GTU - CSE454

DATA MINING
PROJECT REPORT

STUDENT PERFORMANCE CLASSIFICATION ANALYSIS

ELİF GORAL
171044003

Dataset:

This data approach student 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. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

Why this dataset?

I chose this dataset because having 33 attributes meant there was a lot of data and relation for me to examine. I wanted to be able to examine it according to more parameters. Since I am also a student myself, I thought it was very convenient for me to interpret the dataset.

Attribute Information:

index	Attribute	Meaning
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 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
8	Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - 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)
31	G1	first period grade (numeric: from 0 to 20)
32	G2	second period grade (numeric: from 0 to 20)
33	G3	final grade (numeric: from 0 to 20, output target)
34	final_grade	AA,BA,BB,CB,CC,DC,DD,FF

Dataset Analysis:

First of all, I am reading the dataset with the pandas library.

```
data = pd.read_csv("dataset/student performance/student/student-mat.csv",sep=";")
```

G1, G2, G3 values are given between 0-20 on the dataset. So I multiplied the values by 5 to make it similar to the grading at our university. Afterwards, I took the weights of the G1 and G2 values (visa grades) as 0.25 and the value of the G3 (final) exam as 0.5. Then I graded AA, BA,....FF according to the following ranges and added a new column(final_grade). I planned to do the classification according to this attribute.

90 - 100 : AA
80 - 90 : BA
70 - 80 : BB
60 - 70 : CB
50 - 60 : CC
45 - 50 : DC
40 - 45 : DD
0 - 40 : FF

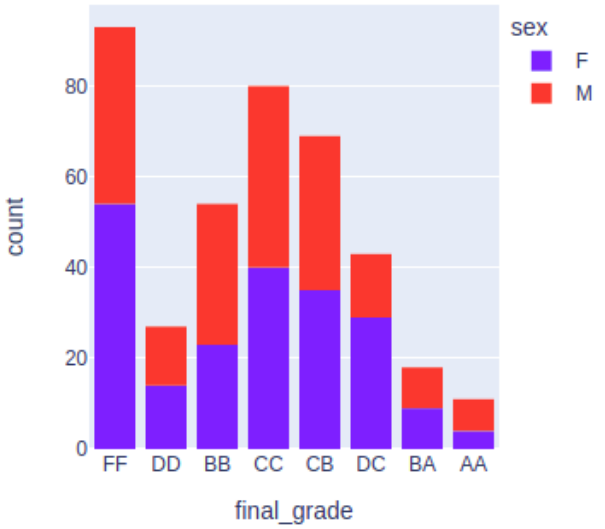
```
data["G1"] = data["G1"].apply(lambda x: x * 5)
data["G2"] = data["G2"].apply(lambda x: x * 5)
data["G3"] = data["G3"].apply(lambda x: x * 5)

data['final_grade'] = 'NA'
data.loc[((data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) >= 90) & ( (data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) <= 100), 'final_grade'] = 'AA'
data.loc[((data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) >= 80) & ( (data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) <= 89), 'final_grade'] = 'BA'
data.loc[((data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) >= 70) & ( (data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) <= 79), 'final_grade'] = 'BB'
data.loc[((data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) >= 60) & ( (data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) <= 69), 'final_grade'] = 'CB'
data.loc[((data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) >= 50) & ( (data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) <= 59), 'final_grade'] = 'CC'
data.loc[((data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) >= 45) & ( (data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) <= 49), 'final_grade'] = 'DC'
data.loc[((data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) >= 40) & ( (data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) <= 45), 'final_grade'] = 'DD'
data.loc[((data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) >= 0) & ( (data.G1 * 0.25 ) + (data.G2 * 0.25) + (data.G3 * 0.5) <= 40), 'final_grade'] = 'FF'
```

Sex:

```
Percentage of female students:  47.34 %  --> (187 student)
Percentage of male students: 52.66  %  --> (208 student)
```

When we examine this output, we see that the numbers of female and male students are very close to each other.



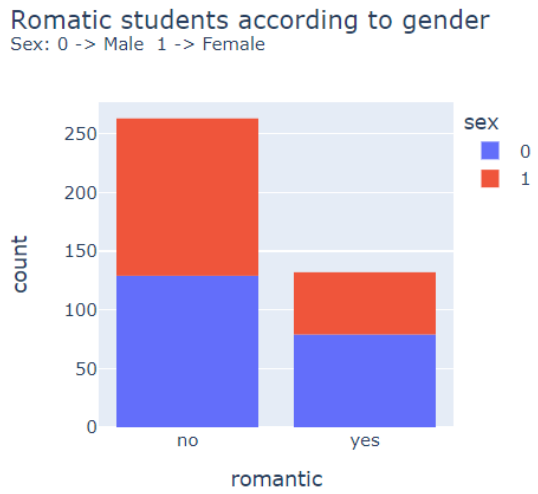

```
fig = px.histogram(data_frame=data, x="final_grade", color="sex", width=400, height=400)
fig.show()
```

When we examine this graph, we see that most of the students receive FF. In general, we see that the gender distribution in letter grades is very close to each other.

State of being Romantic

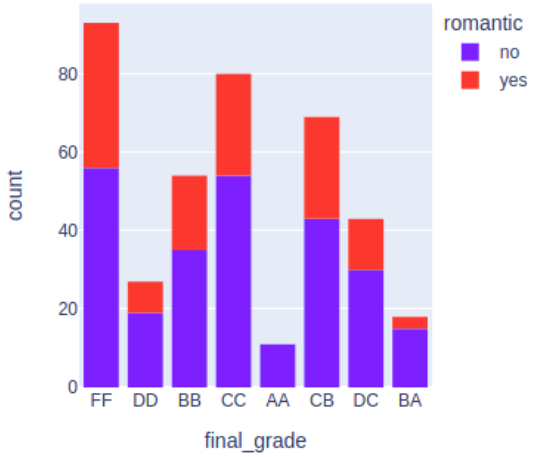
```
Percentage of romantic students:  33.42 %  --> (132 student)
Percentage of not romantic students: 66.58  %  --> (263 student)
```

When we examine this output, we see that there are twice as many romantic students as non-romantic students among young people. From here, we can observe that the young people in the data set are mostly not romantic.



It was observed that the number of boys and girls in the non-romantic people was close to each other, while the number of boys in the romantic ones was higher than the number of girls.

```
fig = px.histogram(data_frame=data, x="final_grade", color="romantic", width=400, height=400)
fig.show()
```

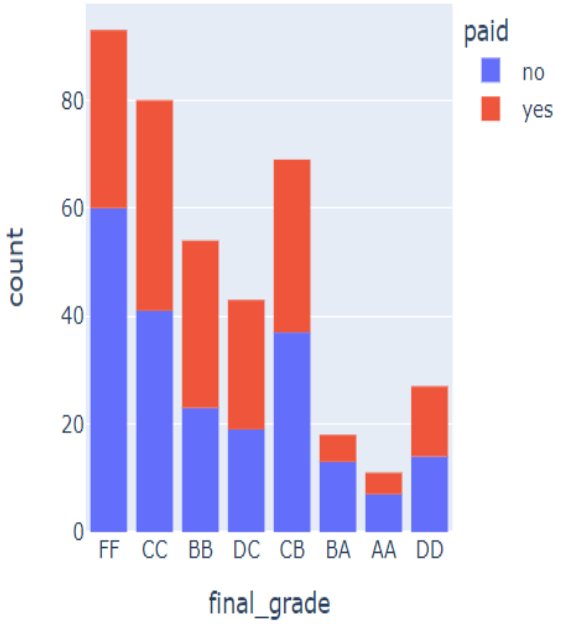


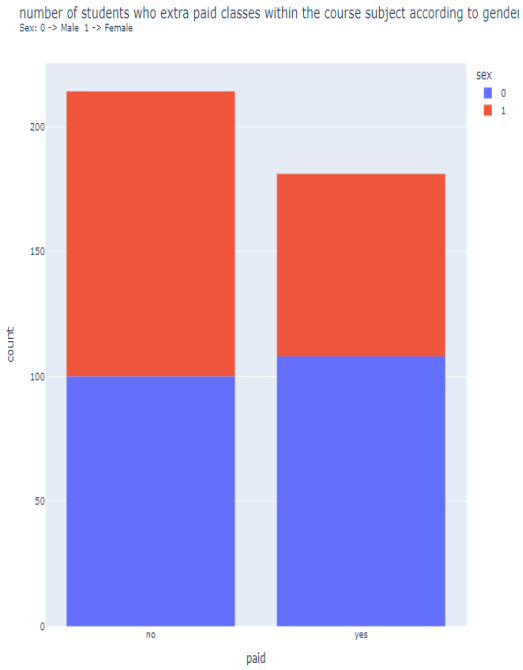
When we examine this histogram, we observe that especially the most successful students who take AA are not romantic at all. In BA, which is the second most successful grade, it is observed that non-romantic students are in the majority. It is observed that as the romanticism rate increases, the grades of the students decrease. Although this rate is relatively low in DC and DD areas, we can deduce that the non-romantic situation increases the success of the students.

State of getting extra paid class:

```
Percentage of students who extra paid classes within the course subject:  45.82 %  --> (181 student)
Percentage of students who not extra paid classes within the course subject: 54.18  %  --> (214 student)
```

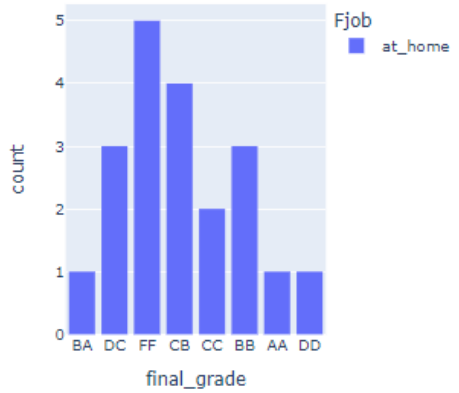
When we examine this output, we see that the number of students who take extra paid classes and those who do not take extra paid classes is close. Here, we see that more than half of the students need it and their financial situation can afford it.

