Projektni rad

Softverski algoritmi u sistemima automatskog upravljanja

Uvod i generalna postavka zadatka

Cilj projekta je predikcija završne ocene **G3** na osnovu dostupnih podataka. Prilikom analize ocene G3, korišćenje promenljivih **G1** i **G2** nema smisla, jer su one u snažnoj linearnoj korelaciji sa završnom ocenom, čime bi sama predikcija bila značajno pojednostavljena. Zbog toga će model biti evaluiran bez uključivanja parametara **G1** i **G2**.

Faze projekta:

- Analiza skupa podataka radi adekvatne pripreme za dalju obradu
- Eksplorativna analiza (uklanjanje anomalija, ispitivanje korelacija među podacima i sl.)
- Odabir odgovarajućeg modela
- Treniranje modela
- Evaluacija i analiza rezultata

Analiza skupa podataka

Ulazni podaci su:

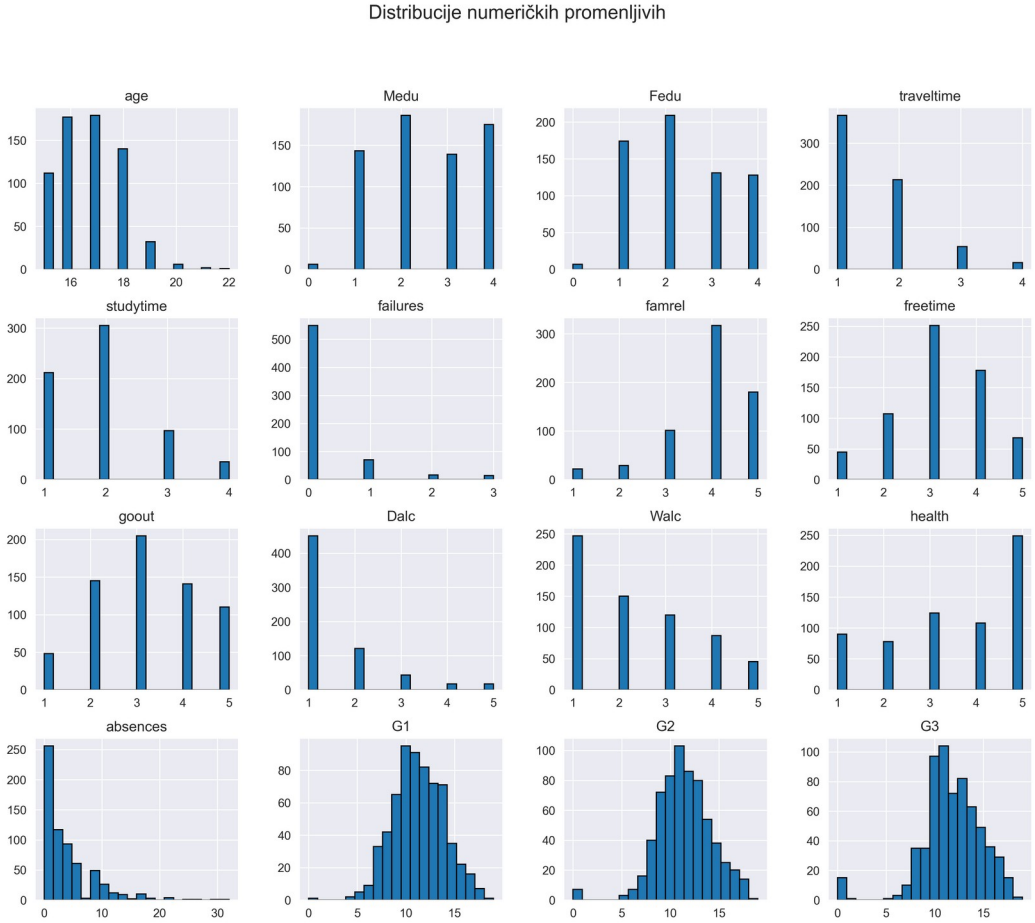
- Informacije o učenicima srednjih škola u Portugalu, ukupno *649 zapisa i* 33 promenljive
- Demografski podaci: *pol (sex), godine (age), adresa (address – urbana ili ruralna), veličina porodice (famsize)* i *status roditelja (Pstatus)*
- Obrazovanje i zanimanje roditelja: *stepen obrazovanja majke i oca (Medu, Fedu)* i *zanimanje (Mjob, Fjob)*
- Obrazovne navike: *vreme putovanja do škole (traveltime), vreme učenja nedeljno (studytime), broj neuspeha (failures)*
- Dodatni faktori: *porodična podrška (famsup), privatni časovi (paid), aktivnosti van škole (activities), prisustvo interneta (internet), romantična veza (romantic)*
- Slobodno vreme i ponašanje: *porodični odnosi (famrel), slobodno vreme (freetime), izlasci (goout), konzumacija alkohola radnim danima i vikendom (Dalc, Walc), zdravstveno stanje (health)* i *izostanci sa nastave (absences)*
- Ocene: *G1* (prva periodična ocena), *G2* (druga periodična ocena) i *G3* (završna ocena), pri čemu je G3 ciljna promenljiva u ovom projektu.<class 'pandas.core.frame.DataFrame'>

