MACHINE LEARNING I REGRESSION ANALYSIS OF STUDENT GRADE DATA

BUĞRA DUMAN

KAMIL MATUSZELAŃSKİ

SCOPE

✓ Description of Dataset

✓EDA

✓Feature selection

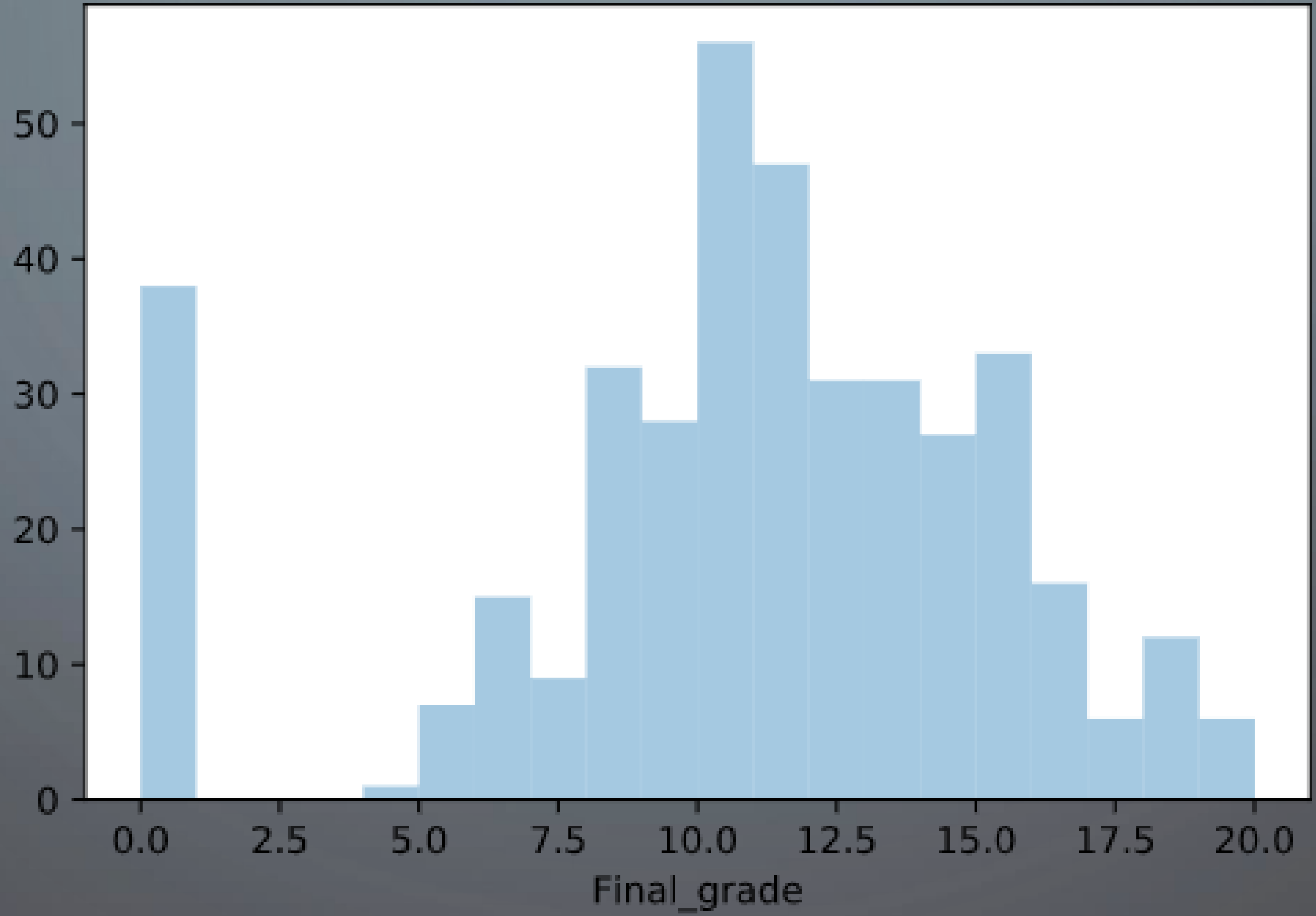
✓Models (Linear,Ridge,Lasso,KNN)

FEATURES

- **MATH FINAL GRADE**
- Father Education
- Failures
- Traveltime
- Studytime
- Weekday Alcohol consumption
- Weekend Alcohol consumption
- Parent Status
- Romantic