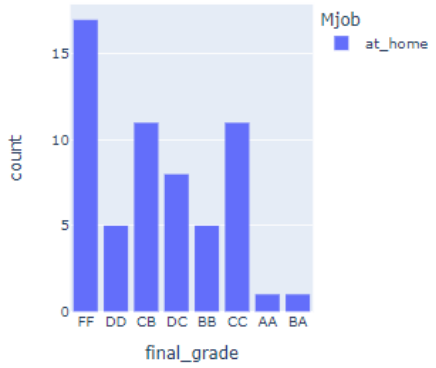




When the success rate goes up to AA, BA values, we observe that the paid class parameter does not increase the success. But in general, most of the rest are students who do not take paid classes. I expect the family's income situation to be good in order to take the Paid class, so when the family business is examined.

Job == at home



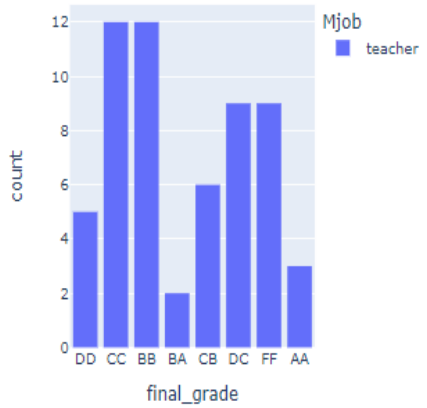


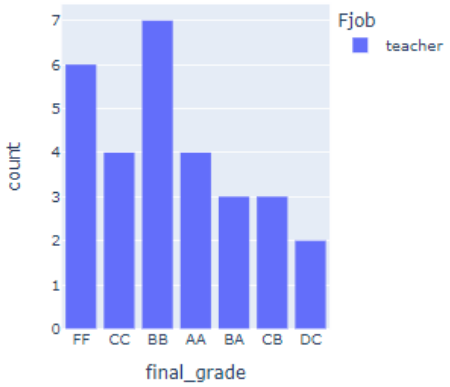

```
grade and family job relation
final_grade=FF          Fjob=at_home num: 5
final_grade=CB          Fjob=at_home num: 4
final_grade=DC          Fjob=at_home num: 3
final_grade=BB          Fjob=at_home num: 3
final_grade=CC          Fjob=at_home num: 2
final_grade=BA          Fjob=at_home num: 1
final_grade=AA          Fjob=at_home num: 1
final_grade=DD          Fjob=at_home num: 1
```

```
final_grade=FF          Mjob=at_home num: 17
final_grade=CB          Mjob=at_home num: 11
final_grade=CC          Mjob=at_home num: 11
final_grade=DC          Mjob=at_home num: 8
final_grade=BB          Mjob=at_home num: 5
final_grade=DD          Mjob=at_home num: 5
final_grade=BA          Mjob=at_home num: 1
final_grade=AA          Mjob=at_home num: 1
```

We can observe that most of the students whose parents are at_home get FF. This means that parents who work at home or do not work at all.

Job=teacher

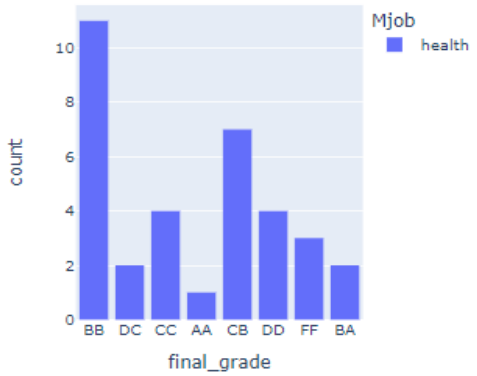


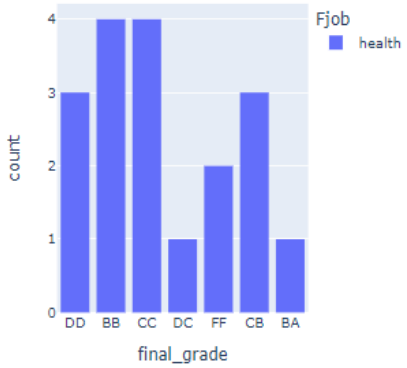



```
grade and family job relation
final_grade=CC          Mjob=teacher num: 12
final_grade=BB          Mjob=teacher num: 12
final_grade=DC          Mjob=teacher num: 9
final_grade=FF          Mjob=teacher num: 9
final_grade=CB          Mjob=teacher num: 6
final_grade=DD          Mjob=teacher num: 5
final_grade=AA          Mjob=teacher num: 3
final_grade=BA          Mjob=teacher num: 2
```

```
grade and family job relation
final_grade=BB          Fjob=teacher num: 7
final_grade=FF          Fjob=teacher num: 6
final_grade=AA          Fjob=teacher num: 4
final_grade=CC          Fjob=teacher num: 4
final_grade=CB          Fjob=teacher num: 3
final_grade=BA          Fjob=teacher num: 3
final_grade=DC          Fjob=teacher num: 2
```

Job = health



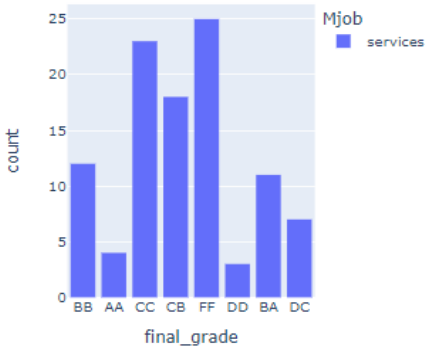


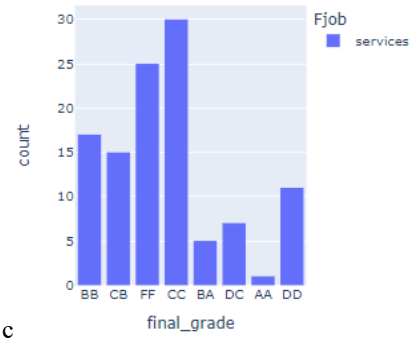

```
final_grade=BB          Mjob=health num: 11
final_grade=CB          Mjob=health num: 7
final_grade=DD          Mjob=health num: 4
final_grade=CC          Mjob=health num: 4
final_grade=FF          Mjob=health num: 3
final_grade=BA          Mjob=health num: 2
final_grade=DC          Mjob=health num: 2
final_grade=AA          Mjob=health num: 1
```

```
final_grade=BB          Fjob=health num: 4
final_grade=CC          Fjob=health num: 4
final_grade=DD          Fjob=health num: 3
final_grade=CB          Fjob=health num: 3
final_grade=FF          Fjob=health num: 2
final_grade=BA          Fjob=health num: 1
final_grade=DC          Fjob=health num: 1
```

It is observed that the majority of the students whose families work in the field of health have a high success in passing the course.

Job = service





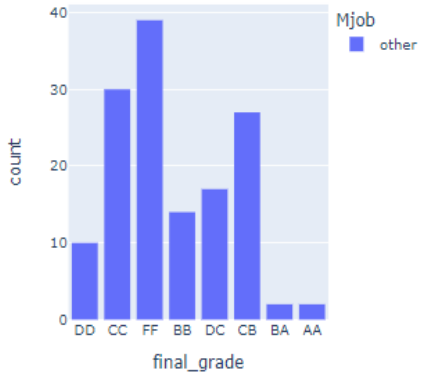

```
final_grade=FF          Mjob=services num: 25
final_grade=CC          Mjob=services num: 23
final_grade=CB          Mjob=services num: 18
final_grade=BB          Mjob=services num: 12
final_grade=BA          Mjob=services num: 11
final_grade=DC          Mjob=services num: 7
final_grade=AA          Mjob=services num: 4
final_grade=DD          Mjob=services num: 3
```

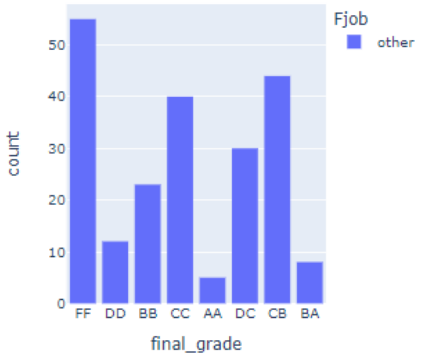
c

```
final_grade=CC          Fjob=services num: 30
final_grade=FF          Fjob=services num: 25
final_grade=BB          Fjob=services num: 17
final_grade=CB          Fjob=services num: 15
final_grade=DD          Fjob=services num: 11
final_grade=DC          Fjob=services num: 7
final_grade=BA          Fjob=services num: 5
final_grade=AA          Fjob=services num: 1
```

It has been observed that the success of those whose families are service workers is not very high.