#	Kolona	Broj ne-nedostajućih vrednosti	Tip podataka
0	school	649	object
1	sex	649	object
2	age	649	int64
3	address	649	object
4	famsize	649	object
5	Pstatus	649	object
6	Medu	649	int64
7	Fedu	649	int64
8	Mjob	649	object
9	Fjob	649	object
10	reason	649	object
11	guardian	649	object
12	traveltime	649	int64
13	studytime	649	int64
14	failures	649	int64
15	schoolsup	649	object
16	famsup	649	object
17	paid	649	object
18	activities	649	object
19	nursery	649	object
20	higher	649	object
21	internet	649	object
22	romantic	649	object
23	famrel	649	int64
24	freetime	649	int64
25	goout	649	int64
26	Dalc	649	int64
27	Walc	649	int64
28	health	649	int64
29	absences	649	int64
30	G1	649	int64
31	G2	649	int64
32	G3	649	int64

Ukupno: 649 unosa, 33 kolone
Tipovi podataka: 16 numeričkih (*int64*), 17 kategorijskih (*object*)

Eksplorativna analiza podataka

Prva stvar koja ce se ovde uraditi je generalno posmatranje podataka iz fajla i pronalazenje bilo kakvih stvari koje odskacu ili nemaju smisla, kako bismo stekli generalni utisak o izgledu podataka i sledecim stvarima koje moramo da uradimo.



Demografske promenljive:

- Većina učenika ima između 15 i 18 godina, distribucija starosti je blago nagnuta ulevo – manji broj učenika starijih od 19 godina.

Obrazovanje roditelja (Medu, Fedu):

- Stepen obrazovanja roditelja je neravnomerno raspoređen, ali se vidi da najveći broj roditelja ima srednju školu (vrednost 2 ili 3), vrlo mali broj roditelja ima univerzitetsko obrazovanje (vrednost 4).

Studijske navike i školski faktori:

- Vreme učenja (studytime) i vreme putovanja (traveltime) imaju snažan disbalans – većina učenika provodi 1–2 sata u učenju nedeljno i ima kratko vreme putovanja do škole.
- Broj neuspeha (failures) je kod ogromne većine 0, što ukazuje da je dataset dominantno sastavljen od učenika koji nisu ponavljali godinu.

Slobodno vreme i ponašanje:

- Većina učenika ima solidne porodične odnose (famrel ≈ 4) i umerenu količinu slobodnog vremena (freetime ≈ 3), distribucija odlazaka u izlazak (goout) pokazuje ravnomernu raspodelu, što znači da se učenici međusobno razlikuju po socijalnim navikama.

Konzumacija alkohola (Dalc, Walc):

