MACHINE LEARNING

19BCE298
MAULIK VIRPARIYA

Data Visulation

Feature Selection

Training and testing the model

DataSet

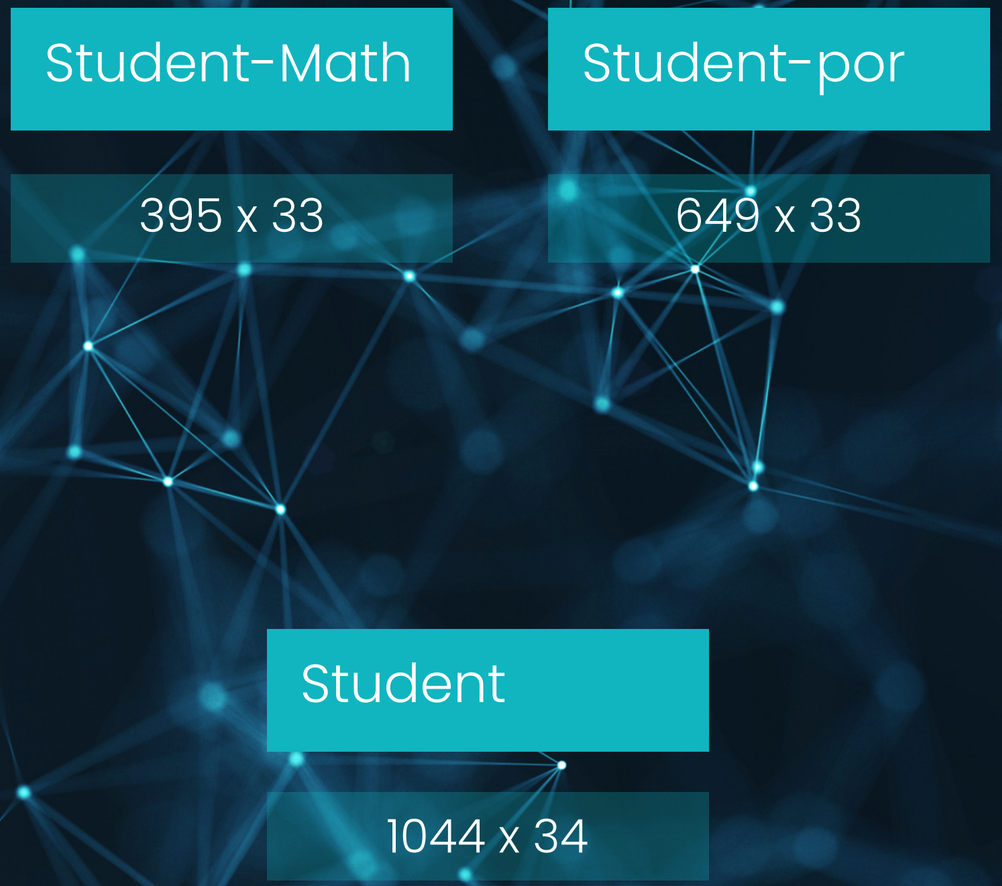
Student Performance

Characteristic	Type	Attribute Characteristic
Multivarite	Regression, Classification	Integer

There are two Dataset so first we marge both

Added new feature named subject :

Math
Portuguese



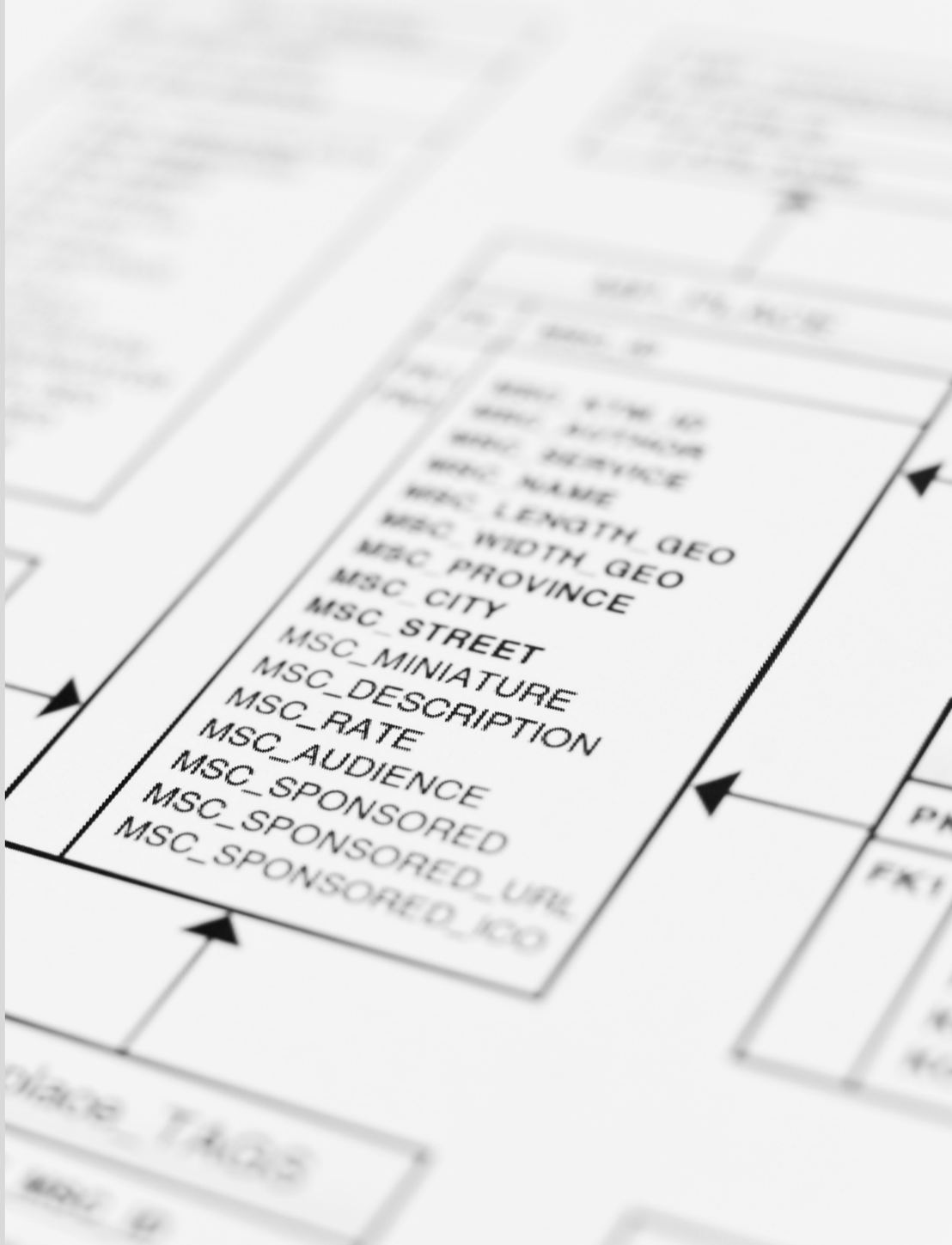
FEATURES

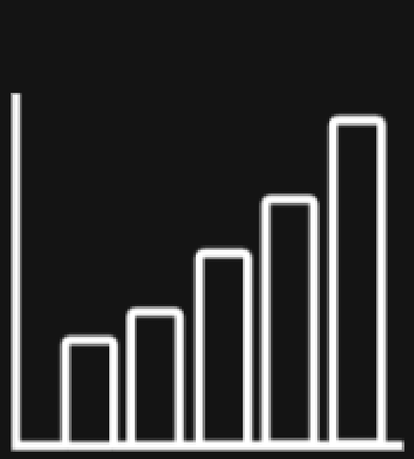
There are two dataset. Both have 29 same features. One has subject Math and one has Portuguese.

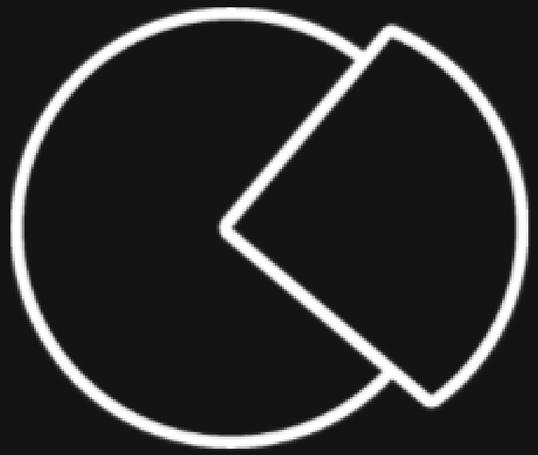
Math data set Contains 395 rows (instances)

Portuguese data set Contains 649 rows (instances)

