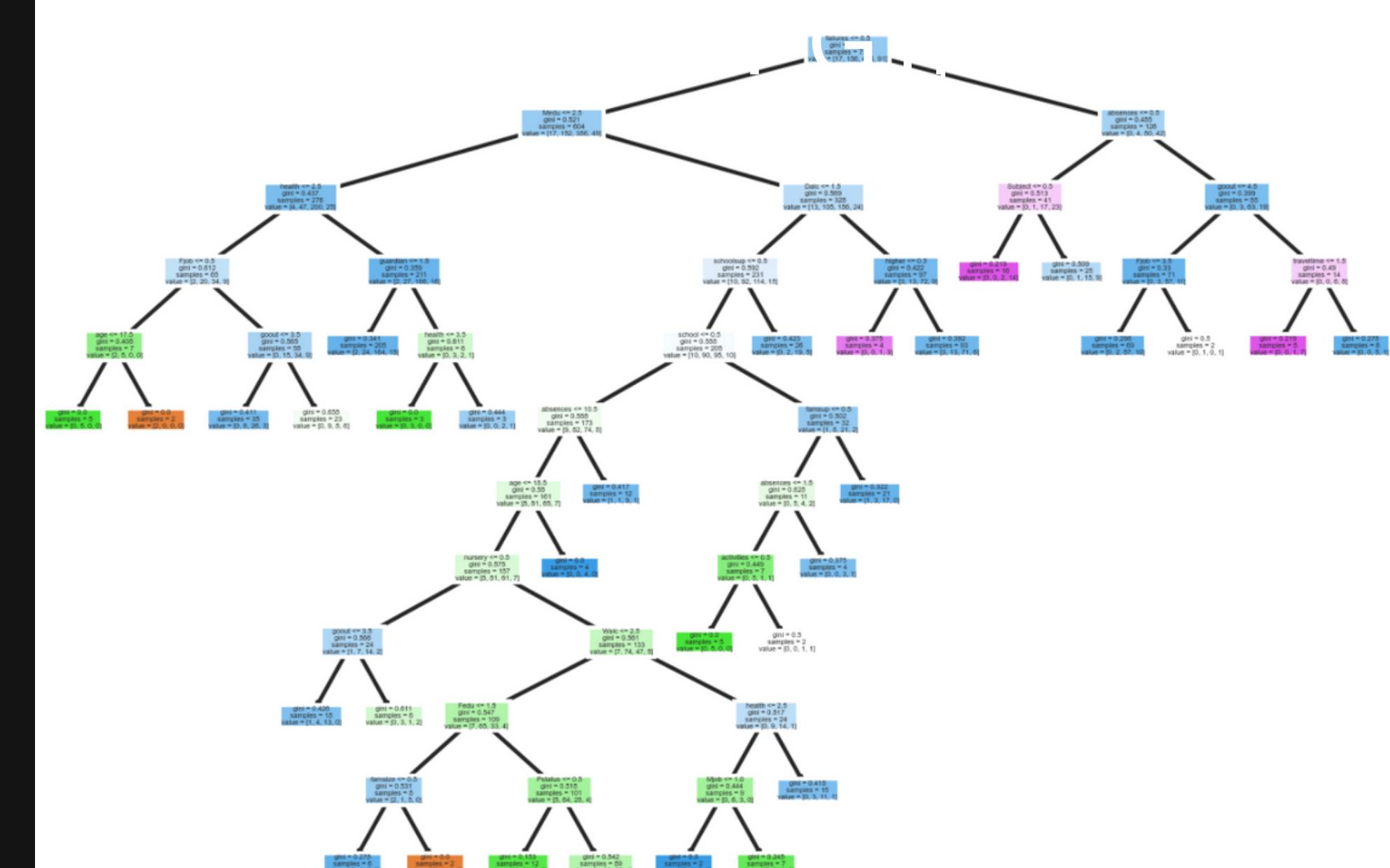
D 0.69 0.28 0.39 4

accuracy

063

314

WITH ALL FEATURES



DECISION TREE

Accuracy: 94.5859872611465

confusion_matrix:

```
[[ 7  1  0  0]
 [ 3 66  2  0]
 [ 0 4187  4]
 [ 0  0  3 37]]
```

Classification report :

	precision	recall	f1-score	support
0	0.70	0.88	0.78	8
1	0.93	0.93	0.93	71
2	0.97	0.96	0.97	195
3	0.90	0.93	0.91	40

accuracy

0.95

314

WITH MOST RELATED FEATURES



DECISION TREE

Accuracy: 62.101910828025474

confusion_matrix:

```
[[ 0  5  3  0]
```

```
[ 0 27 44  0]
```

```
[ 0 32 158  5]
```

```
[ 0  3 27 10]]
```

Classification report :