Job =other






```
final_grade=FF          Mjob=other num: 39
final_grade=CC          Mjob=other num: 30
final_grade=CB          Mjob=other num: 27
final_grade=DC          Mjob=other num: 17
final_grade=BB          Mjob=other num: 14
final_grade=DD          Mjob=other num: 10
final_grade=AA          Mjob=other num: 2
final_grade=BA          Mjob=other num: 2
```

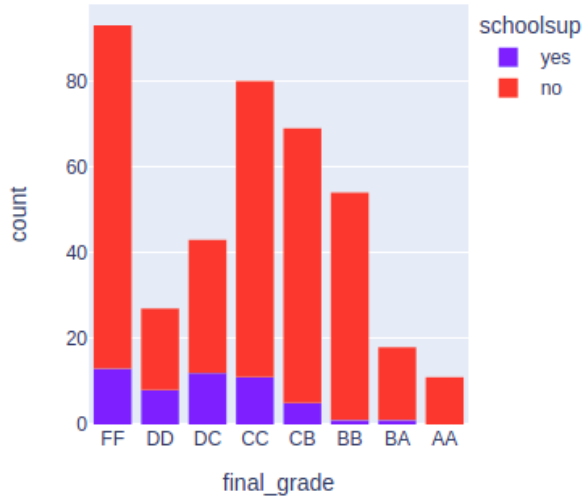
```
final_grade=FF          Fjob=other num: 55
final_grade=CB          Fjob=other num: 44
final_grade=CC          Fjob=other num: 40
final_grade=DC          Fjob=other num: 30
final_grade=BB          Fjob=other num: 23
final_grade=DD          Fjob=other num: 12
final_grade=BA          Fjob=other num: 8
final_grade=AA          Fjob=other num: 5
```

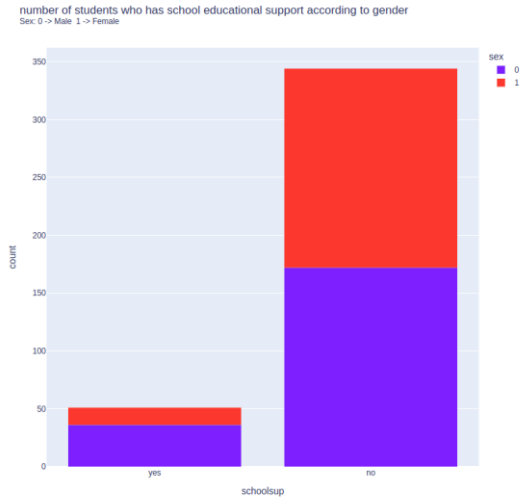
It has been observed that the success of those whose family has other professions is not high.

School Support: extra educational support from school

```
fig = px.histogram(data_frame=data, x="final_grade", color="schoolsup", width=400, height=400)
fig.show()
```

```
Percentage of students who has school educational support:  12.91 %  --> (51 student)
Percentage of students who has not school educational support: 87.09  %  --> (344 student)
```



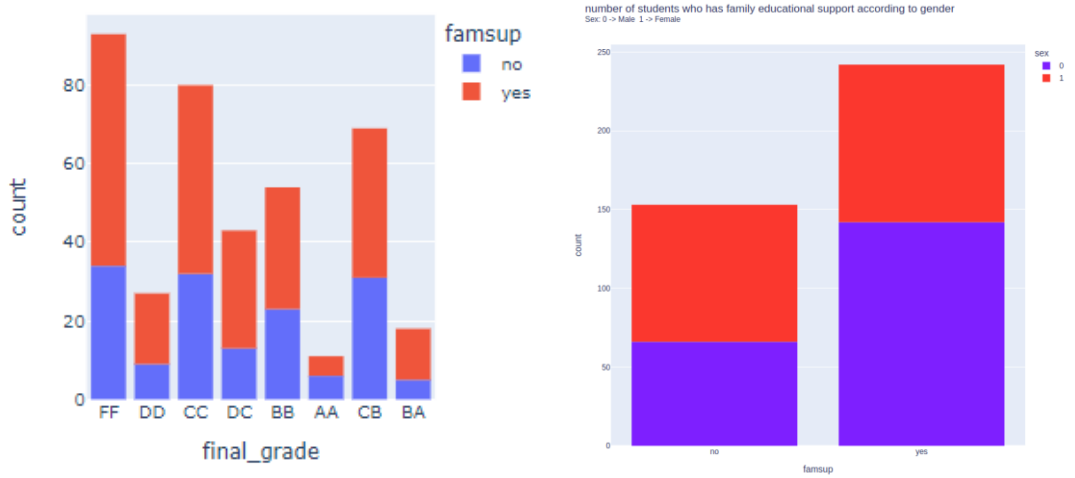


It was observed that 87% of the students did not receive any extra support. It was also noteworthy that the students who took AA did not receive any help at all. In other words, it has been observed that the help does not have a great effect on success. In terms of gender, it was observed that while the gender ratios of the students who did not receive help were very close to each other, it was observed that male students received more help among students who received assistance.

Family Support:

```
fig = px.histogram(data_frame=data, x="final_grade", color="famsup", width=400, height=400)
fig.show()
```

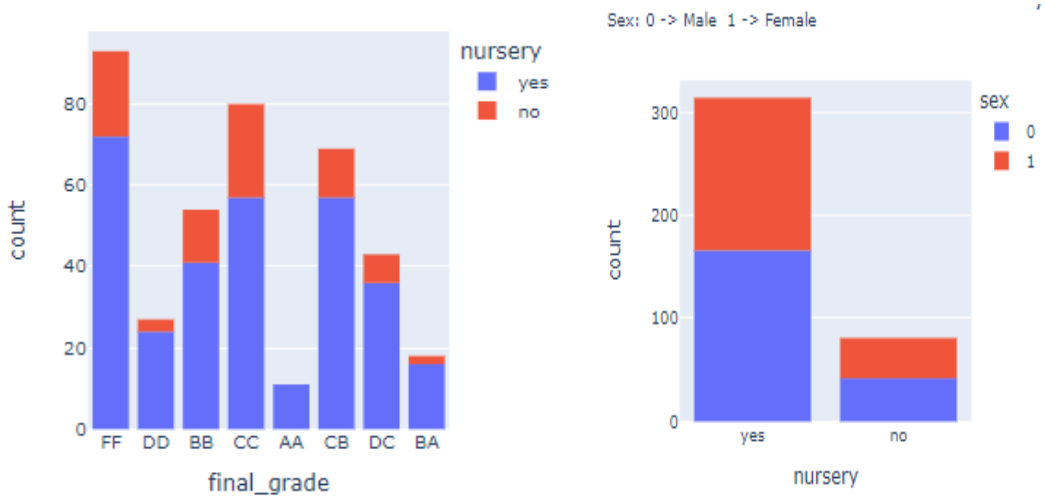
```
Percentage of students who has family educational support :  38.73 %  --> (153 student)
Percentage of students who has not family educational support : 61.27  %  --> (242 student)
```



61.27% of the students do not receive educational support from their families. No difference in distribution by gender was observed.

Nursery: Whether to go to nursery or not.

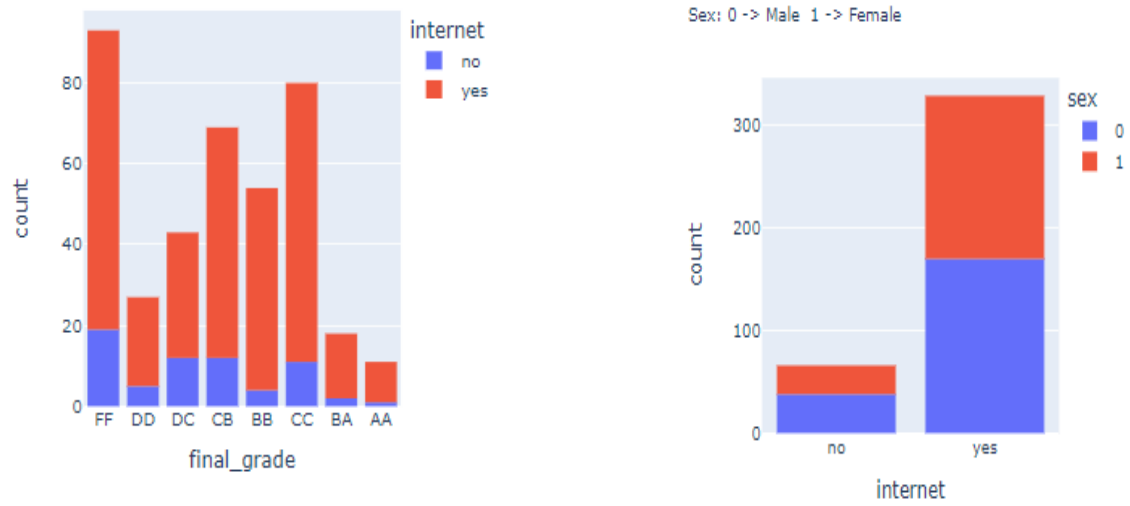
```
Percentage of students who attended nursery school:  20.51 %  --> (81 student)
Percentage of students who did not attend nursery school: 79.49  %  --> (314 student)
```



79.49% of the students went to nursery. Here, too, no different distribution was observed according to gender. But it has been observed that all of the students who took AA went to nursery. A large number of students who received BA went to nursery. From this, we can deduce that nursery increases the success of the student.

Internet Access:

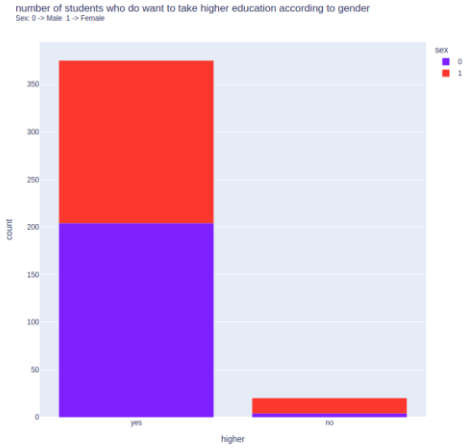
```
Percentage of students who have access internet access at home:  16.71 %  --> (66 student)
Percentage of students who have not access internet access at home: 83.29  %  --> (329 student)
```

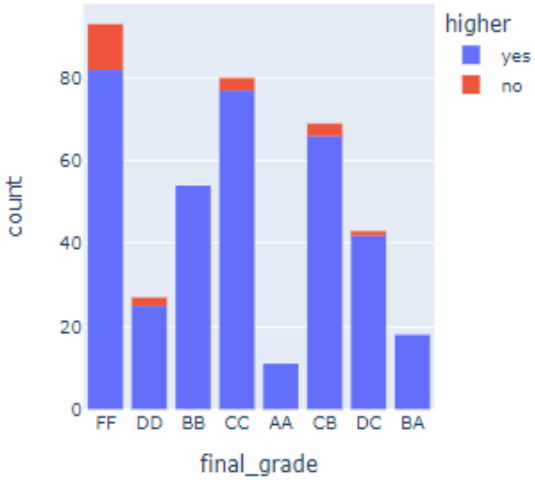


It is seen that 83.29% of the students have internet access. It is observed that in cases where internet access decreases, the success also decreases, that is, there is a correct ratio between them. No gender-related distribution abnormality is observed, the values are close to each other.

