SCHOOL SUCCESS ANALYSIS & MACHINE LEARNING MODEL

BERFİN OKÇU

TARGETS

Examining the factors affecting the general success of students in Gabriel Pereira and Mousinho da Silveira schools based on their Portuguese exam grades.

Selecting successful students for the Olympiads that will be held at end of the semester with a machine learning model that can predict students' final grades

Methodology

1. Accessing open source data with UCI Machine Learning Repository

2. Analyzing open source data with pandas in Jupyter Notebook environment
   Visualizing these analysis with matplotlib and seaborn libraries

3. Determination of the best model using different machine learning algorithms and building a model that predicts the final grade.

UCI Machine Learning Repository

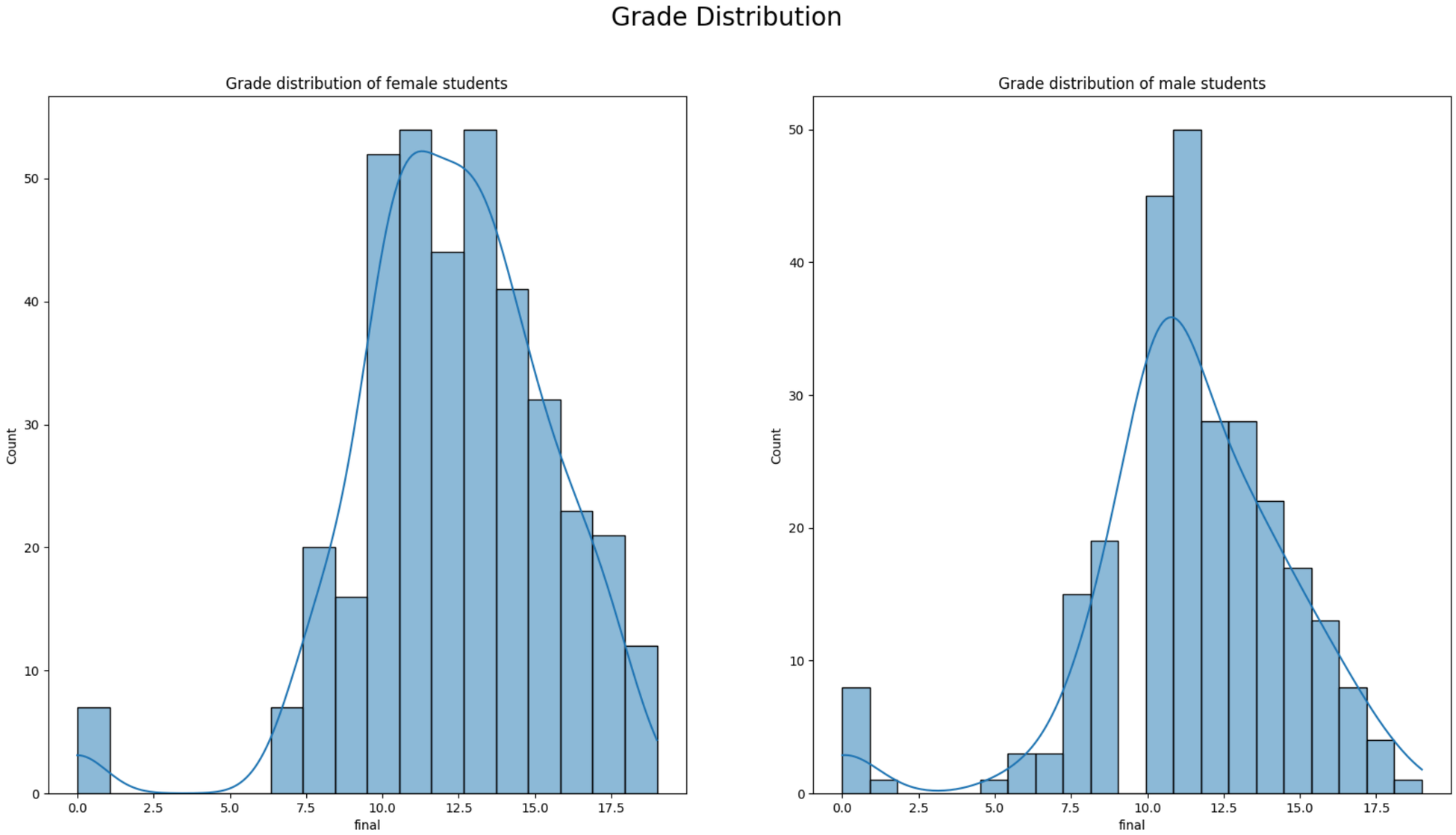
seaborn

jupyter

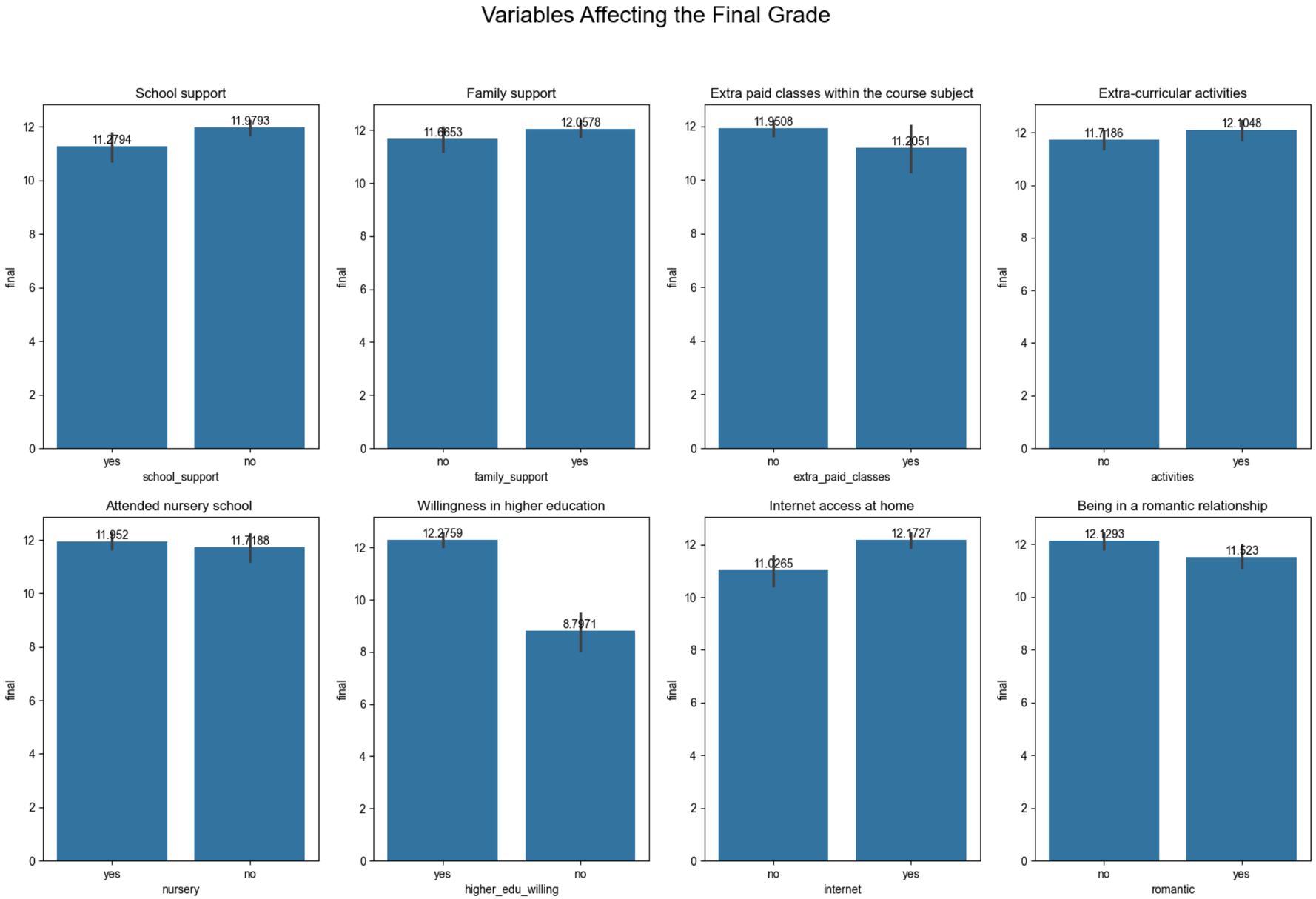
pandas

matplotlib

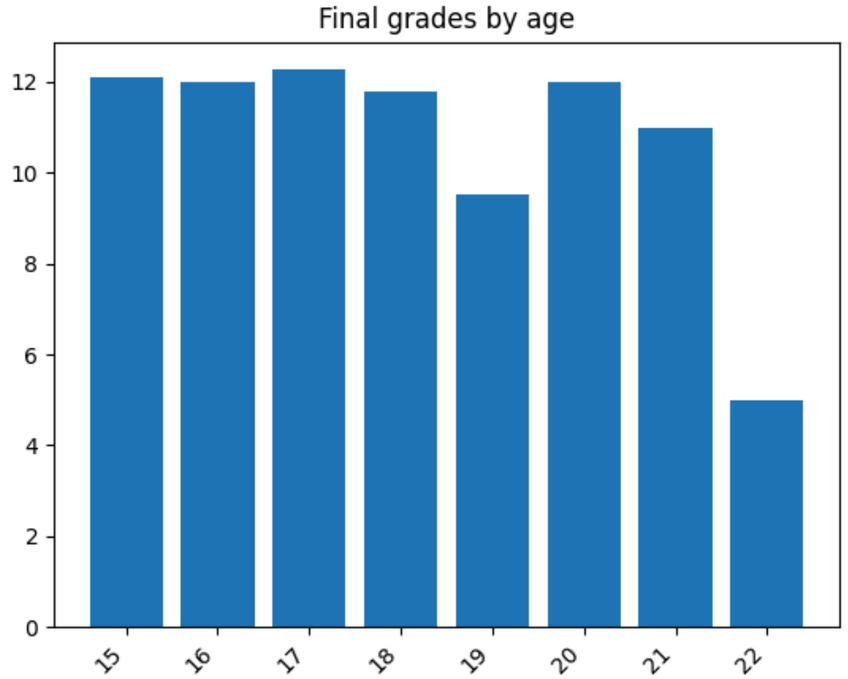
ANALYSIS

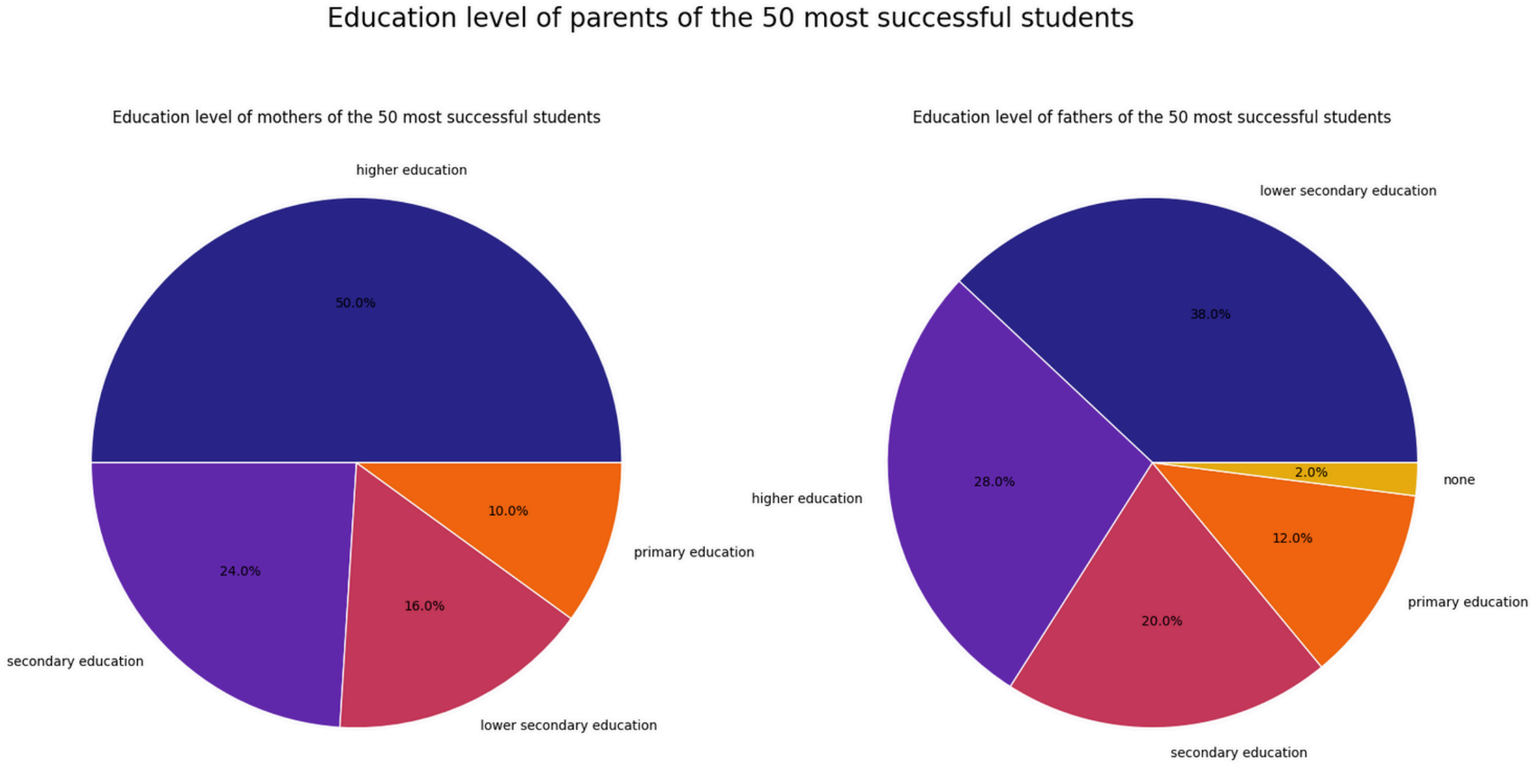

Grade Distribution
Grade distribution of female students
Count
final
0
10
20
30
40
50
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Grade distribution of male students
Count
final
0
10
20
30
40
50
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

