Izveštaj

Bojana Đokić 2023/3101

Predviđanje uspeha učenika u školi

Cilj projekta je bio predviđanje uspehaučenika srednje škole. Podaci su preuzeti sa https://archive.ics.uci.edu/dataset/320/student+performance.

Ukupno ima 1044 primera i 30 prediktora. Od tih 1044 primera 649 su ocene iz portugalskog, a 395 ocene iz matematike. Neki od prediktora su binarni, neki multiklasni, neki nominalni, a neki kontinualni. Ocene se kreću od 0 do 20.

Izlazna promenljiva je broj poena u trećoj godini srednje škole. Radi detaljnije analize izlazna promenljiva je posmatran na 3 načina i to kao:

- Kontinualna vrednost
- Binarna vrednost položio (poeni >= 10) ili pao (poeni < 10)
- Ocene A (16-20), B(14-15), C(12,13), D(10,11), F(0,9)

Neki modeli su bolji u rešavanju klasifikacionih problema, a neki u rešavanju regresionih. Takođe nekada je možda značajnije izvršiti pojednostavljenu predikciju da li je neko pao ili položio, nego koju je ocenu dobio. Ocene se svode na F(pao) i A, B, C, D koje su samo detaljnija analiza skora učenika koji su položili. U zavisnosti od toga u koje svrhe se ispitivanje izvršava različito tumačenje uspeha može biti od koristi.

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
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).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

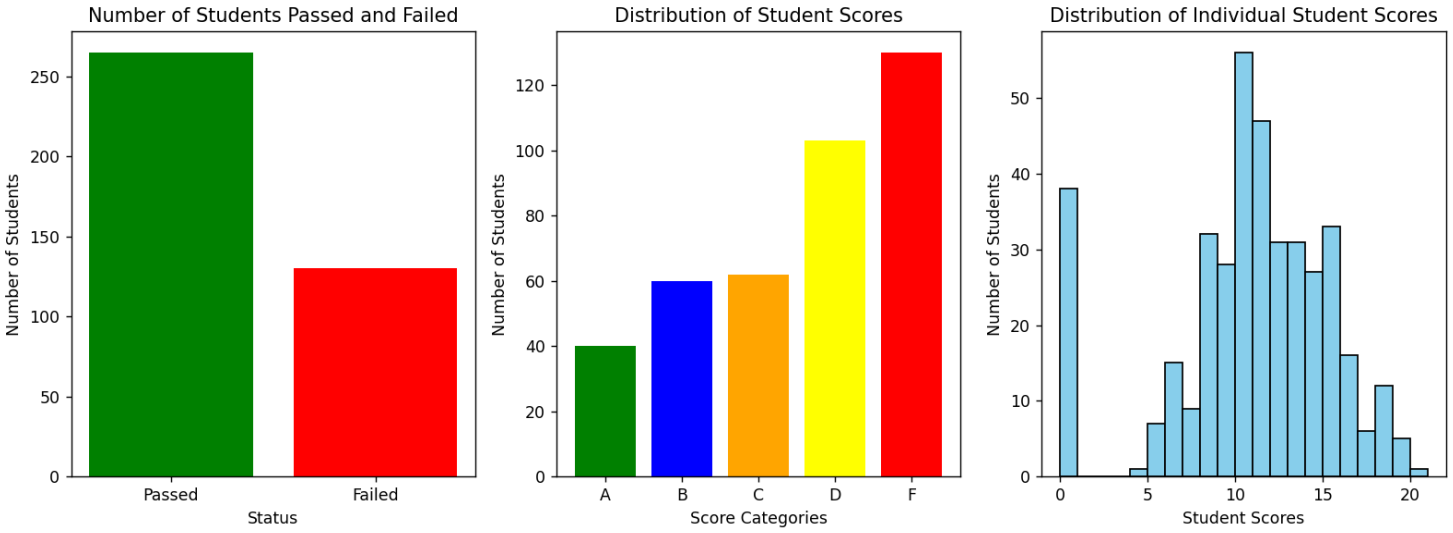
a 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
b teacher, health care related, civil services (e.g. administrative or police), at home or other.

Prikaz prediktora i ciljne promenljive

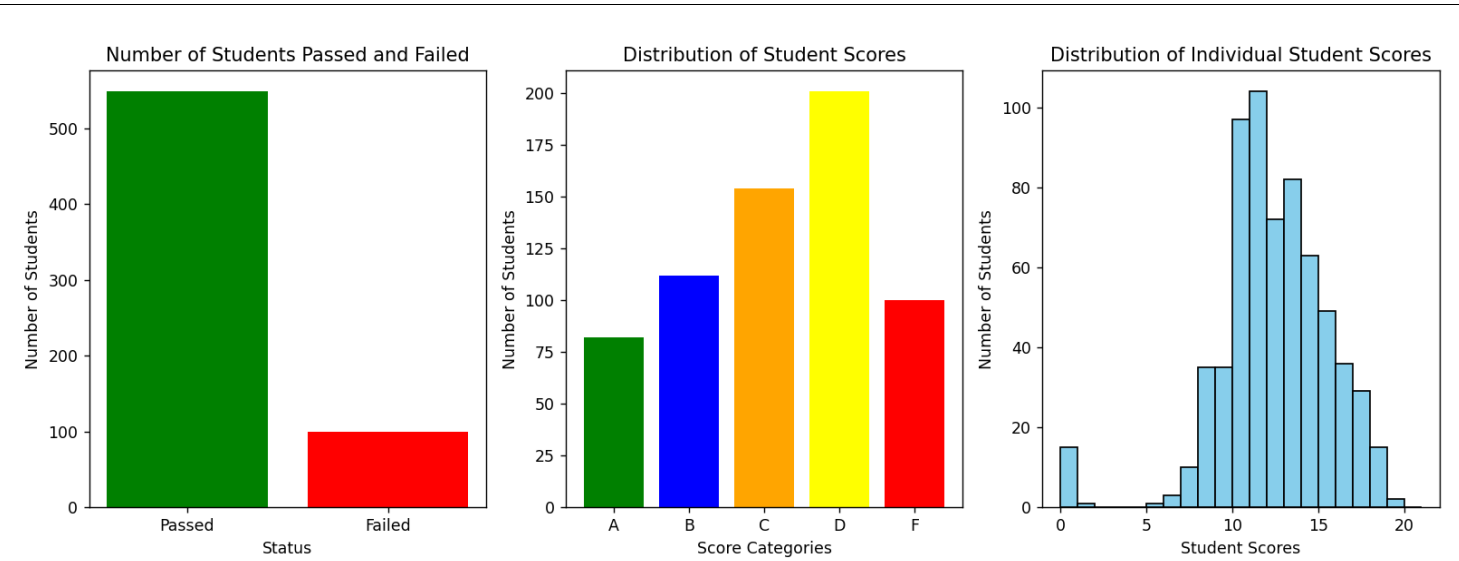
school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	4	0	11	11
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	2	9	11	11
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	0	yes	no	no	no	yes	yes	yes	no	4	3	2	2	3	3	6	12	13	12
GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	no	yes	yes	yes	yes	yes	3	2	2	1	1	5	0	14	14	14
GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	no	no	yes	yes	no	no	4	3	2	1	2	5	0	11	13	13
GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	no	yes	no	yes	yes	yes	yes	no	5	4	2	1	2	5	6	12	12	13
GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	13	12	13
GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no	yes	yes	no	no	4	1	4	1	1	1	2	10	13	13
GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	no	no	yes	yes	yes	no	4	2	2	1	1	1	0	15	16	17
GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no	yes	no	yes	yes	yes	yes	no	5	5	1	1	1	5	0	12	12	13
GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1	2	0	no	yes	no	no	yes	yes	yes	no	3	3	3	1	2	2	2	14	14	14
GP	F	15	U	GT3	T	2	1	services	other	reputation	father	3	3	0	no	yes	no	yes	yes	yes	yes	no	5	2	2	1	1	4	0	10	12	13
GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1	0	no	yes	no	yes	yes	yes	yes	no	4	3	3	1	3	5	0	12	13	12
GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2	2	0	no	yes	no	no	yes	yes	yes	no	5	4	3	1	2	3	0	12	12	13
GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	0	no	yes	no	no	yes	yes	yes	yes	4	5	2	1	1	3	0	14	14	15
GP	F	16	U	GT3	T	4	4	health	other	home	mother	1	1	0	no	yes	no	no	yes	yes	yes	no	4	4	4	1	2	2	6	17	17	17

Primer nekoliko redova ulaznih podataka

Analiza podataka



Prikaz uspeha učenika iz matematike

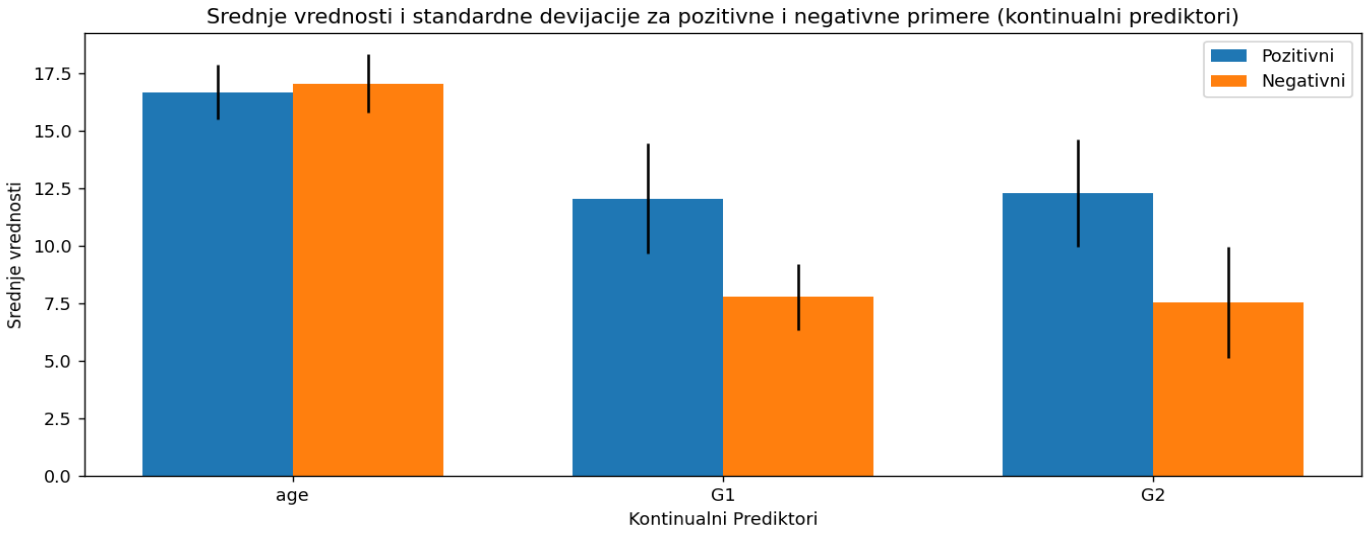


Prikaz uspeha učenika iz portugalskog

Prolaznost za portugalski je 69.64%, a za matematiku 52.91%.