Higher Education:

```
Percentage of students who want to take higher education:  5.06 %  --> (20 student)
Percentage of students who do not want to take higher education: 94.94  %  --> (375 student)
```

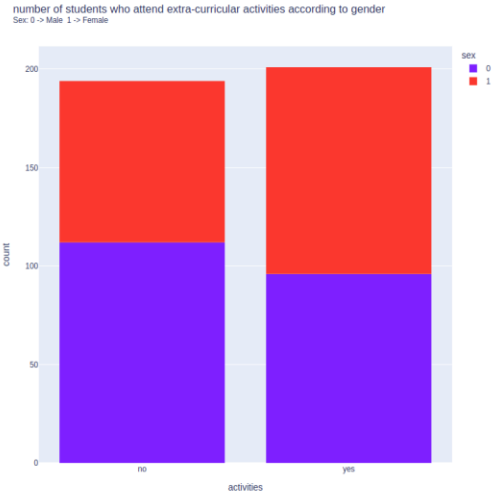


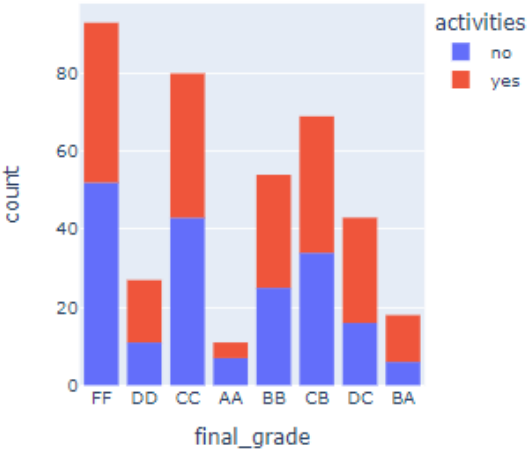


It seems that 94.94% of the students want higher education. In cases where success is high (such as AA, BA), all students want higher education. Demand for higher education decreases as success decreases. There is a correct relationship between them. While the gender distribution of students who want higher education is equal, it is observed that the number of female students among students who do not want higher education is higher than the number of male students.

Attending extra curricular Activities:

```
Percentage of students who attend extra-curricular activities:  49.11 %  --> (194 student)
Percentage of students who do not attend extra-curricular activities: 50.89  %  --> (201 student)
```

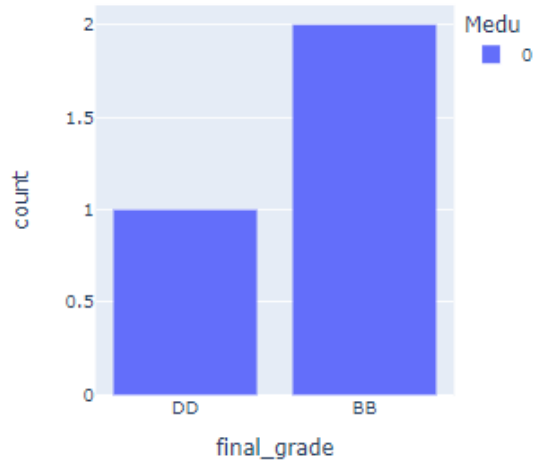


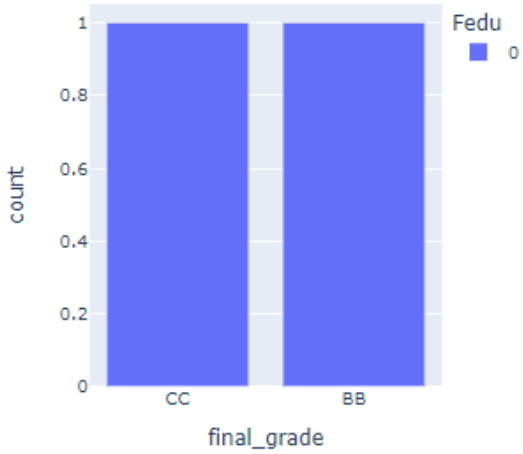


Half of the students participate in extra activities, while the remaining half do not. The distribution in terms of gender shows close values to each other.

Family Education:

Education = 0: None



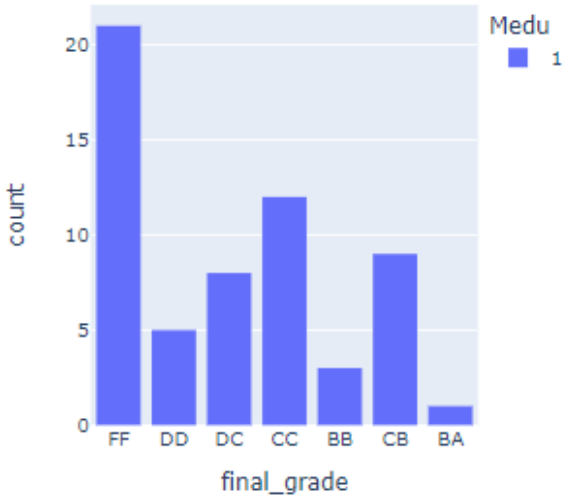


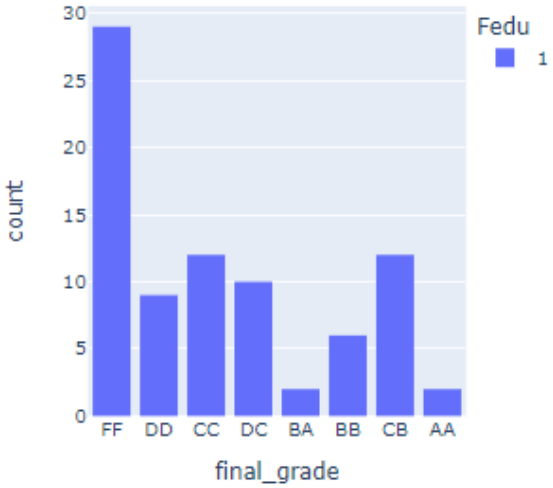

```
final_grade=BB          Medu=none num: 2
final_grade=DD          Medu=none num: 1
```

```
final_grade=CC          Fedu=none num: 1
final_grade=BB          Fedu=none num: 1
```

Although there are not many uneducated parents, it has been observed that their children cannot achieve high success in education.

Education = 1: primary education (4th grade)



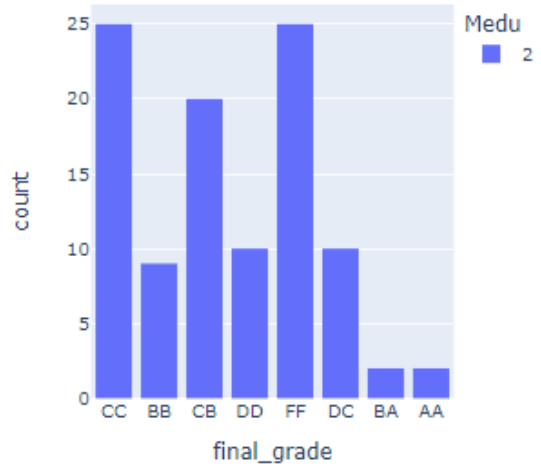


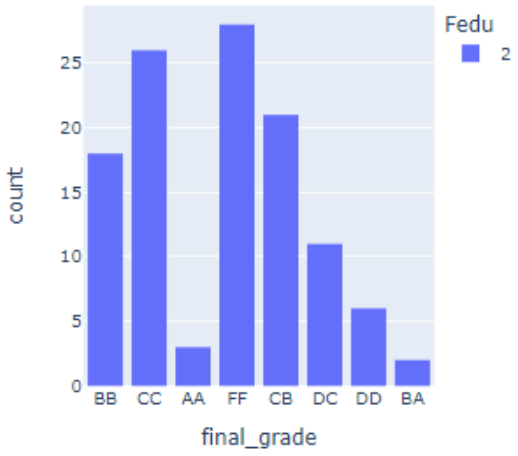

```
final_grade=FF          Medu=primary education (4th grade) num: 21
final_grade=CC          Medu=primary education (4th grade) num: 12
final_grade=CB          Medu=primary education (4th grade) num: 9
final_grade=DC          Medu=primary education (4th grade) num: 8
final_grade=DD          Medu=primary education (4th grade) num: 5
final_grade=BB          Medu=primary education (4th grade) num: 3
final_grade=BA          Medu=primary education (4th grade) num: 1
```

```
final_grade=FF          Fedu=primary education (4th grade) num: 29
final_grade=CB          Fedu=primary education (4th grade) num: 12
final_grade=CC          Fedu=primary education (4th grade) num: 12
final_grade=DC          Fedu=primary education (4th grade) num: 10
final_grade=DD          Fedu=primary education (4th grade) num: 9
final_grade=BB          Fedu=primary education (4th grade) num: 6
final_grade=AA          Fedu=primary education (4th grade) num: 2
final_grade=BA          Fedu=primary education (4th grade) num: 2
```

It has been observed that the success rates of children from families with low education levels are very low.

