INFORMACIJOS TEORIJA IR DUOMENŲ STRUKTŪRA

PROJEKTAS

Robertas Kudlis
2019 m. Lapkritis - Gruodis

1. DUOMENŲ APŽVALGA BEI PROJEKTO TIKSLAS

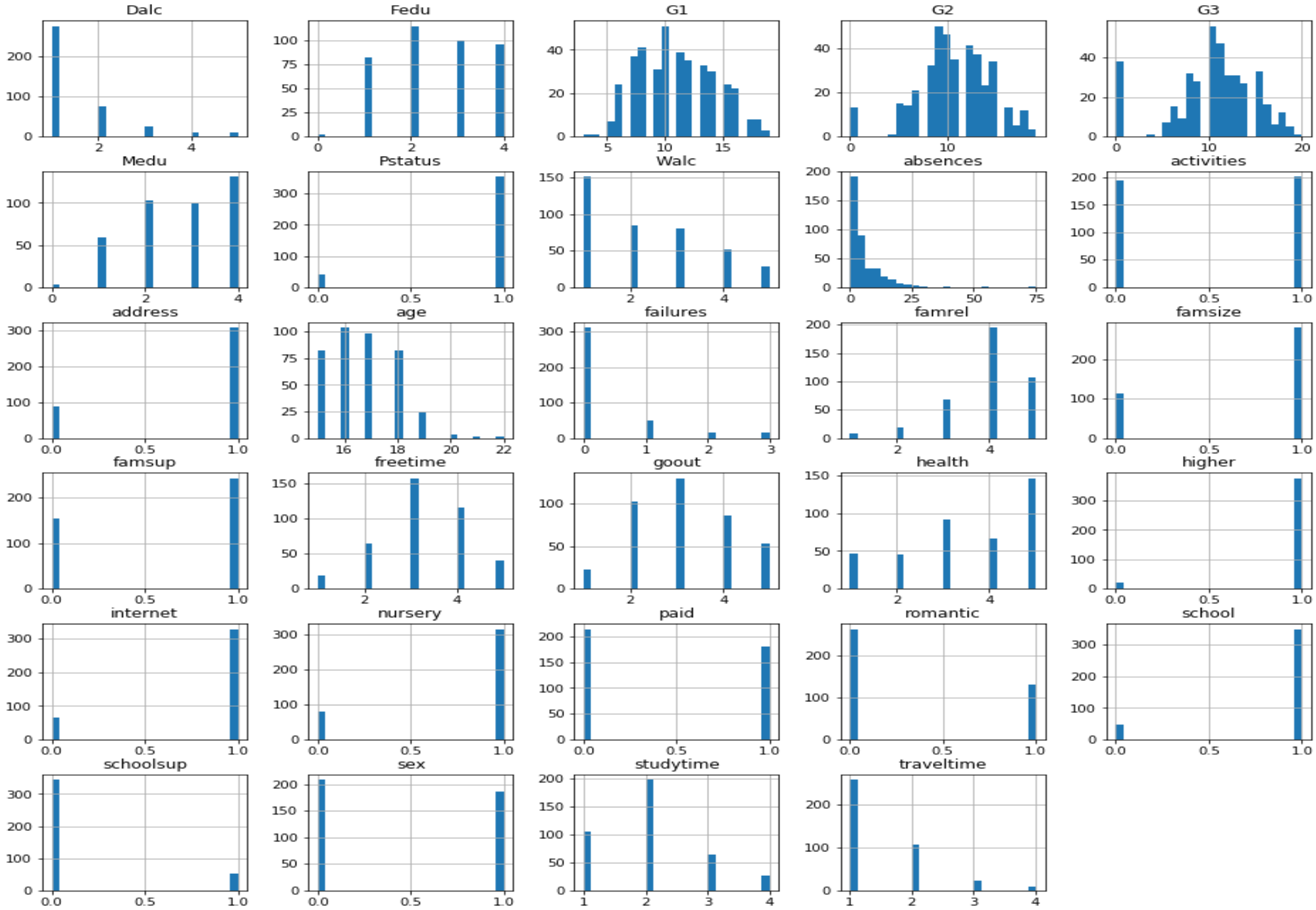
Turime įvairios socialinės, mokslo tematikos bei asmeninės informacijos apie dviejų Portugalijos gimnazijos mokinius.