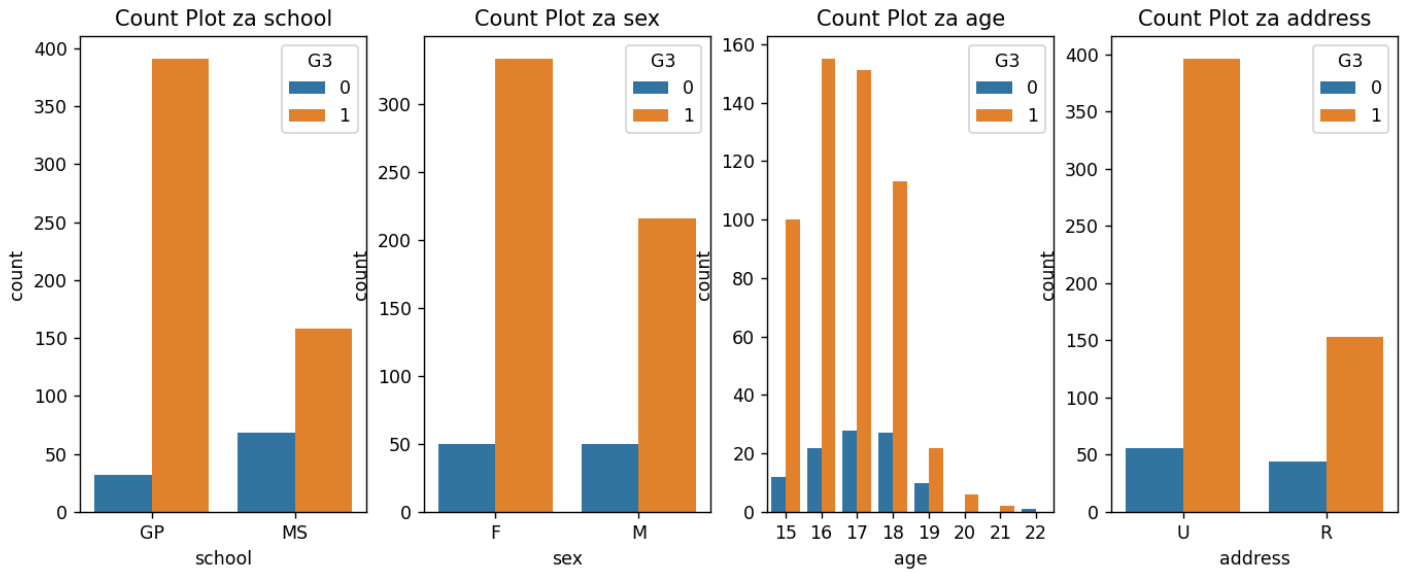
Analiza prediktora

G1 predstavlja proj poena u prvoj godini, a G2 broj poena u drugoj godini srednje škole.

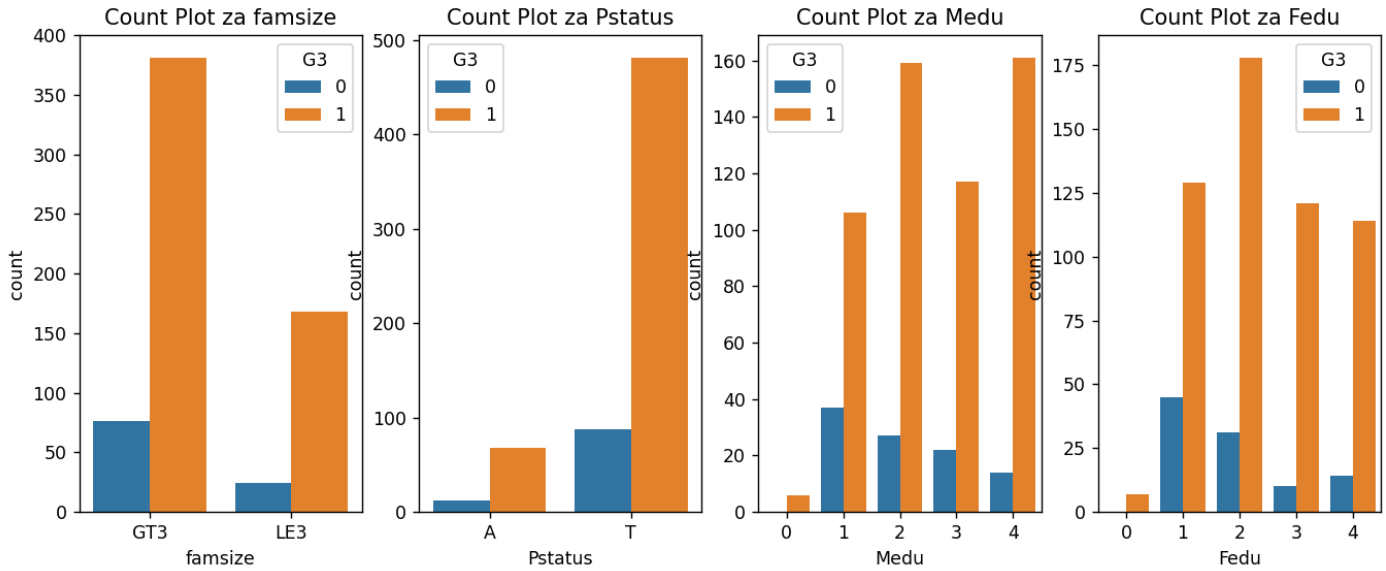


Godine malo variraju. To je jer su one u opsegu od 15 – 22 godina i ne bi trebalo preterno da utiču na izlaznu promenljivu.
G1 i G2 imaju veoma slične srednje vrednosti I standardne devijacije. Videćemo kasnije kako su ova dva skora korelisana međusobno, a I sa izlazom.

Na count plotovima narandžastom bojom su predstavljeni učenici koji su položili, a plavom učenici koji su pali.