ANALYSIS

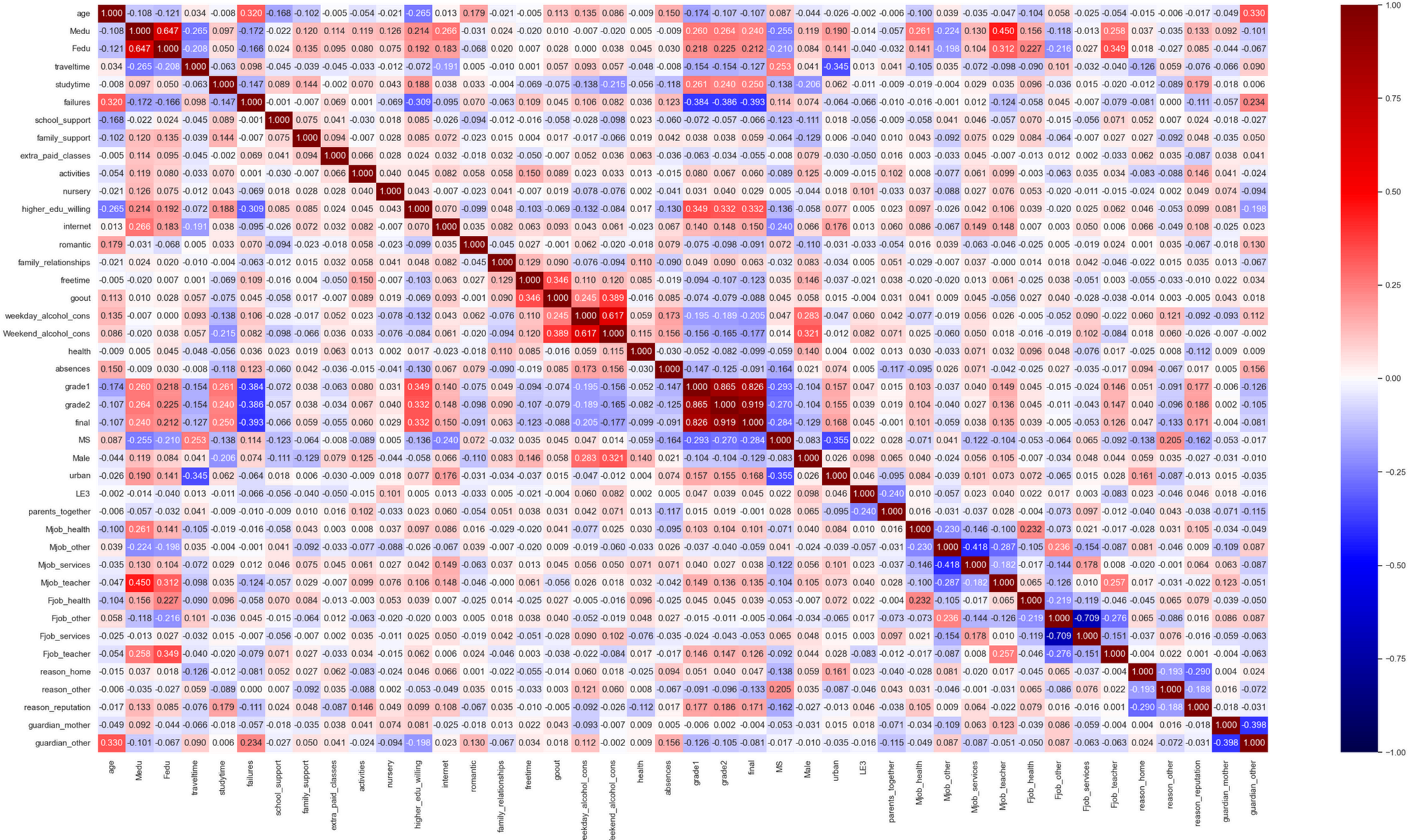

Variables Affecting the Final Grade
School support
11.2794
11.9793
yes
no
school_support
Family support
11.6653
12.0578
no
yes
family_support
Extra paid classes within the course subject
11.9508
11.2051
no
yes
extra_paid_classes
Extra-curricular activities
11.7186
12.1048
no
yes
activities
Attended nursery school
11.952
11.7188
yes
no
nursery
Willingness in higher education
12.2759
8.7971
yes
no
higher_edu_willing
Internet access at home
11.0265
12.1727
no
yes
internet
Being in a romantic relationship
12.1293
11.523
no
yes
romantic
final