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

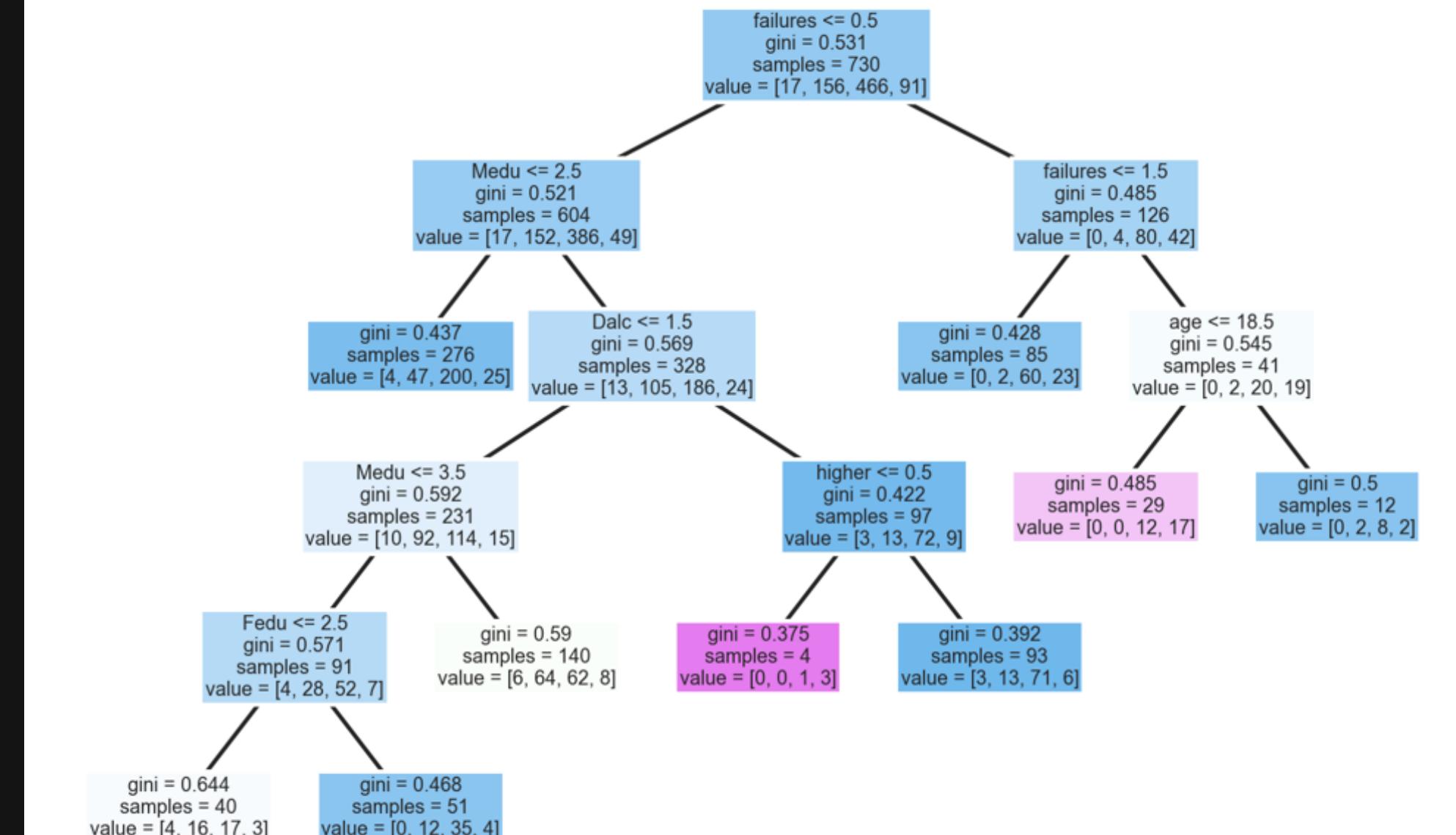
0	0.00	0.00	0.00	8
---	------	------	------	---

1	0.40	0.38	0.39	71
---	------	------	------	----

2	0.68	0.81	0.74	195
---	------	------	------	-----

3	0.67	0.25	0.36	40
---	------	------	------	----

WITH MOST RELATED
FEATURES WITHOUT G1 & G2



DECISION TREE

Accuracy: 95.21531100478468

confusion_matrix:

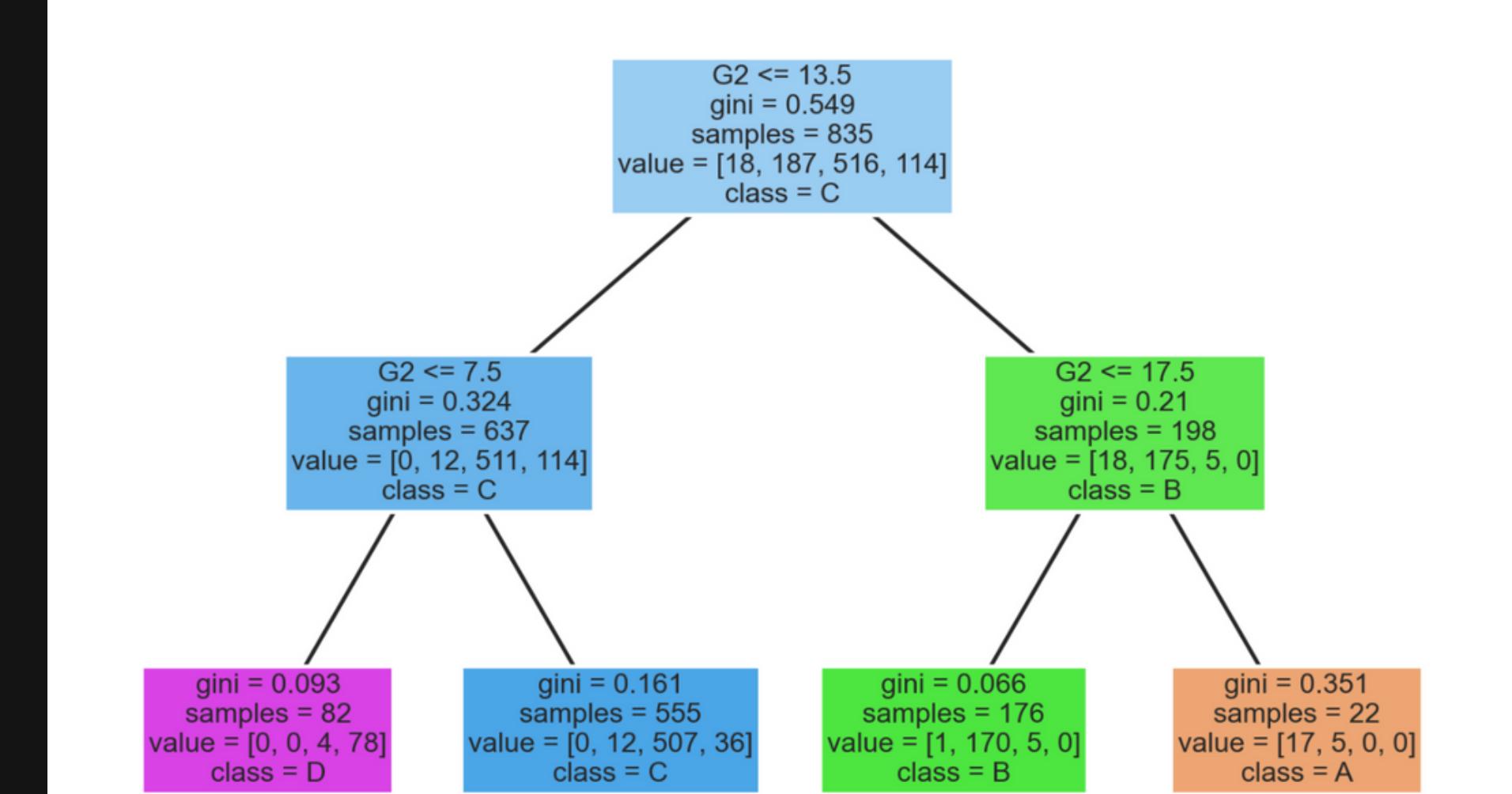
```
[[ 6  1  0  0]
 [ 2 34  4  0]
 [ 0 1144  0]
 [ 0  0  2 15]]
```

Classification report :

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.75	0.86	0.80	7
1	0.94	0.85	0.89	40
2	0.96	0.99	0.98	145
3	1.00	0.88	0.94	17

WITH MOST RELATED
FEATURES WITH LOWEST HEIGHT



LOGISTIC REGRESSION

Using Feature selection Technique

with Chi2

```
from sklearn.feature_selection import SelectKBest, chi2
```

Best Regression

MAE : 0.024

MSE : 0.024

Regression score 97.607%

school Medu Fedu reason failures schoolsup

Dalc Walc absences G1 G2 Subject

confusion matrix:

```
[[ 7  0  0  0]
 [ 1 34  5  0]
 [ 0  2 141  2]
 [ 0  0  0 17]]
```

Classification report :

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

A	0.88	1.00	0.93	7
---	------	------	------	---

B	0.94	0.92	0.89	40
---	------	------	------	----

C	0.97	0.92	0.97	145
---	------	------	------	-----

D	0.89	0.75	0.94	17
---	------	------	------	----

accuracy

0.95 209

SUPPORT VECTOR MACHINE

GRIDSEARCH

by this tuning technique we can to compute the optimum values of hyper parameters.

```
param_grid = {  
    'C': [0.1, 1, 10, 100, 1000],  
    'gamma': [1, 0.1, 0.01, 0.001, 0.0001],  
    'kernel': ['poly', 'rbf', 'sigmoid', 'linear']  
}
```

Regression With All features

```
c      : 010  
gamma : 0.01  
kernel : rbf  
accuracy: 0.945859872611465
```

	precision	recall	f1-score	support
0	0.80	0.44	0.57	9
1	0.88	0.90	0.89	67
2	0.96	0.99	0.97	210
3	1.00	0.93	0.96	28
accuracy				0.95
macro avg				0.81
				314

Thank You