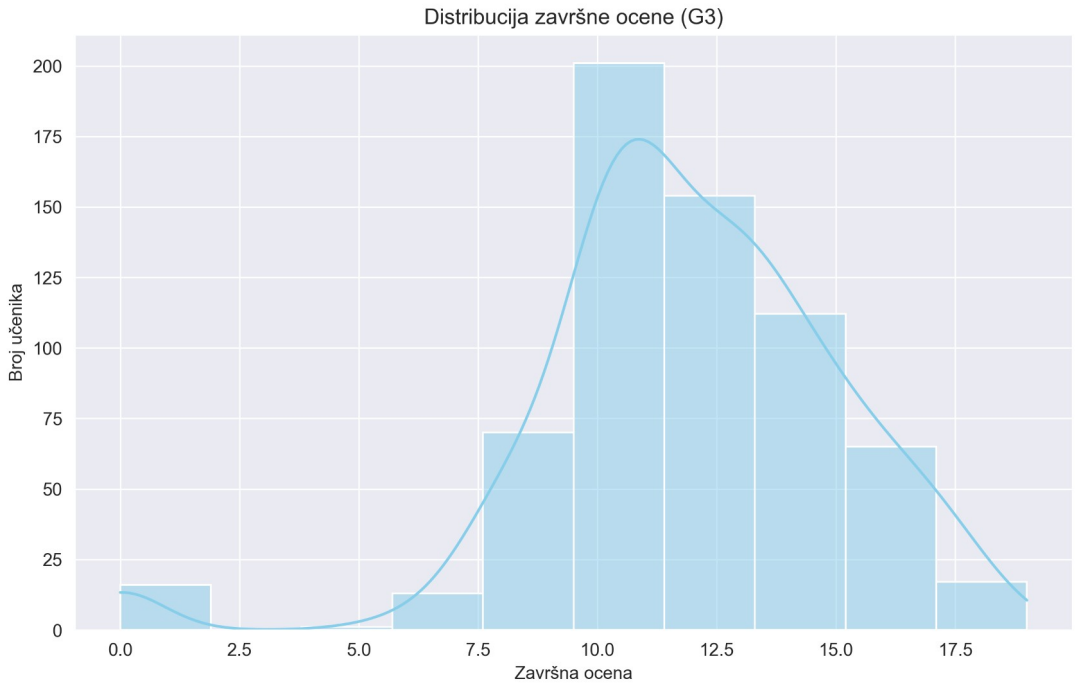
- Konzumacija alkohola radnim danima (Dalc) je vrlo niska – ogromna većina ima vrednost 1 (minimalno), konzumacija vikendom (Walc) je viša i raspodela je ravnomernija – što sugeriše da učenici vikendom češće piju.

Zdravstveno stanje i izostanci:

- Zdravstvena ocena (health) je uglavnom u opsegu 3–5, dakle učenici sebe procenjuju kao umereno do dobro zdrave.
- Izostanci (absences) imaju ekstremno desno nagnutu distribuciju, što znači da većina ima vrlo malo izostanaka, ali postoji mali broj učenika sa izuzetno velikim brojem izostanaka.

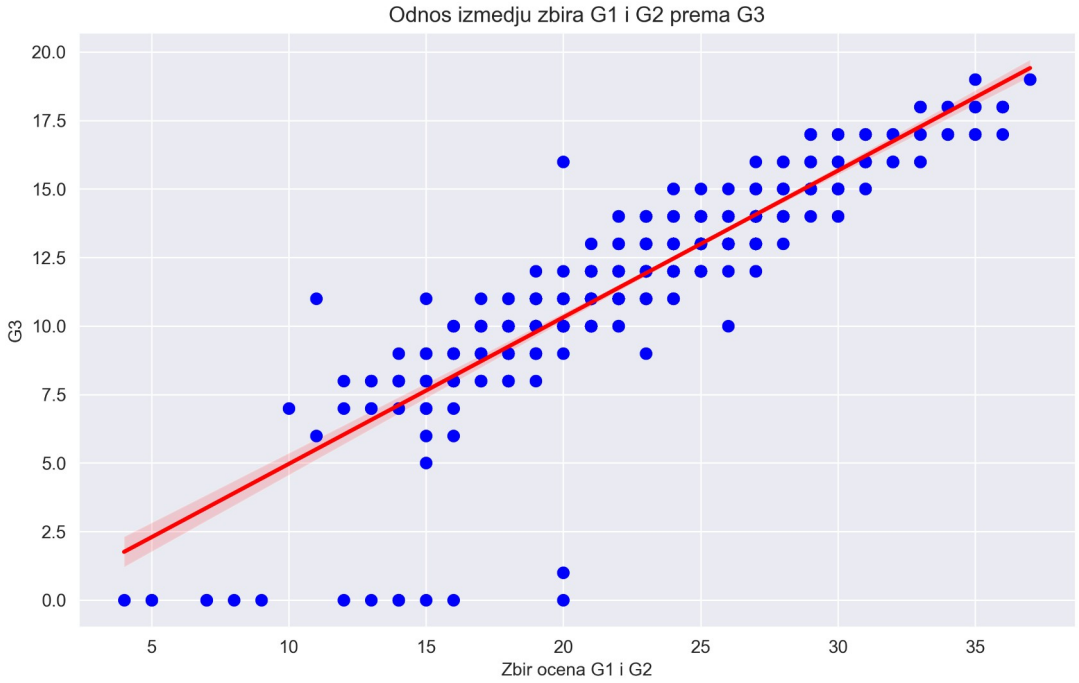
Ocene (G1, G2, G3):

- Sve tri ocene imaju sličnu raspodelu — oblik približno normalan, centriran oko vrednosti 10–12, što je srednji uspeh.
- G1, G2 i G3 su očigledno u snažnoj korelaciji, jer se njihove distribucije gotovo poklapaju.
- Ima vrlo malo učenika sa ocenom 0, što može ukazivati na izuzetne slučajeve (možda nedostatak