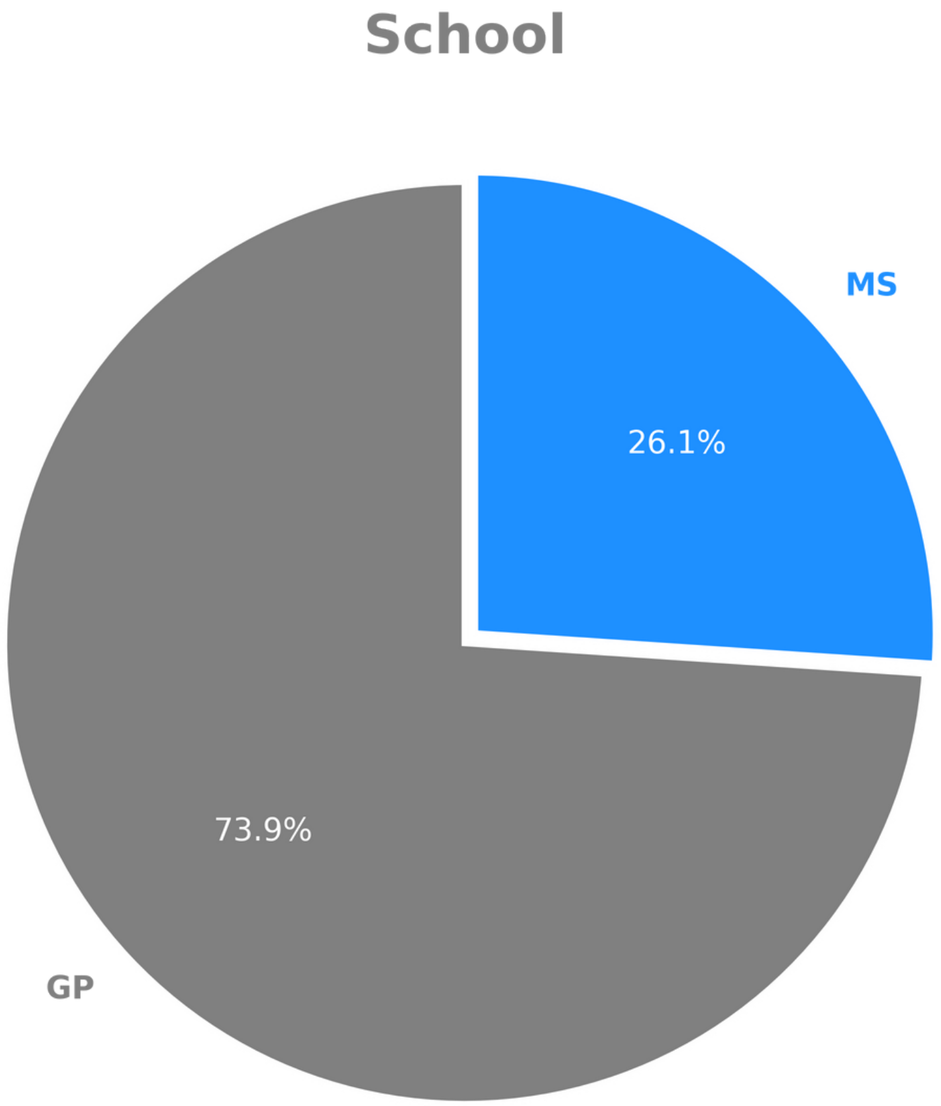
1 school - student's school
2 sex - student's sex
3 age - student's age
4 address - student's home address type
5 famsize - family size
6 Pstatus - parent's cohabitation status
7 Medu - mother's education
8 Fedu - father's education
9 Mjob - mother's job
10 Fjob - father's job
11 reason - reason to choose 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)