ANALYSIS

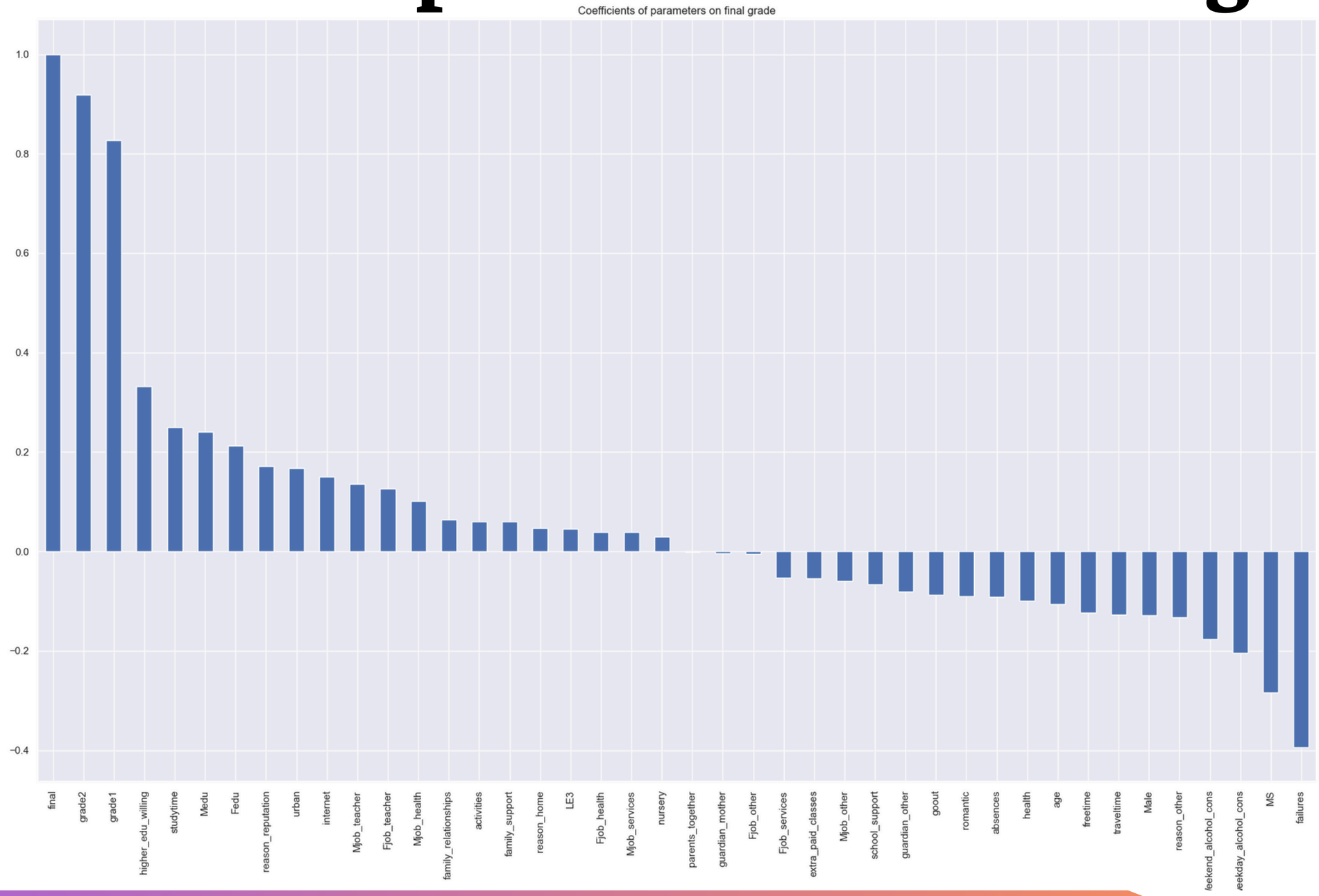

Final grades by age
0
2
4
6
8
10
12
15
16
17
18
19
20
21
22


Education level of parents of the 50 most successful students
Education level of mothers of the 50 most successful students
higher education
50.0%
primary education
10.0%
24.0%
16.0%
secondary education
lower secondary education
Education level of fathers of the 50 most successful students
lower secondary education
38.0%
higher education
28.0%
2.0%
none
12.0%
primary education
20.0%
secondary education

Other features that affect the final grade



Coefficients of parameters on final grade



Optimum Parameters Selected for Machine Learning Model

```
lsm2 = smf.ols( 'final ~ grade2 + grade1 + failures + MS + weekday_alcohol_cons', data=df)
fit2 = lsm2.fit()
fit2.summary()
```

