Matematički fakultet, Univerzitet u Beogradu

Klasifikacija perifernih mononuklearnih krvnih ćelija

Seminarski rad u okviru kursa Istraživanje podataka 2

Profesor:
Prof. dr Nenad Mitić

Autori:
Bogdan Stojadinović
Stefan Kerkoč

Maj 2024.

Sadržaj

1 UVOD

Naš zadatak je da iskoristimo podatke iz datoteka BS1, BS2, GEO i 10x da napravimo klasifikacione modele koji će da predviđaju tipove perifernih krvnih mononuklearnih ćelija. Za početak ćemo napraviti model za svaku pojedinačnu datoteku, a zatim ćemo iskoristiti model treniran na BS1 podacima i primeniti ga na preostala tri skupa. Ulazne podatke ćemo deliti u srazmeri 70:30 (trening:test), a kod primene ćemo koristiti cele skupove BS2, 10X i BS2 kao test. Kako bismo za svaki algoritam našli optimalne parametre, poslužićemo se tehnikom unakrsne provere. To je tehnika koja se koristi za procenu performansi i uopštavanja modela mašinskog učenja deljenjem podataka u više podskupova, ili „preklopa“ (en. folds), i iterativno obučavanjem i proverom modela na ovim preklopima.

Za potrebe našeg projekta korišćen je programski jezik python, konkretno jupyter sveske. Pre pokretanja, neophodno je pokrenuti sledeću komandu kako biste instalirali sve neophodne python biblioteke:

```
pip install pandas numpy xgboost scikit-learn imbalanced-learn matplotlib seaborn joblib
```

Takođe potrebno je skinuti pomenute datoteke BS1, BS2, GEO i 10x, napraviti direktorijum *data* u istom direktorijumu gde se nalazi ostatak projekta i u njega smestiti preuzete datoteke.

2 BIOLOŠKA OSNOVA

Periferna krvna mononuklearna ćelija (en. Peripheral blood mononuclear cell, PBMC) je svaka periferna krvna ćelija koja ima okruglo jedro. Ove ćelije mogu biti limfociti (T ćelije, B ćelije, NK ćelije) i monocita (M ćelije), dok eritrociti i trombociti nemaju jedra, kao i granulociti (neutrofili, bazofili i eozinofili). Kod ljudi, limfociti čine većinu PBMC populacije, zatim slede monociti, a samo mali procenat čine dendritske (D) ćelije.

Ove ćelije se mogu izvući iz krvi korišćenjem ficola, hidrofilnog polisaharida koji razdvaja slojeve krvi, i gradijentnom centrifugacijom, koja razdvaja krv na gornji sloj plazme, zatim sloj PBMC-a (en. buffy coat) i donji sloj polimorfonuklearnih ćelija (kao što su neutrofili i eozinofili) i eritrocita. Polimorfonuklearne ćelije se mogu dodatno izolovati liziranjem crvenih krvnih ćelija. Bazofili se ponekad nalaze i u gušćim i u PBMC frakcijama.

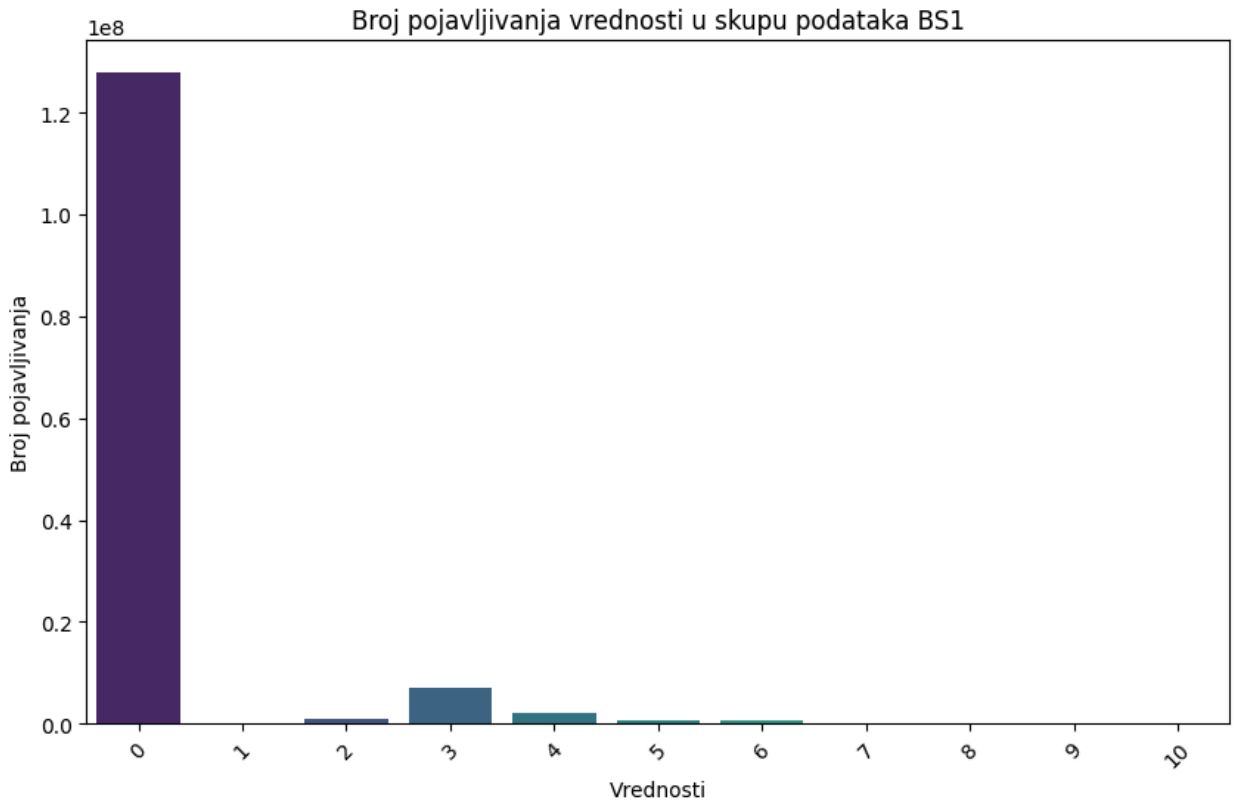
Nedavne studije ukazuju na to da PBMC mogu biti podložne patogenim infekcijama, kao što su Ureaplasma parvum i urealiticum, Mycoplasma genitalium i hominis, i Chlamydia trachomatis infekcije. PBMC mogu biti takođe podložne virusnim infekcijama. Zaista, tragovi JC poliomavirusa i Merkel ćelijskog poliomavirusa su detektovani u PBMC kod trudnica i žena pogođenih spontanim pobačajem.

Mnogi naučnici koji sprovode istraživanja u oblastima imunologije (uključujući autoimune poremećaje), infektivnih bolesti, hematoloških maligniteta, razvoja vakcina, transplantacione imunologije i visoko-protočnog skrininga često koriste PBMC. U mnogim slučajevima, PBMC se dobijaju iz banaka krvi. PBMC frakcija takođe sadrži progenitorske populacije, što je demonstrirano testovima formiranja kolonija baziranim na metilcelulozi.

Smatra se da PBMC predstavljaju važan put za vakcinaciju. PBMC od pacijenata sa rakom mogu se ekstrahovati i kultivisati *in vitro*. Nakon toga, PBMC se izlažu tumorskim antigenima kao što je antigen tumorskih matičnih ćelija. Inflamatorni citokini se obično dodaju kako bi pomogli u preuzimanju i prepoznavanju antigena od strane PBMC.