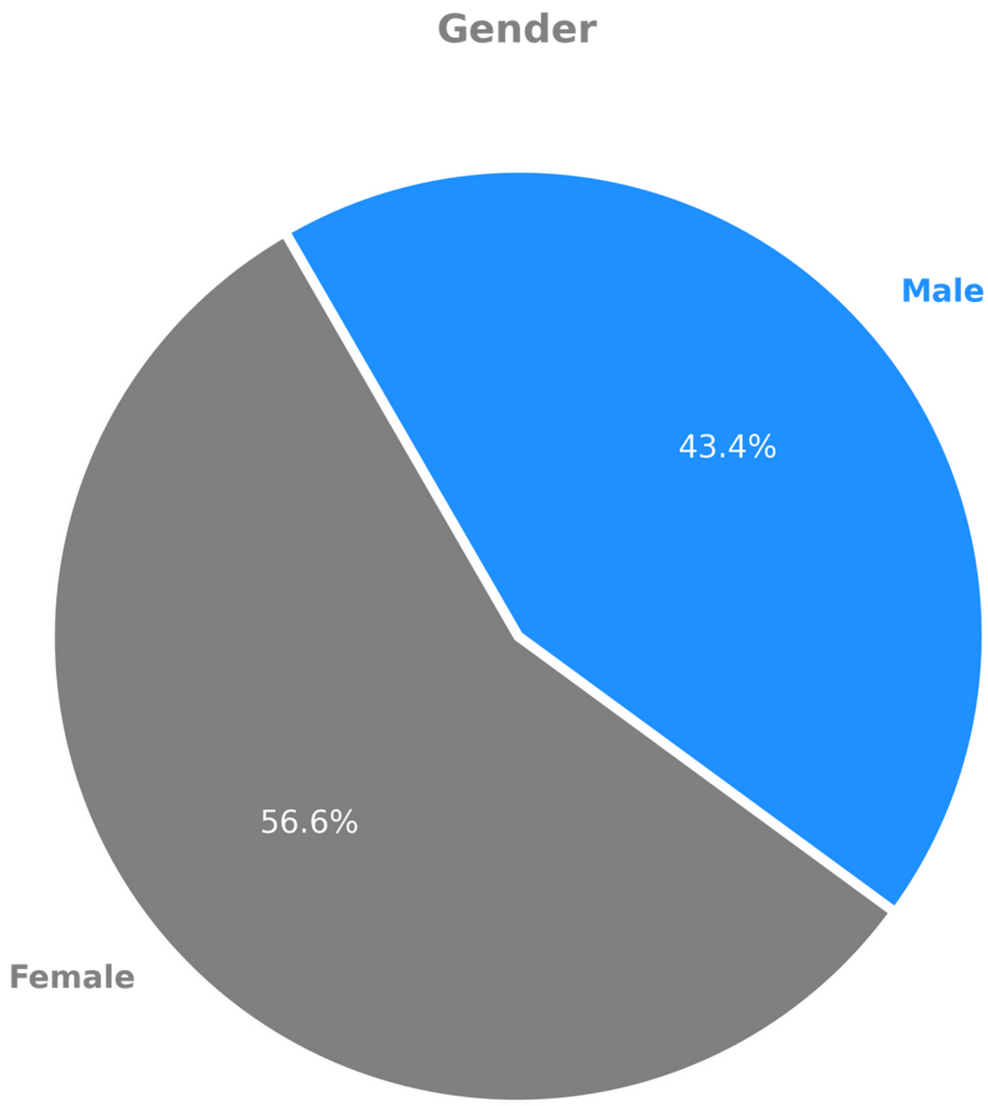




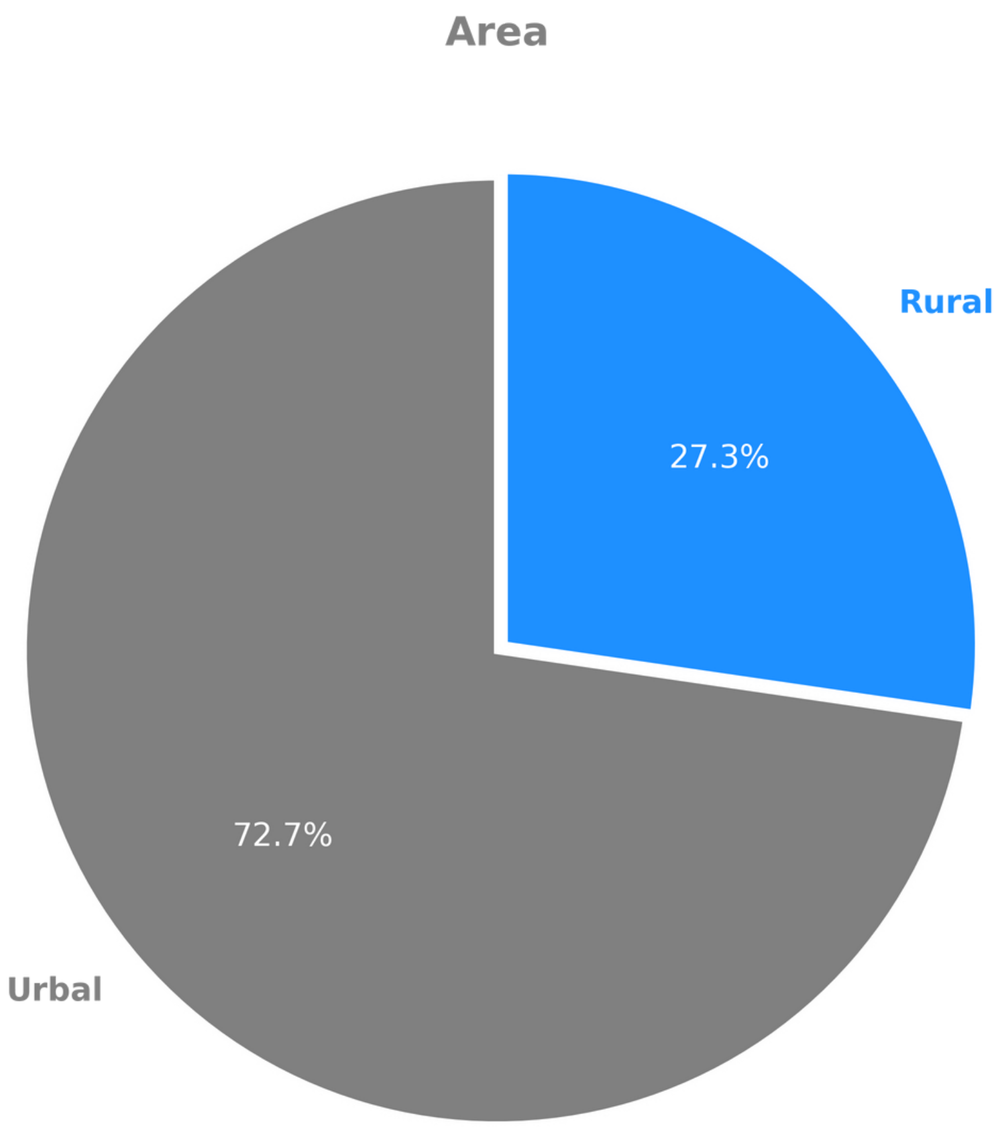
DATA VISUALIZATION



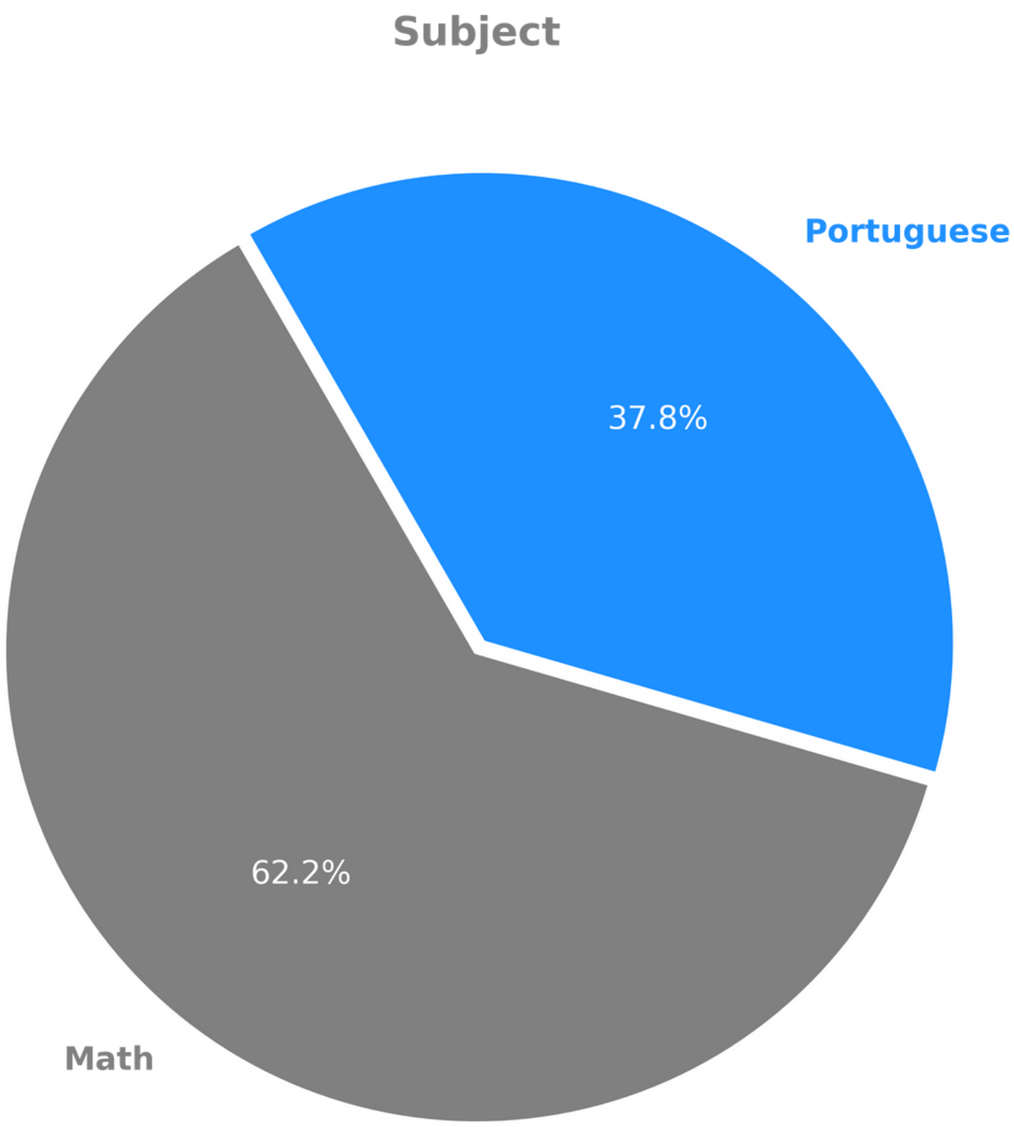
Student count based on school



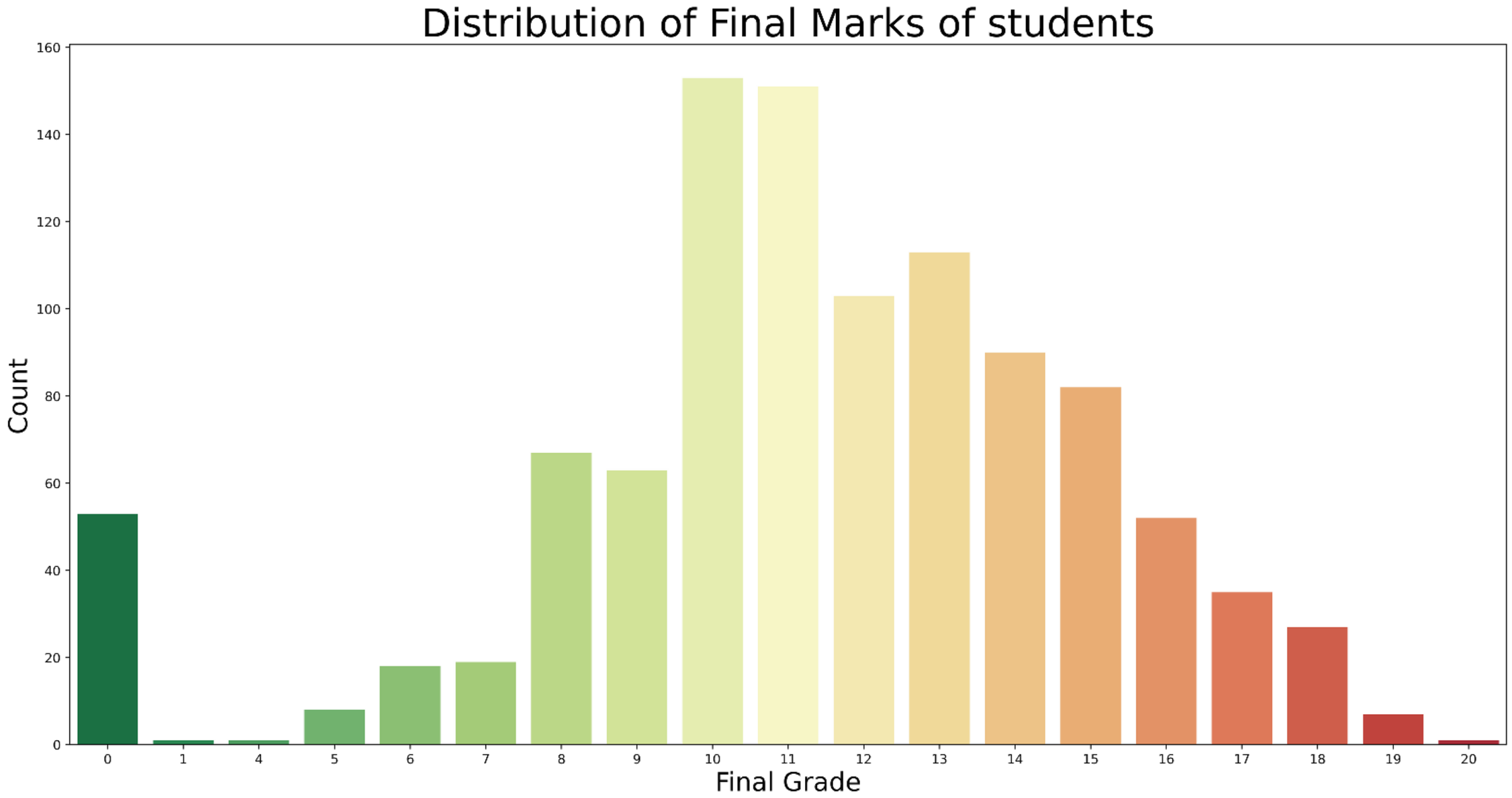
Student count based on Gender