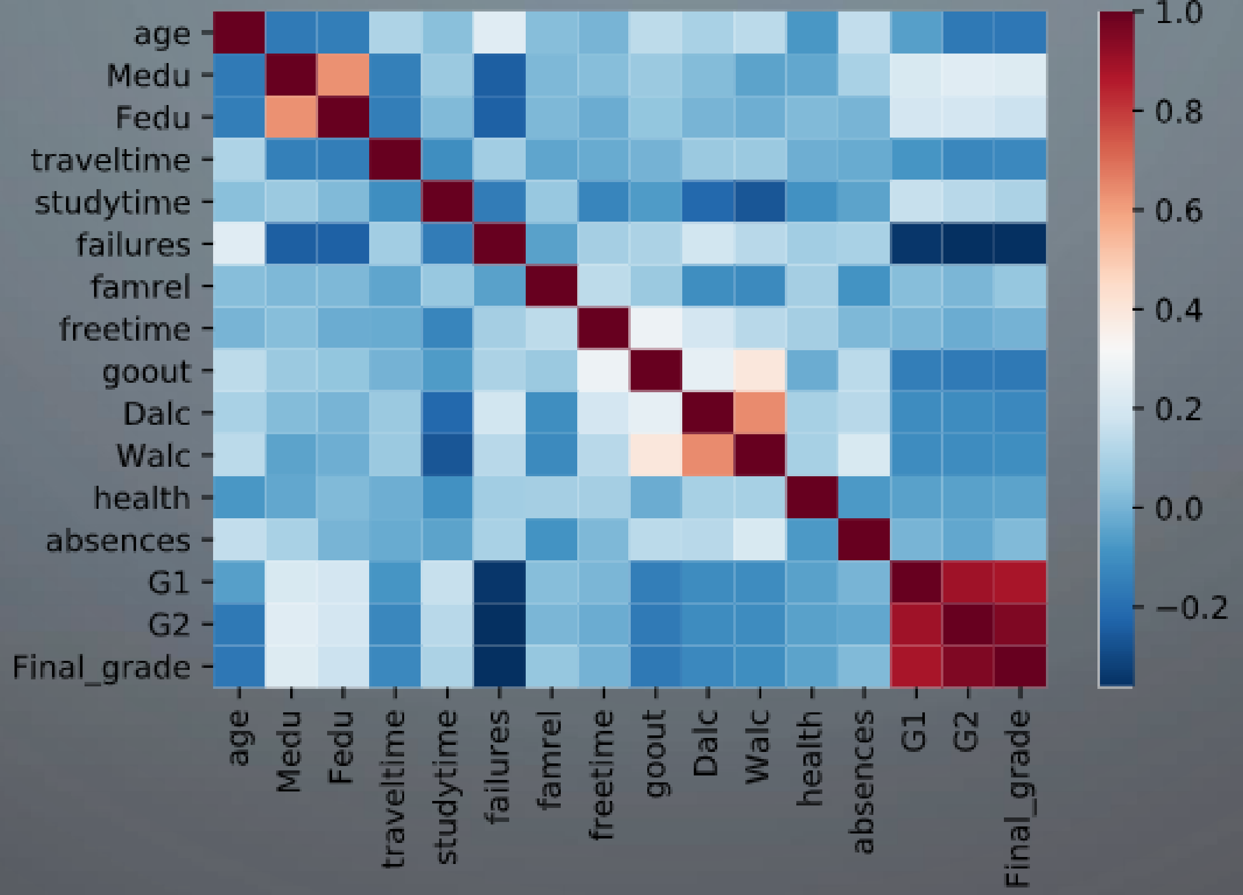
- Absences
- Final Grade
- Go out
- Family relation
- Guardion
- Mother Job
- Father Job
- Sex
- Age