3 ANALIZA I PRETPROCESIRANJE PODATAKA

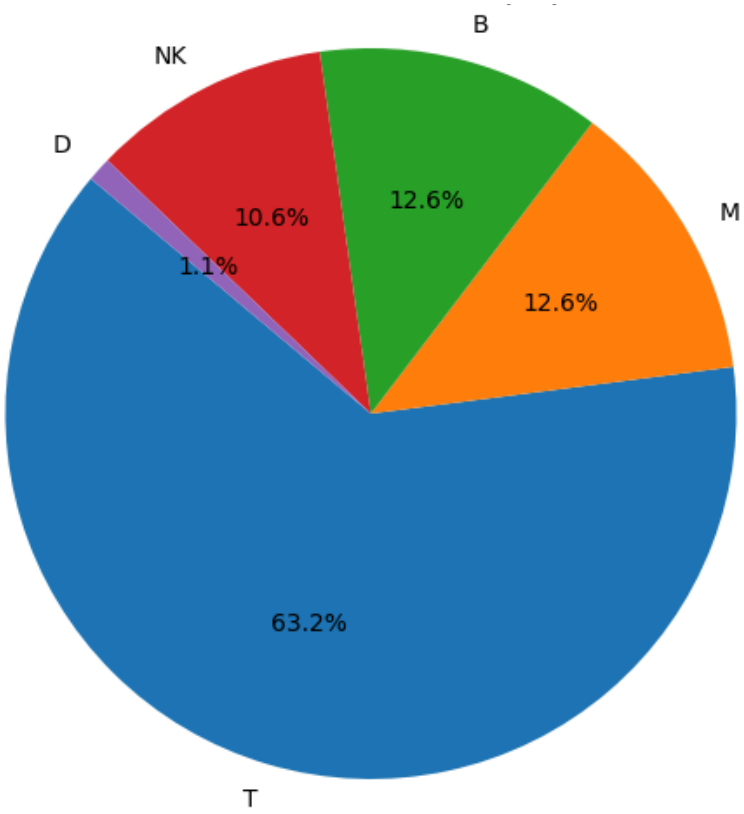
Za početak, zadatak koji je neophodno da uradimo je da izvučemo informacije o klasama iz kolone "`src_file`". To radimo raščlanjivanjem svake vrednosti po karakteru "_" i čitanjem prvog slova trećeg člana. Izvršavanjem *preprocessing.ipynb* iz direktorijuma *preprocessing* kao rezultat dobijamo novonastale datoteke *preprocessed_data_[ime_datoteke].csv*. Nakon toga, pre bilo kakvog pretprocesiranja, potrebno je da analiziramo podatke. Ovo smo radili unutar *eda.ipynb* iz direktorijuma *preprocessing* i prva provera je broj nedostajućih vrednosti, čiji rezultat je da nedostajućih vrednosti nema. Analizom broja pojavljivanja svake od vrednosti u skupu podataka zaključujemo da je u pitanju redak skup podataka. Rezultate ove analize možemo videti na slici 1.



Slika 1: Broj pojavljivanja vrednosti u skupu podataka BS1

Informacija koja nam je takođe potrebna je udeo redova svake klase u skupu podataka. Očekujemo nebalansiran skup, s obzirom da su B i T ćelije znatno brojnije u ljudskom organizmu u odnosu na ostale. Naša pretpostavka će se ispostaviti donekle tačnom, s obzirom da T ćelija ima ubedljivo najviše (slika 2). Bitnije za nas, vidimo da su klase nebalansirane i da ćemo morati da koristimo neke od tehnika za balansiranje klasa ili algoritme koji dobro funkcionišu sa takvim klasama. Tehnike koje biramo su naduzorkovanje i poduzorkovanje, kao i dodavanje težina klasama.

Kako bismo dobili nove podatke nastale primenama gore pomenutih tehnika neophodno je pokrenuti *oversampling.ipynb* iz direktorijuma *preprocessing*, i kao rezultat dobijamo datoteke *oversampled_train_[ime_datoteke].csv* i *oversampled_test_[ime_datoteke].csv* sačuvane u direktorijumu *data*, koje ćemo kasnije koristiti pri izradi i proveri modela.



Slika 2: Udeo redova svake klase u skupu podataka

Zbog prirode problema, čuvanje što većeg broja kolona je jako bitno. Iz tog razloga jedina vrsta pretprocesiranje koja je rađena je uklanjanje kolona sa varijansom 0. Uklonjeno je 239 kolona od početnih 10800.

4 KLASIFIKACIJA

4.1 SLUČAJNA ŠUMA (RANDOM FOREST)

Prvi algoritam koji primenjujemo jesu **Slučajne šume** (en. Random Forest). Ovaj algoritam je svestran i moćan metod učenja ansambla koji se prvenstveno koristi za zadatke klasifikacije i regresije. On funkcioniše tako što konstruiše mnoštvo stabala odlučivanja (en. Decision Trees) tokom treniranja, a onda njihovim glasanjem određuje konačno predviđanje. Time se poboljšava tačnost i kontroliše preprilagođavanje, čime se smanjuje i varijansa. Slučajne šume nam pružaju otpornost na šum u podacima, kao i sposobnost rukovanja velikim skupovima podataka sa većom dimenzionalnošću. Ove osobine ga čine popularnim izborom za različite modele u mašinskom učenju.

Koristimo slučajne šume u kombinaciji sa unakrsnom proverom kako bismo dobili najbolji model po nekom kriterijumu koji odaberemo. Ukoliko za taj kriterijum odaberemo meru tačnosti, dobijeni najbolji model daje sledeće rezultate:

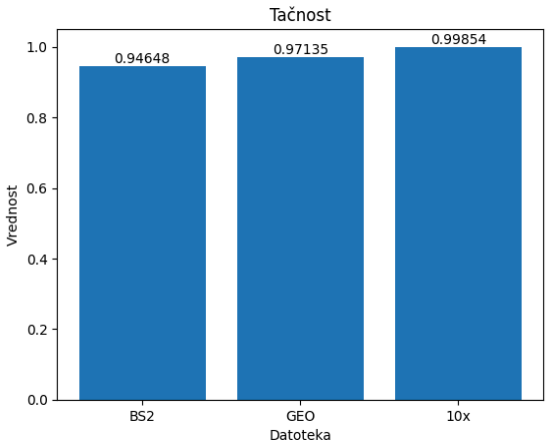
Metrika	Vrednost (%)
Tačnost	96.11
Preciznost	96.03
Odziv	96.11
F1 mera	96.06

Tabela 1: Rezultati klasifikacije za BS1 datoteku koristeći Slučajne šume

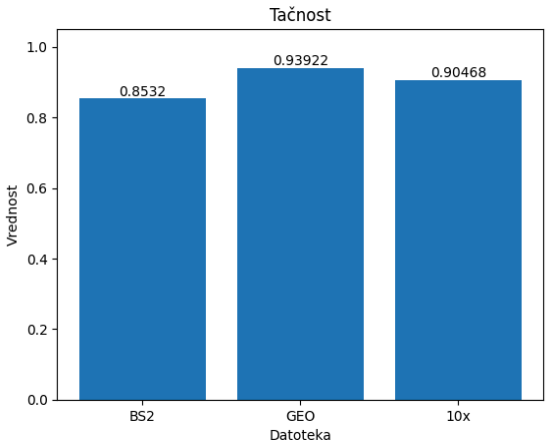
	B	D	M	NK	T
B	1142	0	0	0	0
D	0	98	3	0	0
M	0	0	1165	0	0
NK	0	0	0	977	25
T	0	0	0	2	5816

	B	D	M	NK	T
B	515	1	1	0	1
D	0	36	5	0	0
M	0	1	493	0	2
NK	0	0	0	309	83
T	0	0	1	59	2448

Tabela 2: Matrice konfuzije za trening (levo) i test (desno) pri upotrebi BS1



Slika 3: Rezultati klasifikacije za svaku od datoteka



Slika 4: Rezultati primene BS1 modela na svaku od datoteka

	B	D	M	NK	T
B	1713	0	0	0	164
D	167	0	22	0	81
M	13	0	1459	0	534
NK	0	0	0	46	796
T	4	0	0	2	7145

Tabela 3: Matrica konfuzije pri primeni modela na BS2

	B	D	M	NK	T
B	1360	0	103	25	268
D	0	0	0	0	0
M	3	0	835	0	18
NK	0	0	0	31	278
T	173	0	108	21	13182

Tabela 4: Matrica konfuzije pri primeni modela na GEO

	B	D	M	NK	T
B	9457	0	0	0	268
D	0	0	0	0	0
M	59	0	1617	0	172
NK	0	0	0	855	7324
T	32	0	1	1	62642

Tabela 5: Matrica konfuzije pri primeni modela na 10x

Treniranje ovog modela radimo u *random_forest.ipynb* datoteci iz direktorijuma *models*. U poddirektorijumu *trained_models* čuvamo najbolji model za čije treniranje su korišćeni sledeći parametri:

- `'max_depth': 15,`
- `'min_samples_leaf': 1,`
- `'min_samples_split': 5,`
- `'n_estimators': 300,`
- `'criterion': 'entropy'`

Ukoliko istreniramo modele za svaku od datoteka koristeći pomenute parametre dobijamo rezulate sa slike 3. U slučaju da iskoristimo već kreiran model koji je treniran na BS1 podacima dobijamo rezultate sa slike 4. Tačnost nam je znatno lošija u drugom slučaju. Primećujemo da naš model jako loše predviđa klase D kod BS2, i NK kod sve tri datoteke. Takođe, iz prikazanih matrica konfuzije, vidimo da ima nagon da predviđa klasu T dosta često, što donekle i ima smisla s obzirom da je broj instanci koje pripadaju toj klasi znatno veći od ostalih, ali nam i ukazuje na moguće preprilagođavanje modela.

4.2 XGBOOST

XGBoost (eXtreme Gradient Boosting) je visoko efikasan i skalabilan algoritam mašinskog učenja dizajniran za rešavanje problema u nadgledanom učenju. Slično kao i slučajne šume, zasnovan je na stabalima odlučivanja, ali je više usresređen na smanjivanje pristrasnosti i uklanjanje preprilagođavanja. Koristi gradijentno pojačanje kako bi pronašao najbolji rezultat i gradi modele na sekvencijalni način da bi se ispravile greške napravljene kod prethodnih modela. Ovaj algoritam se ističe zbog svoje brzine, performansi i sposobnosti da rukuje velikim skupovima podataka sa složenim obrascima. Njegove napredne tehnike regularizacije sprečavaju preprilagođavanje, dok funkcije kao što su paralelna obrada i potkresivanje stabala poboljšavaju njegovu računarsku efikasnost i moć predviđanja.

Za kreiranje ovog modela potrebno je pokrenuti svesku *xgb.ipynb* iz direktorijuma *models.* Ponovo koristimo izabrani algoritam zajedno sa unakrsnom proverom kako bismo odredili najbolje parametre. U ovom slučaju korišćeni parametri su:

- `'learning_rate': 0.2,`
- `'max_depth': 4,`
- `'n_estimators': 300`

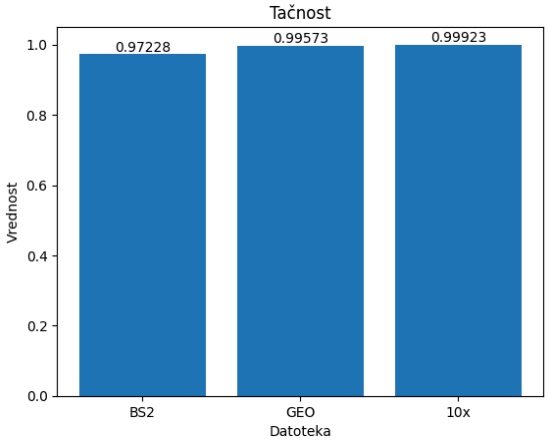
Metrika	Vrednost (%)
Tačnost	96.48
Preciznost	96.58
Odziv	96.48
F1 mera	96.52

Tabela 6: Rezultati klasifikacije za BS1 datoteku koristeći XGBoost

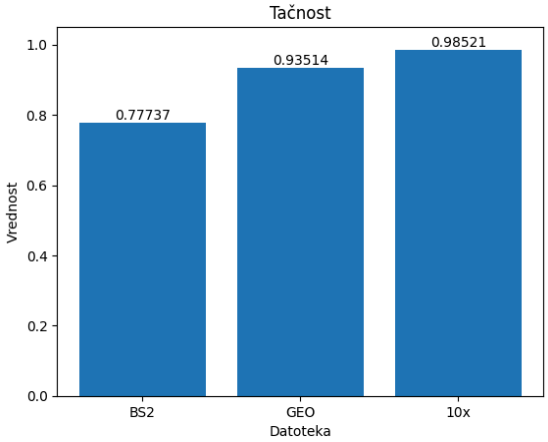
	B	D	M	NK	T
B	1142	0	0	0	0
D	0	101	0	0	0
M	0	0	1165	0	0
NK	0	0	0	1002	0
T	0	0	0	0	5818

	B	D	M	NK	T
B	516	0	1	0	1
D	0	41	0	0	0
M	0	1	493	0	2
NK	1	0	0	338	53
T	1	0	1	78	2428

Tabela 7: Matrice konfuzije za trening (levo) i test (desno) pri upotrebi BS1



Slika 5: Rezultati klasifikacije za svaku od datoteka



Slika 6: Rezultati primene BS1 modela na svaku od datoteka

	B	D	M	NK	T
B	1683	0	0	0	194
D	31	17	44	0	178
M	42	0	520	0	1444
NK	0	0	0	73	769
T	16	0	2	37	7096

Tabela 8: Matrice konfuzije pri primeni modela na BS2

	B	D	M	NK	T
B	1361	1	103	79	212
D	0	0	0	0	0
M	3	13	821	2	17
NK	0	0	0	281	28
T	199	11	107	286	12881

Tabela 9: Matrice konfuzije pri primeni modela na GEO

	B	D	M	NK	T
B	9672	0	0	0	53
D	0	0	0	0	0
M	32	96	1628	3	89
NK	1	0	0	7447	731
T	51	8	6	57	62554

Tabela 10: Matrica konfuzije pri primeni modela na 10x

Kada uporedimo rezultate koje smo dobili koristeći slučajne šume sa rezultatima XGBoost-a, primećujemo slične probleme. Sa tabele 6 i slike 5 vidimo da kada primenimo algoritam na svaku datoteku ponaosob, dobijamo jako dobre rezultate. U slučaju kada na BS2 primenimo model treniran na BS1, rezultati nisu zadovoljavajući, a sa matrica konfuzije možemo da izvedemo zaključak da i ovaj model ima favorizuje predviđanje klase T. Sa druge strane klasifikacija BS2, a pogotovo 10x podataka, je i više nego prihvatljiva. Model i dalje za neke instance kaže da pripadaju klasi D iako u ove dve datoteke nemamo takvih ćelija, ali je broj takvih promašaja zanemarljivo mali.

4.3 LOGISTIČKA REGRESIJA

Logistička regresija je statistički metod koji se, u osnovnoj varijanti koristi za modelovanje binarnih ishoda. U daljem tekstu logistička regresija koja će biti pomenuta je multinomijalna logistička regresija, odnosno ona koja se koristi kada imamo više od dva ishoda. Modeluje verovatnoću da zavisna promenljiva pripada jednoj od nekoliko kategorija. Izlaz su verovatnoće za svaku kategoriju. Logistička regresija koristi logističku funkciju (log-odds) da poveže linearnu kombinaciju ulaza sa verovatnoćom ishoda. Logistička funkcija je log(p/(1-p)), gde je p verovatnoća

događaja. Jedna od kategorija se izabere kao referentna i sve ostale kategorije se porede sa njom. Koeficijenti se procenjuju u odnosu na referentnu kategoriju. Naravno, u našem slučaju, sa više mogućih ishoda, postoji više logističkih funkcija, po jedna za svaku kategoriju (osim referentne kategorije). Formula za kategoriju *k* je:

$$\log\left(\frac{p_k}{p_{ref}}\right) = \beta_{k0} + \beta_{k1}x_1 + \beta_{k2}x_2 + \ldots + \beta_{kn}x_n$$

Gde su $\beta 0$ i βn koeficijenti modela, a xn ulazi, *pk* verovatnoća da ishod pripada kategoriji *k*, a *pref* je verovatnoća referentne kategorije. Koeficijenti modela se obično procenjuju metodom maksimalne verodostojnosti. Koeficijenti (β) u logističkoj regresiji se interpretiraju kao promena u logističke funkcije za prelazak iz referentne kategorije u datu kategoriju za jedinicu promene u ulazu.

Metrika	Vrednost (model bez klasnih težina) (%)	Vrednost (model sa klasnim težinama) (%)
Tačnost	95.72	95.49
Preciznost	95.73	95.78
Odziv	95.72	95.49
F1 mera	95.72	95.58

Tabela 11: Rezultati klasifikacije za BS1 datoteku koristeći logističku regresiju

	B	D	M	NK	T
B	1142	0	0	0	0
D	0	99	2	0	0
M	0	0	1165	0	0
NK	0	0	0	993	9
T	0	0	0	5	5813

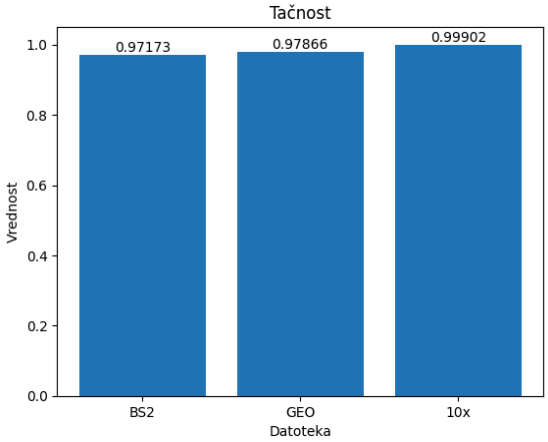
	B	D	M	NK	T
B	510	1	1	0	6
D	0	37	4	0	0
M	1	2	488	0	5
NK	1	0	0	319	72
T	1	0	1	74	2432

Tabela 12: Matrice konfuzije (model bez klasnih težina) za trening (levo) i test (desno) pri upotrebi BS1

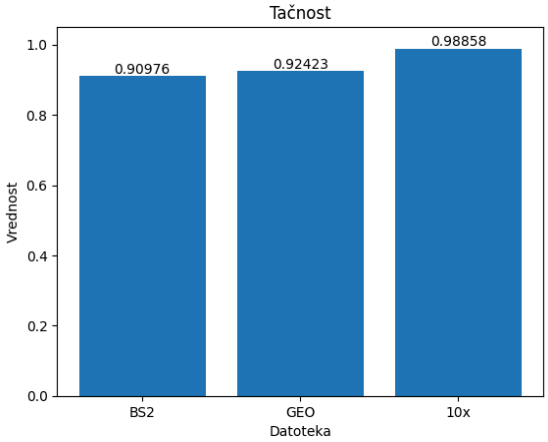
	B	D	M	NK	T
B	1142	0	0	0	0
D	0	101	0	0	0
M	0	0	1165	0	0
NK	0	0	0	1002	0
T	0	0	0	168	5650

	B	D	M	NK	T
B	515	1	1	0	1
D	0	37	4	0	0
M	1	1	494	0	0
NK	1	0	0	363	28
T	2	0	1	125	2380

Tabela 13: Matrice konfuzije (model sa klasnim težinama) za trening (levo) i test (desno) pri upotrebi BS1



Slika 7: Rezultati klasifikacije za svaku od datoteka



Slika 8: Rezultati primene BS1 modela na svaku od datoteka

	B	D	M	NK	T
B	1855	0	1	6	15
D	88	9	167	0	6
M	0	0	2006	0	0
NK	0	0	0	796	46
T	1	0	7	760	6383

Tabela 14: Matrica konfuzije pri primeni modela na BS2

	B	D	M	NK	T
B	1286	0	106	83	281
D	0	0	0	0	0
M	1	0	841	3	11
NK	0	0	0	302	7
T	58	0	110	583	12733

Tabela 15: Matrice konfuzije pri primeni modela na GEO

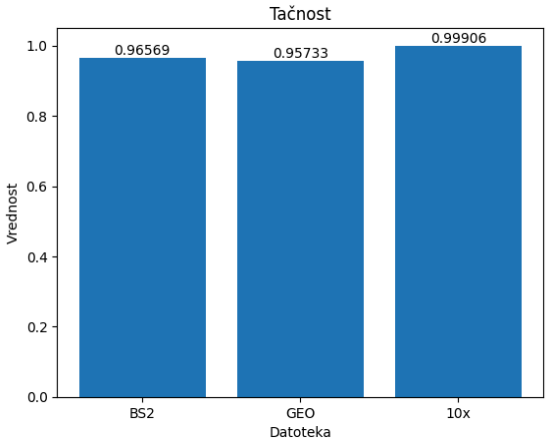
	B	D	M	NK	T
B	9442	0	0	0	283
D	0	0	0	0	0
M	134	0	1683	3	28
NK	0	0	1	8031	147
T	28	0	0	317	62331

Tabela 16: Matrice konfuzije pri primeni modela na 10x

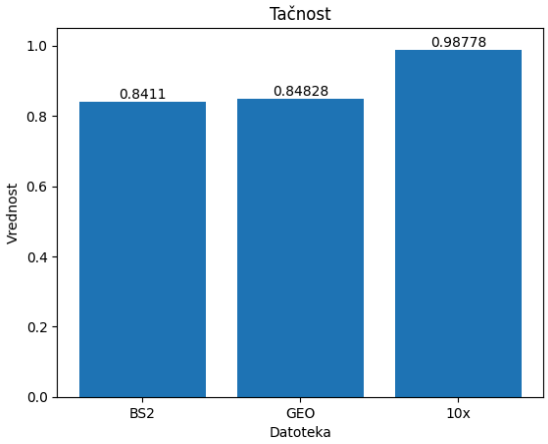
Ovaj model nam pruža i mogućnost da dodelimo težine klasama kako bi se borili sa problemom nebalansiranih klasa. Kao i do sad, unakrsnom proverom pronađeni su najbolji parametri koje smo kasnije koristili da istreniramo model, ali sada smo ih testirali na slučajevima sa i bez klasnih težina, i dobili da je ista kombinacija parametra najbolja na oba slučaja:

- `'C': 0.001,`
- `'penalty': 'l2',`
- `'solver': 'sag',`
- `'class_weight': 'balanced'`

Rezultati dobijeni ovim algoritmom su generalno dobri, bolji nego kod prethodnih algoritama, u oba slučaja. Kada dodamo težine klasama, u većini slučajeva, vrednosti metrika jako blago opadnu, ali dobijemo mnogo bolja predviđanja retkih klasa (D i NK). Preciznost kod većih klasa uglavnom opadne, ali u većini slučajeva je i dalje dobra. Takođe, modeli pravljeni na drugim datotekama, a pogtovo na BS2 i GEO, i testirani na istim, daju dosta bolje rezultate nego oni trenirani na BS1, što nam, zajedno sa rezultatima prethodinh algotritama, dalje ukazuje na različitost podataka u ovim skupovima.



Slika 9: Rezultati klasifikacije za svaku od datoteka



Slika 10: Rezultati primene BS1 modela na svaku od datoteka

	B	D	M	NK	T
B	1862	0	3	9	3
D	84	62	119	0	5
M	0	0	2006	0	0
NK	1	0	0	827	14
T	4	0	11	1801	5335

Tabela 17: Matrica konfuzije pri primeni modela na BS2

	B	D	M	NK	T
B	1352	0	112	83	209
D	0	0	0	0	0
M	1	0	843	5	7
NK	0	0	0	309	0
T	126	0	407	1832	11119

Tabela 18: Matrice konfuzije pri primeni modela na GEO

	B	D	M	NK	T
B	9719	0	3	0	3
D	0	0	0	0	0
M	80	42	1700	6	20
NK	0	0	2	8112	65
T	33	5	8	759	61871

Tabela 19: Matrice konfuzije pri primeni modela na 10x

4.4 NEURONSKE MREŽE

Neuronske mreže su klasa modela mašinskog učenja inspirisanih strukturom i funkcijom ljudskog mozga. Sastoje se od slojeva međusobno povezanih čvorova, ili neurona, koji obrađuju ulazne podatke da bi prepoznali obrasce i napravili predviđanja. Svaka veza ima pridruženu težinu, koja se prilagođava tokom treninga kako bi se minimizirala greška. One su posebno efikasne za zadatke koji uključuju složene i nelinearne odnose, kao što je prepoznavanje slike i govora.
Za potrebe našeg zadatka mi ćemo koristiti višeslojni perceptron (en. Multi-Layer Perceptron, MLP), vrstu neuronske mreže koja uključuje jedan ili više skrivenih slojeva između ulaznog i izlaznog sloja. On koristi propagaciju unazad za prilagođavanje težina kroz gradijentni spust, omogućavajući modelu da uči iz označenih podataka za trening i poboljša svoju tačnost u predviđanju klasa za nove podatke.

Kao i do sad, unakrsnom proverom pronađeni su najbolji parametri koje smo kasnije koristili da istreniramo model, koristili smo *neural_network.ipynb* iz direktorijuma *models* i kao dobijene parametre imamo:

- `'alpha': 1e-05,`
- `'batch_size': 32,`
- `'hidden_layer_sizes': (50,),`
- `'learning_rate_init': 0.001,`
- `'max_iter': 1000`

I u ovom slučaju dobili smo pozitivne rezultate za BS1 (tabela 20). Takođe, i modeli trenirani za preostale tri datoteke deluju jako dobro (slika 11), u nivou sa ostalim algoritmima koje smo koristili. Ono gde se neuronske mreže pokazuju malo bolje od drugih jeste primena BS1 modela na svaku od datoteka (slika 12). Tačnost u svim slučajevima prelazi 92%, i odlično predviđa sve klase osim klase D koja nam i dalje predstavlja problem.

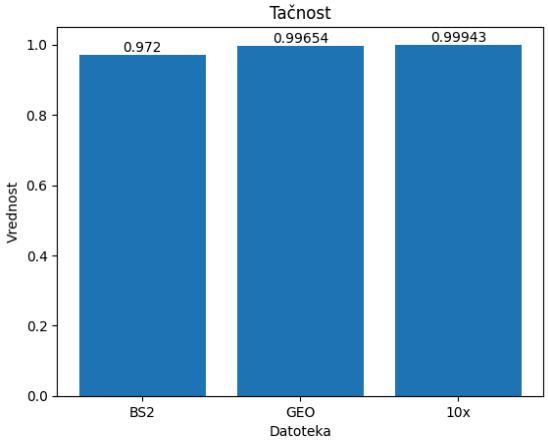
Metrika	Vrednost (%)
Tačnost	95.80
Preciznost	95.87
Odziv	95.80
F1 mera	95.83

Tabela 20: Rezultati klasifikacije za BS1 datoteku koristeći neuronske mreže

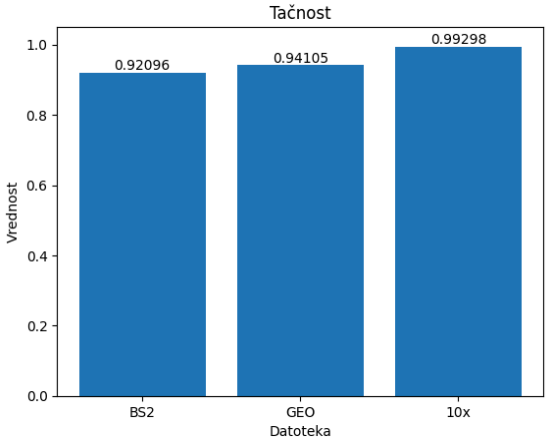
	B	D	M	NK	T
B	1142	0	0	0	0
D	0	101	0	0	0
M	0	0	1165	0	0
NK	0	0	0	1002	0
T	0	0	0	0	5818

	B	D	M	NK	T
B	511	1	1	0	5
D	0	39	2	0	0
M	1	2	490	0	3
NK	1	0	0	321	70
T	1	0	1	81	2425

Tabela 21: Matrice konfuzije za trening (levo) i test (desno) pri upotrebi BS1



Slika 11: Rezultati klasifikacije za svaku od datoteka



Slika 12: Rezultati primene BS1 modela na svaku od datoteka

	B	D	M	NK	T
B	1861	0	1	3	12
D	79	15	172	0	4
M	0	0	2006	0	0
NK	0	0	0	763	79
T	4	0	7	599	6541

Tabela 22: Matrica konfuzije pri primeni modela na BS2

	B	D	M	NK	T
B	1346	0	109	85	216
D	0	0	0	0	0
M	2	0	840	1	13
NK	0	0	0	282	27
T	108	0	114	292	12970

Tabela 23: Matrice konfuzije pri primeni modela na GEO

	B	D	M	NK	T
B	9651	0	0	0	74
D	0	0	0	0	0
M	93	2	1726	2	25
NK	0	0	1	8005	173
T	31	0	0	177	62468

Tabela 24: Matrice konfuzije pri primeni modela na 10x

4.5 POREĐENJE MODELA

Sada kada imamo rezultate klasifikacije za svaki od algoritama, možemo da analiziramo kvalitete modela. Metrike koje koristimo za procene kvaliteta su prvenstveno **tačnost** (en. accuracy), a pored toga ćemo pratiti i **preciznost** (en. precision), **odziv** (en. recall) i **f1 meru** (en. f1 score). Zbog toga što su naši podaci

neuravnoteženi bitno je posmatrati i preostale tri metrike, jer gledanje samo tačnosti u ovakvim slučajevima može da bude obmanjujuće.

U tabelama 25, 26, 27 i 28 možemo da vidimo dobijene rezultate, računate za modele trenirane na BS1 skupu podataka, a primenjene za testiranje na svim datotekama. Primećujemo da ipak nema neke značajne razlike u rezultatima između metrika. Iz tog razloga nema potrebe analizirati svaku za sebe već možemo uopštiti i usresrediti se na posmatranje samo jedne metrike, na primer tačnost. Rezultate tačnosti čitamo sa slike 13.

Model	Tačnost (%)	Preciznost (%)	Odziv (%)	F1 mera (%)
RF	96.18	96.11	96.18	96.14
XGB	96.48	96.58	96.48	96.52
LR	95.72	95.73	95.72	95.73
MLP	95.80	95.87	95.80	95.83

Tabela 25: Upoređivanje modela za podatke iz BS1

Model	Tačnost (%)	Preciznost (%)	Odziv (%)	F1 mera (%)
RF	83.64	84.07	83.64	80.20
XGB	77.73	80.73	77.73	72.73
LR	90.97	93.95	90.97	90.83
MLP	92.09	94.08	92.09	91.75

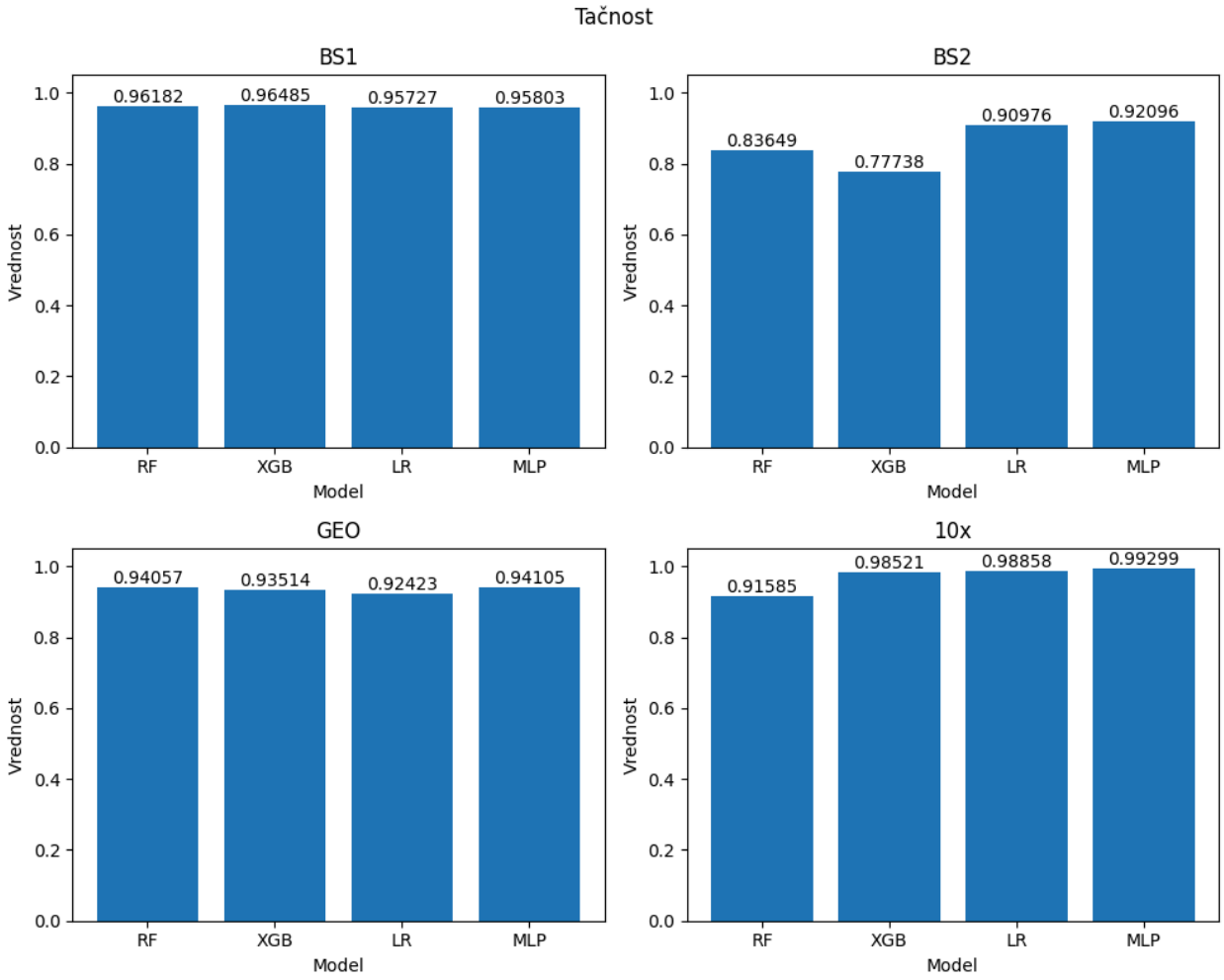
Tabela 26: Upoređivanje modela za podatke iz BS2

Model	Tačnost (%)	Preciznost (%)	Odziv (%)	F1 mera (%)
RF	94.05	93.75	94.05	93.75
XGB	93.51	94.90	93.51	93.96
LR	92.42	95.27	92.42	93.29
MLP	94.10	95.42	94.10	94.46

Tabela 27: Upoređivanje modela za podatke iz GEO

Model	Tačnost (%)	Preciznost (%)	Odziv (%)	F1 mera (%)
RF	91.58	92.38	91.58	89.59
XGB	98.52	98.65	98.52	98.57
LR	98.85	98.86	98.85	98.85
MLP	99.29	99.30	99.29	99.28

Tabela 28: Upoređivanje modela za podatke iz 10x

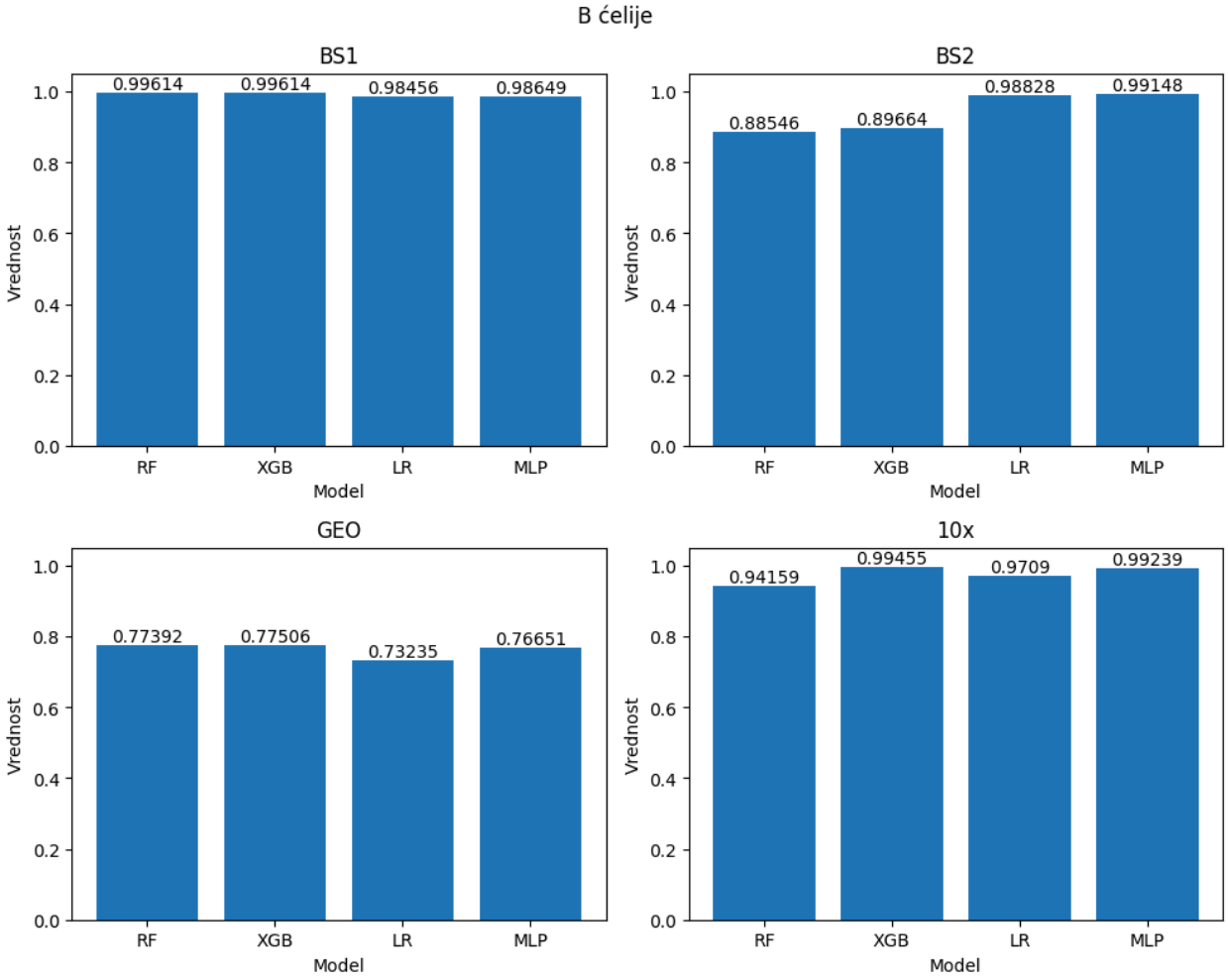


Slika 13: Tačnost istreniranih modela

S obzirom da je naš glavni zadatak bio da napravimo predviđač za BS1, možemo biti zadovoljni dobijenim vrednostima, jer su veoma visoke za svaki model i za svaku metriku. Kao dodatni test kvaliteta, pogledaćemo kako su naši modeli predviđali podatke iz BS2, GEO i 10x. Datoteka BS2 se pokazala kao najteža za predvideti, dok se

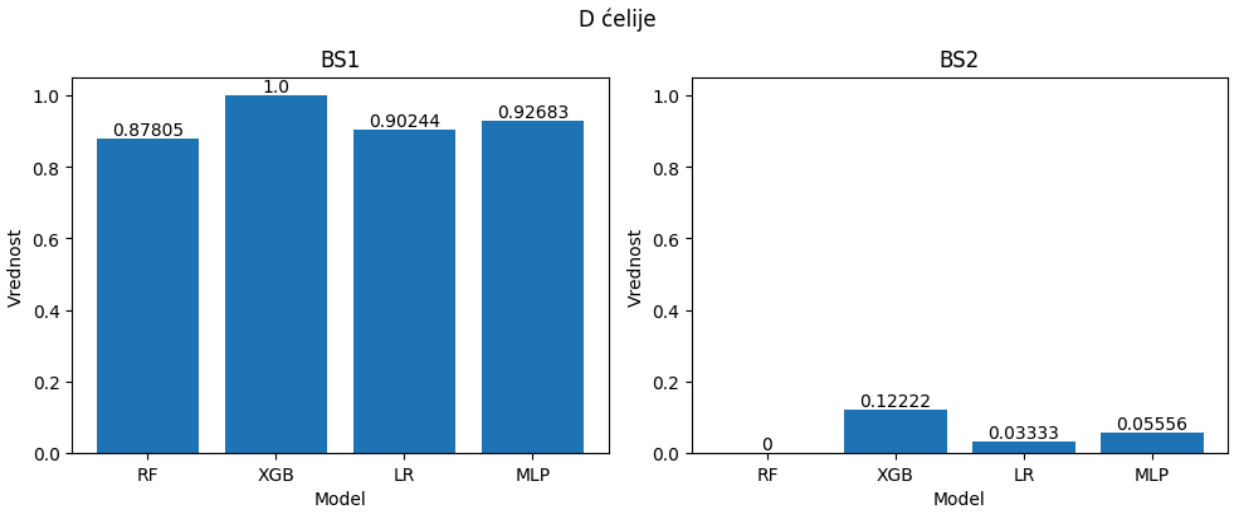
10x ispostavila kao najlakša, i pored toga je najveća. Drvoliki algoritmi poput slučajnih šuma i XGBoost-a deluju najlošije od isprobanih, dok se neuronske mreže sa višeslojnim perceptronom i logistička regresija izdvajaju kao najbolji, sa blagom prednošću na strani neuronskih mreža.

Pored analize kvaliteta celokupnih modela, pogledaćemo i kako se ponašaju kod predviđanja svakog tipa ćelija posebno. Za svaku klasu ćemo izračunati procenat tačno predviđenih instanci. Bitno je naglasiti da skupovi podataka GEO i 10x ne sadrže nijednu instancu koja pripada klasi D, te ćemo zbog toga zanemariti taj slučaj.



Slika 14: Tačnost kod predviđanja B ćelija

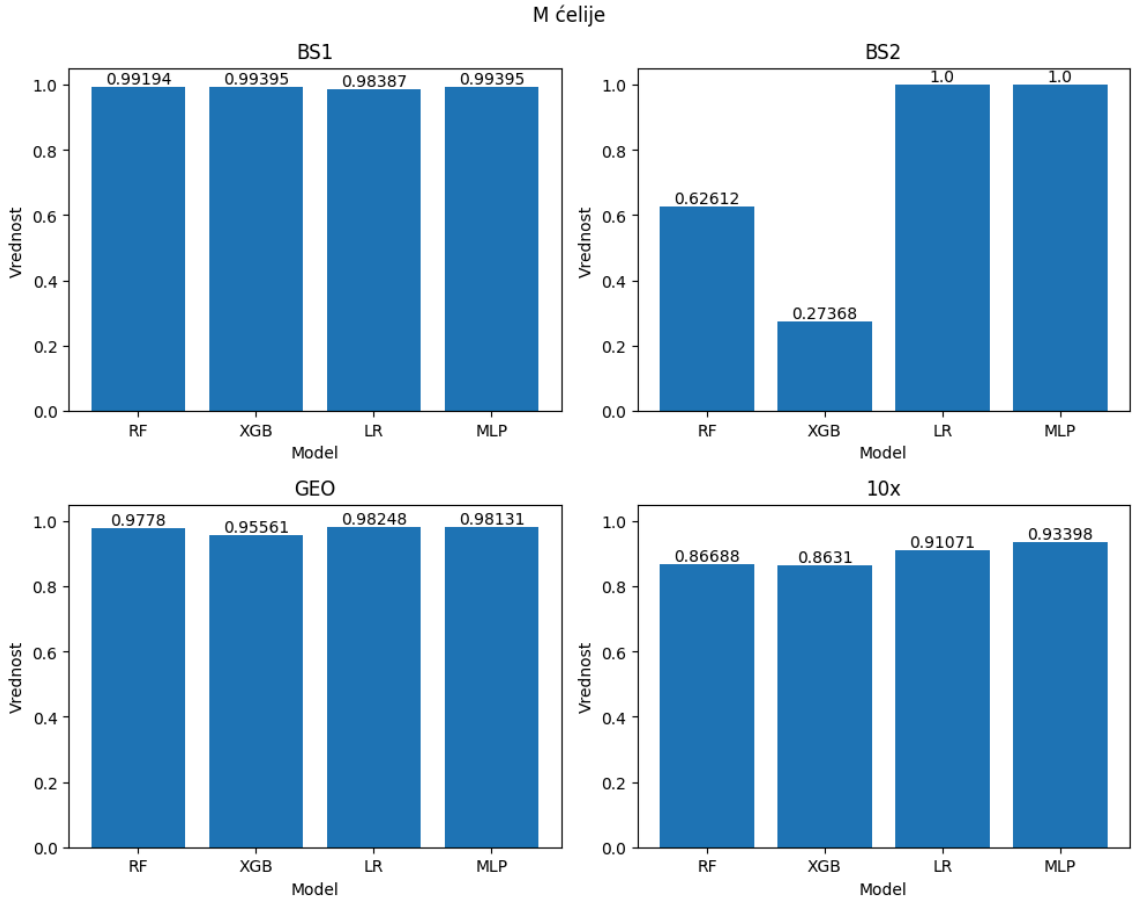
Slika 14 nam opisuje situaciju sa B ćelijama. Vidimo da za BS1 i 10x imamo odlične rezultate, za BS2 nam drvoliki algoritmi ponovo zaostaju sa rezultatima u odnosu na preostale, dok se kod GEO javlja odudaranje. Procenat tačnih predviđanja je dosta niži u odnosu na ostale skupove podataka.



Slika 15: Tačnost kod predviđanja D ćelija

Izazov u radu sa D ćelijama je pre svega veoma mali broj instanci u skupu za treniranje. Za klasifikaciju nad BS1 to nije predstavljalo veliki problem, kao što vidimo na slici 15, međutim kad pogledamo rezultate za BS2 primećujemo da naši modeli skoro pa i ne pogađaju ovu klasu. Ovo se ne vidi u celokupnim ocenama modela, baš zbog tog malog udela ove klase u skupu.

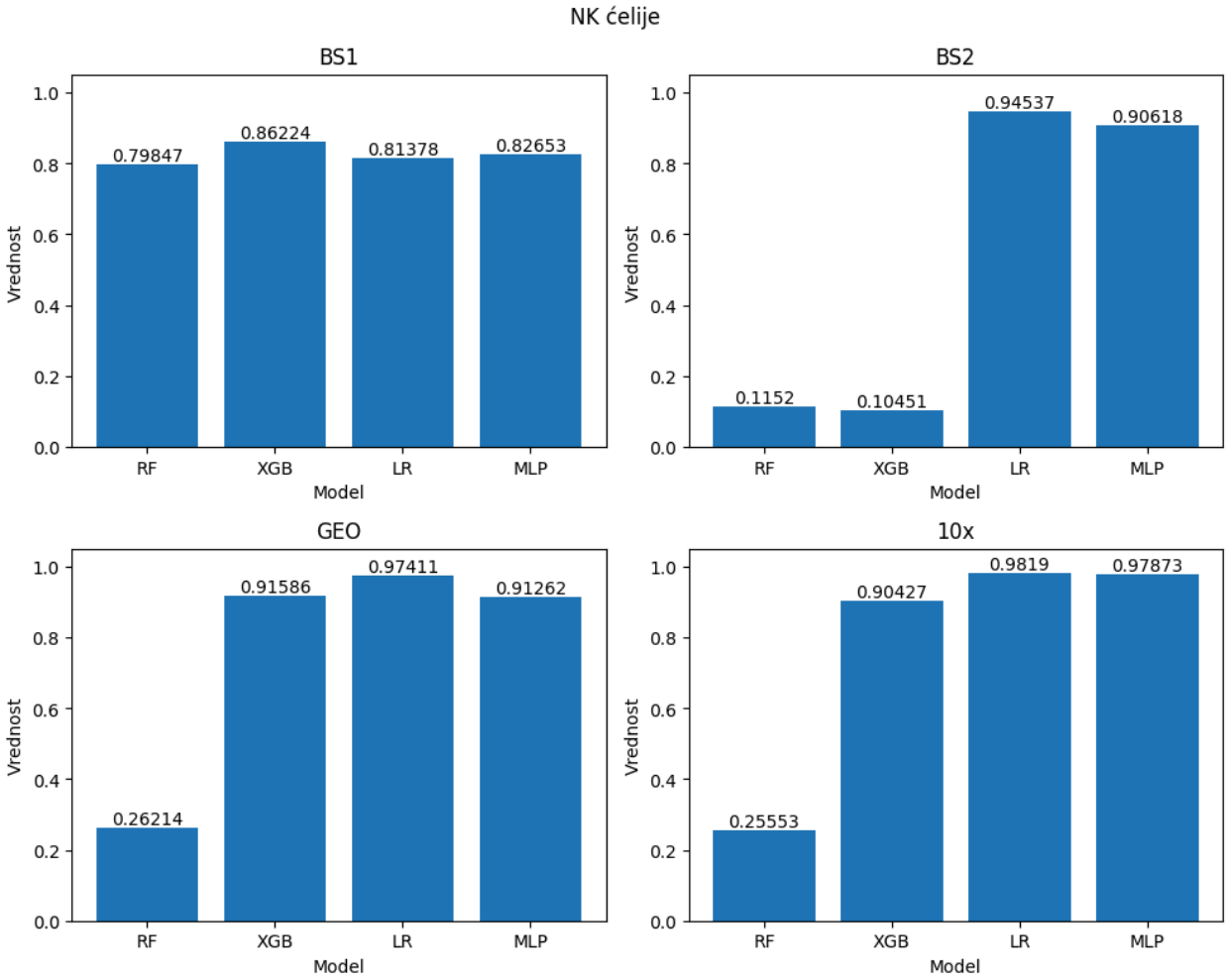
Rezultate za M ćelije možemo pronaći na slici 16 Ovde nam je ponovo najzanimljiviji skup BS2, kod preostalih imamo dobre rezultate jedino malo slabije za 10x. Uočavamo ogromne razlike kod modela slučajnih šuma i XGBoosta-a, u odnosu na modele logističke regresije i neuronskih mreža. Sa jedne strane imamo osrednje, odnosno dosta loše rezultate, dok sa druge strane imamo savršena predviđanja. Ovo nam je samo još jedan pokazatelj koliko su, za naš problem, drvoliki algoritmi lošiji od drugih.



Slika 16: Tačnost kod predviđanja M ćelija

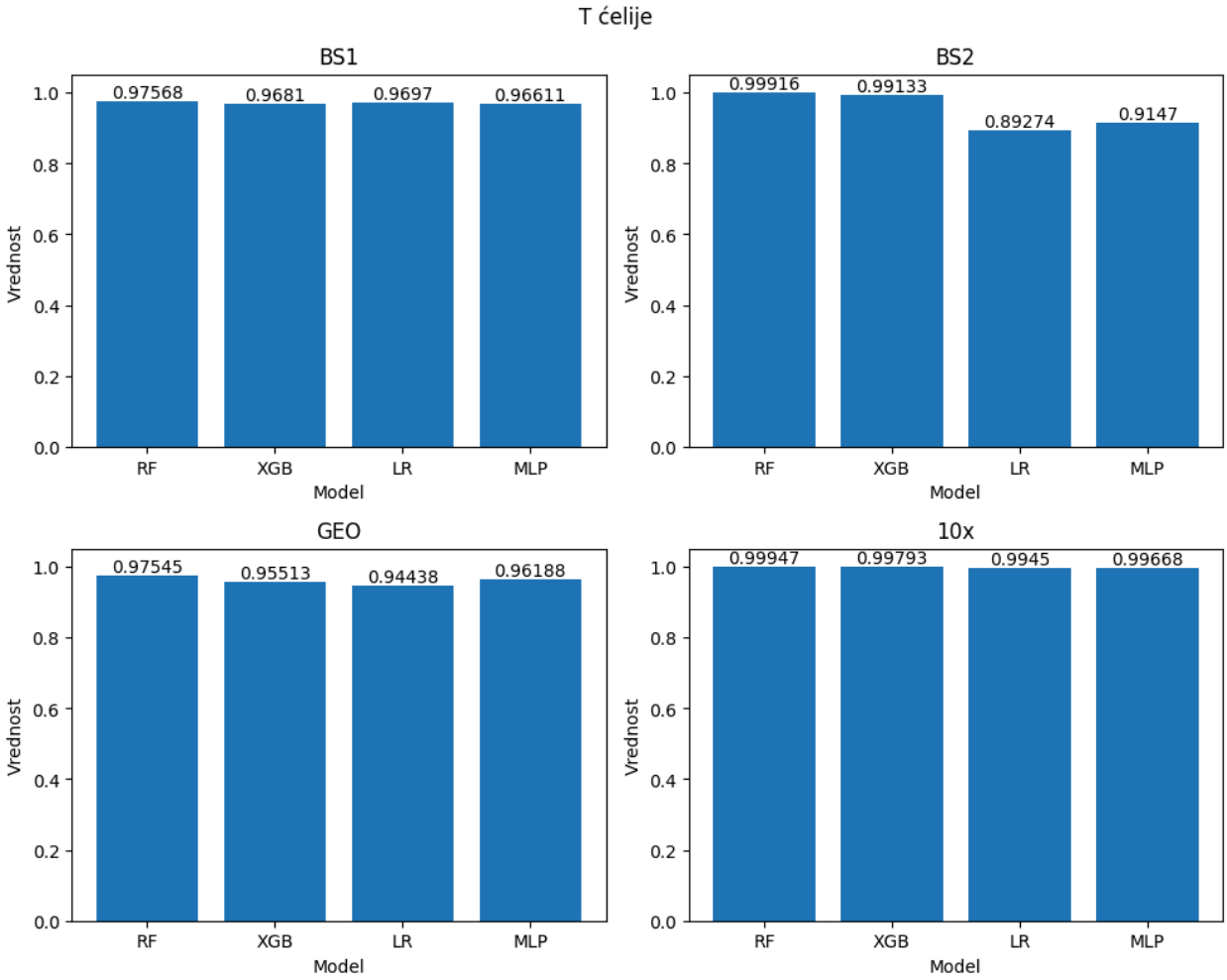
.

Za NK ćelije dobijamo malo lošije rezultate, u poređenju sa standardom koji smo postavili do sad. Ova klasa je druga najmanje zastupljena, posle D, pa je to jedan od razloga zbog čega su rezultati takvi kakvi su. Na slici 17 imamo priliku da primetimo neke već poznate probleme (slabije vrednosti za drvolike algoritme, u ovom slučaju posebno loše za slučajne šume, neuravnotežen BS2...). Ono što možemo posebno da naglasimo jeste najmanji procenat tačno pogođenih instanci u BS1, u poređenju sa ostalim klasama.



Slika 17: Tačnost kod predviđanja NK ćelija

I za kraj nam ostaju još samo T ćelije, ali one su nam možda i najznačajnije zbog najvećeg broja instanci koje im pripadaju. Iz tog razloga možemo biti veoma zadovoljni dobijenim rezultatima sa slike 18. Kroz sve datoteke i sve modele imamo izvanredne rezultate, što nam je samo još jedan pokazatelj da su naši modeli zaista kvalitetni. Ali ovde možemo primetiti još jednu zanimljivost. Naime, kod BS2 skupa vidimo da su slučajne šume i XGBoost sada bolji nego logistička regresija i neuronske mreže, što objašnjava zbog čega su u celokupnoj klasifikaciji veoma slični pri metrikama, iako su kod pojedinačnih klasa zaostajali. To nam i govori da su se ovi modeli možda ipak malo preprilagodili zbog samog broja instanci klase T.



Slika 18: Tačnost kod predviđanja T ćelija

4.6 NADUZORKOVANJE

Iako su navedeni rezultati sasvim prihvatljivi, zbog činjenice da je naš skup podataka nebalansiran, poslužićemo se tehnikom naduzorkovanja, a zatim i poduzorkovanjem (en. oversampling, undersampling). Kao što je ranije napomenuto, pokretanjem *oversampling.ipynb* iz direktorijuma *preprocessing* dobili smo nove skupove podataka za trening (i testiranje) modela za BS1. Ovi modeli su kreirani na iste načine (i u istim sveskama) kao što su i prethodno navedeni modeli, za svaki od algoritama.

Ukoliko bi uporedili rezultate klasifikacije modela treniranih na naduzorkovanim podacima sa modelima treniranim na celokupnim skupovima, na prvi pogled ne primećujemo skoro nikakve razlike. Iz tog razloga ih nećemo tako detaljno analizirati, samo ćemo u tabeli 29 prikazati neposredno poređenje za tačnost predviđanja pre i posle naduzorkovanja.

	BS1		BS2		GEO		10x	
Model	Pre (%)	Posle (%)	Pre (%)	Posle (%)	Pre (%)	Posle (%)	Pre (%)	Posle (%)
RF	96.10	96.16	85.32	86.32	93.92	94.12	90.46	92.75
XGB	96.48	96.41	77.73	85.76	93.51	91.45	98.52	98.90
LR	95.85	96.03	84.10	84.65	84.82	84.60	98.77	98.78
MLP	95.80	95.88	92.09	91.51	94.10	91.68	99.29	98.93

Tabela 29: Poređenje modela treniranih na podacima pre i posle naduzorkovanja

5 ZAKLJUČAK

Konačne rezultate možete pronaći u direktorijumu *results* unutar sveske *results.ipynb*, a izlazne vrednosti modela testiranih nad BS2, GEO i 10x datotekama možete videti u datotekama *results_bs2.csv*, *results_geo.csv* i *results_10x.csv* u istom direktorijumu.

Na osnovu svih izloženih rezultata i statistika, možemo reći da smo uspešno uspeli da napravimo modele klasifikacije koji će predviđati periferne mononuklearne krvne ćelije. Glavni problem koji je bio klasifikacija za skup podataka iz datoteke BS1 je odrađen sa više nego zadovoljavajućim krajnjim ocenama, dok je čak i predviđanje klasa u preostalim datotekama ispalo sasvim korektno. U ponudi imamo 4 modela, od kojih svaki ima svoje prednosti i mane, ali ne bismo pogrešili ukoliko bismo izabrali bilo koji od njih. Algoritmi poput logističke regresije i neuronskih mreža se pokazuju kao najuravnoteženiji, dok slučajne šume i XGBoost deluju kao najbolji izbor ukoliko nam je najbitnije da uspešno predviđamo najveću klasu, tj. T ćelije. Konačan izbor će uvek zavisiti od konkretnog problema na koji ćemo primenjivati ove modele, kao i raznih drugih parametara.

6 REFERENCE

1. https://github.com/bogdans55/ip2-projekat
2. https://en.m.wikipedia.org/wiki/Peripheral_blood_mononuclear_cell
3. https://scikit-learn.org/stable/supervised_learning.html
4. https://xgboost.readthedocs.io/en/stable/
5. https://imbalanced-learn.org/stable/index.html