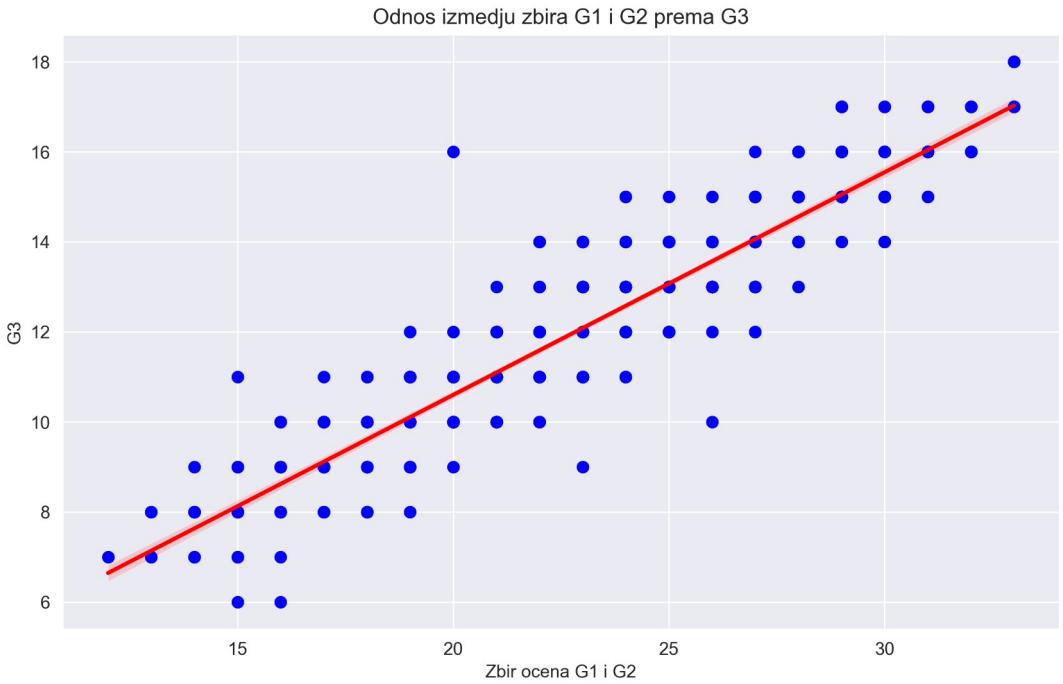


podataka ili učenike koji nisu prisustvovali završnom ispitu).

Napomena: svi podaci nisu nuzno bitni za dalju analizu, ali je i dalje veoma korisno videti kako izgledaju podaci kako bismo stekli utisak o generalnom stanju dataset-a