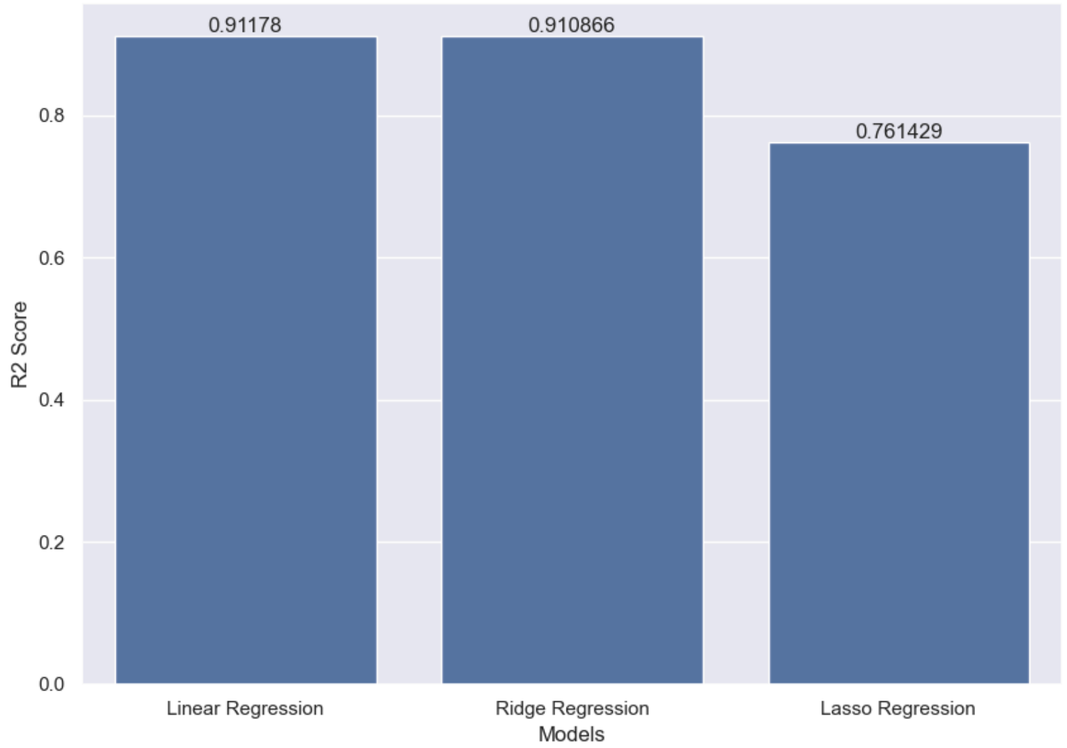
OLS Regression Results

Dep. Variable:	final	**R-squared:**	0.851
Model:	OLS	**Adj. R-squared:**	0.849
Method:	Least Squares	**F-statistic:**	732.5
Date:	Sun, 07 Apr 2024	**Prob (F-statistic):**	1.09e-262
Time:	12:49:51	**Log-Likelihood:**	-1064.4
No. Observations:	649	**AIC:**	2141.
Df Residuals:	643	**BIC:**	2168.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5011	0.288	1.740	0.082	-0.064	1.067
grade2	0.8834	0.034	26.022	0.000	0.817	0.950
grade1	0.1271	0.036	3.501	0.000	0.056	0.198
failures	-0.2066	0.091	-2.282	0.023	-0.384	-0.029
MS	-0.2180	0.108	-2.017	0.044	-0.430	-0.006
weekday_alcohol_cons	-0.0948	0.054	-1.745	0.081	-0.202	0.012

-First exam grade
-Second exam grade
-Number of past class failures
-Student's school
-Weekday alcohol consumption of the student

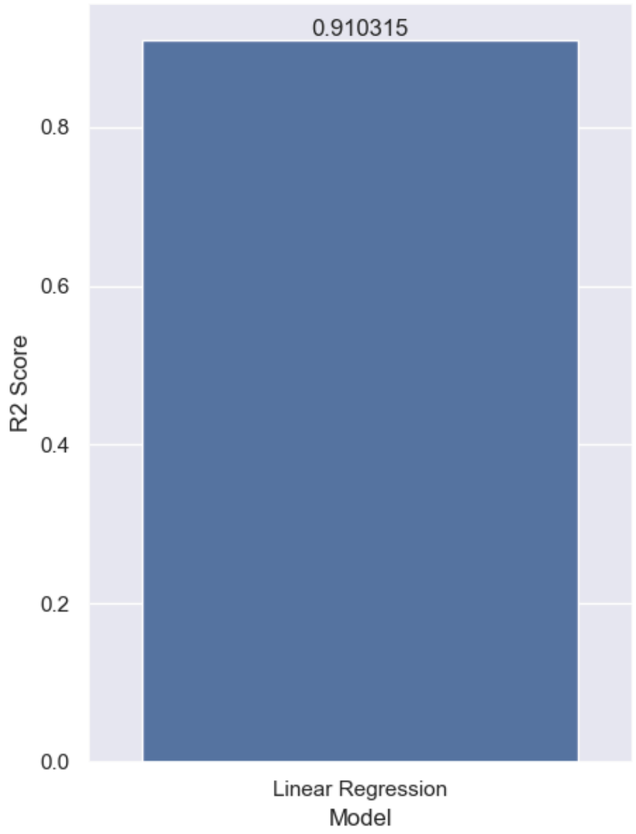
Model Selection



-Linear Regression
-Ridge Regression
-Lasso Regression

Final Model Decision with Cross Validation

Model	Cross Validated Score (Mean)	Cross Validated Score (Std)
Linear Regression	0.865621	0.105227
Ridge Regression	0.865593	0.105219



The success of model that built with the Linear Regression algorithm is 91%