Education = 2: 5th to 9th grade



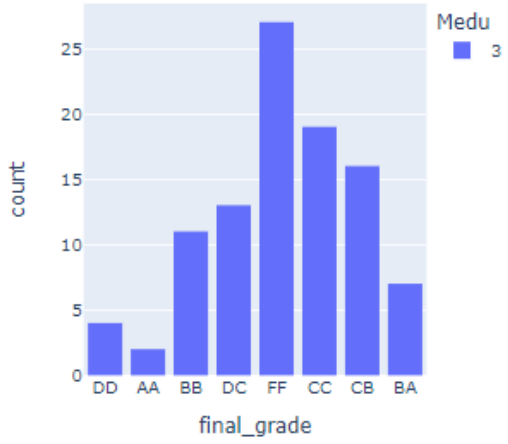


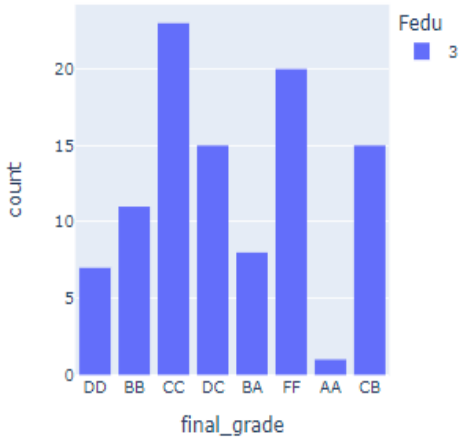

```
final_grade=CC          Medu=5th to 9th grade num: 25
final_grade=FF          Medu=5th to 9th grade num: 25
final_grade=CB          Medu=5th to 9th grade num: 20
final_grade=DD          Medu=5th to 9th grade num: 10
final_grade=DC          Medu=5th to 9th grade num: 10
final_grade=BB          Medu=5th to 9th grade num: 9
final_grade=AA          Medu=5th to 9th grade num: 2
final_grade=BA          Medu=5th to 9th grade num: 2
```

```
final_grade=FF          Fedu=5th to 9th grade num: 28
final_grade=CC          Fedu=5th to 9th grade num: 26
final_grade=CB          Fedu=5th to 9th grade num: 21
final_grade=BB          Fedu=5th to 9th grade num: 18
final_grade=DC          Fedu=5th to 9th grade num: 11
final_grade=DD          Fedu=5th to 9th grade num: 6
final_grade=AA          Fedu=5th to 9th grade num: 3
final_grade=BA          Fedu=5th to 9th grade num: 2
```

It has been observed that the success rates of children from families with low education levels are very low.

Education = 3: Secondary Education



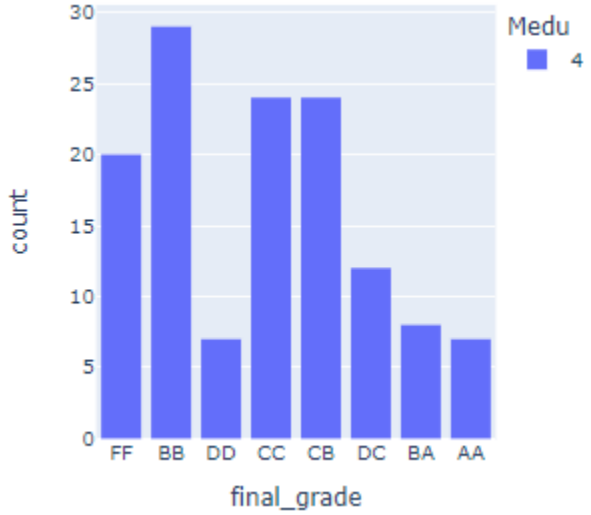


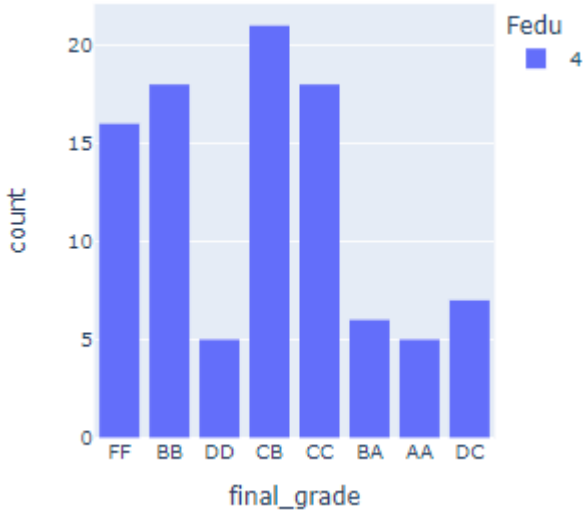

```
final_grade=FF          Medu=secondary education num: 27
final_grade=CC          Medu=secondary education num: 19
final_grade=CB          Medu=secondary education num: 16
final_grade=DC          Medu=secondary education num: 13
final_grade=BB          Medu=secondary education num: 11
final_grade=BA          Medu=secondary education num: 7
final_grade=DD          Medu=secondary education num: 4
final_grade=AA          Medu=secondary education num: 2
```

```
final_grade=CC          Fedu=secondary education num: 23
final_grade=FF          Fedu=secondary education num: 20
final_grade=CB          Fedu=secondary education num: 15
final_grade=DC          Fedu=secondary education num: 15
final_grade=BB          Fedu=secondary education num: 11
final_grade=BA          Fedu=secondary education num: 8
final_grade=DD          Fedu=secondary education num: 7
final_grade=AA          Fedu=secondary education num: 1
```

It has been observed that as the education level increases, the number of middle-level successful (BB, BA) students also increases.

Education = 4: Higher Education





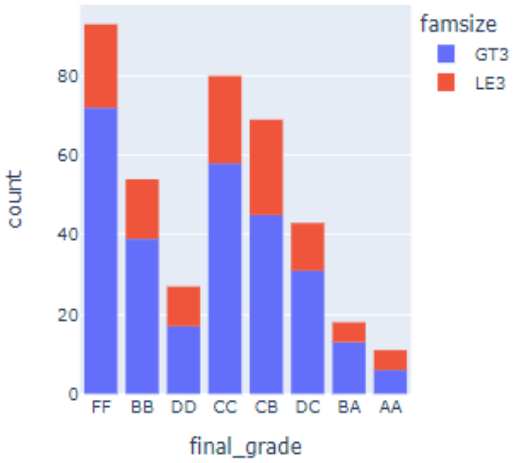

```
final_grade=BB          Medu=higher education num: 29
final_grade=CC          Medu=higher education num: 24
final_grade=CB          Medu=higher education num: 24
final_grade=FF          Medu=higher education num: 20
final_grade=DC          Medu=higher education num: 12
final_grade=BA          Medu=higher education num: 8
final_grade=DD          Medu=higher education num: 7
final_grade=AA          Medu=higher education num: 7
```

```
final_grade=CB          Fedu=higher education num: 21
final_grade=BB          Fedu=higher education num: 18
final_grade=CC          Fedu=higher education num: 18
final_grade=FF          Fedu=higher education num: 16
final_grade=DC          Fedu=higher education num: 7
final_grade=BA          Fedu=higher education num: 6
final_grade=DD          Fedu=higher education num: 5
final_grade=AA          Fedu=higher education num: 5
```

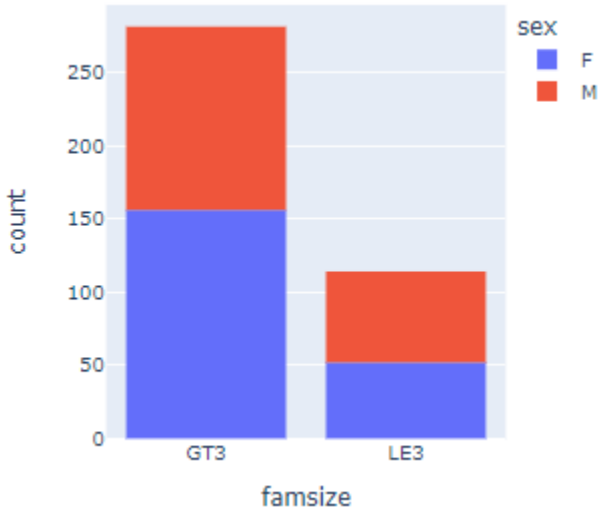
As the education level of the family increases, the number of high-achieving students increases and the number of low-successful/unsuccessful learners also decreases.

Family size relations:

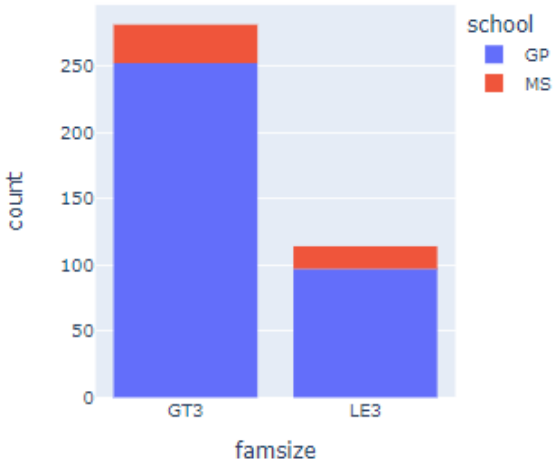
GT3: greater than 3
LT3: less than 38



```
gender and family size relation
A student, whose famsize is LE3, has a probability of 0.46 of being female and 0.54 of being male.
A student, whose famsize is GT3, has a probability of 0.56 of being female and 0.44 of being male.
```

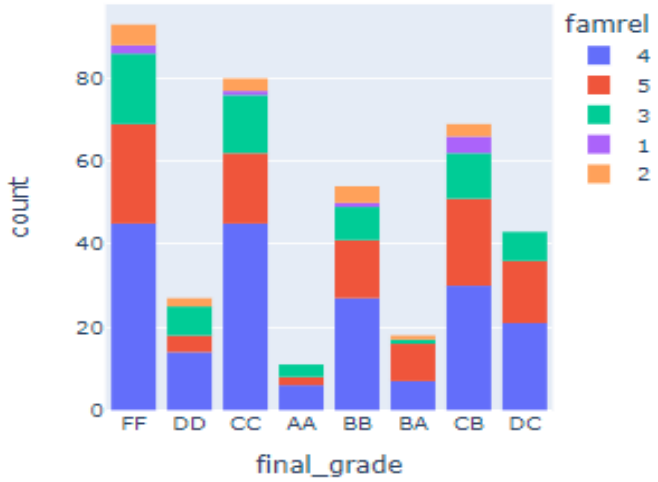



```
school and family size relation
A student, whose famsize is LE3, has a probability of 0.85 of being GP and 0.15 of being MS.
A student, whose famsize is GT3, has a probability of 0.9 of being GP and 0.1 of being MS.
```




```
A student, whose gender is male, has very low quality family relationship: 0.02
A student, whose gender is male, has low quality family relationship: 0.05
A student, whose gender is male, has medium quality family relationship: 0.16
A student, whose gender is male, has high quality family relationship: 0.47
A student, whose gender is male, has very high quality family relationship: 0.3
```