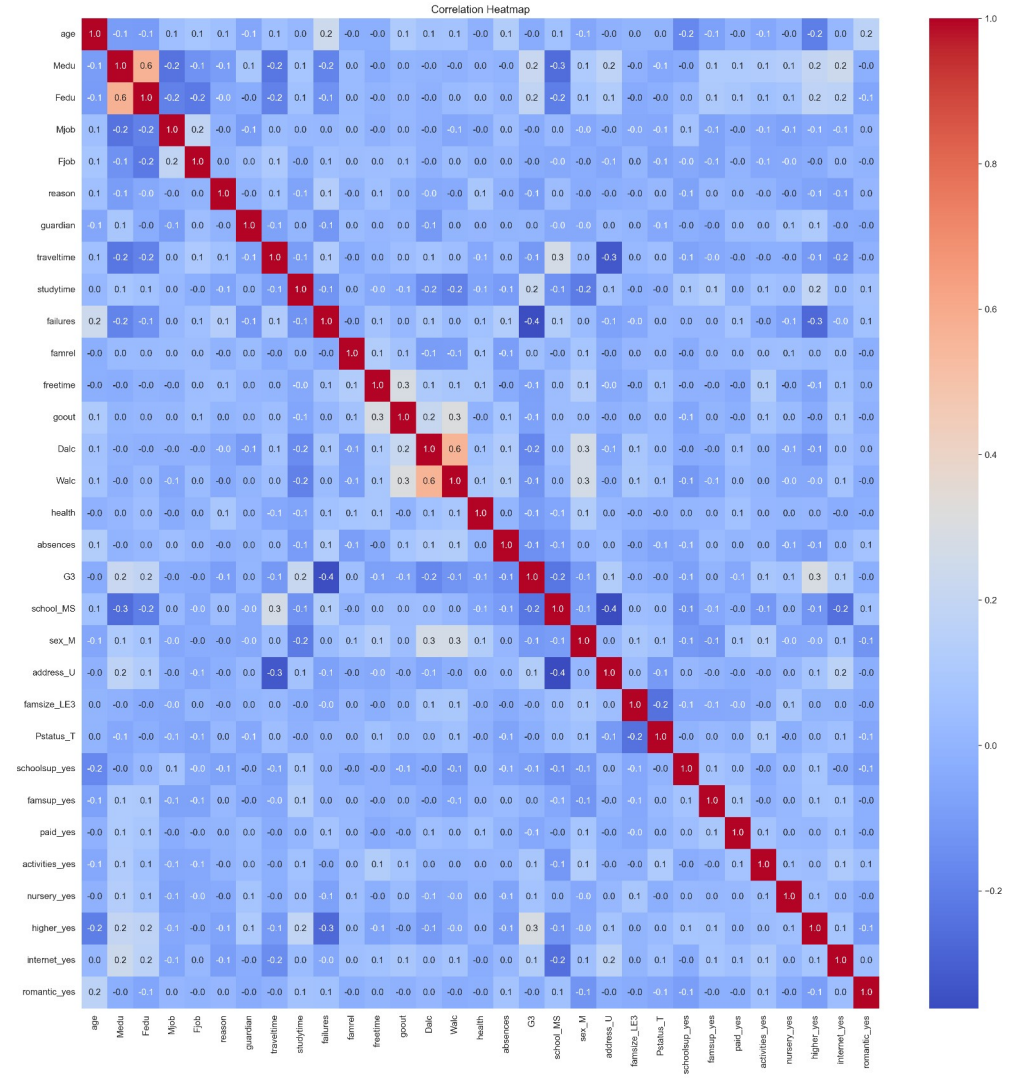
Iz prethodne analize nije uoceno odstupanja u drugim ulaznim parametrima sem u G1,G2 i G3. Anomalije uklanjamo koriscenjem z-score-a i koriscenjem IQR metode za uklanjanje anomalija, takodje uklanjamo neke podatke koji nemaju smisla poput toga da je jedna od ocena jednaka 0 a finalna ocena G3 razlicita od nule…
Nakon obrade anomalija u G1, G2 i G3, grafika izgleda ovako:



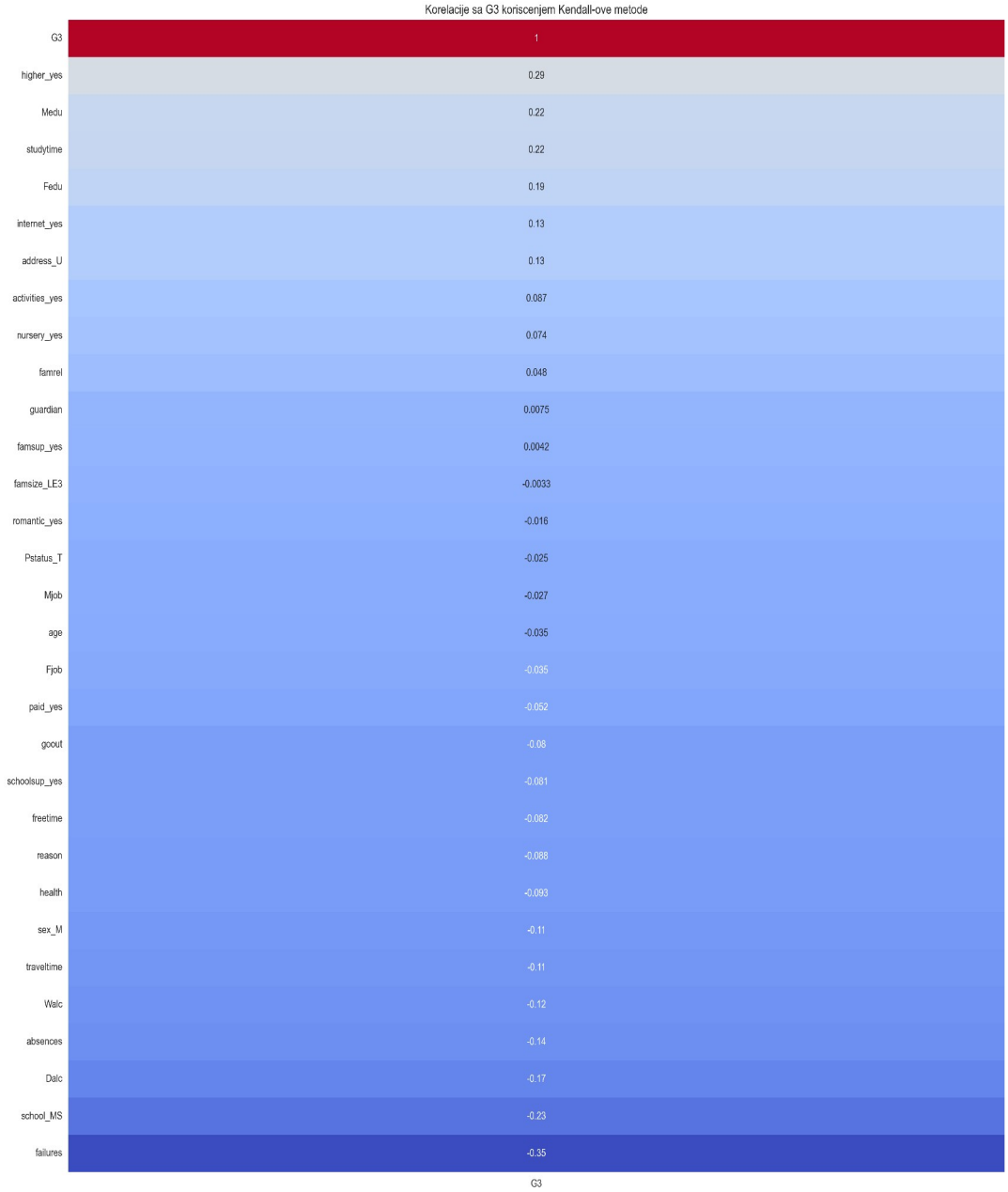
Uocavamo mnogo bolji izgled podataka G1, G2 i G3 gde su oni prakticno u linearnom odnosu. Linearnom regresijom ova cinjenica postaje ocevidna.