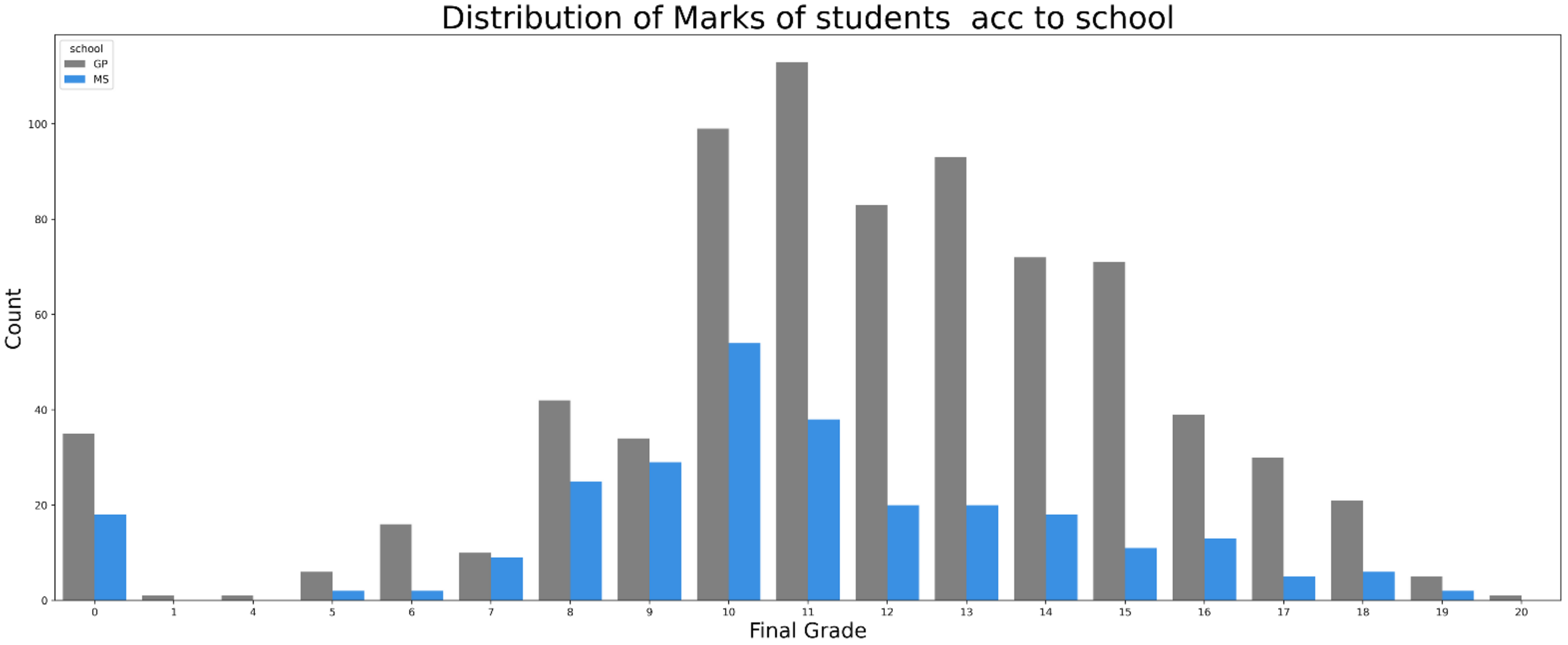
Student count based on Area



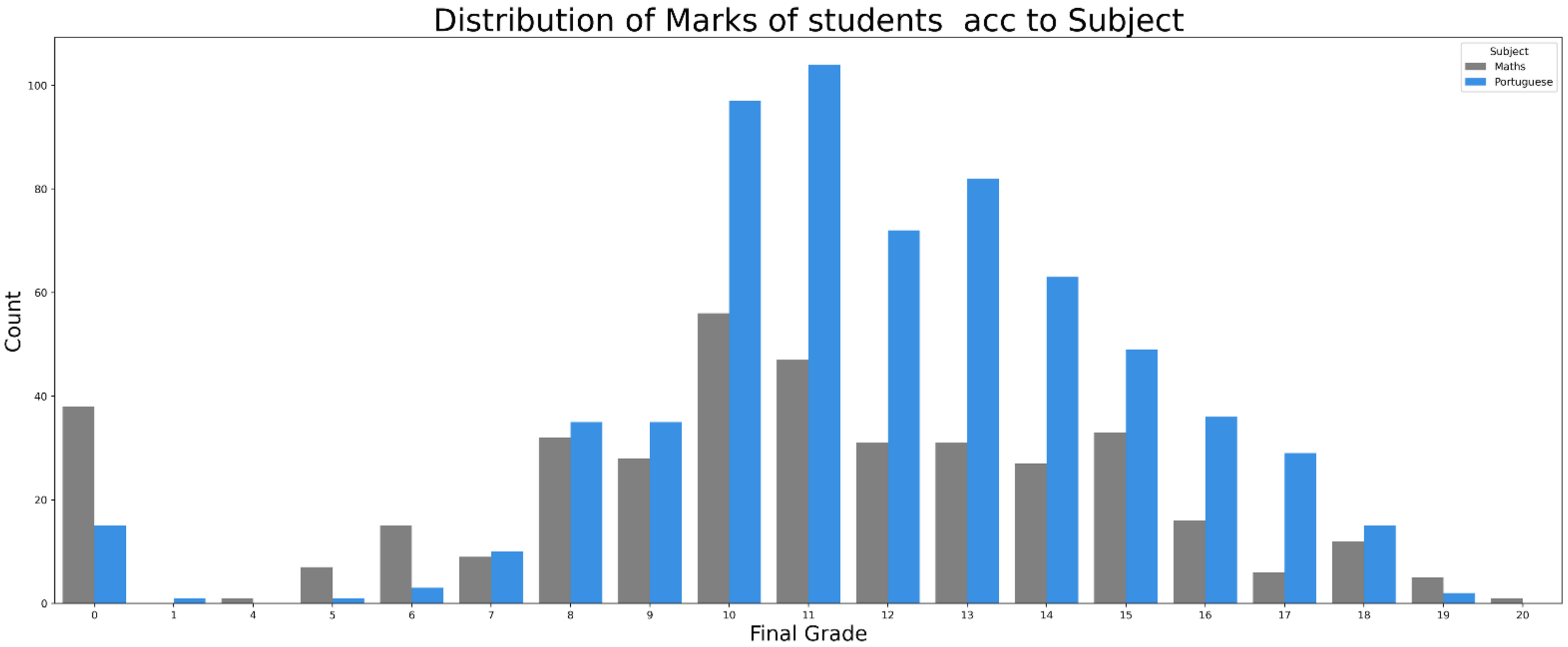
Student count based on Selected



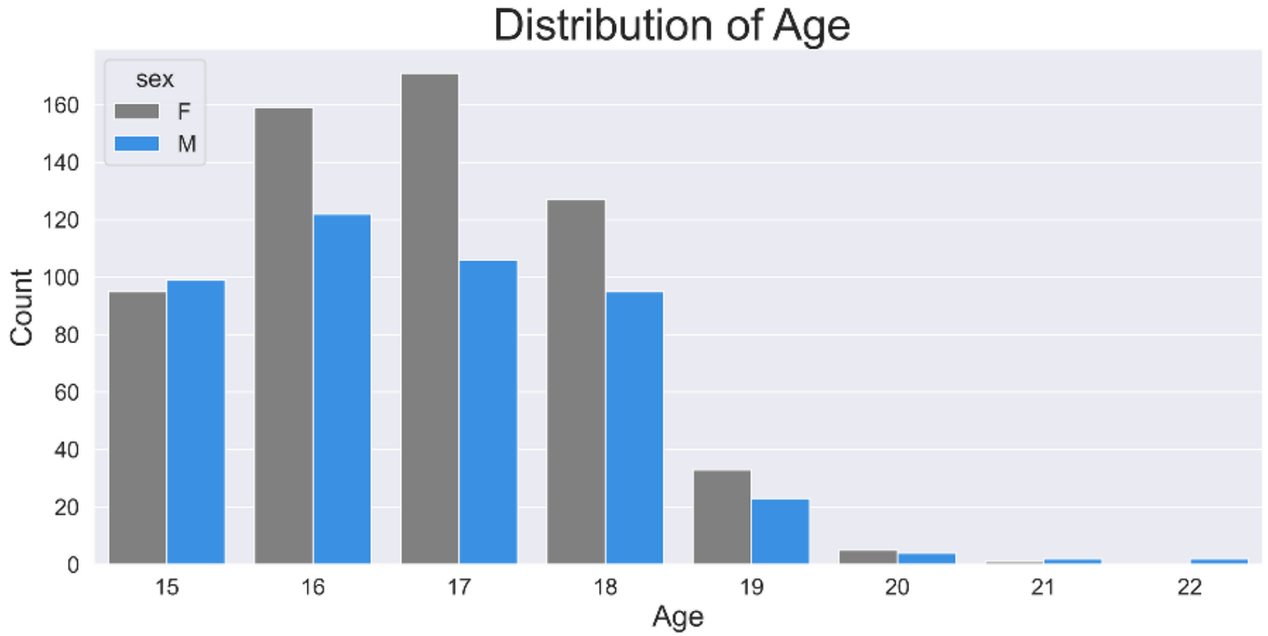
Distributions According to their Marks in G3