FINAL GRADE

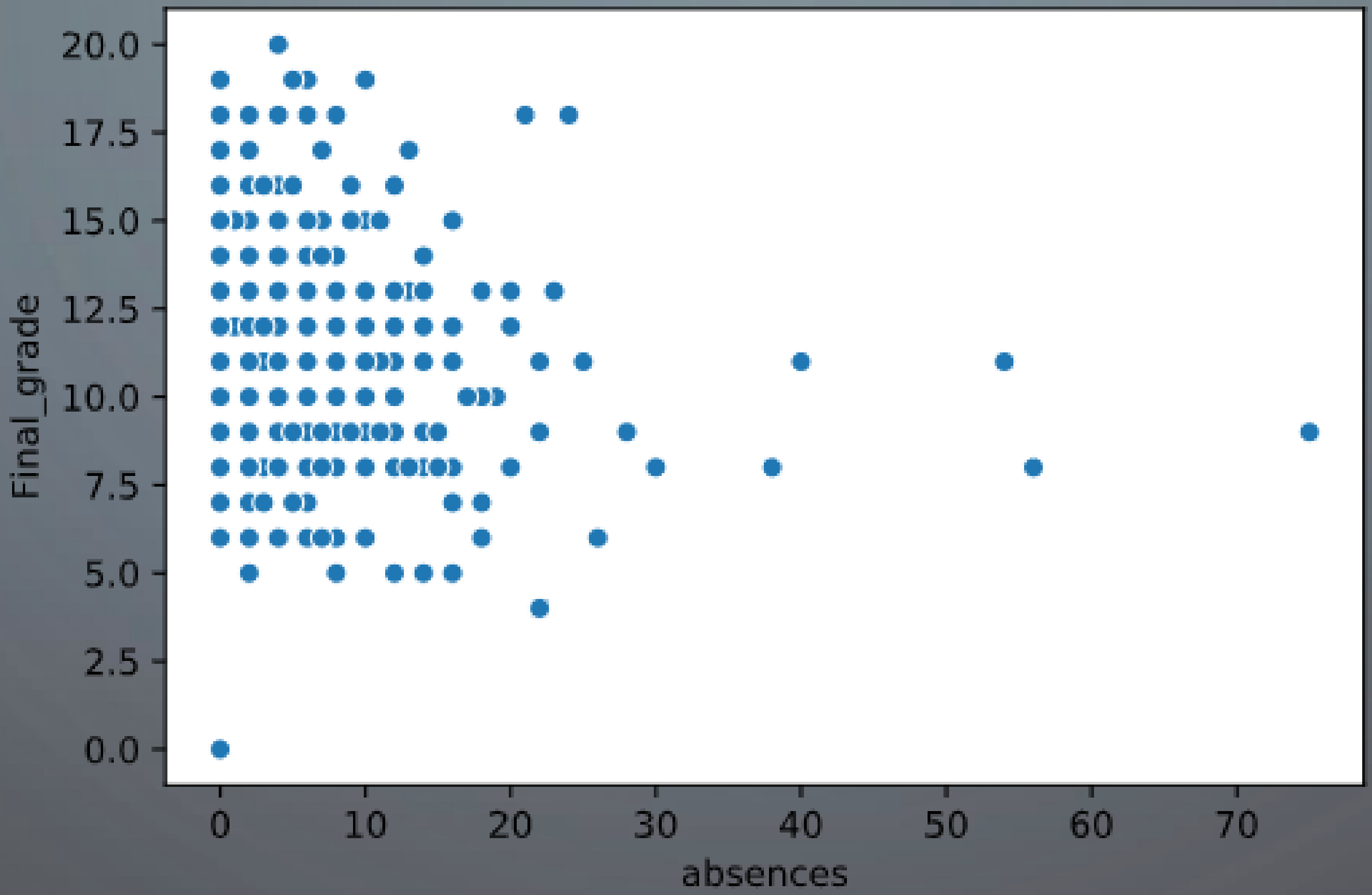


CORRELATION MATRIX

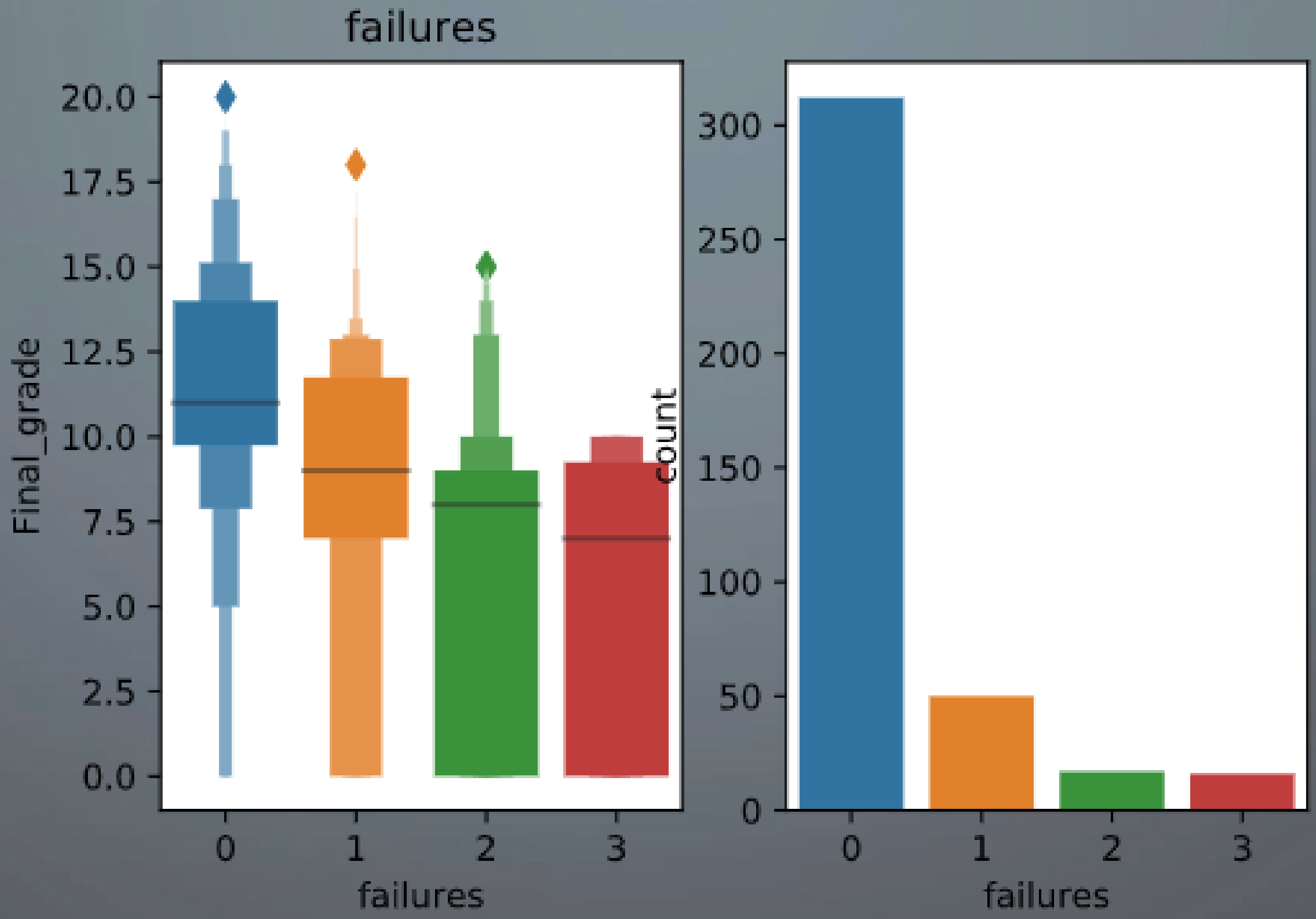
Correlation matrix based on Spearman



ABSENCES

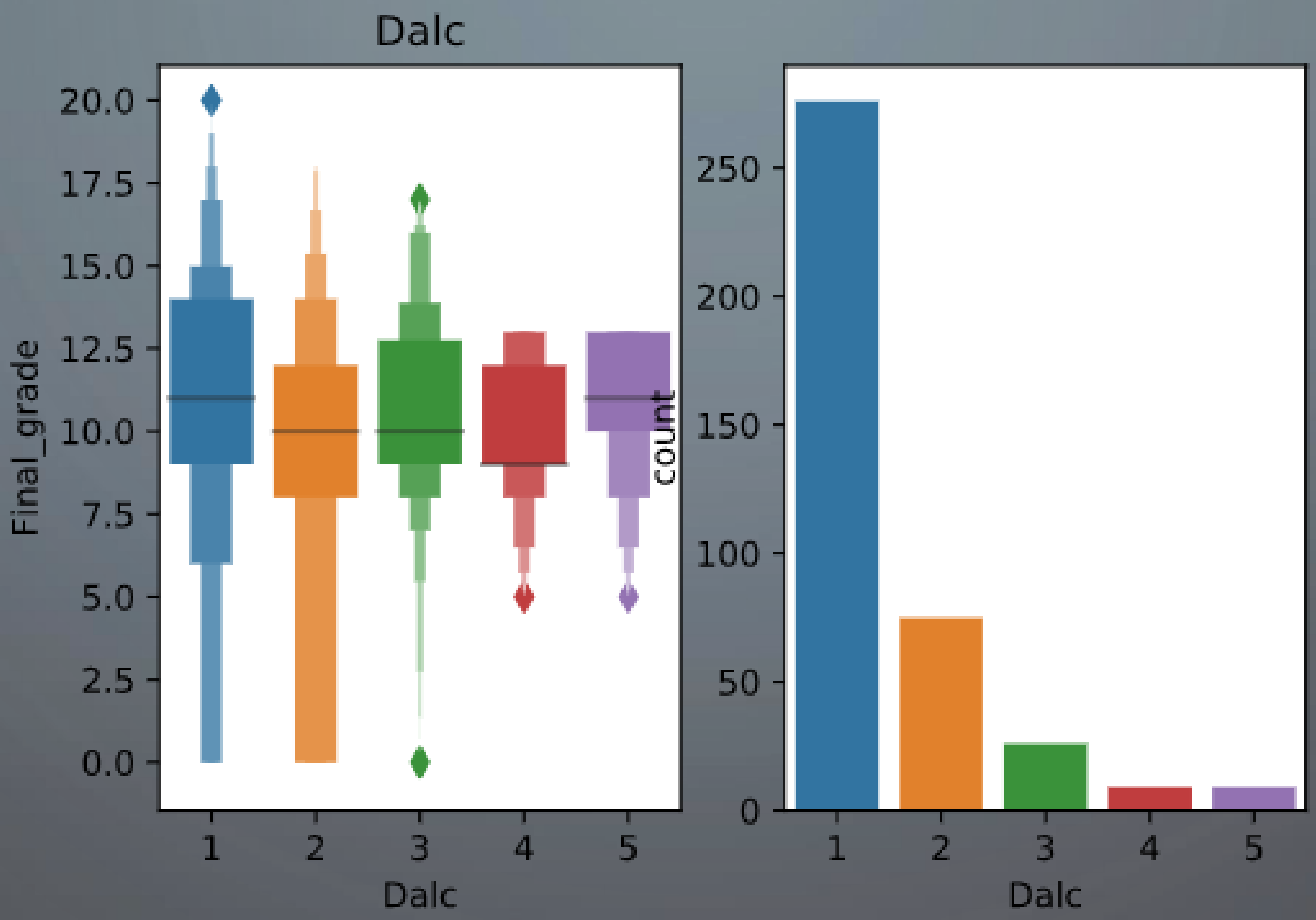

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Final_grade
0
10
20
30
40
50
60
70
absences

FAILURES

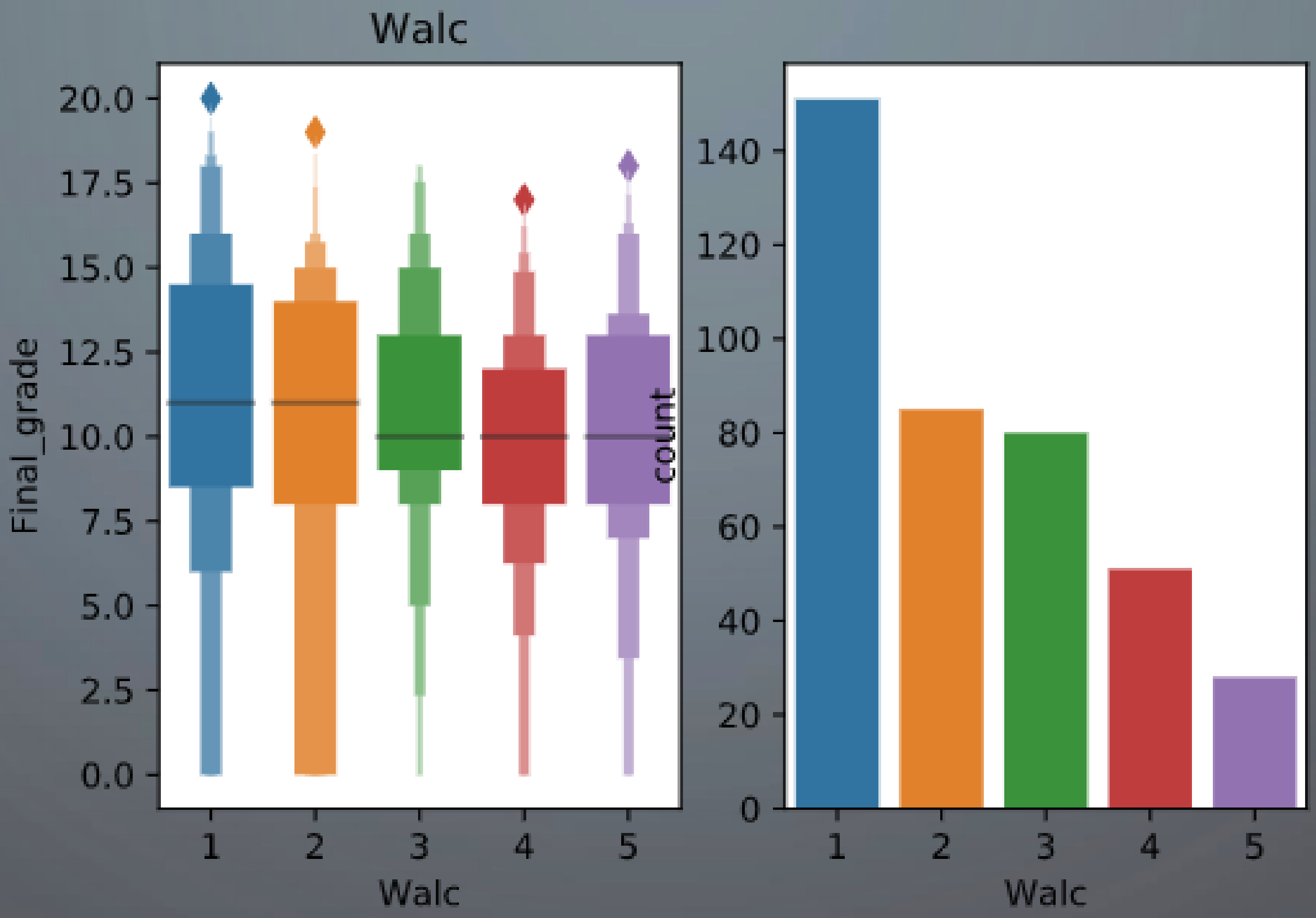


failures
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Final_grade
0
1
2
3
failures
300
250
200
150
100
50
0
count
0
1
2
3
failures

DAILY ALCOHOL CONSUMPTION



WEEKLY ALCOHOL CONSUMPTION



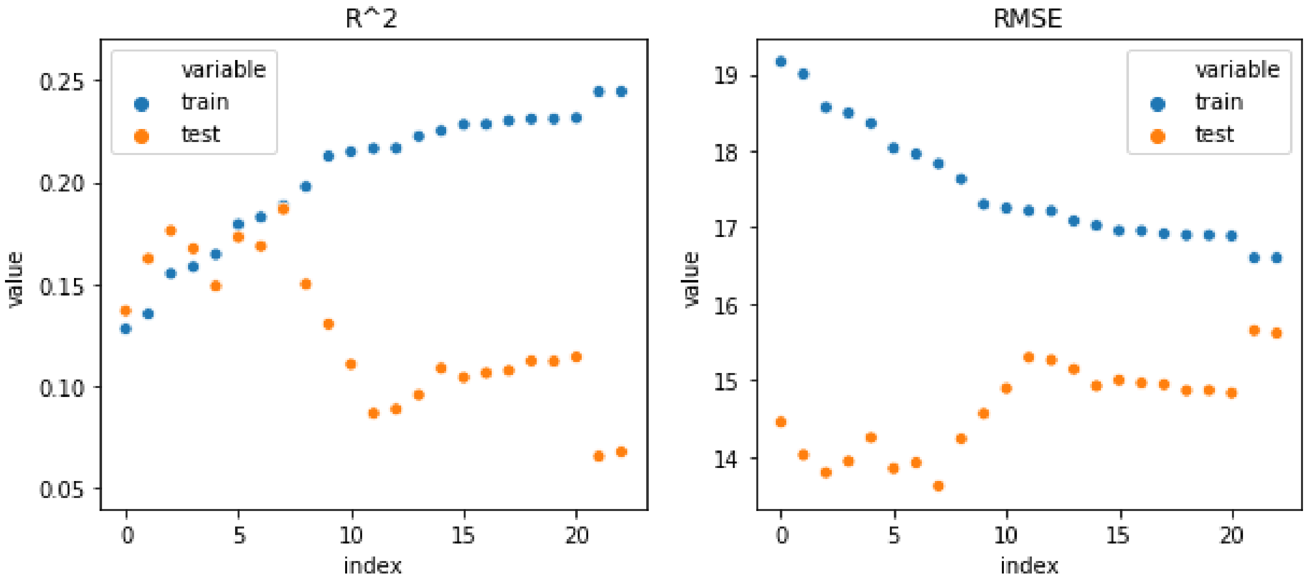
METHODS

- Linear model
- SVM
- KNN
- Ridge and Lasso
- Hurdle model

LINEAR MODEL

- With all features
- R2 train: 0.24
- R2 test: 0.06
- Overfitting

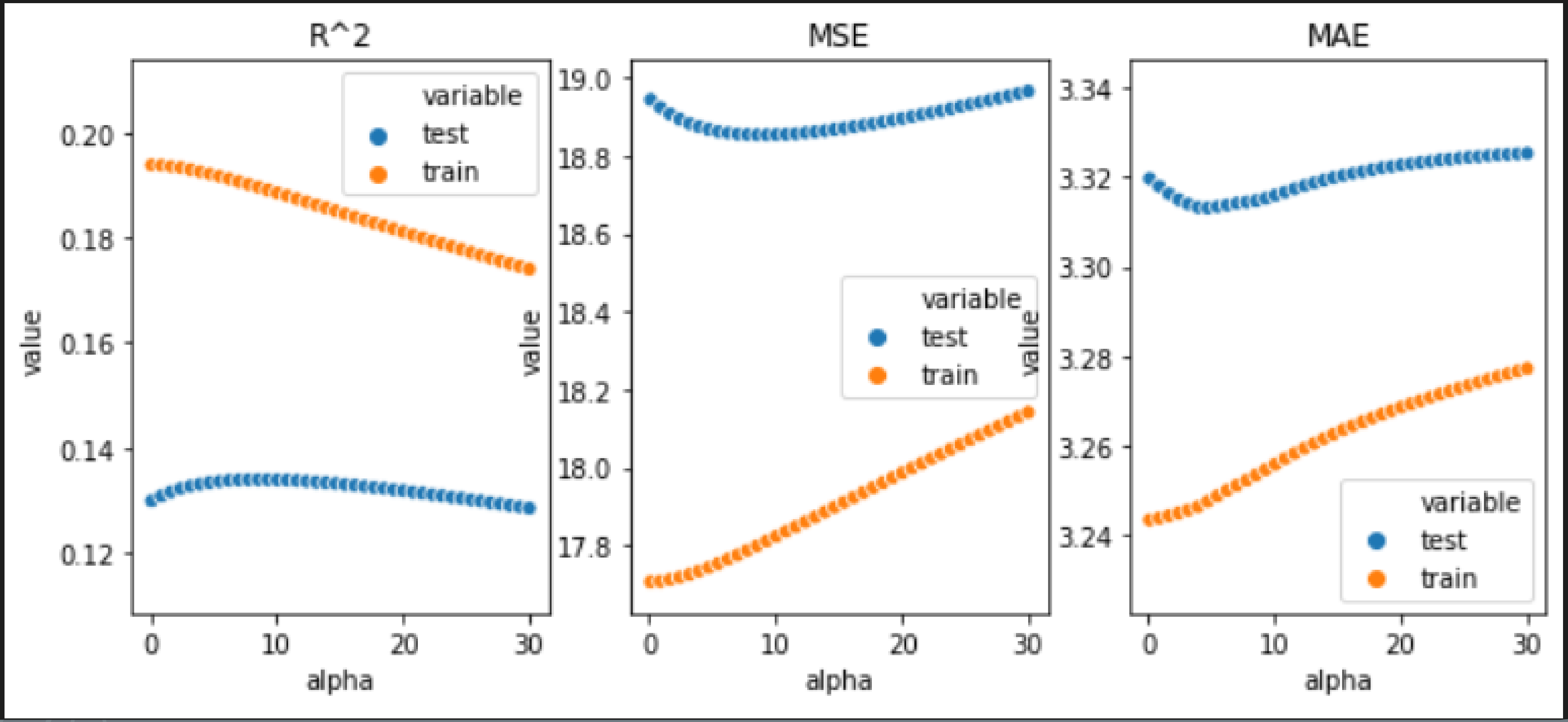
FIGHTING WITH OVERFITTING – RFE



R2 TRAIN: 0.1886

R2 TEST: 0.1868

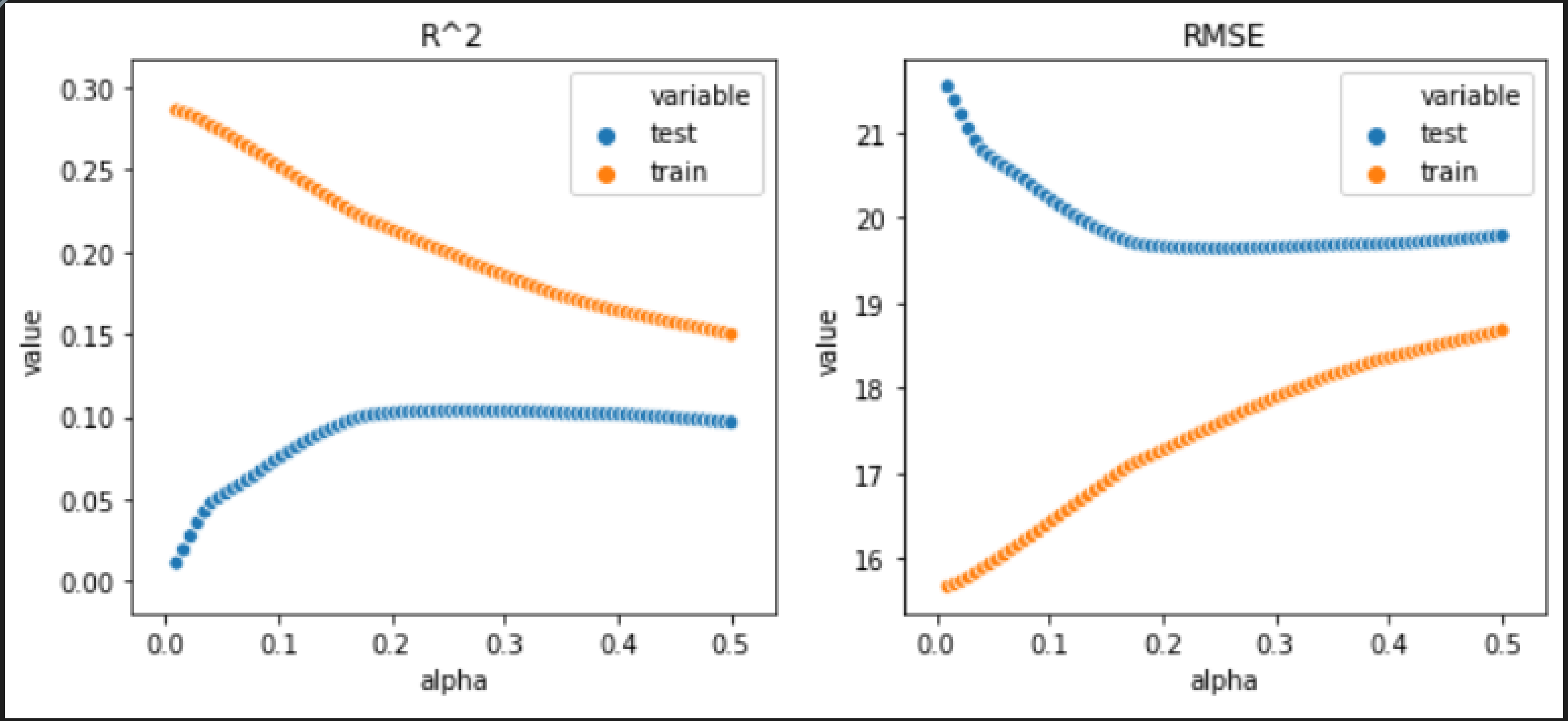
FIGHTING WITH OVERFITTING -RIDGE



R2 TRAIN: 0.184

R2 TEST: 0.195

FIGHTING WITH OVERFITTING -LASSO

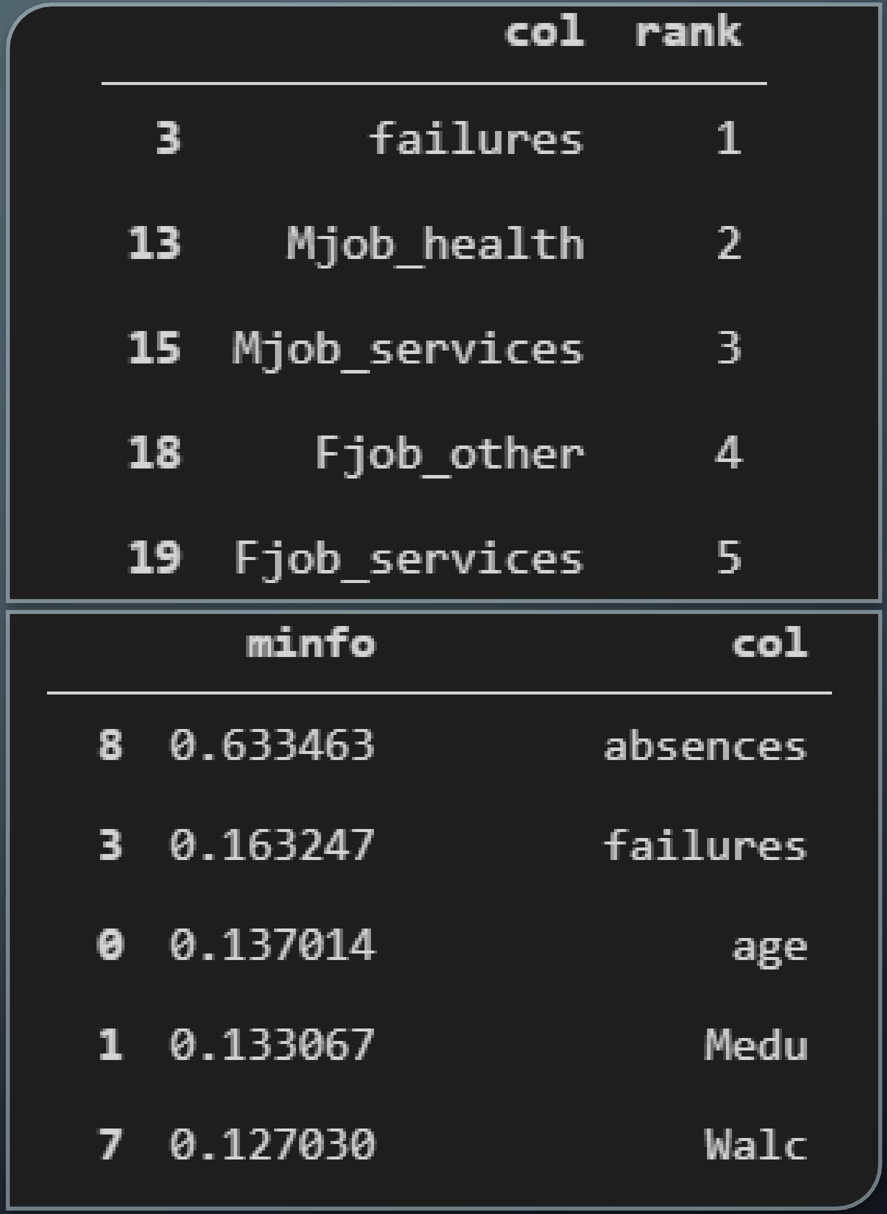


- R2 train: 0.182
- R2 test: 0.100

FEATURE SELECTION

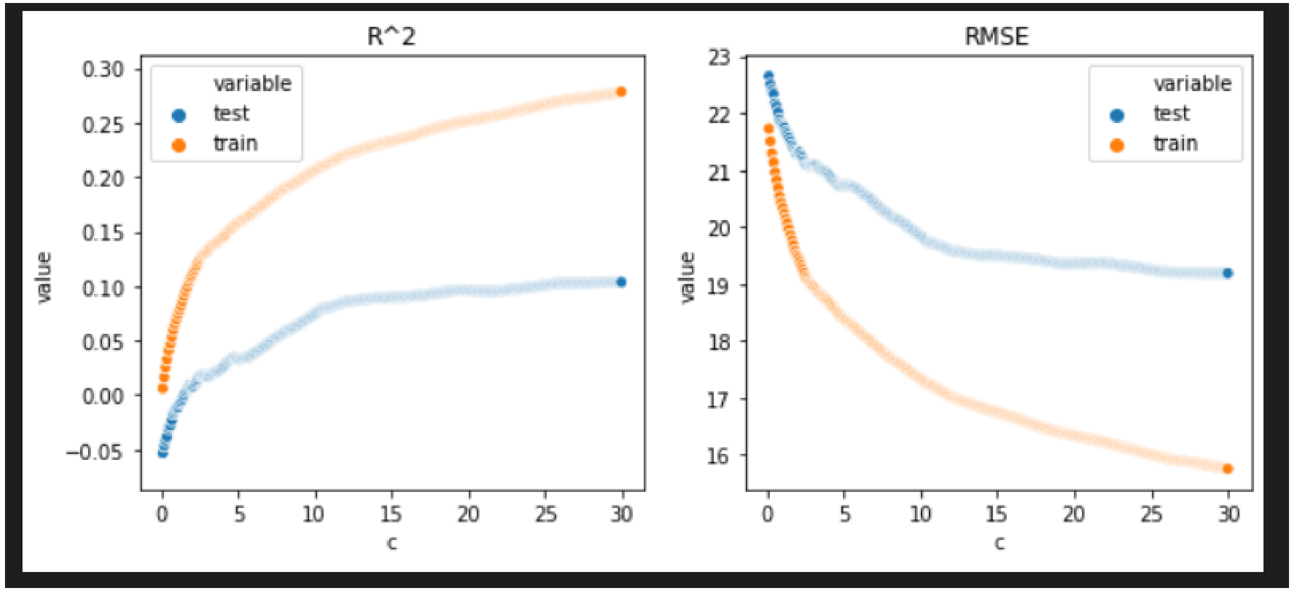
- RFE

- Mutual Info



	col	rank
3	failures	1
13	Mjob_health	2
15	Mjob_services	3
18	Fjob_other	4
19	Fjob_services	5

	minfo	col
8	0.633463	absences
3	0.163247	failures
0	0.137014	age
1	0.133067	Medu
7	0.127030	Walc



SVM

- R2 train: 0.2270
- R2 test: 0.1984

OTHER MODELS

WITH ALL FEATURES

- KNN
- Hurdle model

CONCLUSIONS

- R^2 around 20% is not good, but pretty standard on social data
- SVM was the best model, with Ridge being almost as good
- Feature generation, KNN, Hurdle models did not improve performance