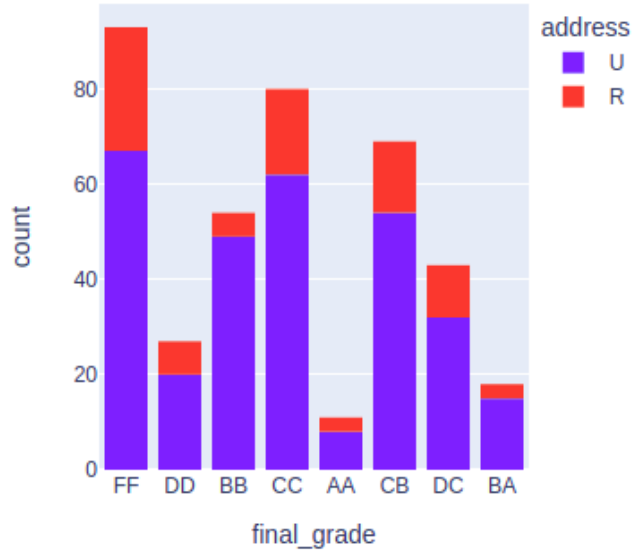
```
A student, whose gender is female, has very low quality family relationship: 0.02
A student, whose gender is female, has low quality family relationship: 0.04
A student, whose gender is female, has medium quality family relationship: 0.18
A student, whose gender is female, has high quality family relationship: 0.51
A student, whose gender is female, has very high quality family relationship: 0.24
```



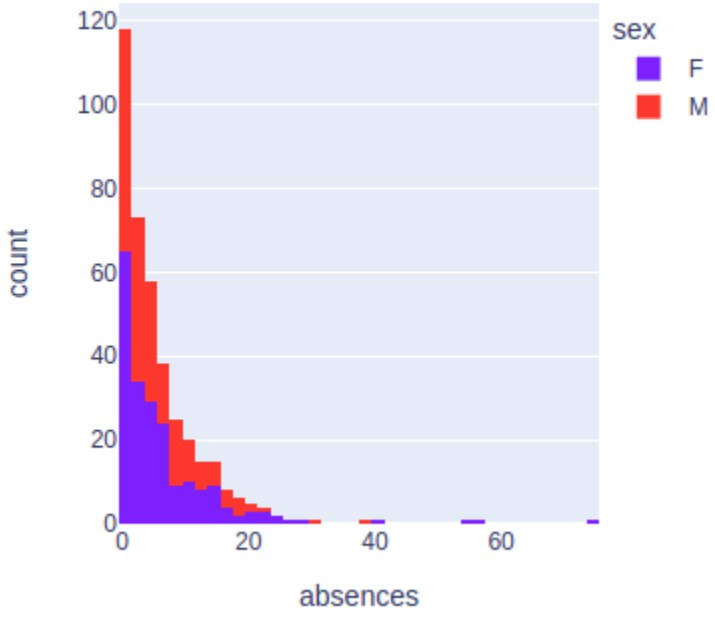
Address:

student's home address type (binary: 'U' - urban or 'R' - rural)

```
fig = px.histogram(data_frame=data, x="final_grade", color="address", width=400, height=400)
fig.show()
```



Absences:



I assign 0 for female and 1 for male so that the sex value is comparable.

```
data_1 = data.copy()
data_1["sex"] = data_1["sex"].map(lambda x: 0 if x=="F" else 1)
df = data_1.copy()
```

Shape of dataset:

```
print(df.shape)
```

-> `(395, 34)`

Finds how many of the same rows. In this dataset, the result is None.

```
print(df.duplicated().sum())
```

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 395 entries, 0 to 394
Data columns (total 34 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   school       395 non-null    object
 1   sex          395 non-null    int64 
 2   age          395 non-null    int64 
 3   address      395 non-null    object
 4   famsize      395 non-null    object
 5   Pstatus      395 non-null    object
 6   Medu         395 non-null    int64 
 7   Fedu         395 non-null    int64 
 8   Mjob         395 non-null    object
 9   Fjob         395 non-null    object
 10  reason       395 non-null    object
 11  guardian     395 non-null    object
 12  traveltime   395 non-null    int64 
 13  studytime    395 non-null    int64 
 14  failures     395 non-null    int64 
 15  schoolsup    395 non-null    object
 16  famsup       395 non-null    object
 17  paid         395 non-null    object
 18  activities   395 non-null    object
 19  nursery      395 non-null    object
 20  higher       395 non-null    object
 21  internet     395 non-null    object
 22  romantic     395 non-null    object
 23  famrel       395 non-null    int64 
 24  freetime     395 non-null    int64 
 25  goout        395 non-null    int64 
 26  Dalc         395 non-null    int64 
 27  Walc         395 non-null    int64 
 28  health       395 non-null    int64 
 29  absences     395 non-null    int64 
 30  G1           395 non-null    int64 
 31  G2           395 non-null    int64 
 32  G3           395 non-null    int64 
 33  final_grade  395 non-null    object
dtypes: int64(17), object(17)
memory usage: 105.0+ KB
```

missing numbers:

	Missing_Number	Missing_Percent
school	0	0.0
goout	0	0.0
nursery	0	0.0
higher	0	0.0
internet	0	0.0
romantic	0	0.0
famrel	0	0.0
freetime	0	0.0
Dalc	0	0.0
sex	0	0.0
Walc	0	0.0
health	0	0.0
absences	0	0.0
G1	0	0.0
G2	0	0.0
G3	0	0.0
activities	0	0.0
paid	0	0.0
famsup	0	0.0
schoolsup	0	0.0
failures	0	0.0
studytime	0	0.0
traveltime	0	0.0
guardian	0	0.0
reason	0	0.0
Fjob	0	0.0
Mjob	0	0.0
Fedu	0	0.0
Medu	0	0.0
Pstatus	0	0.0
famsize	0	0.0
address	0	0.0
age	0	0.0
final_grade	0	0.0

Numerical columns:

```
Numerical Columns:  Index(['age', 'Medu', 'Fedu', 'traveltime', 'studytime', 'failures', 'famrel',
       'freetime', 'goout', 'Dalc', 'Walc', 'health', 'absences', 'G1', 'G2',
       'G3'],
      dtype='object')
```

Categorical Columns:

```
Categorical Columns:  Index(['school', 'address', 'famsize', 'Pstatus', 'Mjob', 'Fjob', 'reason',
       'guardian', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery',
       'higher', 'internet', 'romantic', 'final_grade'],
      dtype='object')
```

Categorical columns' unique value size

```
Categorical unique:  school         2
address         2
famsize         2
Pstatus         2
Mjob            5
Fjob            5
reason          4
guardian        3
schoolsup       2
famsup          2
paid            2
activities      2
nursery         2
higher          2
internet        2
romantic        2
final_grade     8
dtype: int64
```

First, I converted all categorical columns to numbers. Then I made the classification according to final_grade (AA, BA,...FF).