Dalje enkodiramo i skaliramo podatke kako bismo mogli uravnotezeno da vidimo korelacija izmedju istih.
Za enkodiranje podataka poput: school, sex, address, famsize… koristimo one hot encoding sa drop_true kako ne bismo koristili bepotrebne podatke, dok za ostale podatke sa vise mogucih stanja (pod ovime se podrazumeva npr. Mjob koji moze da ima 3 stanja) koristimo frequency encoding, koji gleda koliko se puta pojavljuje neki podatak i njega zameni sa frekvencijom njegovog pojavljivanja.
Za skaliranje podataka koristimo StandardScaler() iz sklearn.preprocessing biblioteke.

Dalje gledamo korelacije izmedju podataka:



Za pronalazenje korelacija koristimo Kendall-ovu metodu, koja je dosta slican Pearson-ovoj, ali je dosta fleksibilnija za nas dataset.
Zbog enkodiranja podataka heatmap-a ima puno parametara i izgleda poprilicno necitljivo, ali ovo nece pretstavljati problem jer je nama samo bitna korelacija izmedju G3 i ostalih podataka, ne korelacija izmedju svih podataka jer ostali podaci nemaju relevantne korelacije kako bismo ih mogli ikako zameniti, tako da dalje samo to mozemo da posmatramo:



Odavde se vidi da najveci uticaj na G3 ima sama G3, sto je trivijalno, ali pored nje najveci uticaj imaju aspiracije ucenika za upisivanjem viseg obrazovanja, kao i broj neuspeha ucenika, dok malu korelaciju imaju podaci poput romantic, age… (manju korelaciju od praga (0.05 po apsolutnoj vrednosti)) tako da te podatke necemo uvazavati u daljoj analizi.
Pred obucavanje konkretnih modela trebala bi se uraditi PCA metoda za smanjivanje dimenzionalnosti podataka, ali iz razloga sto ovo lose utice na performanse modela, ovo nece biti uradjeno.

Odabir modela

U cilju predviđanja završne ocene učenika (**G3**) na osnovu skupa osobina (demografske informacije, navike u učenju, porodično okruženje i prethodne ocene), korišćeni su različiti regresioni modeli.
Izabrani modeli pokrivaju i linearne i nelinearne pristupe, kako bi se ispitalo koji tip bolje opisuje zavisnosti u podacima.
Korišćeni modeli su:

- **Linear Regression** – kao osnovni model radi poređenja;
- **Ridge i Lasso Regression** – kao regularizovane verzije linearne regresije radi sprečavanja prenaučenosti (overfittinga);
- **Decision Tree Regressor** – kao predstavnik nelinearnih modela koji omogućava interpretaciju granica odlučivanja;
- **Random Forest Regressor** i **Gradient Boosting Regressor** – kao ansambl metode za unapređenje performansi i smanjenje varijanse modela.