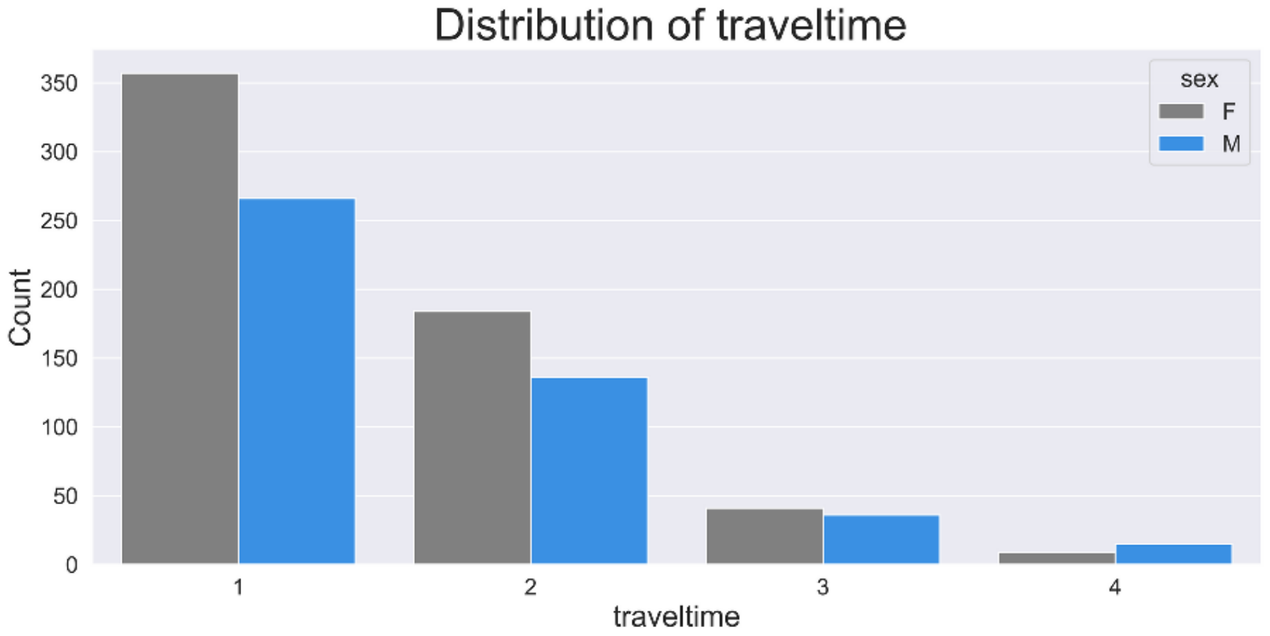


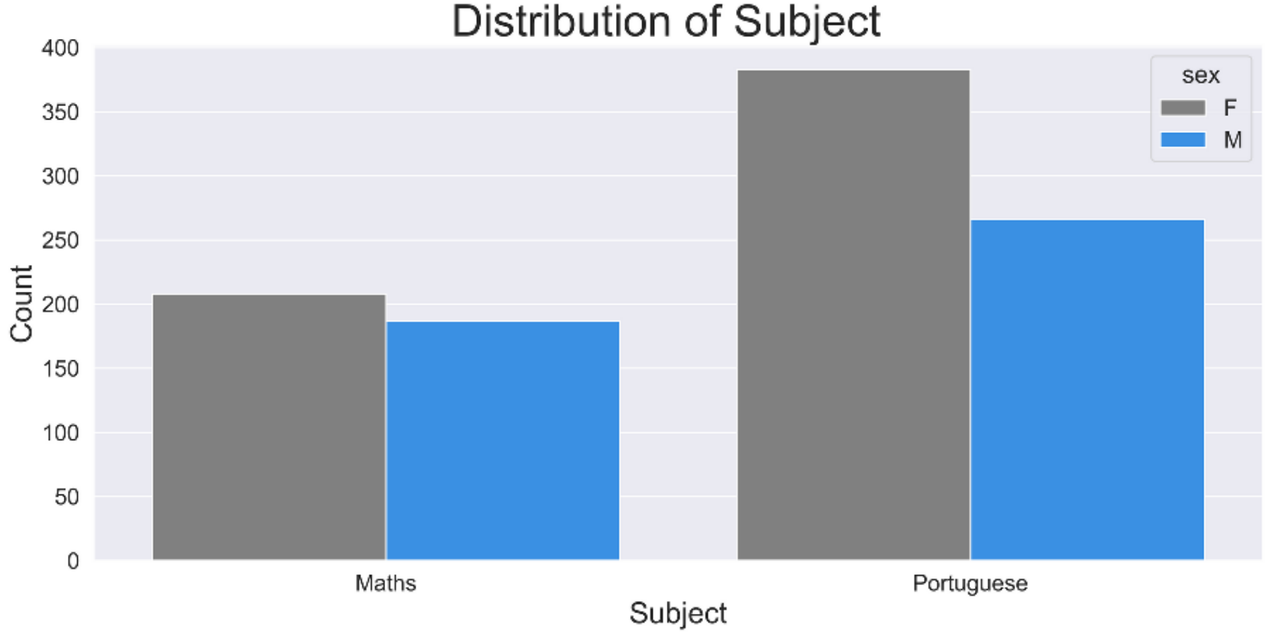
Distributions According to their School

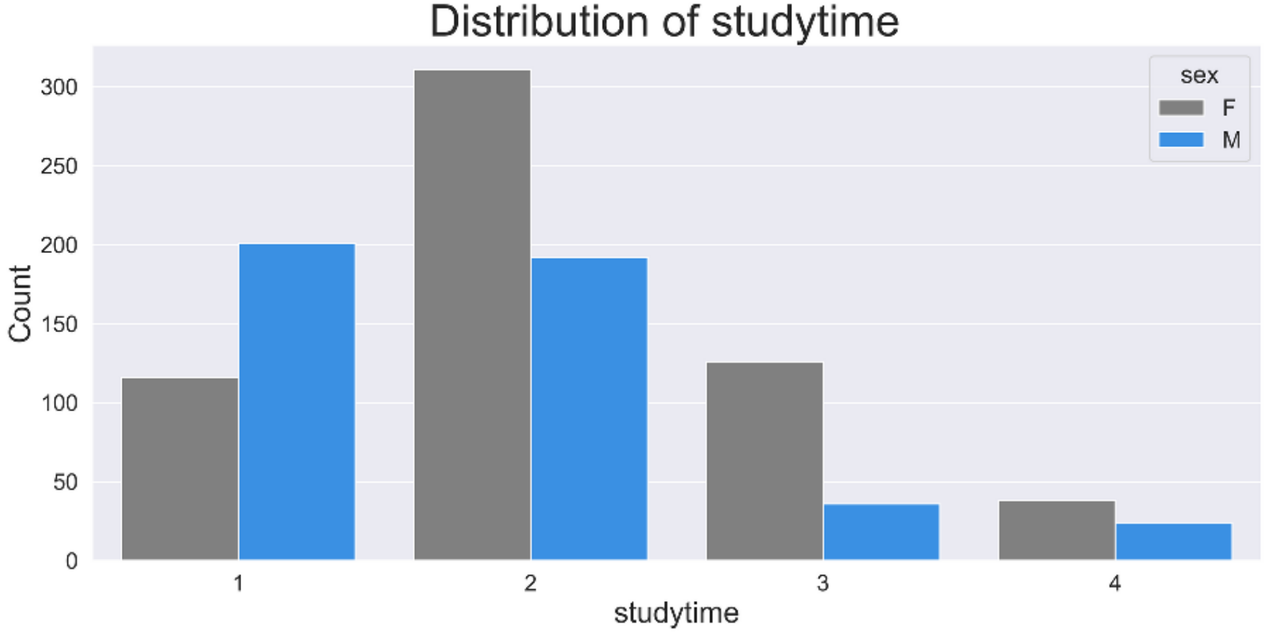


Distributions According to their Subject





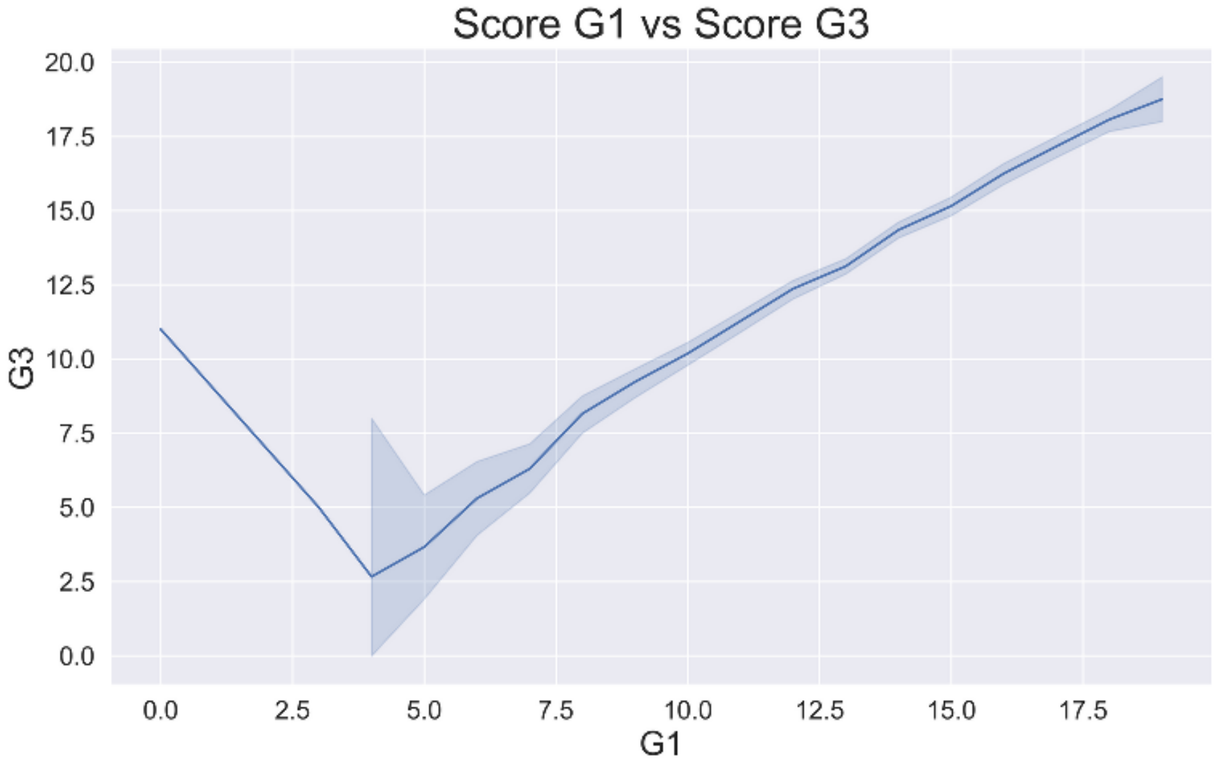


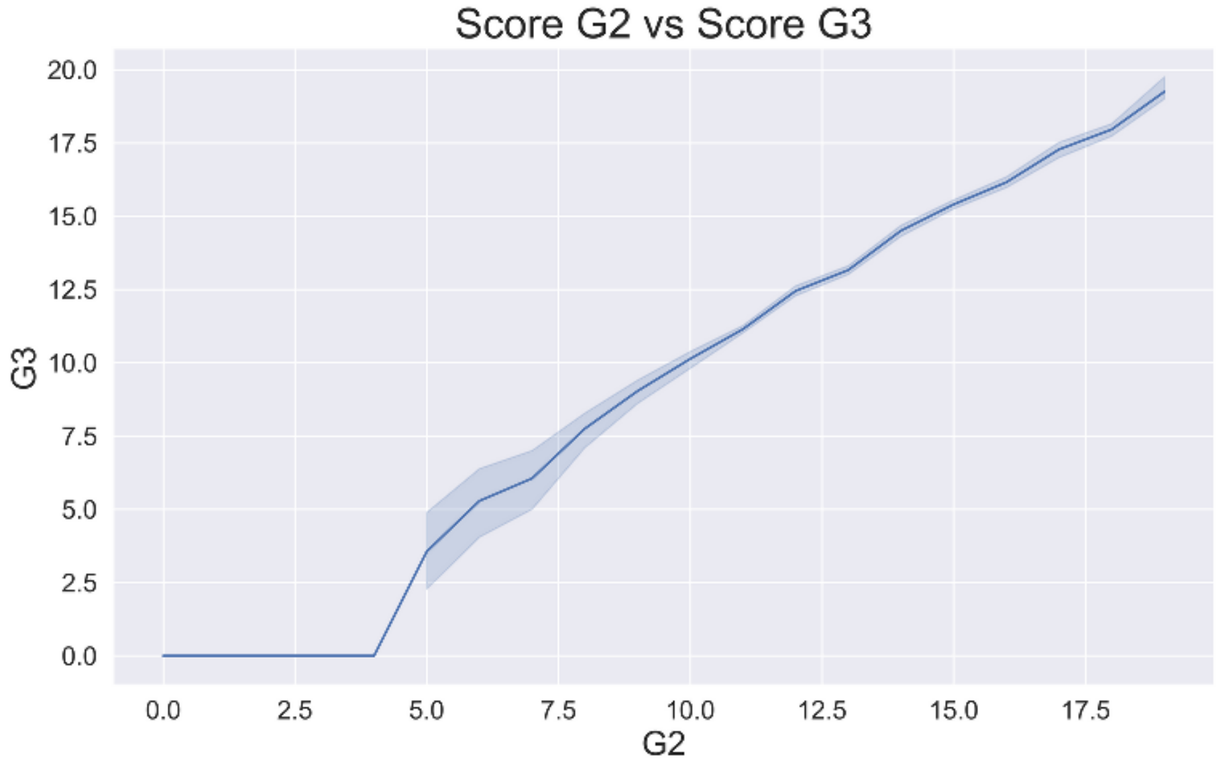


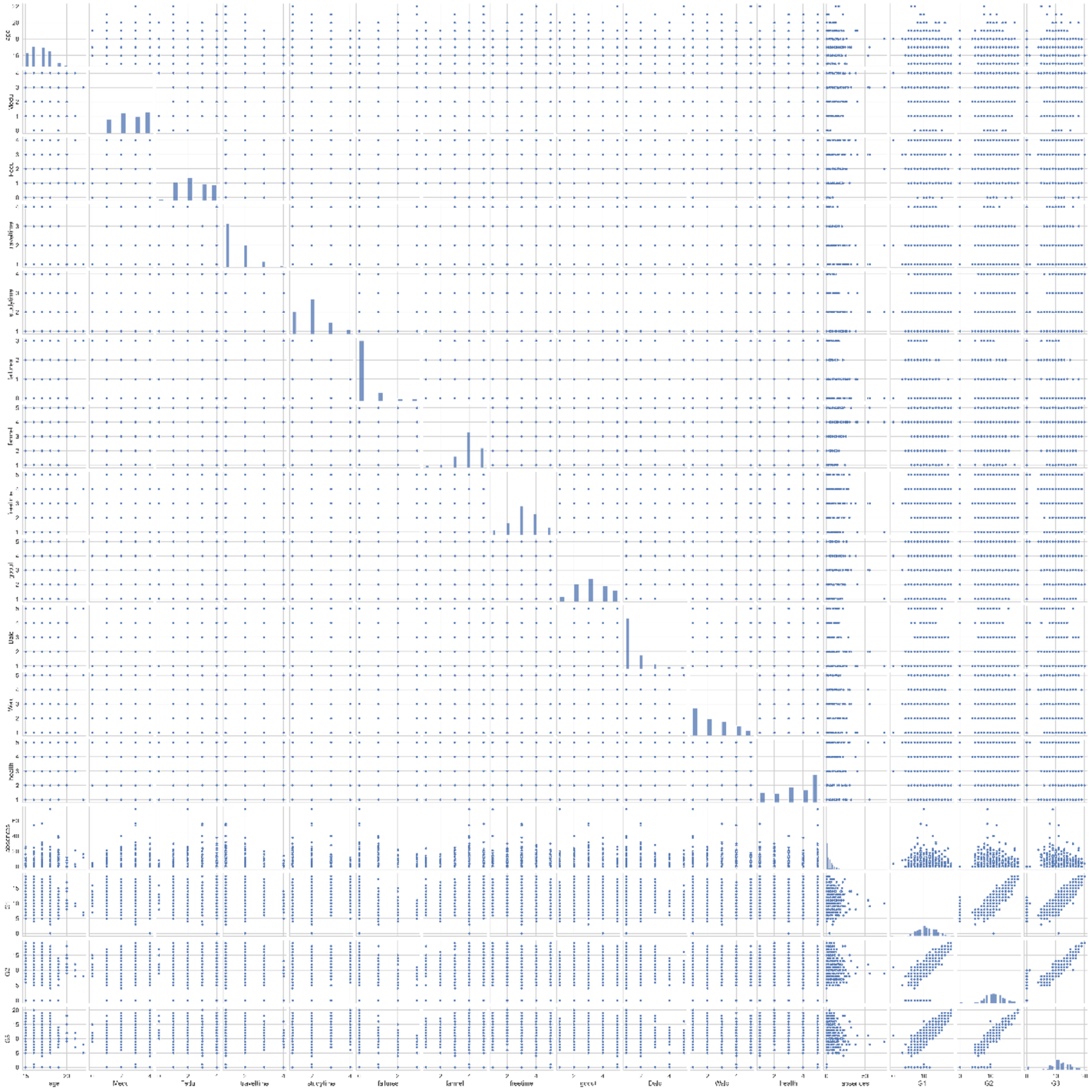
Distributions According to Gender

FEATURE SELECTION

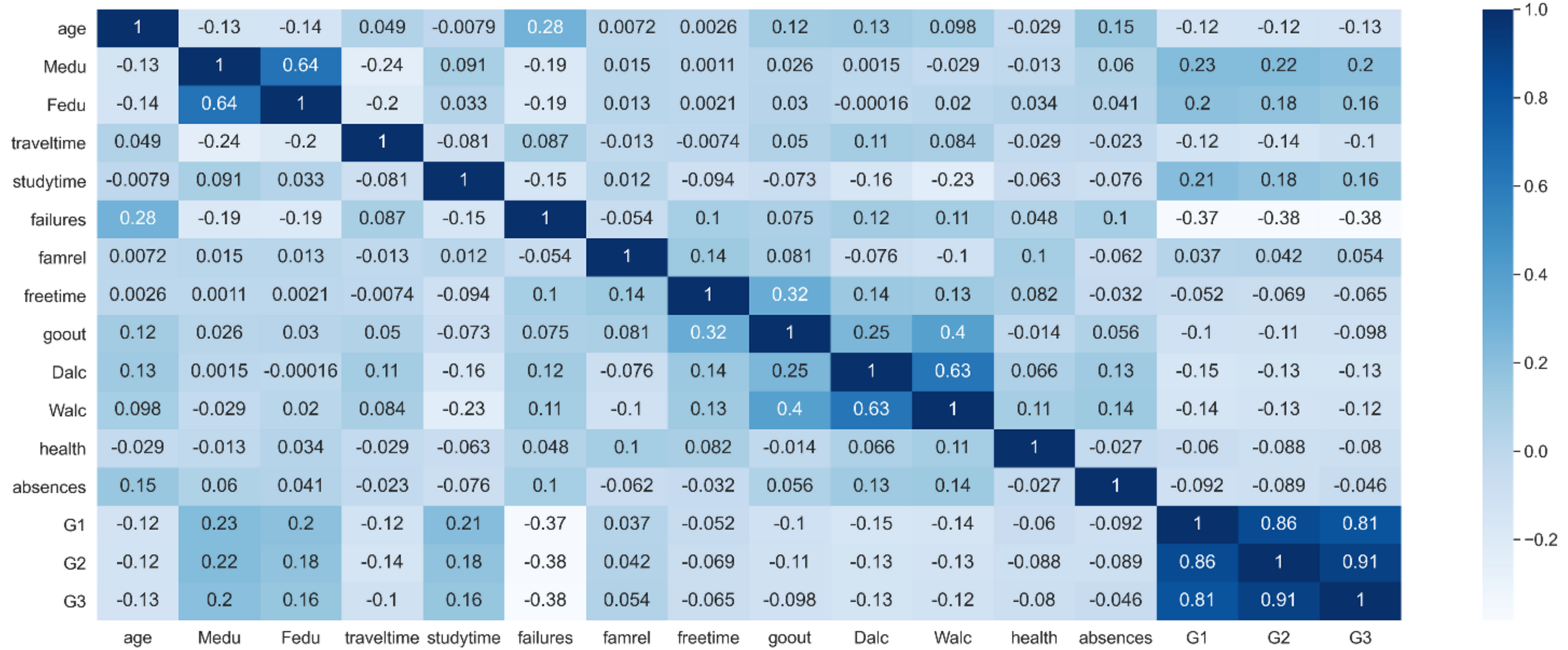
Co-Relation of G3 with Feature G1 and 2







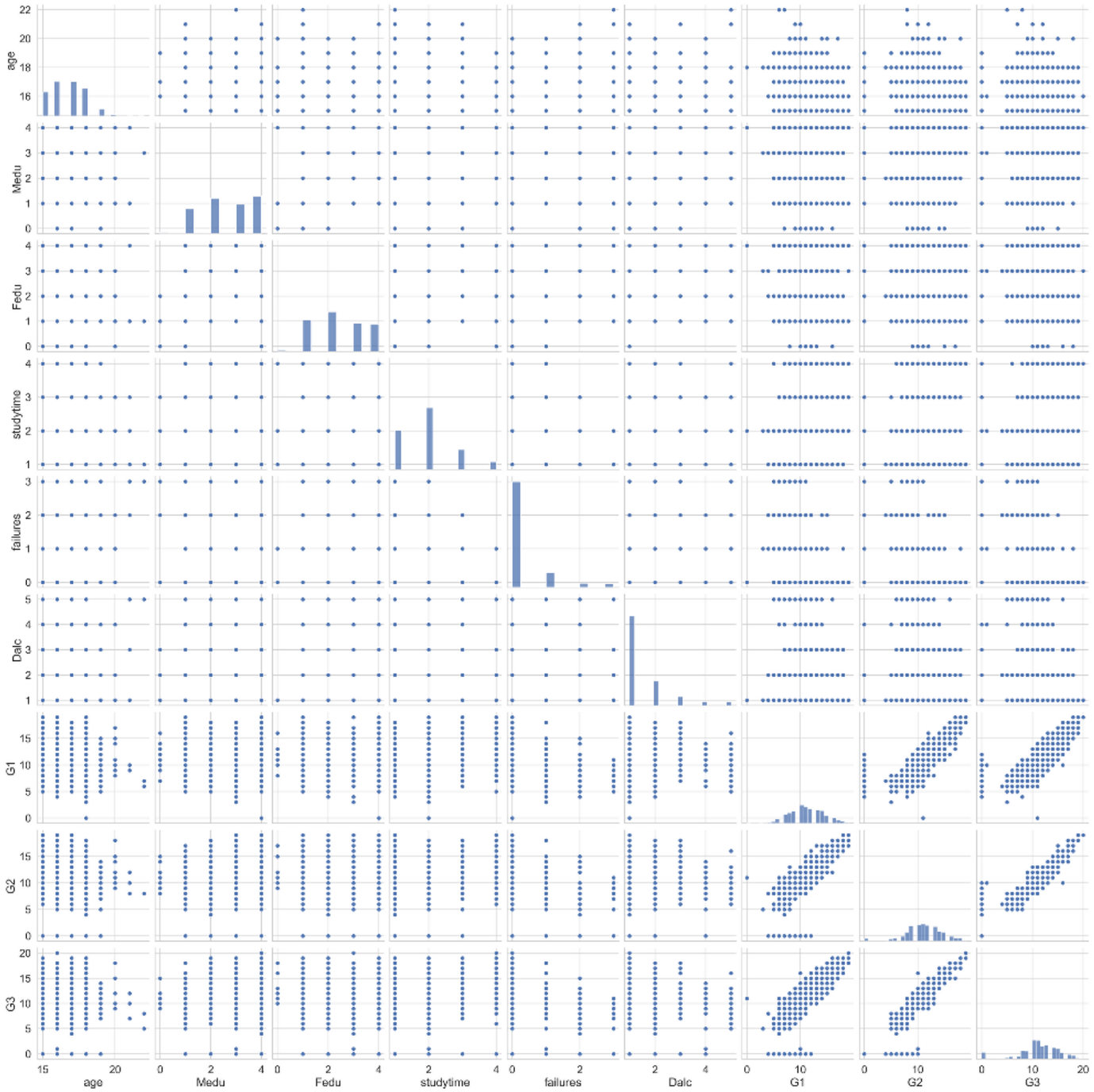
Pair plot with all Feature



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

```
student[top_corr_features]
```

	age	Medu	Fedu	studytime	failures	Dalc	G1	G2	G3
0	18	4	4	2	0	1	5	6	6
1	17	1	1	2	0	1	5	5	6
2	15	1	1	2	3	2	7	8	10
3	15	4	2	3	0	1	15	14	15
4	16	3	3	2	0	1	6	10	10
...	...	...	...	...	...	...	...	...	...
644	19	2	3	3	1	1	10	11	10
645	18	3	1	2	0	1	15	15	16
646	18	1	1	2	0	1	11	12	9
647	17	3	1	1	0	3	10	10	10
648	18	3	2	1	0	3	10	11	11

Training and testing the model

LINEAR REGRESSION

Regression With All features

MAE : 0.829

MSE : 1.501

Regression score 89.031%

Regression With Most related features

MAE : 0.801

MSE : 1.5

Regression score 89.043%

Best Regression

MAE : 0.699

MSE : 1.063

Regression score 91.660%

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

DECISION TREE

Accuracy: 89.80891719745223

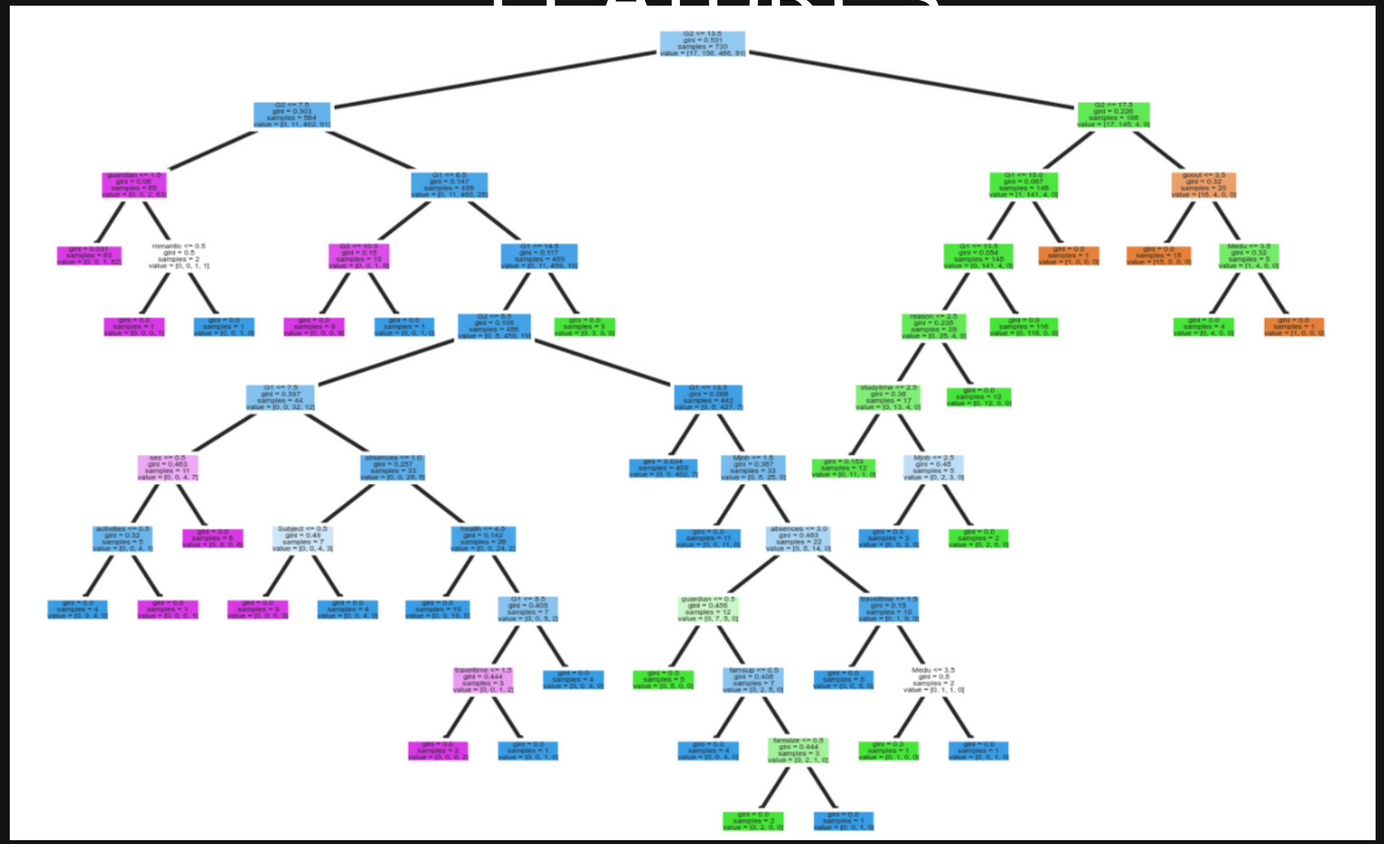
confusion_matrix:

```
[[  7   1   0   0]
 [  3  65   3   0]
 [  0   7 180   8]
 [  0   0  10  30]]
```

Classification report :

	precision	recall	f1-score	support
A	0.70	0.88	0.78	8
B	0.89	0.92	0.90	71
C	0.93	0.92	0.93	195
D	0.79	0.75	0.77	40
accuracy			0.90	314

WITH ALL FEATURES



DECISION TREE

Accuracy: 63.05732484076433

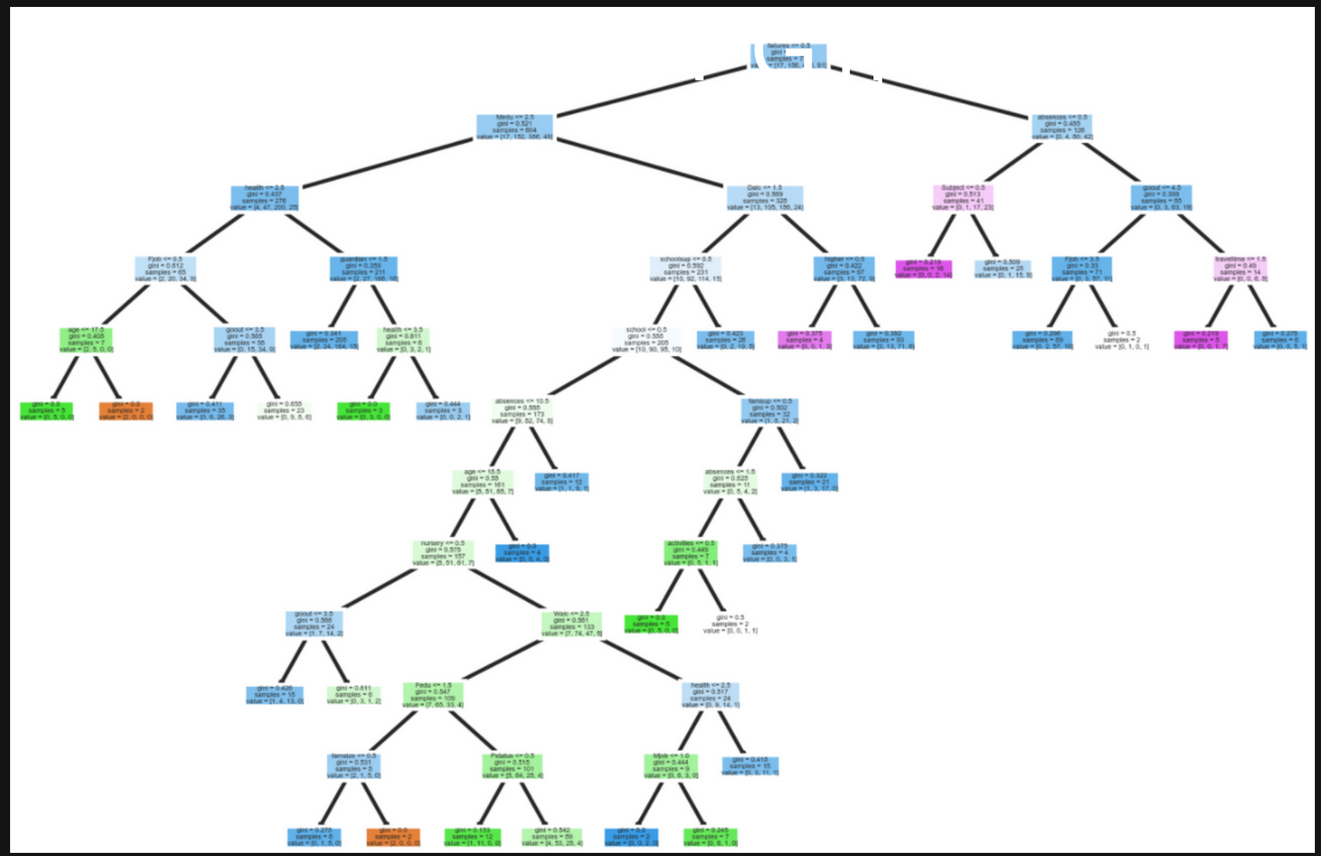
confusion_matrix:

```
[[  0   4   4   0]
 [  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	71
C	0.68	0.85	0.75	195
D	0.69	0.28	0.39	40
accuracy			0.63	314

WITH ALL FEATURES



DECISION TREE

Accuracy: 94.5859872611465

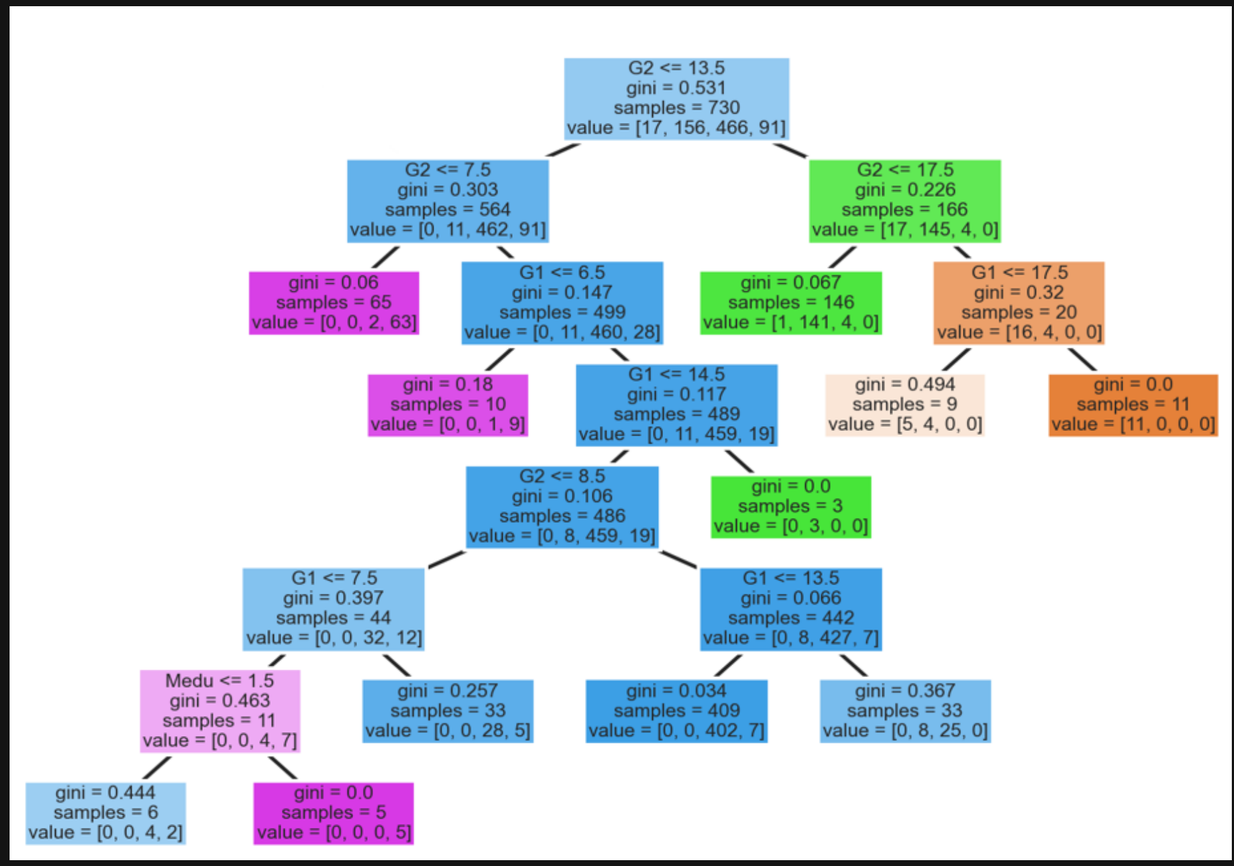
confusion_matrix:

```
[[  7   1   0   0]
 [  3  66   2   0]
 [  0   4 187   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 RELEATED FEATURS



DECISION TREE

Accuracy: 62.101910828025474

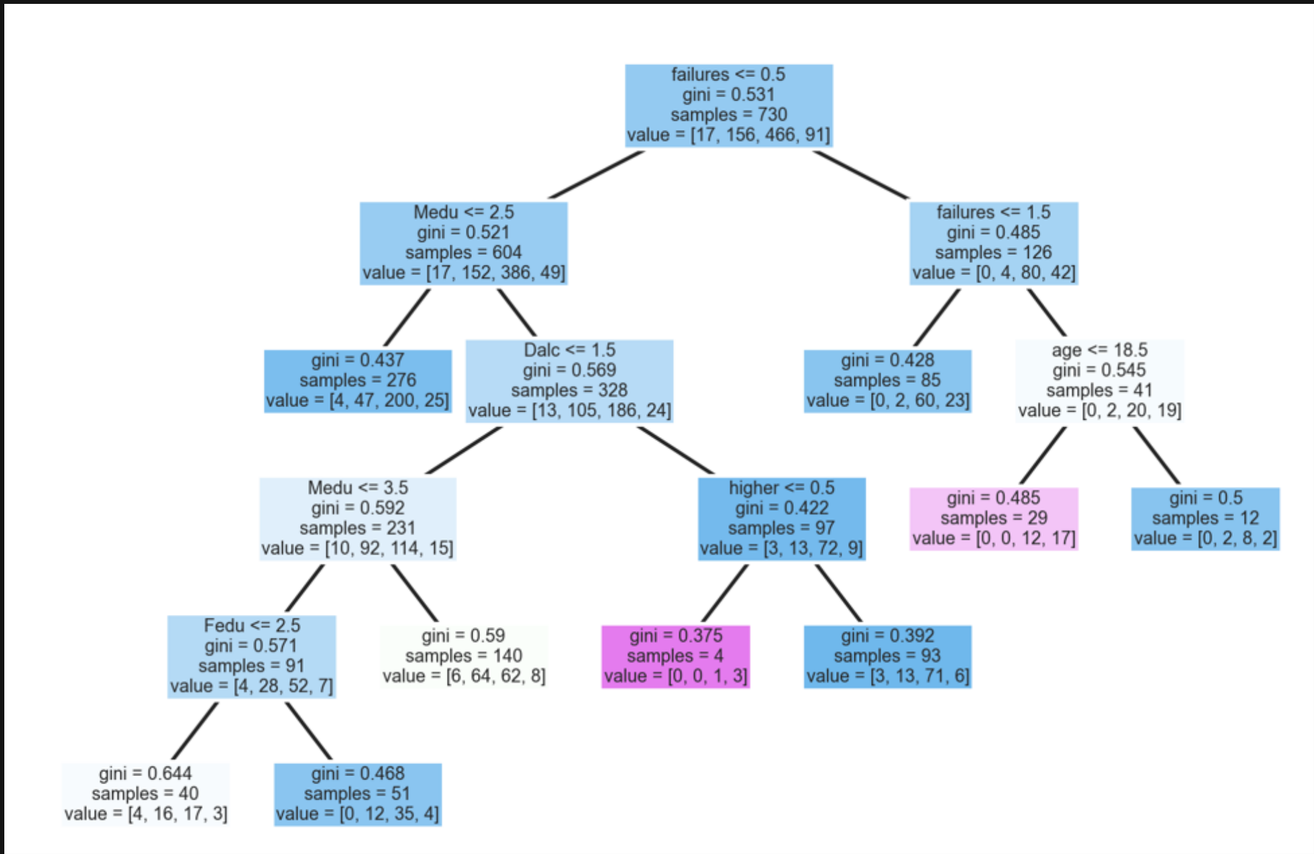
confusion_matrix:

```
[[ 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
accuracy			0.62	314

WITH MOST RELEATED FEATURS WITHOUT G1 & G2



DECISION TREE

Accuracy: 95.21531100478468

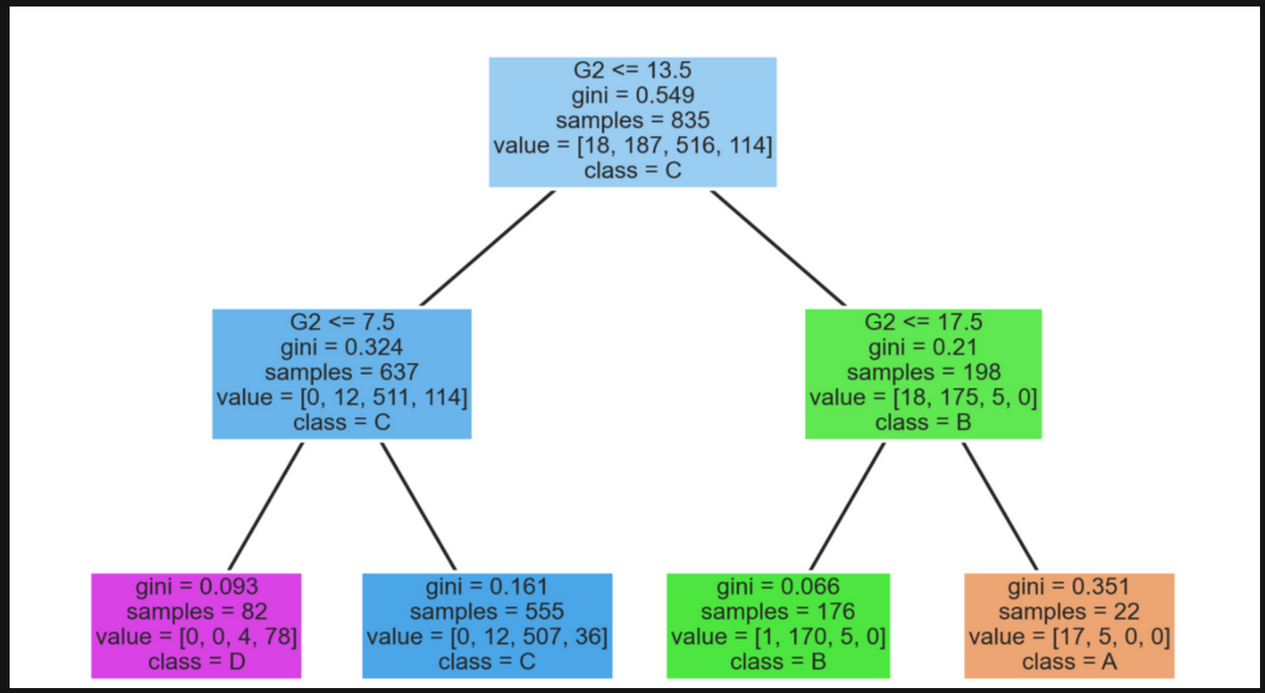
confusion_matrix:

```
[[ 6  1  0  0]
 [ 2 34  4  0]
 [ 0  1 144  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
accuracy			0.95	209

WITH MOST RELEATED FEATURS 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	314
macro avg	0.91	0.81	0.85	314

Thank You