```
data_converted = convert_categorical_to_binary(data)
X = data_converted.values
y = data_converted["final_grade"].values
```

That's why I removed the final_grade from the data frame I used. Then I separated 0.3 as test-train.

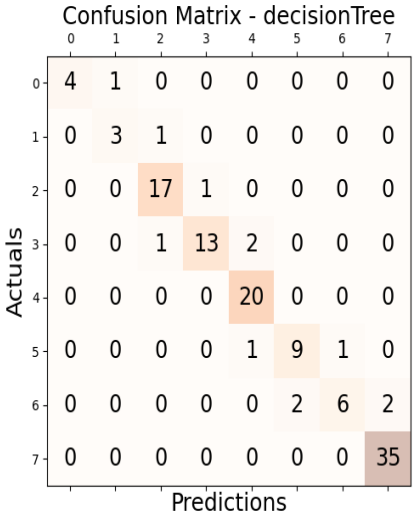
```
X = np.delete(X,[33],axis=1)
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3,random_state=0)
```

I tried and observed the classifiers in the sklearn library. I have their classification reports and confusion matrixes.

Decision Tree:

Classification Report of decisionTree

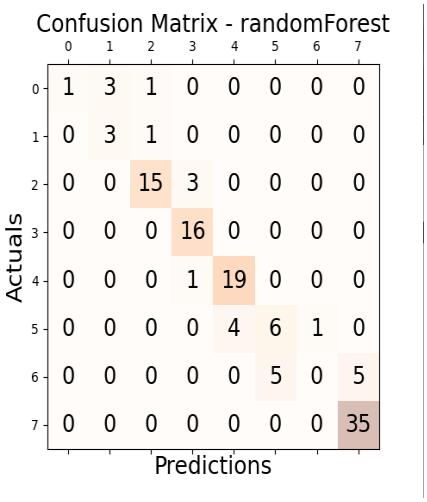
	precision	recall	f1-score	support
0	1.000	0.800	0.889	5
1	0.750	0.750	0.750	4
2	0.895	0.944	0.919	18
3	0.929	0.812	0.867	16
4	0.870	1.000	0.930	20
5	0.818	0.818	0.818	11
6	0.800	0.400	0.533	10
7	0.897	1.000	0.946	35
accuracy			0.882	119
macro avg	0.870	0.816	0.832	119
weighted avg	0.880	0.882	0.873	119



Random Forest

Classification Report of randomForest

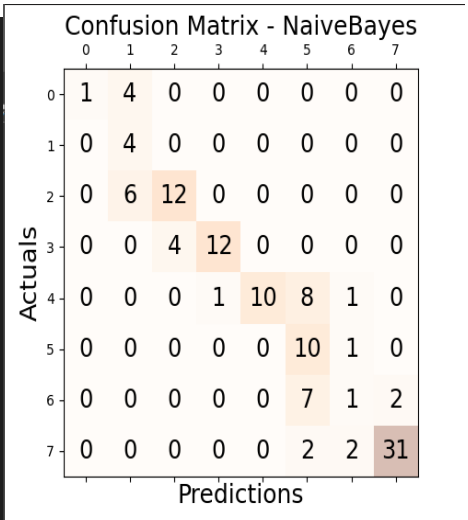
	precision	recall	f1-score	support
0	1.000	0.200	0.333	5
1	0.400	0.500	0.444	4
2	0.842	0.889	0.865	18
3	0.842	1.000	0.914	16
4	0.792	0.950	0.864	20
5	0.667	0.545	0.600	11
6	1.000	0.100	0.182	10
7	0.854	1.000	0.921	35
accuracy			0.807	119
macro avg	0.800	0.648	0.640	119
weighted avg	0.826	0.807	0.769	119



Naive Bayes

Classification Report of NaiveBayes

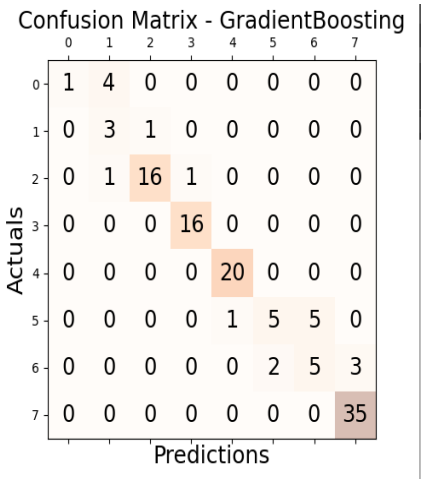
	precision	recall	f1-score	support
0	1.000	0.200	0.333	5
1	0.286	1.000	0.444	4
2	0.750	0.667	0.706	18
3	0.923	0.750	0.828	16
4	1.000	0.500	0.667	20
5	0.370	0.909	0.526	11
6	0.200	0.100	0.133	10
7	0.939	0.886	0.912	35
accuracy			0.681	119
macro avg	0.684	0.626	0.569	119
weighted avg	0.785	0.681	0.687	119



Gradient Boosting

Classification Report of GradientBoosting

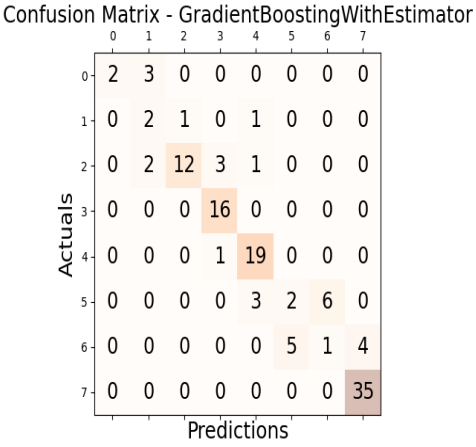
	precision	recall	f1-score	support
0	1.000	0.200	0.333	5
1	0.375	0.750	0.500	4
2	0.941	0.889	0.914	18
3	0.941	1.000	0.970	16
4	0.952	1.000	0.976	20
5	0.625	0.455	0.526	11
6	0.444	0.400	0.421	10
7	0.921	1.000	0.959	35
accuracy			0.840	119
macro avg	0.775	0.712	0.700	119
weighted avg	0.850	0.840	0.830	119



Gradient Boosting with 3 estimators

Classification Report of GradientBoosting with 3 estimator

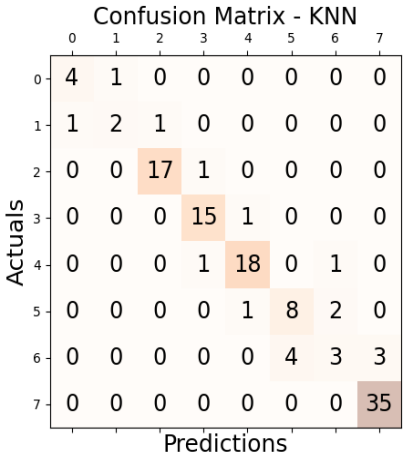
	precision	recall	f1-score	support
0	0.000	0.000	0.000	5
1	0.250	0.500	0.333	4
2	0.923	0.667	0.774	18
3	0.800	1.000	0.889	16
4	0.760	0.950	0.844	20
5	0.286	0.182	0.222	11
6	0.143	0.100	0.118	10
7	0.897	1.000	0.946	35
accuracy			0.731	119
macro avg	0.507	0.550	0.516	119
weighted avg	0.686	0.731	0.698	119



KNN

Classification Report of KNN

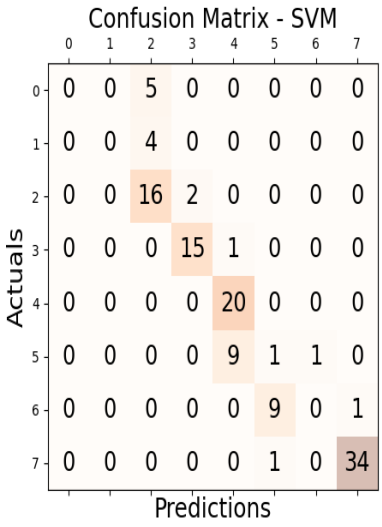
	precision	recall	f1-score	support
0	0.800	0.800	0.800	5
1	0.667	0.500	0.571	4
2	0.944	0.944	0.944	18
3	0.882	0.938	0.909	16
4	0.900	0.900	0.900	20
5	0.667	0.727	0.696	11
6	0.500	0.300	0.375	10
7	0.921	1.000	0.959	35
accuracy			0.857	119
macro avg	0.785	0.764	0.769	119
weighted avg	0.843	0.857	0.847	119



SVM

Classification Report of SVM

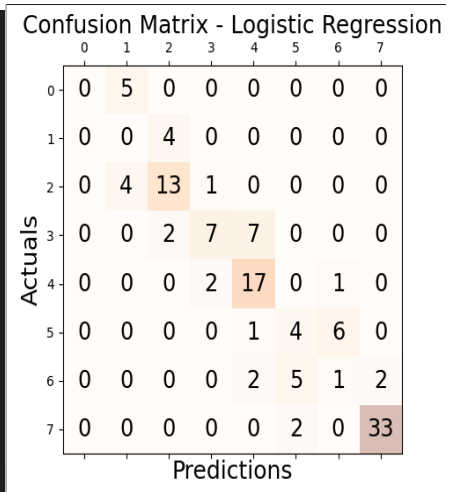
	precision	recall	f1-score	support
0	0.000	0.000	0.000	5
1	0.000	0.000	0.000	4
2	0.640	0.889	0.744	18
3	0.882	0.938	0.909	16
4	0.667	1.000	0.800	20
5	0.091	0.091	0.091	11
6	0.000	0.000	0.000	10
7	0.971	0.971	0.971	35
accuracy			0.723	119
macro avg	0.406	0.486	0.439	119
weighted avg	0.622	0.723	0.663	119



Logistic Regression

Classification Report of Logistic Regression

	precision	recall	f1-score	support
0	0.000	0.000	0.000	5
1	0.000	0.000	0.000	4
2	0.684	0.722	0.703	18
3	0.700	0.438	0.538	16
4	0.630	0.850	0.723	20
5	0.364	0.364	0.364	11
6	0.125	0.100	0.111	10
7	0.943	0.943	0.943	35
accuracy			0.630	119
macro avg	0.431	0.427	0.423	119
weighted avg	0.625	0.630	0.621	119



Algorithm	Accuracy Score	AUC Score
decisionTree	0.8823529411764706	0.9348011363636364
randomForest	0.8067226890756303	0.9646010487528346
NaiveBayes	0.680672268907563	0.8844196750669965
GradientBoosting	0.8403361344537815	0.9596437912286128
GradientBoostingWithEstimator	0.7310924369747899	0.9258934208152958
KNN	0.8571428571428571	0.926517857142857
SVM	0.7226890756302521	0.9772563163522986
logisticRegression	0.6302521008403361	0.8941868944547516

Comparison of F1 Score Values:

Algorithm/class	0	1	2	3	4	5	6	7
Decision Tree	0,89	0,75	0,91	0,86	0,93	0,81	0,53	0,94
Random Forest	0,33	0,44	0,86	0,91	0,86	0,60	0,18	0,92
Naïve Bayes	0,33	0,44	0,70	0,82	0,66	0,52	0,13	0,91
Gradient Boosting	0,33	0,5	0,91	0,97	0,97	0,52	0,42	0,95
KNN	0,8	0,57	0,94	0,90	0,90	0,69	0,37	0,95
SVM	0	0	0,74	0,90	0,80	0,09	0	0,97
Logistic Regression	0	0	0,70	0,53	0,72	0,36	0,11	0,94

- 0: AA
- 1: BA
- 2: BB
- 3: CB
- 4: CC
- 5: DC
- 6: DD
- 7: FF

When we compare according to the F1 score;

- The algorithm that gave the best f1 score for AA students was Decision Tree, while the worst result was SVM and Logistic Regression.
- Decision Tree was the algorithm that gave the best f1 score for BA students, while SVM and Logistic Regression gave the worst results.
- The algorithm that gave the best f1 score for students who received BB was KNN, while the worst result was Naïve Bayes and Logistic Regression.
- The algorithm that gave the best f1 score for students who took CB was Gradient Boosting, while Logistic Regression gave the worst result.
- The algorithm giving the best f1 score for CC students was Gradient Boosting, while Logistic Regression gave the worst result.
- Decision Tree was the algorithm that gave the best f1 score for students who took DC, while SVM gave the worst result.
- Decision Tree was the algorithm that gave the best f1 score for students who took DD, while SVM gave the worst result. But overall, the results were very bad for this class.
- While the algorithm that gave the best f1 score for FF students was SVM, there was no algorithm that gave bad results.

Comparison of Precision Values:

Algorithm/class	0	1	2	3	4	5	6	7
Decision Tree	1	0,75	0,89	0,92	0,87	0,81	0,80	0,89
Random Forest	1	0,40	0,84	0,84	0,79	0,66	1	0,85
Naïve Bayes	1	0,28	0,75	0,92	1	0,37	0,20	0,93
Gradient Boosting	1	0,37	0,94	0,94	0,95	0,62	0,44	0,92