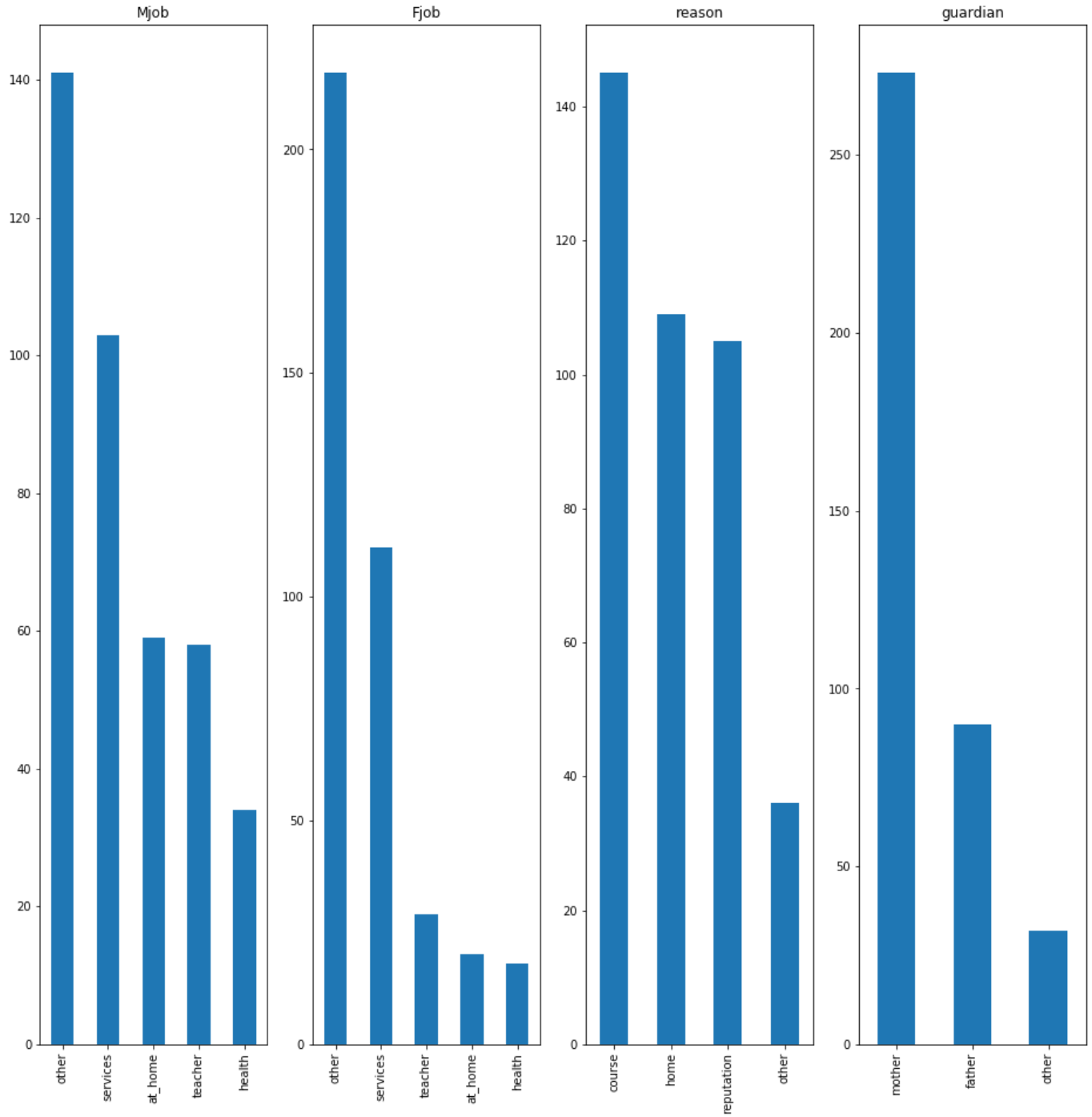
Duomenyse nėra praleistų reikšmių bei „outlier‘ių“.

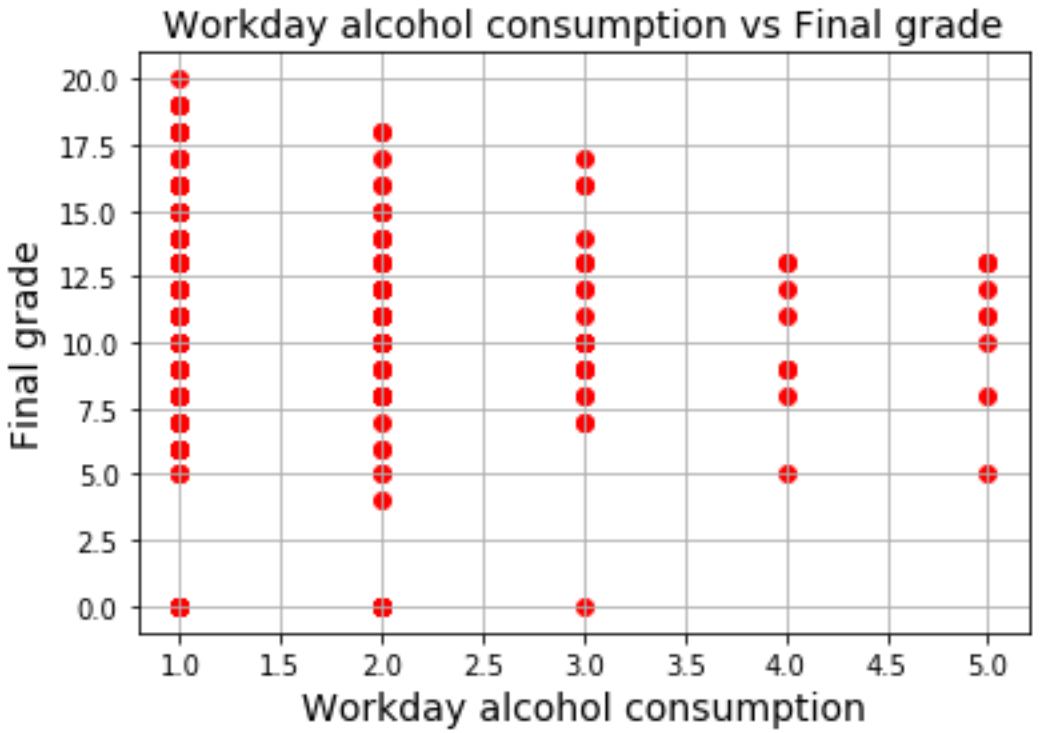
395 stebėjimai, 33 kintamieji (17 kategorinių kintamųjų).