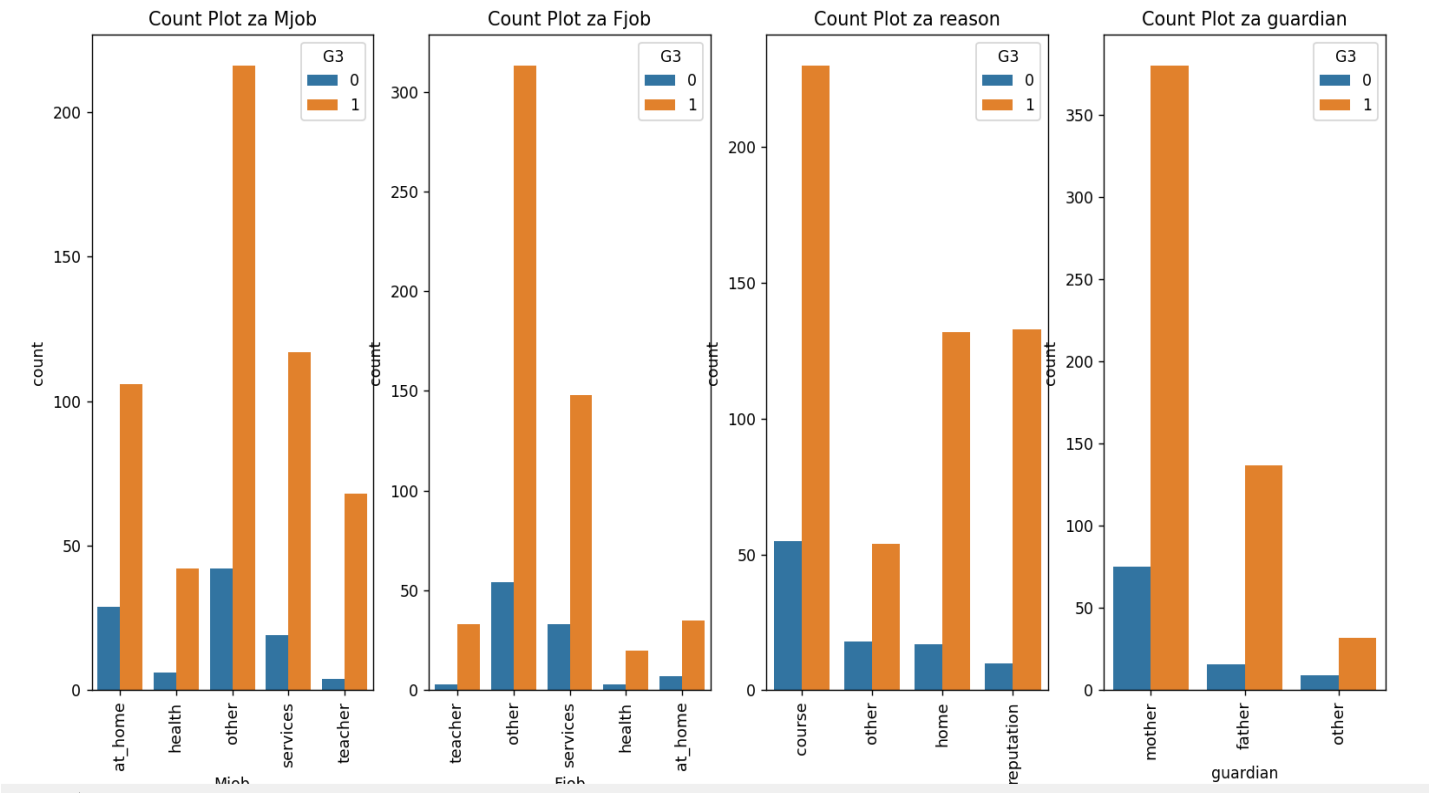


Svi učenici imaju od 15-22 godina.
Adresa je U – urbana ili R – ruralna. Ovaj podatak može uticati na izlaz jer deci iz ruralne sredine možda treba više vremena do škole ili ih možda roditelji ne podstiču na školovanje što može rezultovati manjim brojem poena.
GP(Gabriel Pereira) i MS(Mousinho da Silveira) su skraćenice za dve škole iz kojih su podaci prikupljani. U suštini ova informacija nam ništa ne znači jer su rezultati posmatrani u globalu – da li su učenici položili i koju su ocenu dobili.



Famsize je prediktor koji je LE3 – ako porodica ima 3 ili manje članova i GT3 u suprotnom.
Pstatus nam govori da li su roditelji razvedeni ili iz nekog razloga žive odvojeno (A – apart) ili su zajedno (T – together). Ovo može biti značajan parametar koji može biti tumačen na više načina. Deca čiji roditelji ne žive zajedno možda dolaze iz disfunkcionalne porodice što može uzrokovati gore ocene, ali i ne mora da znači.

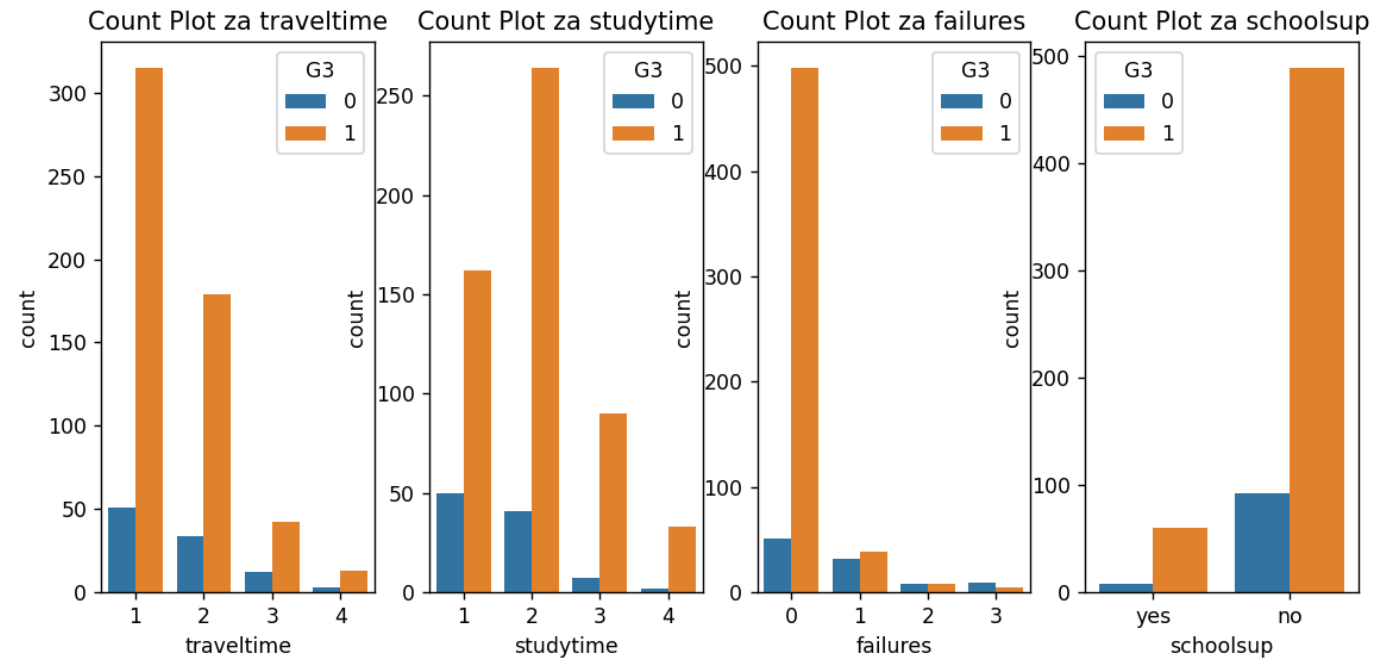
Medu i Fedu predstavljaju nivo edukacije roditelja : 0 – None, 1 – osnovna škola (prva 4 razreda), 2 – osnovna škola (5 – 9. razred), 3 – srednja škola, 4 – više obrazovanje.



Count plotovi za Mjob i Fjob predstavljaju broj roditelja koji spadaju u neku od prikazanih kategorija poslova.

Reason prediktor nam govori koji je razlog za odabir baš te škole : da li je blizu kuće(home) ili je u pitanju reputacija škole(reputation) ili usmerenje škole(course) ili je u pitanju nešto drugo(other).

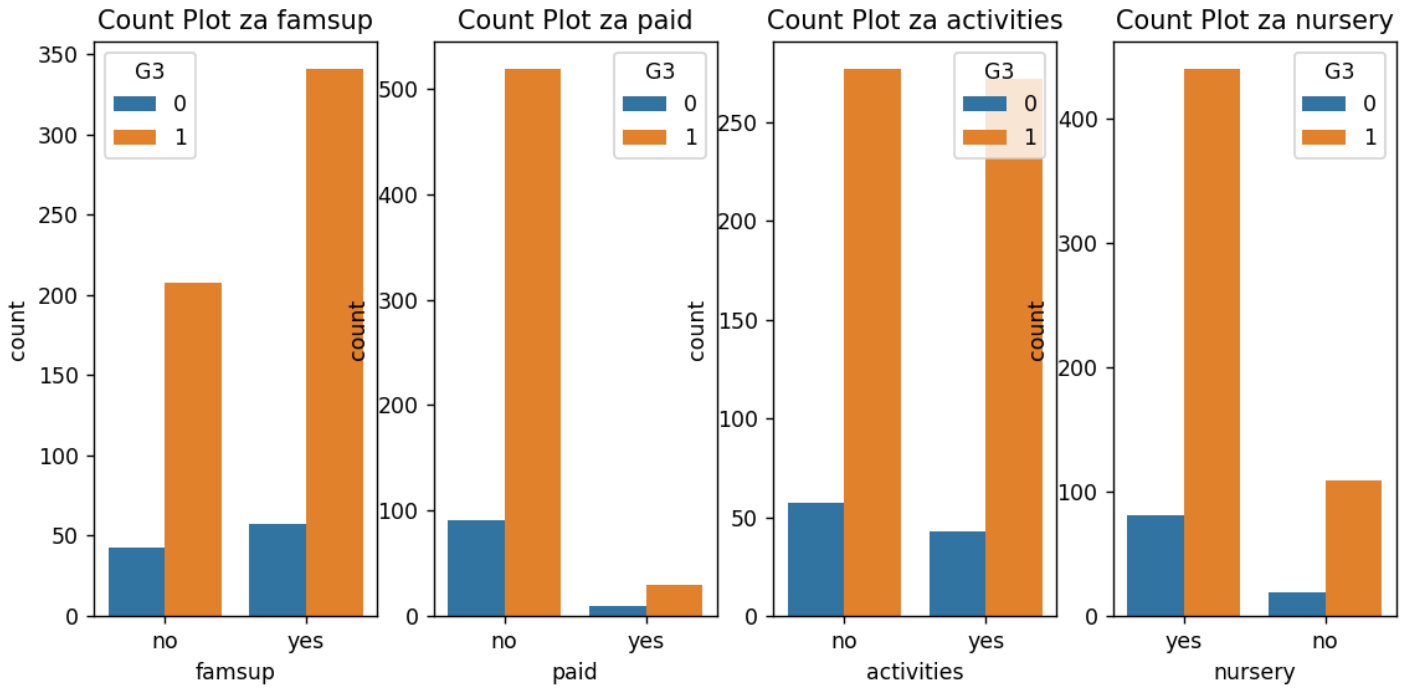
Guardian nam govori ko je staratelj deteta.



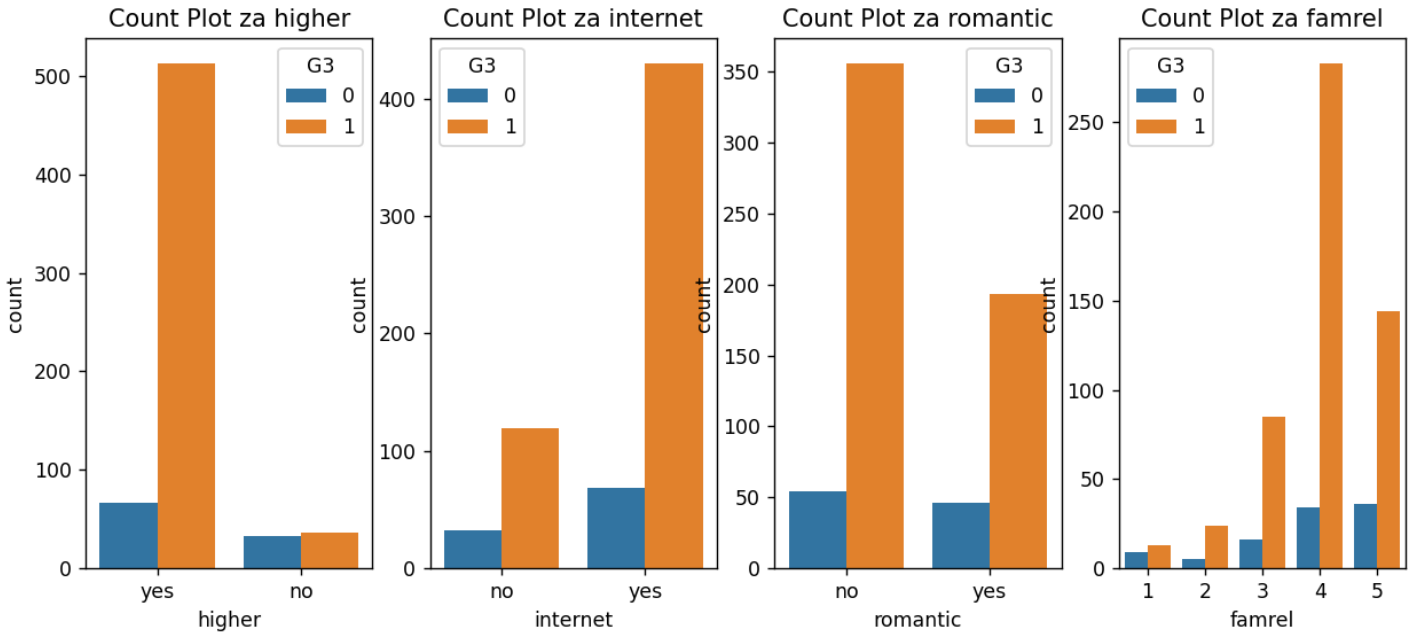
Traveltime predstavlja vreme putovanja do škole: 1 - manje od 15 inuta, 2- od 15 do 30 minuta, 3 – od 30 minuta do sat vremena, 4 – više od sat vremena.

Studytime predstavlja broj sati provedenih učeći na nedeljnom nivou: 1 – manje od 2h, 2 – 2 do 5h, 3 – 5 do 10h, 4 – više od 10h.

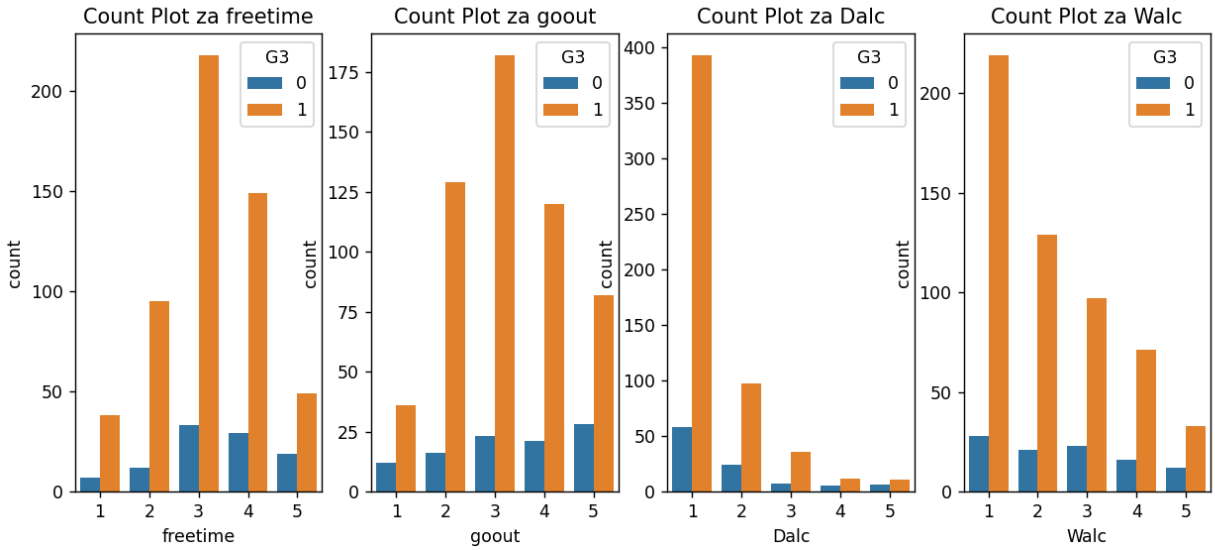
Schoolsup je da li imaju dodatne časove.
Failures je broj godina koji su učenici pali. Ako je 0 <= n < 3, onda je n prikazan, a ako je n > 3, onda je prikazan 4.



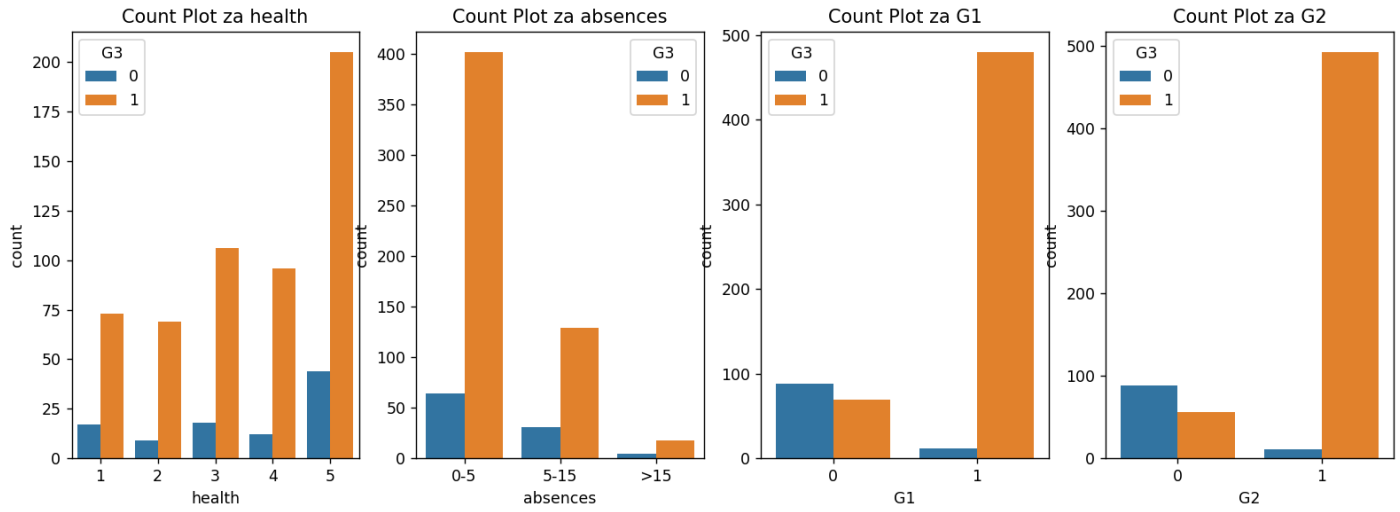
Paid je predictor koji govori da li učenik uzima dopunske časove, a activities da li se bavi nekim dodatnim aktivnostima.
Nursery je da li su pohađali medicinsku školu.



Higher - da li učenik želi da nastavi sa školovanjem.
Internet – da li učenik ima pristup internetu kod kuće.
Romantic – da li je učenik u romantičnoj vezi.
Famrel – kakvi su odnosi u porodici od 1 – veoma loši do 5 – veoma dobri.

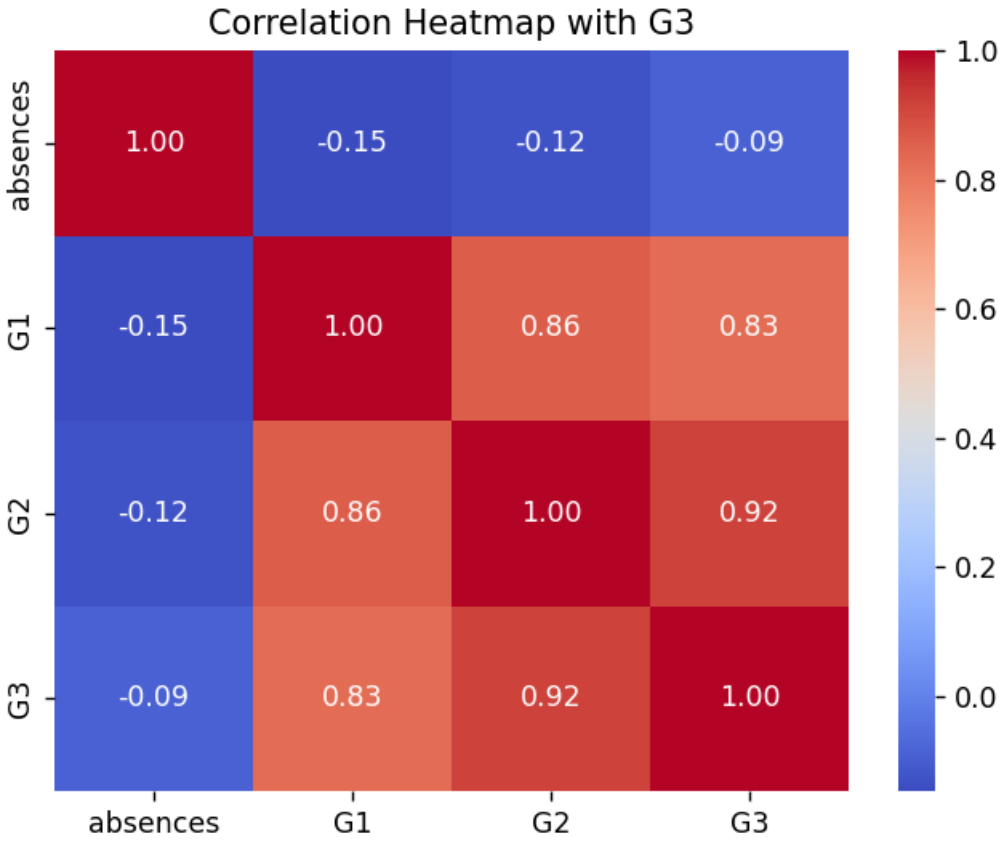


Freetime – Koliko slobodnog vremena učenik ima posle škole od 1 – veoma malo do 5 – mnogo.
Goout – Koliko često učenik izlazi sa prijateljima od 1 – retko do 5 – često.
Dalc - dnevna konzumacija alkohola od 1 - veoma malo ili ništa do 5 – dosta.
Walc – konzumacija alkohola na nedeljnom nivou od 1 - veoma malo ili ništa do 5 – dosta.



Health predstavlja zdravstveno stanje učenika koje se kreće od 1 – veoma lose do 5 – veoma dobro.
Absences je broj izostajanja koji inače prestavlja tačan broj izostanaka, ali je u count plotu grupisan radi lepšeg prikaza.
G1 je prolaznost učenika u prvoj godini, a G2 u drugoj godini srednje škole. Učenik je položio ako na završnom testu sakupi >= 10 poena.

Prediktori gde je velika razlika između onih koji su položili i onih koji nisu bi trebalo da budu od većeg značaja.

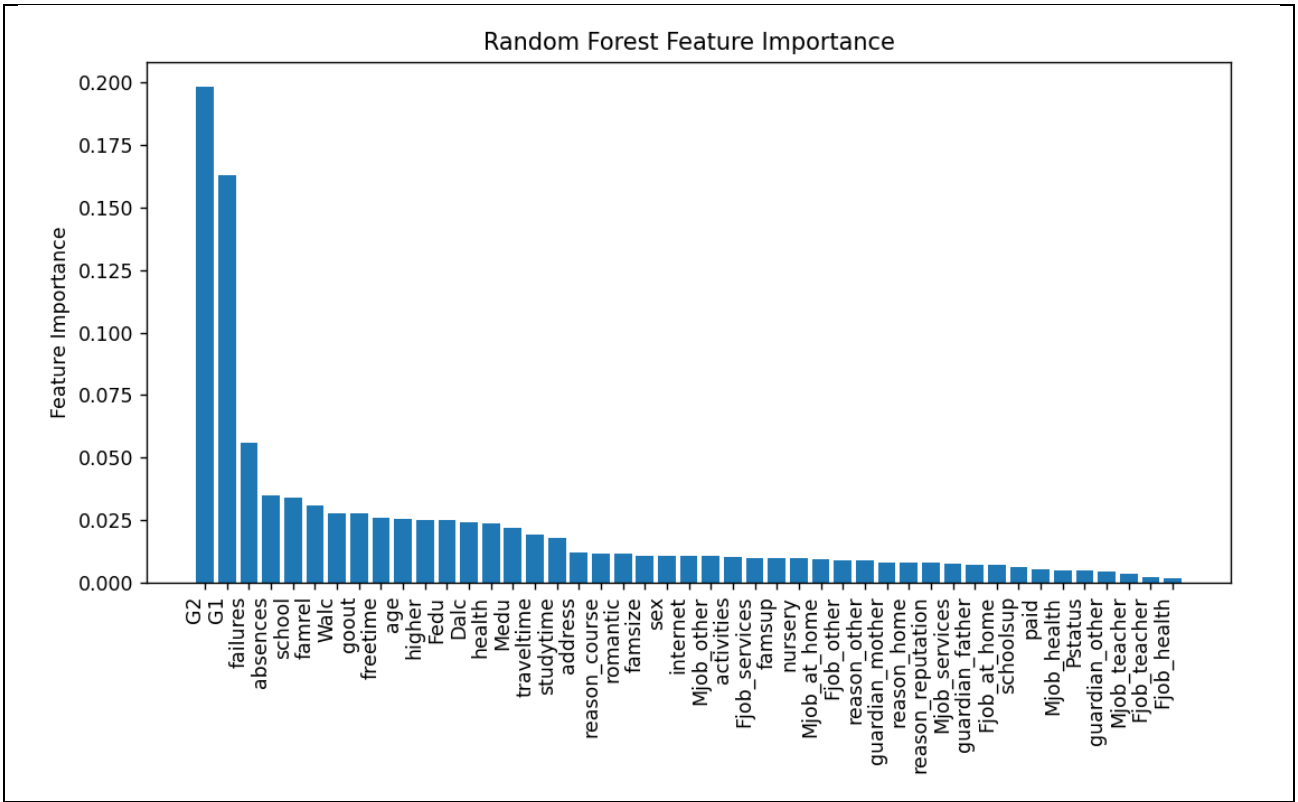


Iz linearne korelacije kontinualnih prediktora vidimo da su ocene iz drugog razreda(G2) najkorelisanije sa izlazom, a zatim ocene iz prvog razreda(G1) i tek onda broj izostanaka. Kasnije ćemo videti da su ocene iz prethodnih razreda najviše rangirani po značaju što se tiče Random Forest modela.

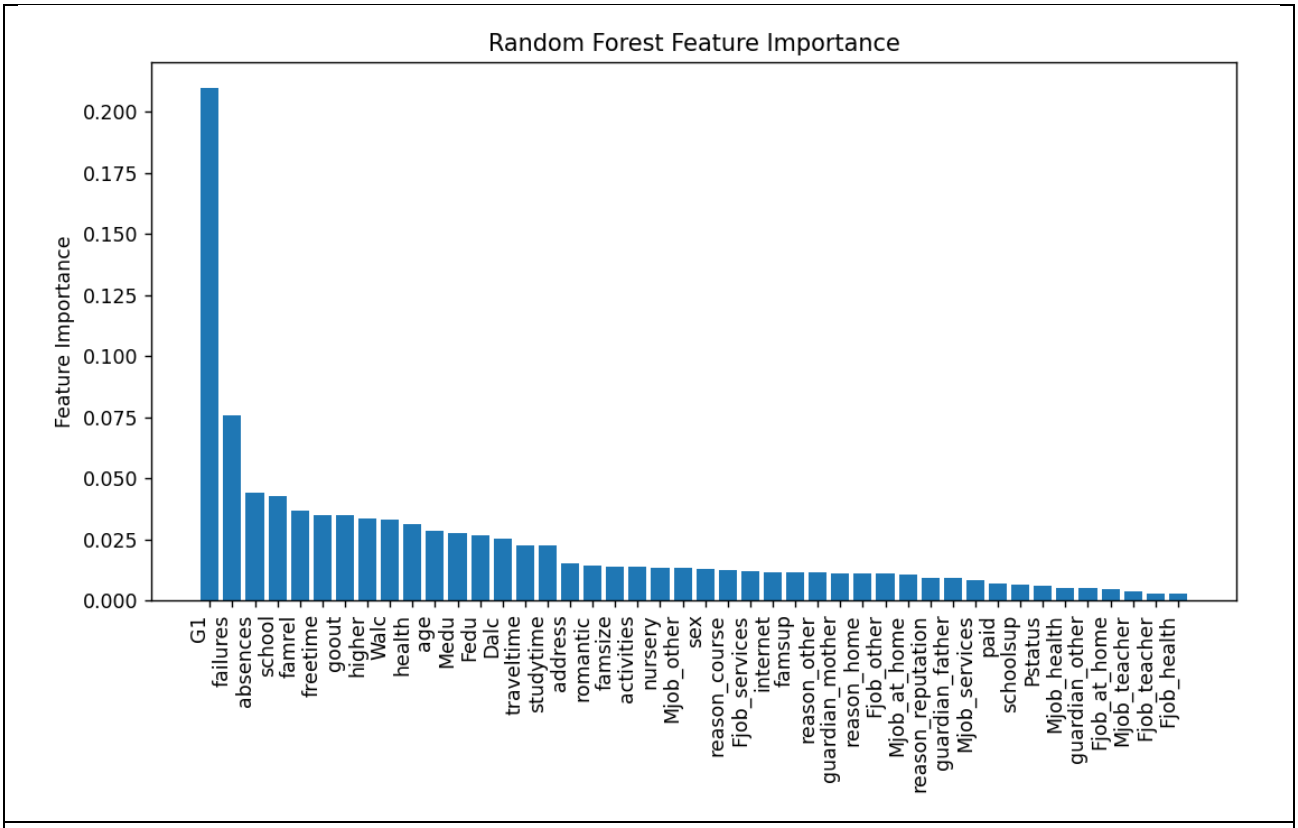
Pošto su ocene iz prošlih razreda dosta korelisane sa izlazom razmatrane su 3 konviguracije:

- A - uključuje sve prediktore
- B - uključuje sve prediktore osim G2 (skor iz druge godine)
- C - slično kao B samo još bez G1 (skor iz prve godine)

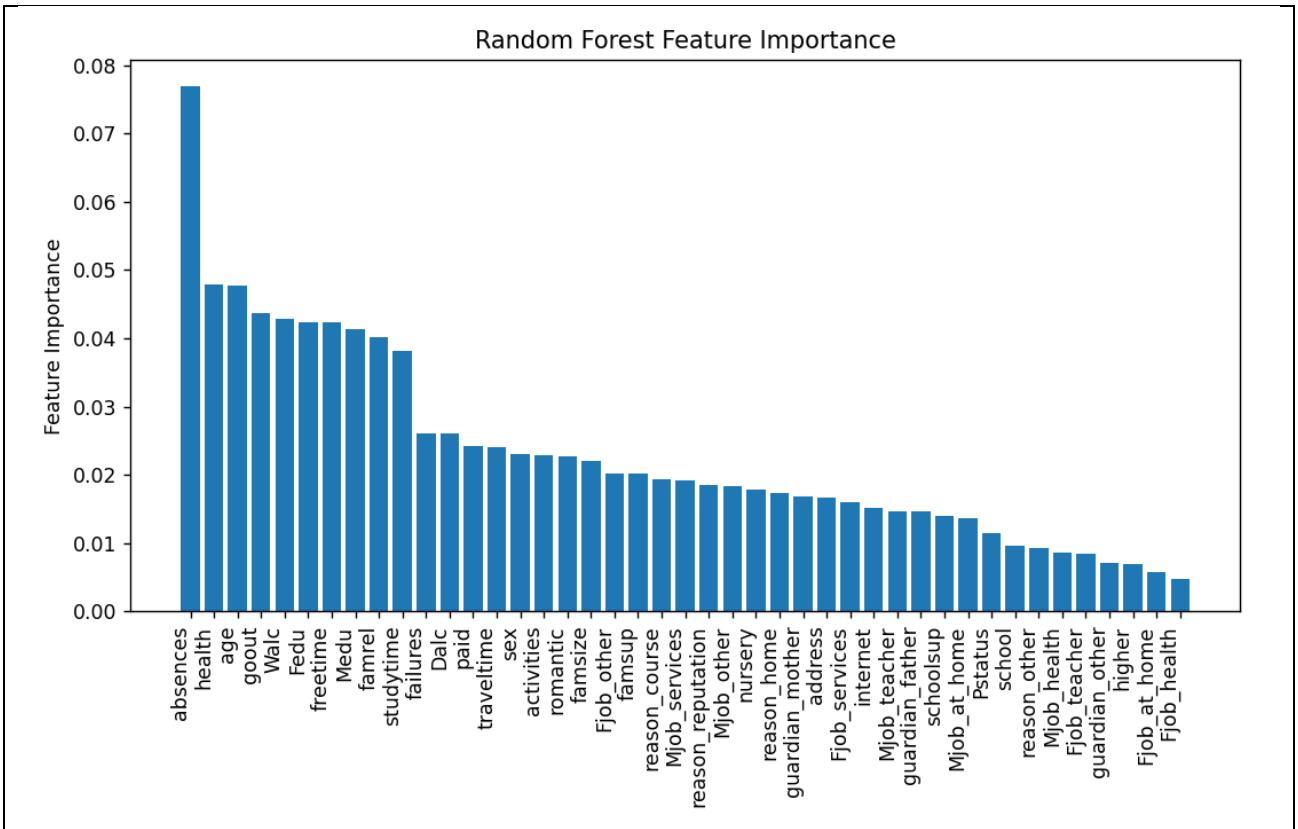
U narednim graficima predtsvljano je rangiranje prediktora po značaju Random Forest modela.
Iz grafika se može zaključiti da je dosta predikotra manje bitno kada su dostupne ocene iz prethodnih godina.
Do povećanja značaja nekih drugih prediktora kao što su broj prethodnih neuspeha, vreme provedeno u izlascima, slobodno vreme, vreme učenja itd. dolazi tek pri C konfiguraciji.



Značaj prediktora konfiguracije A



Značaj prediktora konfiguracije B



Značaj prediktora konfiguracije C

Rezultati

Pre treninga nad skupom podataka nominalni prediktori su pretvarani u binarne pomoću one hot encodinga pa na primer Mjob postaje: 'Mjob_at_home', 'Mjob_health', 'Mjob_services', 'Mjob_teacher', 'Mjob_other'. Umesto 30 prediktora na kraju imamo ukupno 45.

Kao najosnovniji model za predikcije uzeto je:

- Konfiguracija A : Vrednost G2 skora učenika
- Konfiguracija B : Vrednost G1 skora učenika
- Konfiguracija C : Za klasifikacione probleme uzeta je najčešća klasa, u suprotnom srednja vrednost G2 skora

Performanse predviđanja su merene tako što je primenjeno 20 iteracija desetostruke unakrsne validacije za svaku konfiguraciju.
Metika za klasifikacione probleme bila je Percentage of Correct Classifications (PCC), a za regresiju Root Mean Squared (RMSE).

$$\Phi(i) = \begin{cases} 1 & , \quad \text{if } y_i = \widehat{y}_i \\ 0 & , \quad \text{else} \end{cases}$$
$$PCC = \sum_{i=1}^{N} \Phi(i)/N \times 100\,(\%)$$
$$RMSE = \sqrt{\sum_{i=1}^{N} (y_i - \widehat{y}_i)^2/N}$$

Modeli koji su korišćeni su: Osnovni prediktor(NV), Support Vector Machines(SVM), Decision Tree(DT) i Random Forest(RF).
Broj estimatora za Random Forest je 500. Uzeta je ta vrednost zbog literature kako bi upoređivanje rezultata bilo verodostojnije.
Za hiperparametre SVMa grid searchem je određeno najbolje gama i korišćen je gausov kernel.
Ostali parametri nisu specificirani u literaturi, pa su uzeti defaultni

```python
def find_best_param(X_train, Y_train, model):

    # Define the parameter grid for SVM with RBF kernel
    param_grid = {'gamma': [2 ** (-9), 2 ** (-7), 2 ** (-5), 2 ** (-3), 2 ** (-1)]}

    if model == 'regression':
        svm_model = SVR(kernel='rbf')
    else:
        svm_model = SVC(kernel='rbf')

    # Instantiate the GridSearchCV object
    grid_search = GridSearchCV(svm_model, param_grid, cv=10, n_jobs=-1)

    # Fit the grid search to the data (use only training data for internal grid search)
    grid_search.fit(X_train, Y_train)

    # Get the best hyperparameters
    best_params = grid_search.best_params_

    return best_params
```

Grid Search za parametre SVMa

Procenat ispravne klasifikacije za binarne klase

- A - uključuje sve prediktore
- B - uključuje sve prediktore osim G2 (skor iz druge godine)
- C - slično kao B samo još bez G1 (skor iz prve godine)

Matematika

	NV	SVM	DT	RF
A	**91.89**	90.2	84.95	91.31
B	83.79	**85.31**	75.44	81.31
C	67.08	70.09	62.13	**70.44**

Portugalski

	NV	SVM	DT	RF
A	89.67	**92.72**	88.81	91.92
B	87.51	**89.42**	84.82	87.09
C	84.59	84.59	80.31	**84.74**

Input Setup	Mathematics					Portuguese				
	NV	NN	SVM	DT	RF	NV	NN	SVM	DT	RF
A	**91.9**$^{\dagger}{}_{\pm 0.0}$	88.3$_{\pm 0.7}$	86.3$_{\pm 0.6}$	90.7$_{\pm 0.3}$	91.2$_{\pm 0.2}$	89.7$_{\pm 0.0}$	90.7$_{\pm 0.5}$	91.4$_{\pm 0.2}$	**93.0**$^{\dagger}{}_{\pm 0.3}$	92.6$_{\pm 0.1}$
B	**83.8**$^{\dagger}{}_{\pm 0.0}$	81.3$_{\pm 0.5}$	80.5$_{\pm 0.5}$	83.1$_{\pm 0.5}$	83.0$_{\pm 0.4}$	87.5$_{\pm 0.0}$	87.6$_{\pm 0.4}$	88.0$_{\pm 0.3}$	88.4$_{\pm 0.3}$	**90.1**$^{\dagger}{}_{\pm 0.2}$
C	67.1$_{\pm 0.0}$	66.3$_{\pm 1.0}$	**70.6**$^{*}{}_{\pm 0.4}$	65.3$_{\pm 0.8}$	70.5$_{\pm 0.5}$	84.6$_{\pm 0.0}$	83.4$_{\pm 0.5}$	84.8$_{\pm 0.3}$	84.4$_{\pm 0.4}$	**85.0**$^{*}{}_{\pm 0.2}$

Performanse modela iz literature za binarni model

Procenat ispravne klasifikacije za multiklase (ocene A-F)

Matematika

	NV	SVM	DT	RF
A	**78.48**	70.98	66.17	70.82
B	**60.5**	58.42	51.94	49.52
C	32.91	**34.81**	28.92	34.24

Portugalski

	NV	SVM	DT	RF
A	**72.88**	72	64	72
B	**58.7**	57.53	48	52.82
C	30.97	36.19	29.52	**36.48**

Input Setup	Mathematics NV	NN	SVM	DT	RF	Portuguese NV	NN	SVM	DT	RF
A	**78.5**$^{\dagger}{\pm 0.0}$	$60.3_{\pm 1.6}$	$59.6_{\pm 0.9}$	$76.7_{\pm 0.4}$	$72.4_{\pm 0.4}$	$72.9_{\pm 0.0}$	$65.1_{\pm 0.9}$	$64.5_{\pm 0.6}$	**76.1**$_{\pm 0.1}$	$73.5_{\pm 0.2}$
B	**60.5**$^{\dagger}{\pm 0.0}$	$49.8_{\pm 1.2}$	$47.9_{\pm 0.7}$	$57.5_{\pm 0.8}$	$52.7_{\pm 0.6}$	$58.7_{\pm 0.0}$	$52.0_{\pm 0.6}$	$51.7_{\pm 0.6}$	**62.9**$_{\pm 0.2}$	$55.3_{\pm 0.4}$
C	$32.9_{\pm 0.0}$	$30.4_{\pm 1.0}$	$31.0_{\pm 0.7}$	$31.5_{\pm 0.6}$	**33.5**$_{\pm 0.6}$	$31.0_{\pm 0.0}$	$33.7_{\pm 0.6}$	$34.9_{\pm 0.5}$	$32.8_{\pm 0.6}$	**36.7**$^{\dagger}{\pm 0.6}$

Performanse modela iz literature za multiklasni model

Root Mean Square Error za regresioni problem

Matematika

	NV	SVM	DT	RF
A	2.01	2.4	2.12	**1.55**
B	2.8	3.24	3.13	**2.28**
C	4.58	5.51	5.4	**3.85**

Portugalski

	NV	SVM	DT	RF
A	1.32	1.89	1.81	**1.266**
B	1.89	2.15	2.47	**1.76**
C	3.23	3	3.82	**2.69**

I. S.	Mathematics NV	NN	SVM	DT	RF	Portuguese NV	NN	SVM	DT	RF
A	$2.01_{\pm 0.00}$	$2.05_{\pm 0.02}$	$2.09_{\pm 0.02}$	$1.94_{\pm 0.04}$	**1.75**$^{\dagger}{\pm 0.01}$	**1.32**$_{\pm 0.00}$	$1.36_{\pm 0.04}$	$1.35_{\pm 0.01}$	$1.46_{\pm 0.03}$	**1.32**$_{\pm 0.00}$
B	$2.80_{\pm 0.00}$	$2.82_{\pm 0.02}$	$2.90_{\pm 0.02}$	$2.67_{\pm 0.04}$	**2.46**$^{\dagger}{\pm 0.01}$	$1.89_{\pm 0.00}$	$1.88_{\pm 0.02}$	$1.87_{\pm 0.01}$	**1.78**$^{*}{\pm 0.03}$	$1.79_{\pm 0.01}$
C	$4.59_{\pm 0.00}$	$4.41_{\pm 0.03}$	$4.37_{\pm 0.03}$	$4.46_{\pm 0.04}$	**3.90**$^{\dagger}{\pm 0.01}$	$3.23_{\pm 0.00}$	$2.79_{\pm 0.02}$	$2.76_{\pm 0.02}$	$2.93_{\pm 0.02}$	**2.67**$^{\dagger}{\pm 0.01}$

Performanse modela iz literature za kontinualne vrednosti

Pošto su rezultati za Decision Tree bili nešto gori u odnosu na one iz literature, testiranje je ponovljeno samo što se prethodno obavilo traženje najboljih parametara Grid Searchom.

```
# Define the parameter grid to search
param_grid = {
    'criterion': ['squared_error', 'absolute_error', 'poisson', 'friedman_mse'],
    # 'criterion': ['gini', 'entropy'],
    'max_depth': [None, 5, 10, 15],
    'min_samples_split': [5, 10, 20, 30, 50, 70, 100],
    'min_samples_leaf': [1, 2, 4, 5, 10, 20]
}

# Create the Decision Tree classifier
scoring = 'neg_root_mean_squared_error'
classifier = DecisionTreeRegressor()

# Create the GridSearchCV object
grid_search = GridSearchCV(classifier, param_grid, cv=5, scoring=scoring)

# Perform the grid search on the training data
grid_search.fit(X_train, Y_train)

# Print the best parameters found
best_params = grid_search.best_params_

return best_params
```

Parametri testirani grid searchem za Decision tree

Ponovljena merenja za Decision Tree

Zelenom bojom predstavljene su nove vrednosti, a crvenom stare.

Procenat ispravne klasifikacije za binarne klase

Matematika

	DT	DT
A	90.47	84.95
B	84.82	75.44
C	68.42	62.13

Portugalski

	DT	DT
A	91.04	88.81
B	91.66	84.82
C	84.33	80.31

Procenat ispravne klasifikacije za multiklase (ocene A-F)

Matematika

	DT	DT
A	74.98	66.17
B	56.96	51.94
C	30.93	28.92

Por

	DT	DT
A	72.13	64
B	61	48
C	32.86	29.52

Root Mean Square Error za regresioni problem

Matematika

	DT	DT
A	2.05	2.12
B	2.64	3.13
C	4.14	5.4

Portugalski

	DT	DT
A	1.39	1.81
B	1.7	2.47
C	2.84	3.82

Dobijamo bolje rezultate za Decision Tree kada koristimo pogodne parametre.

Osim pomenutih modela primenjeni su još i model linearne i logističke regresije. Pored linearne regresije pokušano je i da se primeni polinomijalna regresija, tj. da se nađe stepen polinoma koji daje bolju tačnost od linearne regresije, međutim većim usložnjavanjem modela dobijali su se sve gori rezultati. Verovatno ima previše parametara u odnosu na broj primera, pa model postane preobučen.

```
1 usage
def find_best_degree(X, y):
    degrees = [1, 2, 3, 4, 5]

    # Create a parameter grid with different degrees
    param_grid = {'polynomialfeatures__degree': degrees}

    # Create a pipeline explicitly and include PolynomialFeatures and LinearRegression
    poly_reg = make_pipeline( *steps: PolynomialFeatures(), LinearRegression())

    # Create a GridSearchCV instance
    grid_search = GridSearchCV(poly_reg, param_grid, cv=5, scoring='neg_mean_squared_error')

    # Fit the grid search to the data
    grid_search.fit(X, y)
    best_degree = grid_search.best_params_['polynomialfeatures__degree']
    print(best_degree)
    return best_degree
```

Grid Search parametara za polinomijalnu regresiju

Pošto su prediktori većinski binarni ili multiklasni za klasifikacione probleme primenjen je još i Naive Bayes(NB).
Prediktori su razvrstani u multiklasne, binarne i kontinualne.
Za kontinualne korišćen je Gaussian Naive Bayes (GNB), za multiklasne Multinomial Naive Bayes(MNB), a za binarne Binomial Naive Bayes(BNB).
Pretpostavka je da će GNB da radi bolje u slučaju kada su poznate vrednosti skorova iz prethodnih godina, a da će MNB i GNB da budu bolji u suprotnom. Predikcije su računate odvojeno za svaki model, kao i zajednike pomoću majority votinga. Za različite konfiguracije uzimana je ona koja je ostvarila najbolji skor, pa je tako uglavnom za konfiguraciju uzimana vrednost GNBa, dok je za ostale konfiguracije to uglavnom bila kombinovana.

U nastavku su prikazane obuhvaćene vrednosti kako bi se svi modeli uporedili.

Procenat ispravne klasifikacije za binarne klase

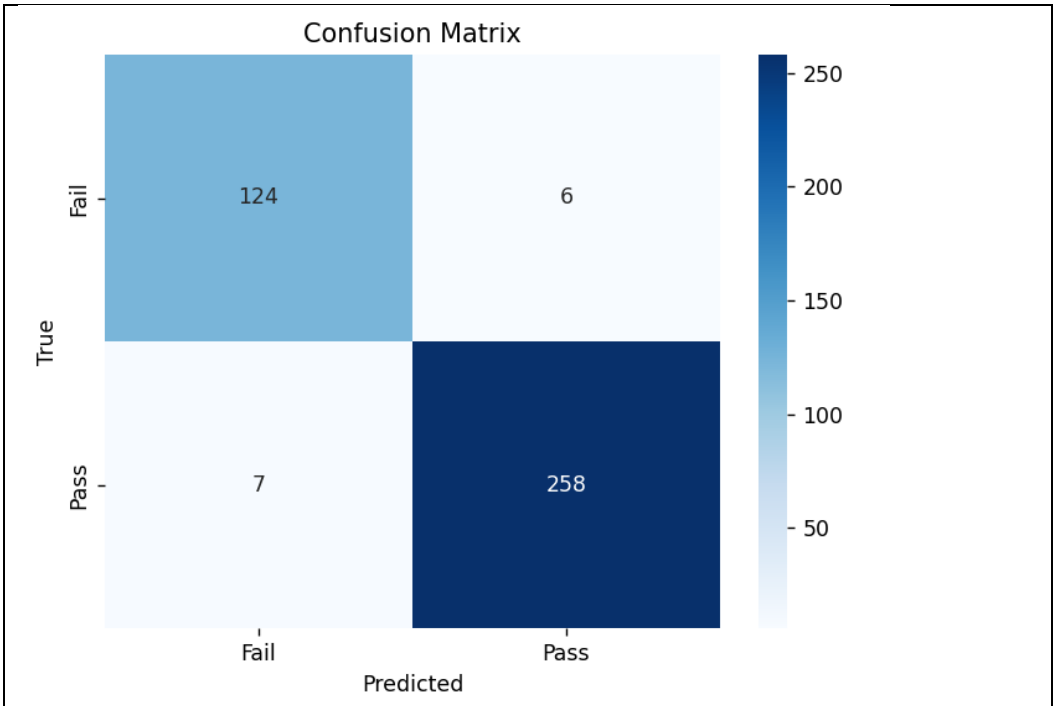
Matematika

	NV	SVM	DT pon	RF	LR	NB
A	91.89	90.2	90.47	91.31	92.9	74.68
B	83.79	85.31	84.82	81.31	83.87	74.68
C	67.08	70.09	68.42	70.44	69.2	74.68

Portugalski

	NV	SVM	DT pon	RF	LR	NB
A	89.67	92.72	91.04	91.92	92.74	90
B	87.51	89.42	91.66	87.09	89.36	90
C	84.59	84.59	84.33	84.74	84.82	90

Što se tiče binarne klasifikacije deluje da Naive Bayes najbolje radi kada nam nisu poznate pređašnje ocene. Za A konfiguraciju je najbolja logistička regresija, a za B konfiguraciju Decision Tree ili SVM. Procenat pogađanja je sličan, a treba i uzeti u obzir da je skup ocena iz matematike dosta manji u odnosu na skup ocena iz portugalskog.



Matrica konfuzije za logističku regresiju konfiguracije A ocean iz matematike

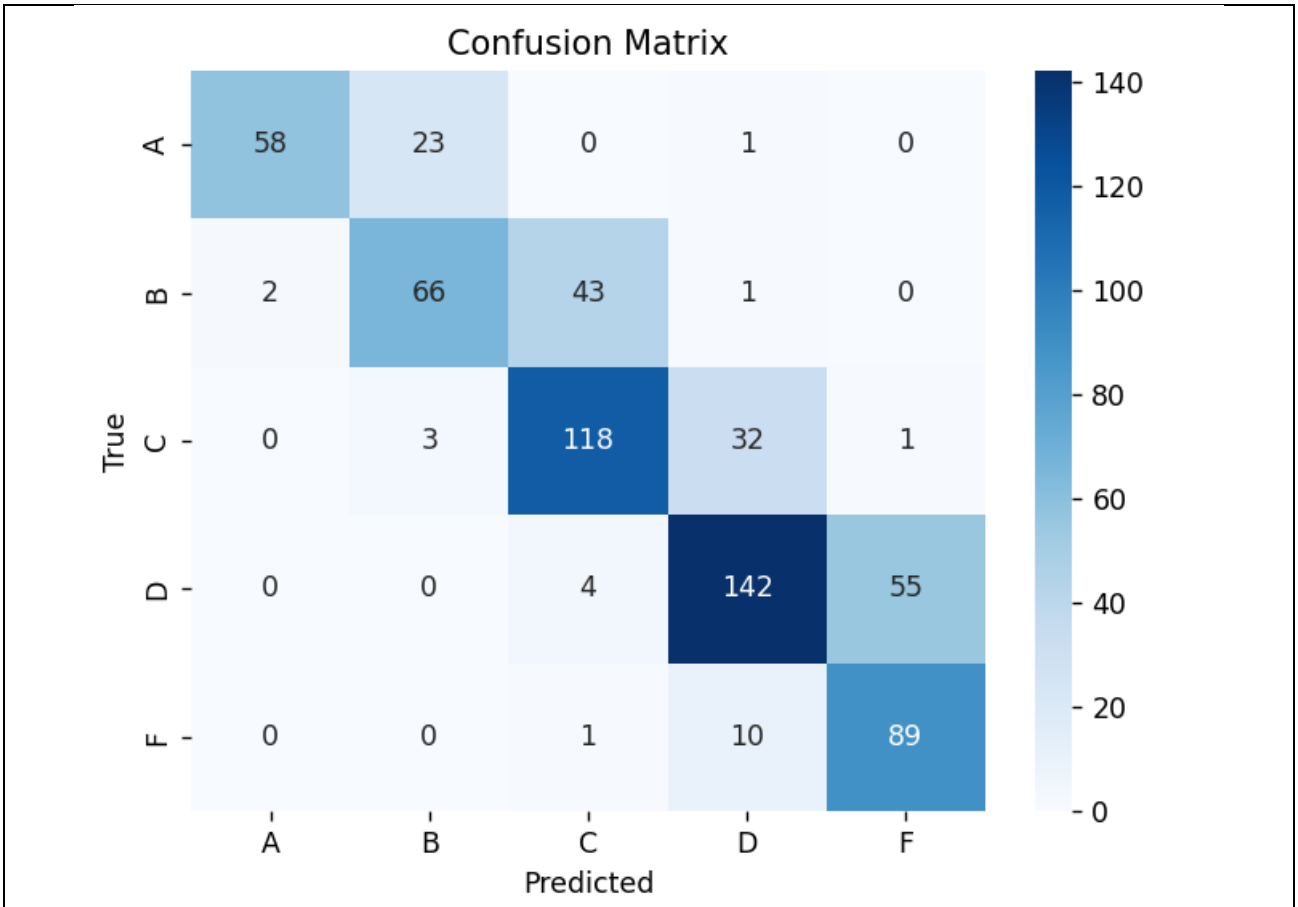
Procenat ispravne klasifikacije za multiklase (ocene A-F)

Matematika

	NV	SVM	DT pon	RF	LR	NB
A	78.48	70.98	74.98	70.82	71.6	34.17
B	60.5	58.42	56.96	49.52	55.2	34.17
C	32.91	34.81	30.93	34.24	32.3	34.17

Portugalski

	NV	SVM	DT pon	RF	LR	NB
A	72.88	72	72.13	72	69.65	33.84
B	58.7	57.53	61	52.82	54.83	33.84
C	30.97	36.19	32.86	36.48	36.26	33.84



Matrica konfuzije konfiguracije A ocean iz matematike

Na ovom primeru vidimo da su većinski ocene ispravno klasifikovane, a čak i kad nisu pomešane su sa susednom ocenom. To su verovatno ocene na samoj granici sa poenima koje je teško sa sigurnošću klasifikovati.

Za multinominalnu klasifikaciju model najbrojnijih klasa je najbolji za konfiguracije A i B što nam i odgovara jer je dovoljno samo prebrojati klase i uraditi predikciju pa nema bespotrebnog usložnjavanja sistema.

Root Mean Square Error za regresioni problem

Matematika

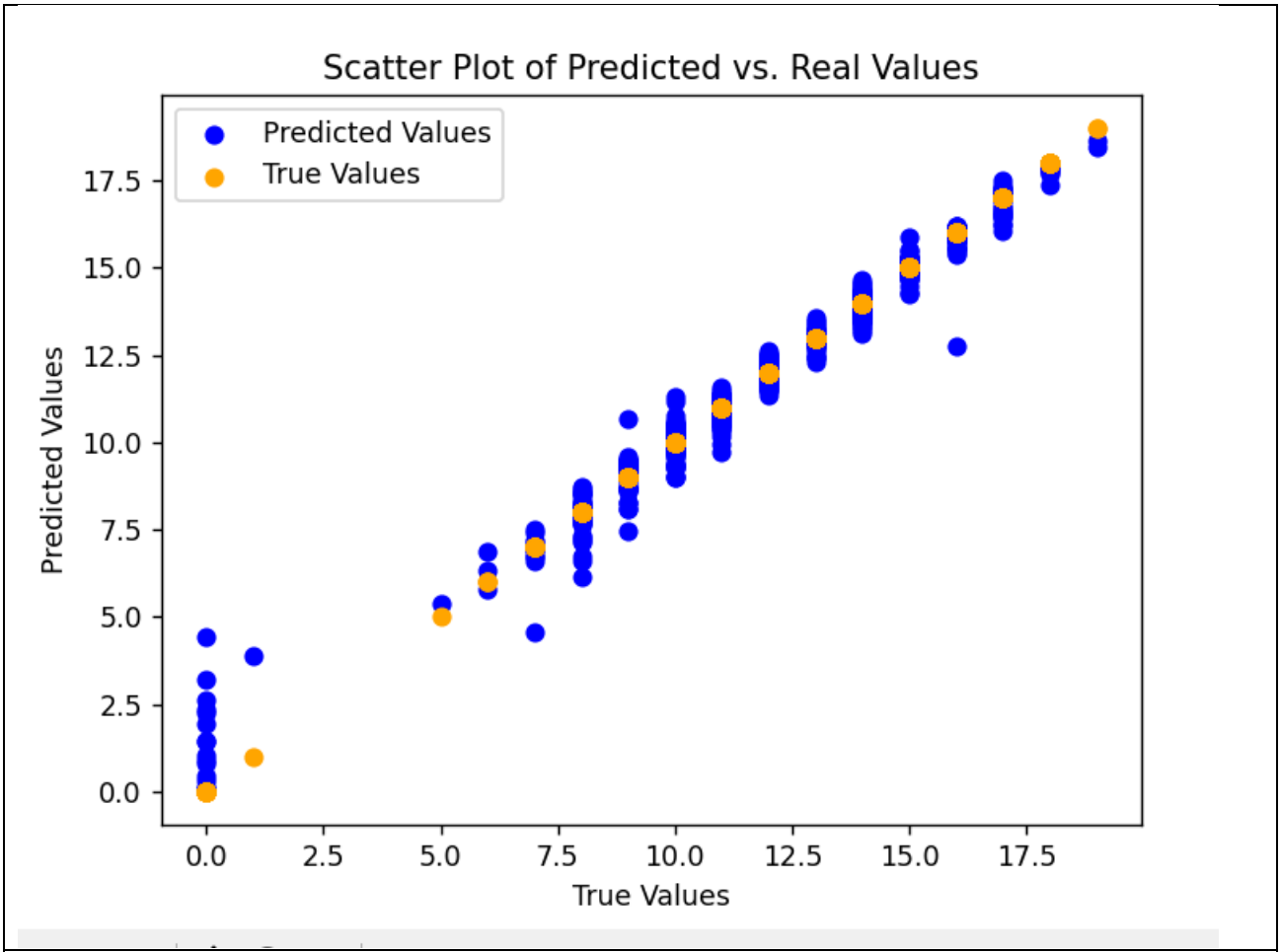
	NV	SVM	DT pon	RF	LR
A	2.01	2.4	2.05	1.55	4.09
B	2.8	3.24	2.64	2.28	7.88
C	4.58	5.51	4.14	3.85	19.2

Portugalski

	NV	SVM	DT pon	RF	LR
A	1.32	1.89	1.39	1.266	1.7
B	1.89	2.15	1.7	1.76	3.42
C	3.23	3	2.84	2.69	7.6

Za regresioni problem Random Forest ostvaruje najmanju grešku.

Za sve ove probleme treba uzeti u obzir složenost modela. Mnogi skorovi su prilično slični i treba sračunati da li se uopšte isplati imati složeniji model radi povećanja tačnosti za nekoliko procenata.



Predikcije Random Forest modela za konfiguraciju A ocena iz portugalskog