KNN	0,8	0,66	0,94	0,88	0,90	0,66	0,50	0,92
SVM	0	0	0,64	0,88	0,66	0,09	0	0,97
Logistic Regression	0	0	0,68	0,70	0,63	0,36	0,12	0,94

- 0: AA
- 1: BA
- 2: BB
- 3: CB
- 4: CC
- 5: DC
- 6: DD
- 7: FF

• Decision Tree, Random Forest, Naïve Bayes and Gradient Boosting were the best predictors of AA students.
• Decision Tree and KNN were the best predictors of BA students.
• Gradient Boosting and KNN were the best predictors of BB students. But in general, students who took BB had high precision values by all algorithms.
• Gradient Boosting, Decision Tree, and Naïve Bayes predicted students who received CB best. But precision values are high by all algorithms.
• Naïve Bayes algorithm gave the best results for students who took CC.
• Among the students who took DC, Decision Tree gave the best result and SVM gave the worst result.
• Random Forest gave the best result for the students who took DD, while SVM gave the worst result.
• While SVM gave the best results for FF students, other algorithms also gave very high precision values.

Comparison of Classification Algorithms:

Decision Tree algorithm gives the highest accuracy result. The lowest accuracy result is given by the Logistic Regression algorithm. On the other hand, as AUC score, we get the highest value with SVM and the lowest value with Naïve Bayes.

Decision Tree vs Logistic Regression:

Decision Trees bisect the space into smaller and smaller regions, whereas Logistic Regression fits a single line to divide the space exactly into two. For higher-dimensional data, these lines would generalize to planes and hyperplanes. A single linear boundary can sometimes be limiting for Logistic Regression. In this example where the two classes are separated by a decidedly non-linear boundary, we see that trees can better capture the division, leading to superior classification performance. However, when classes are not well-separated, trees are susceptible to overfitting the training data, so that Logistic Regression's simple linear boundary generalizes better.[1] Decision Tree's accuracy score (0.88) is greater than Logistic Regression's accuracy score (0.63). In AUC score, Decision Tree's AUC score (0.93) is less than Logistic Regression's AUC score (0.89). Decision Tree is better than Logistic Regression for this dataset.

K-NN vs Naïve Bayes:

Naive Bayes is a linear classifier while K-NN is not; It tends to be faster when applied to big data. In comparison, K-NN is usually slower for large amounts of data, because of the calculations required for each new step in the process. If speed is important, choose Naive Bayes over K-NN. Naive Bayes can suffer from the zero-probability problem; when a particular attribute's conditional probability equals zero, Naive Bayes will completely fail to produce a valid prediction. This could be fixed using a Laplacian estimator, but K-NN could end up being the easier choice. A decision tree will almost certainly prune those important classes out of your model. If you have any rare occurrences, avoid using decision trees [2]. KNN's accuracy score (0.85) is greater than Naive Bayes's accuracy score (0.68). In AUC score, KNN's AUC score (0.92) is greater than Naive Bayes's AUC score (0.88). K-NN is better than Naïve Bayes for this dataset.

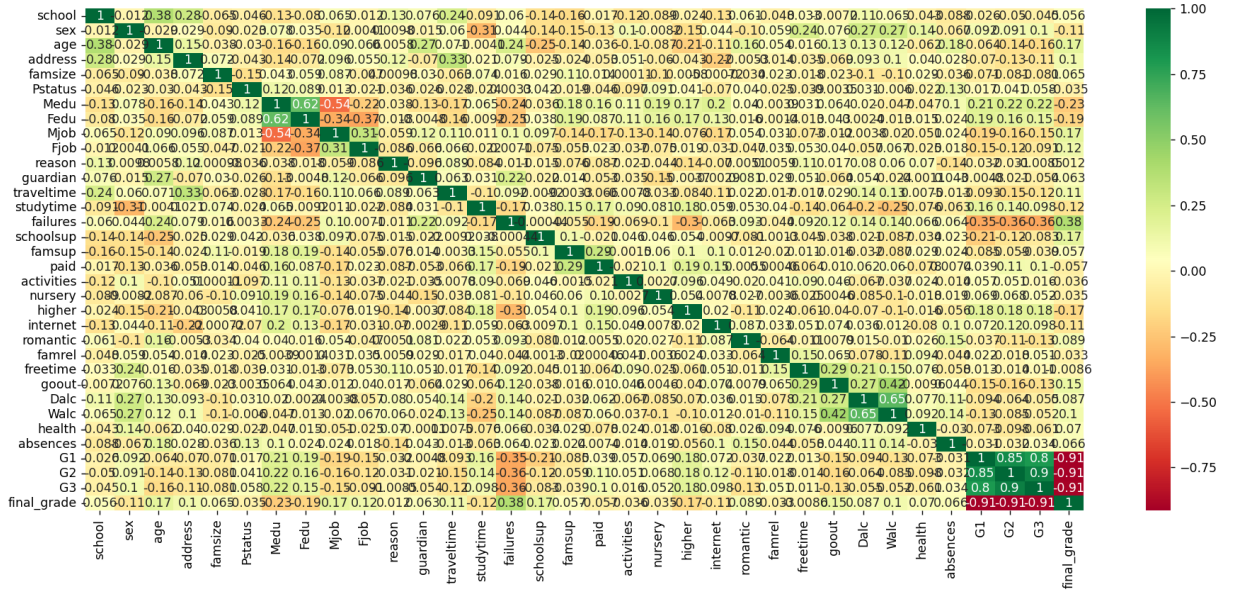
Decision Tree- Random Forest:

A decision tree combines some decisions, whereas a random forest combines several Decision Trees.[3] I expected Random Forest's accuracy is greater than Decision Tree. But Decision Tree's accuracy score (0.88) is greater than Random Forest's accuracy score (0.80). In AUC score, Decision Tree's AUC score (0.93) is less than Random Forest's AUC score (0.96).

SVM-Random Forest:

Random Forest is intrinsically suited for multiclass problems, while SVM is intrinsically two-class. For multiclass problem you will need to reduce it into multiple binary classification problems. Random Forest works well with a mixture of numerical and categorical features.

Correlation Table:



A negative correlation is a relationship between two variables moving in opposite directions. In other words, when variable A increases, variable B decreases. Negative correlation is also known as inverse correlation. a positive correlation is a relationship between two variables moving in the same directions. In other words, when variable A increases, variable B also increases. Two variables can have different strengths of negative correlation. Variable A can be strongly negatively correlated with B and the correlation coefficient can be -0.9. This means that for every positive change in the unit of variable B, variable A experiences a decrease of 0.9.

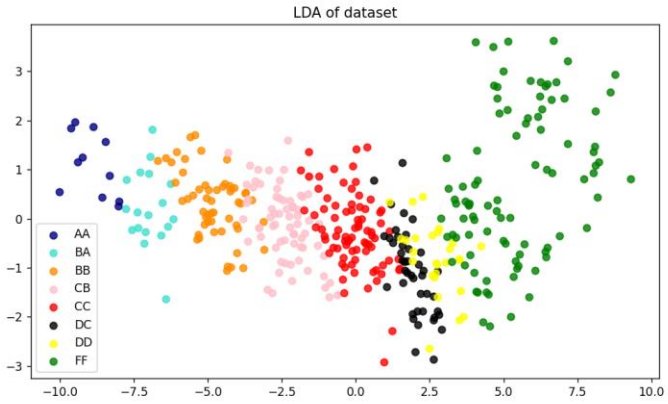
Dimensional Reduction Algorithms:

1) *PCA (Principal Component Analysis)*
2) *LDA (Linear Discriminant Analysis)*

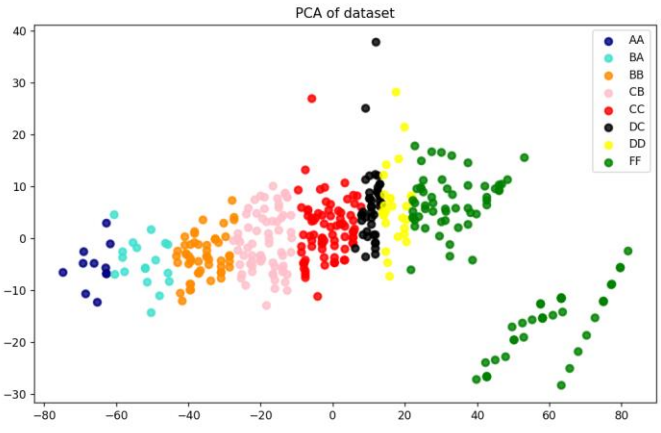
Comparison of PCA and LDA:

- PCA is an unsupervised learning algorithm. LDA is a supervised learning algorithm.
- PCA tries to maximize the distance between data points. LDA tries to maximize the distance between classes.
- PCA is used in clustering problems while LDA is used in classification problems.
- There is no concept of class in PCA. It removes the concept of properties from data. All data is treated as one type.

LDA (Linear Discriminant Analysis)



PCA(Principle Component Analysis)



I expected LDA to be more appropriate since I was doing classification in this project. When I examined the plots, I observed that they were grouped better. So, my expectation was confirmed.

References:

1. https://blog.bigml.com/2016/09/28/logistic-regression-versus-decision-trees/
2. https://www.datasciencecentral.com/comparing-classifiers-decision-trees-knn-naive-bayes/
3. https://www.upgrad.com/blog/random-forest-vs-decision-tree/
4. https://techbigdatacloud.medium.com/veri-biliminde-pca-ve-lda-kavramlar%C4%B1-6039e3aa34de
5. Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018). *Random forest for credit card fraud detection. 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC).*