Projekto tikslai:

1. taikant tiesinės regresijos modelį nuspėti metinį pažymį;
2. įvertinti alkoholio vartojimo įtaką mokslo rezultatams.


Dalc
Fedu
G1
G2
G3
Medu
Pstatus
Walc
absences
activities
address
age
failures
famrel
famsize
famsup
freetime
goout
health
higher
internet
nursery
paid
romantic
school
schoolsup
sex
studytime
traveltime


Mjob
Fjob
reason
guardian
other
services
at_home
teacher
health
other
services
teacher
at_home
health
course
home
reputation
other
mother
father
other



Iš grafikų ne visiškai aišku, kaip (ar) mokslo rezultatai priklauso nuo alkoholio vartojimo.

2. TIESINĖS REGRESIJOS MODELIS $\min_{\omega} ||X\omega - y||_2^2$

2.1 Modelis be kategorinių kintamųjų

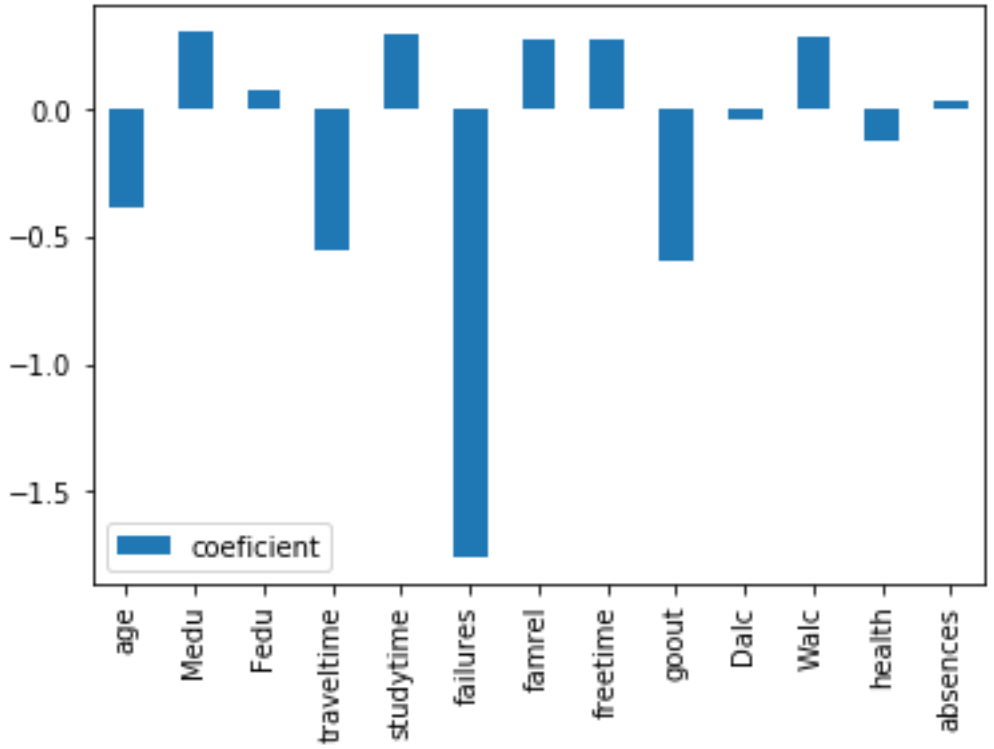
Su „Scikit-learn“:

```
Intercept:
 15.819972157120654
Coefficients:
 [-0.37880415  0.31208936  0.07500229 -0.55448176  0.29928308 -1.76279197
  0.2733752   0.27337528 -0.59621525 -0.03262039  0.2911628  -0.11937624
  0.0408884 ]
```

RMSE = 3.97

Alkoholio vartojimas nėra pats svarbiausias kintamasis.

Ypač pažymį lemia nesėkmės įvairiuose atsiskaitymuose mokslo metų eigoje.



Su „StatsModels“:

OLS Regression Results

Dep. Variable:	G3	R-squared:	0.172
Model:	OLS	Adj. R-squared:	0.136
Method:	Least Squares	F-statistic:	4.830
Date:	Mon, 16 Dec 2019	Prob (F-statistic):	1.16e-07
Time:	13:52:35	Log-Likelihood:	-901.35
No. Observations:	316	AIC:	1831.
Df Residuals:	302	BIC:	1883.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	15.8200	3.748	4.220	0.000	8.444	23.196
age	-0.3788	0.200	-1.890	0.060	-0.773	0.016
Medu	0.3121	0.295	1.057	0.291	-0.269	0.893
Fedu	0.0750	0.289	0.259	0.796	-0.494	0.644
traveltime	-0.5545	0.353	-1.571	0.117	-1.249	0.140
studytime	0.2993	0.314	0.952	0.342	-0.319	0.918
failures	-1.7628	0.355	-4.960	0.000	-2.462	-1.063
famrel	0.2734	0.279	0.981	0.328	-0.275	0.822
freetime	0.2734	0.266	1.028	0.305	-0.250	0.797
goout	-0.5962	0.253	-2.354	0.019	-1.095	-0.098
Dalc	-0.0326	0.368	-0.089	0.929	-0.757	0.692
Walc	0.2912	0.279	1.044	0.297	-0.258	0.840
health	-0.1194	0.178	-0.672	0.502	-0.469	0.230
absences	0.0409	0.030	1.385	0.167	-0.017	0.099

2.2 Modelis su kategoriniais kintamaisiais

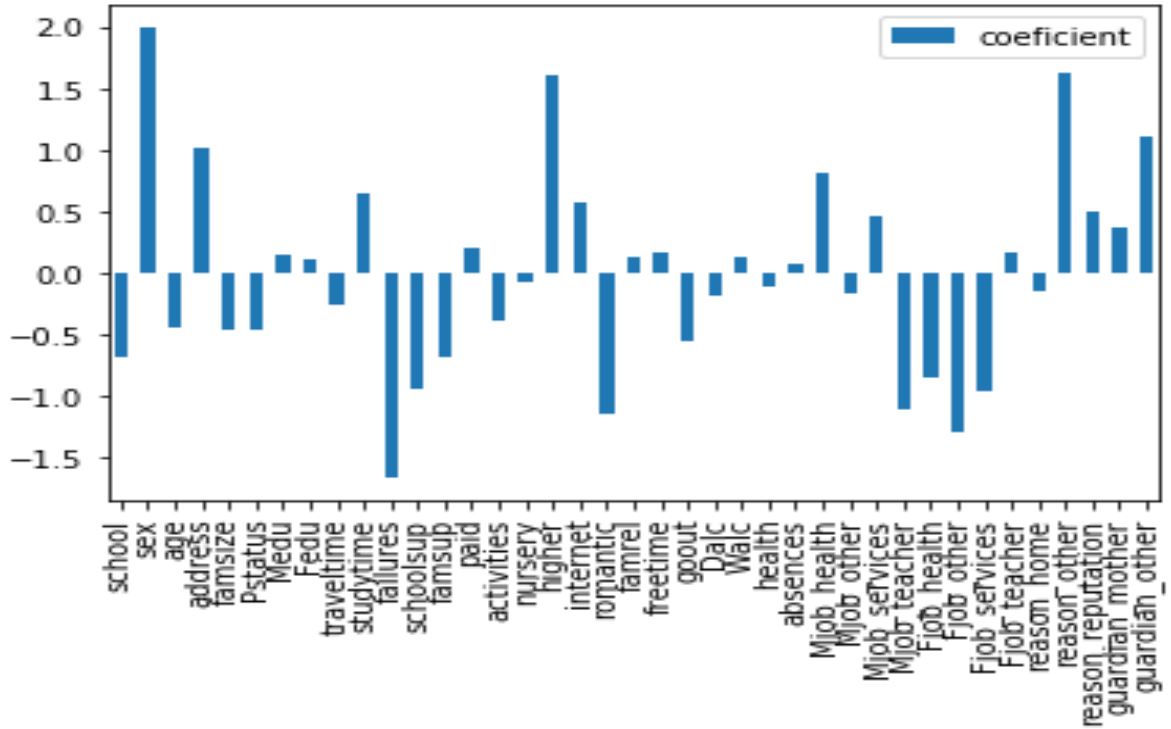
Su „Scikit-learn“:

```
Intercept:
 16.515547690962336
Coefficients:
 [-0.6879941   1.98806008 -0.44858168  1.02377142 -0.4594246  -0.46249174
  0.15556825  0.10874881 -0.25653459  0.64556608 -1.65953604 -0.94042591
 -0.69082961  0.21141089 -0.38445894 -0.08279663  1.61339245  0.57459392
 -1.14352202  0.12471082  0.15892007 -0.55238316 -0.17739516  0.13775681
 -0.10665675  0.06565028  0.80648221 -0.16244308  0.45798483 -1.11082653
 -0.85327885 -1.29540106 -0.95306067  0.16368931 -0.14646815  1.62150839
  0.49171218  0.36837628  1.11422649]
```

RMSE = 4.08

Alkoholio vartojimas vis dar nėra pats svarbiausias kintamasis. Spėjame, kad toks ir nebus.

Ypač pažymį lemia nesėkmės įvairiuose atsiskaitymuose mokslo metų eigoje. Išaugo lyties bei noro studijuoti aukštąjį mokslą kintamieji. Nemažai įtakos turi x_other kintamieji.



Su „StatsModels“:

OLS Regression Results

Dep. Variable:	G3	R-squared:	0.281
Model:	OLS	Adj. R-squared:	0.179
Method:	Least Squares	F-statistic:	2.761
Date:	Mon, 16 Dec 2019	Prob (F-statistic):	8.16e-07
Time:	14:57:56	Log-Likelihood:	-879.15
No. Observations:	316	AIC:	1838.
Df Residuals:	276	BIC:	1989.
Df Model:	39		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	16.5155	5.327	3.100	0.002	6.029	27.002
school	-0.6880	0.909	-0.757	0.450	-2.477	1.101
sex	1.9881	0.573	3.469	0.001	0.860	3.116
age	-0.4486	0.243	-1.844	0.066	-0.927	0.030
address	1.0238	0.685	1.495	0.136	-0.324	2.372
famsize	-0.4594	0.579	-0.793	0.429	-1.600	0.681
Pstatus	-0.4625	0.796	-0.581	0.562	-2.030	1.105
Medu	0.1556	0.379	0.411	0.681	-0.590	0.901
Fedu	0.1087	0.318	0.341	0.733	-0.518	0.736
traveltime	-0.2565	0.380	-0.675	0.500	-1.005	0.492
studytime	0.6456	0.335	1.928	0.055	-0.014	1.305
failures	-1.6595	0.386	-4.300	0.000	-2.419	-0.900
schoolsup	-0.9404	0.769	-1.223	0.222	-2.454	0.573
famsup	-0.6908	0.562	-1.229	0.220	-1.797	0.415
paid	0.2114	0.549	0.385	0.700	-0.869	1.292

2.3 Modelis su pašalintais x_other kintamaisiais

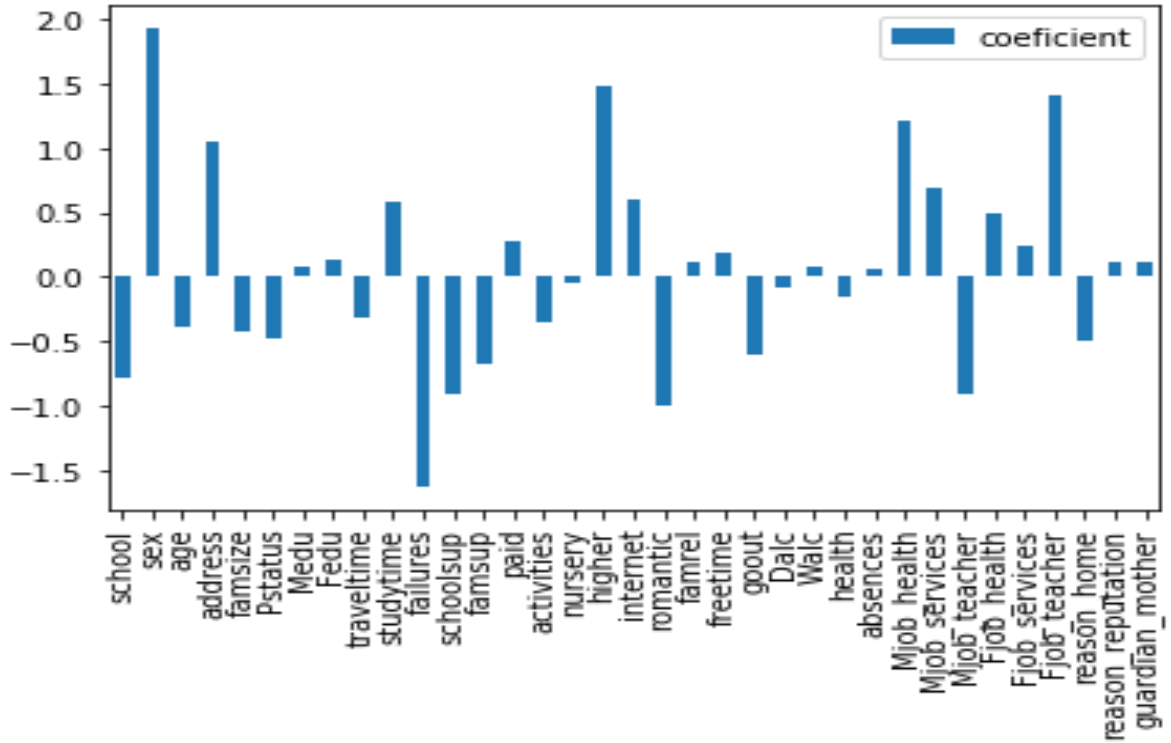
Su „Scikit-learn“:

```
Intercept:
 15.168860451123779
Coefficients:
 [-0.77815019  1.92621504 -0.37808336  1.04245448 -0.42162897 -0.48024593
  0.08695737  0.13252828 -0.31454759  0.5869862  -1.62511441 -0.90807696
 -0.68036271  0.27142816 -0.34735577 -0.05269362  1.4880803   0.60136687
 -1.0009617   0.11705264  0.19516531 -0.59671207 -0.07756903  0.07761958
 -0.16121046  0.06983654  1.20972992  0.6981021  -0.90256273  0.50226413
  0.23988597  1.41713774 -0.49899013  0.11604207  0.1096533 ]
```

RMSE = 3.96

Sprendimas pašalinti x_other kintamuosius pasiteisino.

Alkoholio įtaka nesikeičia, padidėja kintamųjų – motinos bei tėvo darbo sričių ir mokiniu santykių statuso – įtaka.



Su „StatsModels“:

OLS Regression Results

Dep. Variable:	G3	R-squared:	0.267
Model:	OLS	Adj. R-squared:	0.175
Method:	Least Squares	F-statistic:	2.907
Date:	Mon, 16 Dec 2019	Prob (F-statistic):	5.62e-07
Time:	15:07:09	Log-Likelihood:	-882.22
No. Observations:	316	AIC:	1836.
Df Residuals:	280	BIC:	1972.
Df Model:	35		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	15.1689	5.045	3.007	0.003	5.238	25.100
school	-0.7782	0.900	-0.865	0.388	-2.550	0.993
sex	1.9262	0.572	3.366	0.001	0.800	3.053
age	-0.3781	0.232	-1.632	0.104	-0.834	0.078
address	1.0425	0.685	1.522	0.129	-0.306	2.391
famsize	-0.4216	0.577	-0.731	0.466	-1.558	0.714
Pstatus	-0.4802	0.793	-0.606	0.545	-2.041	1.081
Medu	0.0870	0.365	0.238	0.812	-0.631	0.805
Fedu	0.1325	0.314	0.422	0.674	-0.486	0.751
traveltime	-0.3145	0.375	-0.839	0.402	-1.053	0.424
studytime	0.5870	0.334	1.759	0.080	-0.070	1.244
failures	-1.6251	0.375	-4.339	0.000	-2.362	-0.888
schoolsup	-0.9081	0.770	-1.179	0.239	-2.424	0.608
famsup	-0.6804	0.559	-1.217	0.225	-1.781	0.420
paid	0.2714	0.546	0.497	0.620	-0.804	1.347

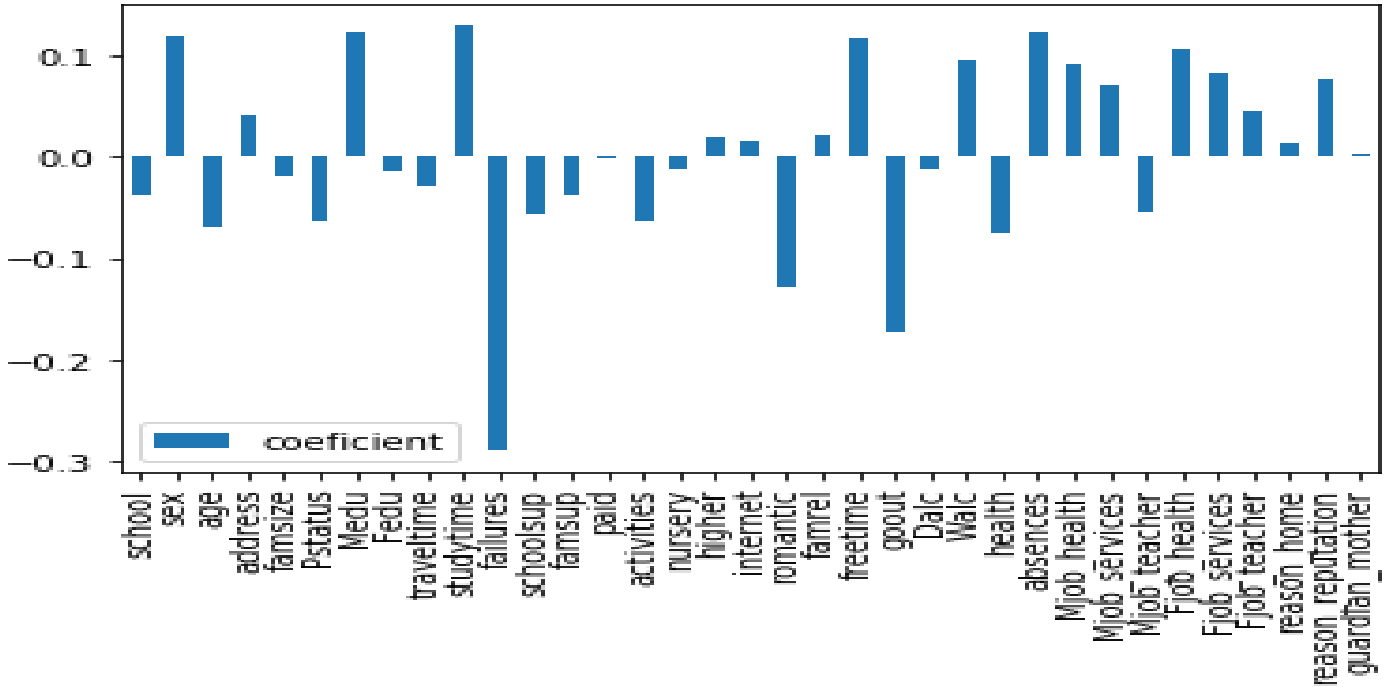
2.4 Modelis, kuriam pritaikytas „Standard Scaler“

Su „Scikit-learn“:

```
Train set shape (294, 42), test set shape (101, 42)
RMSE = 11.12
Intercept:
 -5.3008030818581925e-17
Coefficients:
 [-0.03675064  0.11868185 -0.06959185  0.04058038 -0.01863582 -0.06215937
  0.12416723 -0.01325734 -0.02784664  0.12931203 -0.28902182 -0.05719149
 -0.03792588 -0.0019099  -0.06353105 -0.01296219  0.01961907  0.01619854
 -0.12808763  0.02215389  0.1177681  -0.17241606 -0.01279758  0.09672989
 -0.07568837  0.12283222  0.09175773  0.07012407 -0.05427066  0.1062533
  0.08397164  0.0449209   0.01299928  0.07757926  0.00367748]
```

Gal nereikėjo? Dauguma kintamųjų ir taip užkoduoti 1 ir 0.

Tačiau labiau nuspėjama kintamųjų reikšmė – padidėja mokslams skiriamo laiko bei laiko, skiriamo socializavimuisi.



2.6 Apibendrintieji tiesiniai modeliai. Lasso regresija

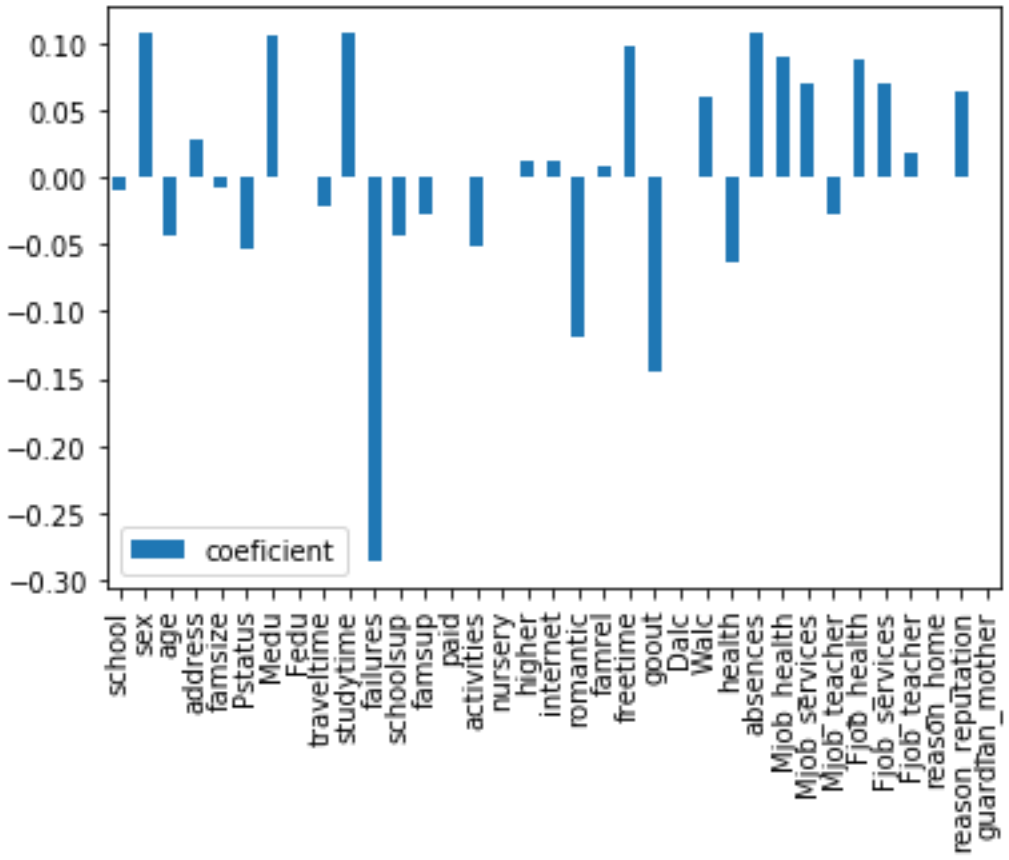
Vietoj paprastos tiesionės regresijos, apibrėžtos pagal formulę

$$\min_{\omega} ||X\omega - y||_2^2,$$

bandysime naudoti Lasso apibendrintąjį tiesinį modelį, apibrėžtą pagal formulę

$$\min_{\omega} \frac{1}{2n_{samples}} ||X\omega - y||_2^2 + \alpha ||\omega||_1.$$

$\alpha = 0.001$:



Bandysime iš modelio šalinti šiuos kintamuosius: *Fedu*, *paid*, *guardian_mother*, *nursery*, *Dalc*, *reason_home*.

Alkoholio vartojimas darbo dienomis šiuo atveju laikomas nereikšmingu.

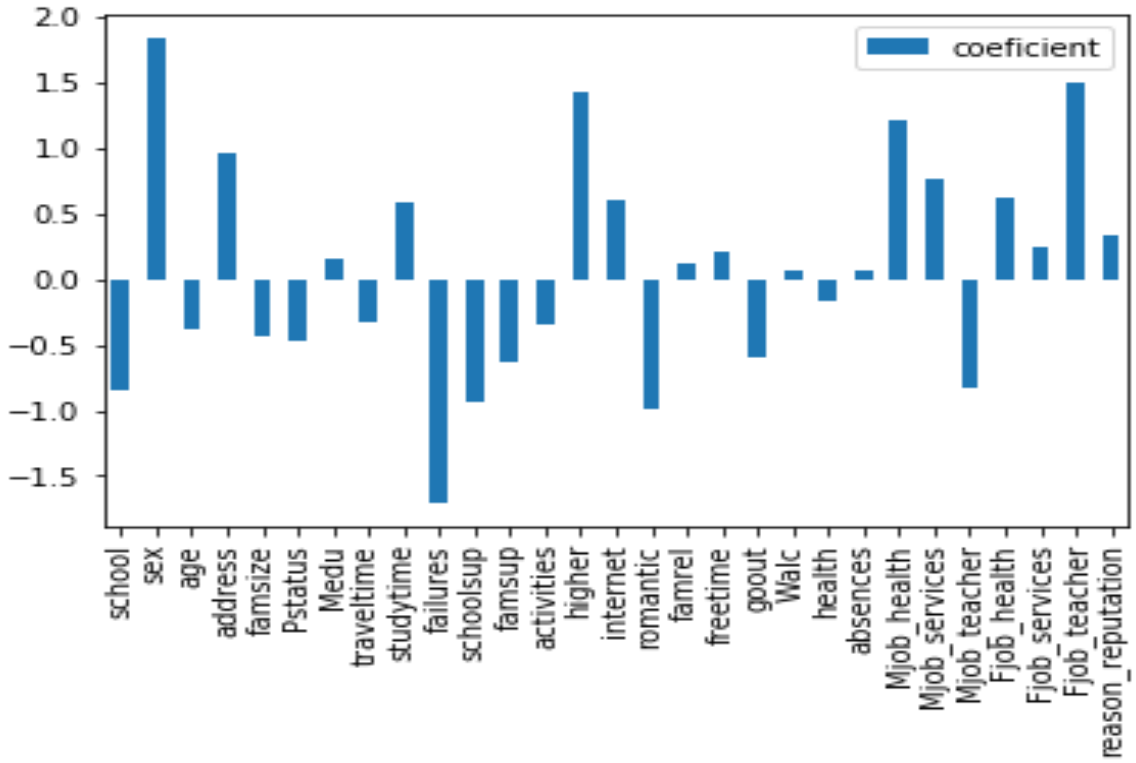
Su „Scikit-learn“:

```
Intercept:
 15.601282554385442
Coefficients:
 [-0.84066487  1.83028224 -0.38830728  0.95111107 -0.42745784 -0.47572851
  0.15616455 -0.33225252  0.58045226 -1.70432262 -0.92821835 -0.62467198
 -0.35092167  1.42217364  0.59912991 -0.98493584  0.11236723  0.19912489
 -0.59252359  0.06890831 -0.16418158  0.06455032  1.20939415  0.76472811
 -0.82874507  0.61765171  0.23685882  1.48985133  0.33803638]
```

RMSE = 3.91

Alkoholio vartojimas darbo dienomis liko nereikšmingas, savaitgaliais – arti nereikšmingumo.

Svarbiausiais kintamaisiais išliko lytis, atsiskaitymų nesėkmės mokslo metų eigoje, noras studijuoti aukštąjį mokslą bei tėvų darbų sritis. Šį modelį ir tvirtiname.



2.5 Tiesinės regresijos modelio išvados

RMSE be jokių duomenų transformacijų – 3.97.

Be kategorinių kintamųjų – 4.08.

Be x_other kintamųjų – 3.96.

Po Lasso regresijos nunulintų kintamųjų – 3.91

Darbo dienomis vartojamo alkoholio įtaka galutiniam metiniam pažymiui yra nereikšminga.

Savaitgaliais vartojamo alkoholio įtaka galutiniam metiniui arti nereikšmingumo.

Maksimali galutinio metinio pažymio reikšmė yra 20. Mūsų gautas RMSE ~4.

3. RANDOM FOREST

Random Forest modelis geresnių rezultatų neduoda.

Kadangi jis yra atsitiktinis dėl savo kilmės, tai jo pateikiama RMSE varijuoja, kartais net šokteli aukščiau tiesinės regresijos modelio RMSE.

Geriausius rezultatus davė šie parametrai:

n_estimators=1500,

max_depth=5,

max_leaf_nodes=10.

Geriausias gautas RMSE: 3.86


YOU THERE!
DO YOU HAVE A QUESTION?