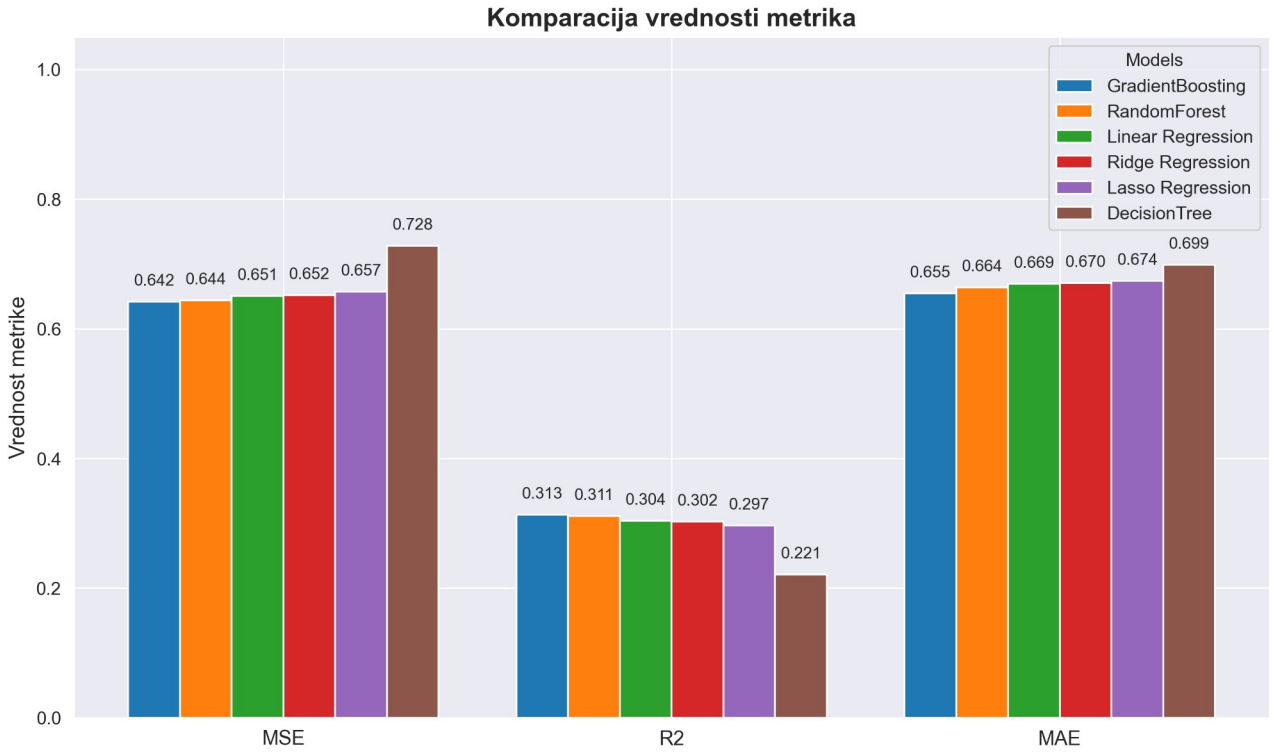
Treniranje modela

Podaci su podeljeni na **trening** (80%) i **test** (20%) skup pomoću funkcije train_test_split sa zadatim random state = 42 radi ponovljivosti rezultata.
Za Ridge i Lasso modele korišćena je unakrsna validacija (cross_val_score, RandomSearchCV...) kako bi se pronašle optimalne vrednosti hiperparametara (alpha).
Kod ansambl modela (Random Forest i Gradient Boosting), izvršena je optimizacija parametara poput broja stabala i maksimalne dubine, kako bi se postigao balans između tačnosti i vremena izvršavanja.
Za **treniranje linearne** regresije koriscena je najbazicnija funkcija linearne regresije iz biblioteke sklearn.
Za **ridge regresiju** koriscena je optimizacija ridge regresije sa RidgeCV metodom, kojoj je bio prosledjen parametar alpha sa logaritamskim korakom od -3 do 3, cv koji oznacave sa koliko modela ce biti obavljena cross validacija I scoring neg_mean_squared_error koji oznacava da treba da se pronadje model za najmanjom kvadratnom greskom. Za najbolji alpha dobijena je vrednost: 130.953.
Za **lasso regresiju** koriscnea je optimizacije LassoCV metodom, kojoj su proslednjeni slicni parametri kao i kod RidgeCV metode. Za najbolji alpha izabran je: 0.031
Za **Decision Tree** parametri koji su optimizovani su: criterion': ['squared_error','absolute_error'],'min_samples_split': np.arange(2, 15),'min_samples_leaf': np.arange(1, 10), 'max_leaf_nodes': [None] + list(np.arange(5, 100, 5)),'ccp_alpha': np.linspace(0.0, 0.05, 30). Za optimizaciju parametara koriscena je RandomSearchCV metoda radi smanjivanja vremena izvrsavanja (moguce je koristiti i GridSearchCV, bez vecih izmena u kodu). Za najbolje parametre dobijeni su: {'min_samples_split': np.int64(9), 'min_samples_leaf': np.int64(8), 'max_leaf_nodes': np.int64(5), 'criterion': 'squared_error', 'ccp_alpha': np.float64(0.032758620689655175)}
Za **RandomForest** optimizovani parametri su: 'criterion': ['squared_error', 'absolute_error'], 'min_samples_split': np.arange(2, 15),'min_samples_leaf': np.arange(1, 10),'max_leaf_nodes': [None] + list(np.arange(5, 100, 5)),'ccp_alpha': np.linspace(0.0, 0.05, 30). Slicno kao I kod DecisionTree-a koriscena je RandomSearchCV metoda. Najbolji parametri su: {'min_samples_split': np.int64(12), 'min_samples_leaf': np.int64(7),'max_leaf_nodes':np.int64(5),'criterion':'squared_error','ccp_alpha':np.float64(0.024137931034482762)}
Za **GradientBoosting** optimizovani su: 'n_estimators': np.linspace(100, 1000, 10, dtype=int),'learning_rate': np.linspace(0.01, 0.3, 10),'min_samples_split': [2, 5, 10],'min_samples_leaf': [1, 2, 4],'subsample': np.linspace(0.6, 1.0, 5),'max_features': [None, 'sqrt', 'log2']. Najbolji parametri su: {'subsample': np.float64(0.8), 'n_estimators': np.int64(400), 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'log2', 'learning_rate': np.float64(0.01)}

Rezultati izvrsavanja

Za prikaz performanse modela koriscene su: Mean Squared Error, Mean Absolute Error i R^2 score. Vrednosti su prikazane u tabeli ispod:

	MSE	R2	MAE
Linear Regression	0.651	0.304	0.669
Ridge Regression	0.652	0.302	0.67
Lasso Regression	0.657	0.297	0.674
DecisionTree	0.728	0.221	0.699
RandomForest	0.644	0.311	0.664
GradientBoosting	0.642	0.313	0.655



Analiza rezultata

Na osnovu prikazane komparacije metrika (MSE, R^2 i MAE) za šest različitih regresionih modela — **Gradient Boosting**, **Random Forest**, **Linear Regression**, **Ridge Regression**, **Lasso Regression** i **Decision Tree** — može se zaključiti sledeće:

- **Decision Tree** model pokazuje **najveću vrednost MSE (0.728)** i **najmanju vrednost R^2 (0.221)**, što ukazuje na to da se loše prilagođava podacima i ima najveću grešku predviđanja.
- S druge strane, **Gradient Boosting** i **Random Forest** daju **najbolje ukupne rezultate**, sa **najnižim MSE (≈0.642–0.644)** i **najvišim R^2 (≈0.313–0.311)**, što znači da imaju najveću tačnost u predviđanju ciljne promenljive.
- Linearni modeli (**Linear Regression**, **Ridge Regression** i **Lasso Regression**) imaju slične performanse – R^2 vrednosti se kreću oko **0.30**, dok su MSE i MAE vrlo blizu vrednostima koje postižu Random Forest i Gradient Boosting, što ukazuje da linearni pristup i dalje uspešno opisuje deo varijabilnosti u podacima.
- Vrednosti **MAE** (srednje apsolutne greške) dodatno potvrđuju da su **Gradient Boosting** i **Random Forest** najstabilniji, sa najmanjim prosečnim odstupanjem predikcija od stvarnih vrednosti (~0.655–0.664).

Na osnovu svih metrika može se zaključiti da:

- **Gradient Boosting** i **Random Forest** predstavljaju **najbolji izbor modela** za dati skup podataka, jer ostvaruju **najmanje greške** i **najveći koeficijent determinacije (R^2)**.
- Klasični linearni modeli su solidni, ali ne uspevaju da uhvate složenije nelinearne odnose u podacima.
- **Decision Tree** kao pojedinačno stablo pokazuje značajno slabije performanse i nije pogodan za krajnji model, ali njegovi rezultati opravdavaju upotrebu **ensemble metoda** poput Random Foresta i Gradient Boostinga koji značajno poboljšavaju tačnost.

Zavisnost izlaza od promenljivih za najbolji model: