Dijagnostika bolesti na temelju podataka o ekspresiji gena

Perković, Marija

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1419

DIJAGNOSTIKA BOLESTI NA TEMELJU PODATAKA O EKSPRESIJI GENA

Marija Perković

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1419

DIJAGNOSTIKA BOLESTI NA TEMELJU PODATAKA O EKSPRESIJI GENA

Marija Perković

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

ZAVRŠNI ZADATAK br. 1419

Pristupnica:	**Marija Perković (0036543892)**
Studij:	Elektrotehnika i informacijska tehnologija i Računarstvo
Modul:	Računarstvo
Mentor:	doc. dr. sc. Krešimir Križanović

Zadatak: **Dijagnostika bolesti na temelju podataka o ekspresiji gena**

Opis zadatka:

Istraživanje još iz 1999. godine pokazalo je da je bolesti poput raka moguće klasificirati na temelju podataka o ekspresiji gena. Pristupi temeljeni na tom istraživanju koriste se i danas. Potrebno je preuzeti podatke iz navedenog istraživanja (Golub et al. 1999.) te dodatne skupove podataka CuMiDa i METABRIC (internetska stranica Kaggle). Podatke je potrebno analizirati tehnikama nadziranog strojnog učenja kao što je logistička regresija i stroj potpornih vektora. Inicijalne podatke potrebno je i grupirati nekom od tehnika nenadziranog učenja. Dodatno, podatke je potrebno analizirati jednostavnim modelom dubokog učenja s dva potpuno povezana sloja. Usporediti rezultate dobivene različitim tehnikama. Rješenje mora biti napisano koristeći programski jezik Python i okruženje Jupyter notebook. Koristiti programske biblioteke sklearn i keras. Kako bi se osigurala ponovljivost eksperimenta, napisati upute za instalaciju i izvođenje.

Rok za predaju rada: 14. lipnja 2024.

Zahvaljujem svojem mentoru doc. dr. sc. Krešimiru Križanoviću na korisnim savjetima i stručnom vođenju tijekom izrade ovog rada. Također, zahvaljujem se svojoj obitelji i prijateljima na pruženoj podršci i razumijevanju tijekom studiranja.

Sadržaj

1. Strojno učenje 1
1.1. Logistička regresija 3
1.2. Stroj potpornih vektora 3
1.3. Duboko učenje 3
1.4. Grupiranje 4
1.5. Random Forest 4
2. Strojno učenje u medicini 5
3. Klasifikacija leukemije 6
3.1. Priprema podataka 6
3.2. Analiza algoritama i rezultati 8
3.3. Usporedba rezultata 15
4. Klasifikacija raka dojke 17
4.1. Priprema podataka 17
4.2. Analiza algoritama i rezultati 18
4.3. Usporedba rezultata 23
5. Klasifikacija preživljavanja raka dojke 25
5.1. Priprema podataka 25
5.2. Analiza algoritama i rezultati 29
5.3. Usporedba rezultata 33
6. Zaključak 35
Literatura 36

Uvod

U posljednjim desetljećima, razvoj tehnologija analize ekspresije gena unaprijedio je naše razumijevanje bioloških procesa, omogućujući detaljnu analizu molekularnih uzroka bolesti poput karcinoma. Ekspresija gena, složen proces pretvaranja genetskog koda u funkcionalne proteine, ključna je za regulaciju staničnih procesa i održavanje homeostaze u organizmu.

Jedno od ključnih područja primjene analize ekspresije gena je u klasifikaciji i dijagnostici karcinoma, uključujući leukemiju i karcinom dojke. Kroz analizu uzoraka ekspresije gena, identificirani su specifični genetski uzorci povezani s različitim tipovima tumora, što omogućuje bolje razumijevanje bolesti i individualizirani pristup liječenju.

Istraživanje Goluba i suradnika (1999) predstavilo je revolucionarnu metodu klasifikacije tumora temeljenu na ekspresiji gena, posebno za razlikovanje akutne limfoblastične leukemije (ALL) i akutne mijeloične leukemije (AML). Ovo istraživanje postavilo je temelje za daljnja istraživanja u području onkologije, potičući istraživače diljem svijeta da istraže potencijal analize ekspresije gena u dijagnostici i terapiji karcinoma.

U sklopu ovog istraživanja, koristimo podatke o ekspresiji gena kako bismo razlikovali AML i ALL leukemiju, kao i pet različitih vrsta karcinoma dojke. Analiziramo skup podataka koji sadrži informacije o ekspresiji gena, skup podataka CuMiDa koji pruža informacije o ekspresiji gena pacijenata s različitim vrstama karcinoma dojke te dodatni skup METABRIC koji se koristi u istraživanju raka dojke i sadrži detaljne informacije o kliničkim karakteristikama pacijenata, njihovim genetskim profilima, te podatke o preživljavanju i liječenju.

Koristeći metode strojnog učenja, kao što su logistička regresija, stroj potpornih vektora (*engl. Support vector machine, SVM*), algoritme grupiranja i duboko učenje, cilj nam je razviti modele klasifikacije koji mogu točno klasificirati tipove leukemije i karcinoma dojke. Kroz primjenu naprednih tehnika analize gena i strojnog učenja, nadamo se pružiti korisne uvide za unapređenje kliničke prakse i individualiziranog pristupa liječenju karcinoma.

1. Strojno učenje

Strojno učenje (*engl. Machine learning, ML)* je grana računalne znanosti koja se bavi programiranjem računala kako bi optimizirala određene kriterije uspješnosti na temelju podataka ili prethodnog iskustva. Kroz strojno učenje, modeli se razvijaju i optimiziraju na temelju podataka kako bi predvidjeli svojstva novih, još neviđenih podataka. Početak strojnog učenja se povezuje s radom britanskog matematičara Alana Turinga, čiji su radovi postavili temelje za umjetnu inteligenciju i potaknuli interes za razvoj strojeva koji mogu učiti.

Tijekom godina, istraživači su razvili različite algoritme strojnog učenja, svaki sa svojim prednostima i primjenama. Od ranih modela poput perceptrona, koji je postavio temelje za koncept umjetnih neuronskih mreža, do naprednijih tehnika, razvoj strojnog učenja donio je brojne inovacije i napretke. U 21. stoljeću, razvoj dubokog učenja, podskupa strojnog učenja koji koristi složene neuronske mreže za obradu podataka, označio je revolucionaran napredak u računalnom vidu, obradi prirodnog jezika i drugim područjima.

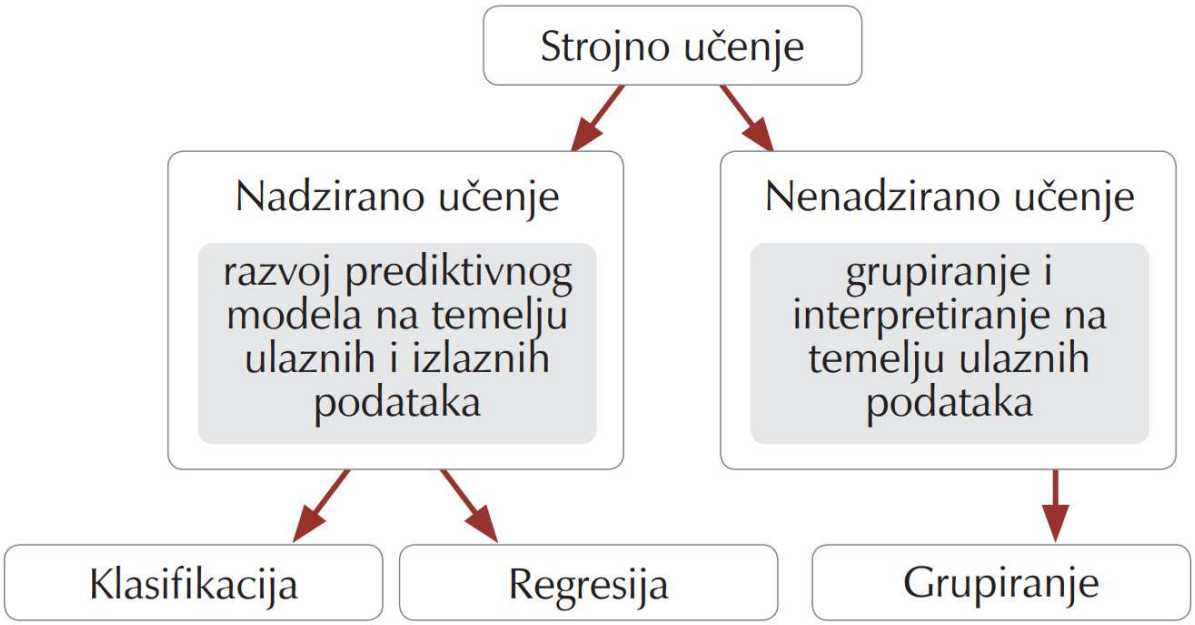
Postoje tri glavne vrste strojnog učenja:

1. Nadzirano učenje: Modeli se treniraju na označenim podacima, što znači da su ulazni podaci povezani s odgovarajućim izlaznim oznakama. Cilj je naučiti model predviđati izlaznu oznaku na temelju ulaznih podataka. Primjeri uključuju logističku regresiju, SVM, i neuronske mreže.
2. Nenadzirano učenje: U ovom slučaju, modeli se treniraju na neoznačenim podacima, što znači da nema odgovarajućih izlaznih oznaka. Cilj je otkriti skrivene strukture ili obrasce u podacima. Primjeri uključuju grupiranje, analizu glavnih komponenti (PCA) i samoorganizirajuće karte.
3. Pojačano učenje (*engl. Reinforcement learning*): Ova tehnika učenja omogućuje agentu (modelu) da uči interaktivno iz svog okruženja putem eksperimentiranja i primanja povratnih informacija. Cilj je maksimizirati neki oblik nagrade prema zadanom cilju.

Strojno učenje ima ključnu ulogu u analizi i obradi velikih skupova podataka. U svijetu gdje se svakodnevno generira ogromna količina podataka, strojno učenje pruža alate i tehnike za ekstrakciju korisnih informacija iz tih podataka. Jedan od glavnih razloga za korištenje strojnog učenja je rješavanje problema koji su previše složeni da bi ih se riješilo tradicionalnim algoritmima. Primjerice, prepoznavanje govora ili obrada slika zahtijeva složene obrasce koji su teško definirani algoritamski, ali se mogu naučiti kroz strojno učenje.

U ovom radu, koristimo strojno učenje kako bismo analizirali i klasificirali skupove podataka ekspresije gena za različite vrste karcinoma, uključujući leukemiju i karcinom dojke.

Kroz primjenu algoritama strojnog učenja, kao što su logistička regresija, SVM, grupiranje i duboko učenje, imamo mogućnost istražiti skrivene obrasce u našim podacima i razviti modele koji mogu točno klasificirati različite tipove karcinoma.



Slika 1.1: Prikaz podjele algoritama strojnog učenja (Portal hrvatskih znanstvenih i stručnih časopisa, 2021)

1.1. Logistička regresija

Logistička regresija je tehnika nadziranog učenja koja se koristi za klasifikaciju i predikciju binarnih ishoda. Kao nadzirani model, logistička regresija koristi označene podatke za učenje, gdje su ulazni podaci (nezavisne varijable) povezani s poznatim ishodima (zavisna varijabla). Cilj je naučiti model koji može točno predvidjeti ishod za nove, nepoznate podatke. Osnovni princip logističke regresije temelji se na korištenju sigmoidne funkcije (logističke funkcije) koja transformira linearne kombinacije ulaznih varijabli u vrijednosti koje predstavljaju vjerojatnosti binarnih ishoda.

1.2. Stroj potpornih vektora

Stroj potpornih vektora je tehnika nadziranog učenja koja se koristi za klasifikaciju i regresiju, s naglaskom na binarne klasifikacijske zadatke. Cilj SVM-a je pronaći hiperravninu koja najbolje razdvaja podatke u dvije klase. SVM koristi funkciju potpore (*engl. support function)* koja identificira podatke koji definiraju marginu. Ovi podaci su nazvani potporni vektori. Oni su ključni za određivanje hiperravnine i definiranje granica između klasa. U slučaju kada podaci nisu linearno razdvojivi u izvornoj dimenziji, SVM koristi funkcije jezgre kako bi preslikao podatke u prostor u kojem su linearno razdvojivi. To omogućuje SVM-u da radi dobro čak i s nelinearnim podacima.

1.3. Duboko učenje

Duboko učenje predstavlja složene matematičke modele inspirirane strukturom ljudskog mozga. Glavna karakteristika dubokog učenja je sposobnost modeliranja apstraktnih i složenih funkcija putem višeslojnih arhitektura neuronskih mreža. Ove arhitekture sastoje se od više slojeva neurona, uključujući ulazni sloj koji prima podatke, skriveni slojevi koji obrađuju te podatke, te izlazni sloj koji generira odgovarajuće izlazne vrijednosti. Učenje dubokih neuronskih mreža odvija se kroz proces optimizacije težina veza između neurona. Ovi algoritmi prilagođavaju težine tako da minimiziraju grešku između stvarnih i predviđenih izlaza mreže.

1.4. Grupiranje

Grupiranje je tehnika u strojnom učenju koja se koristi za segmentaciju skupova podataka u slične grupe ili klastere. Grupiranje ima za cilj otkriti skrivene obrasce i strukture u skupovima podataka, bez potrebe za prethodno definiranim oznakama. Osnovni cilj grupiranja je podijeliti skup podataka u homogene grupe ili klastere, tako da su objekti unutar istog klastera što sličniji, dok su objekti iz različitih klastera što različitiji.

1.5. Random Forest

Random Forest je tehnika u strojnom učenju koja se koristi za klasifikaciju i regresiju. Temelji se na više stabala odlučivanja te kombinirajući njihove predikcije, poboljšava se točnost modela. Osnovni cilj Random Foresta je otkriti kompleksne obrasce i interakcije između značajki podataka, bez potrebe za prethodno definiranim pravilima ili oznakama. Slično grupiranju, Random Forest nastoji grupirati podatke tako da objekti sličnih karakteristika ili značajki budu klasificirani zajedno, što omogućuje pouzdanije predviđanje novih primjera i bolje razumijevanje podataka.

2. Strojno učenje u medicini

Strojno učenje u medicini postaje sve važnije s rastućim potrebama zdravstvenog sektora. S napretkom u tehnologiji medicinskih uređaja i prikupljanju velikih količina podataka, algoritmi strojnog učenja postaju neophodni alati za analizu, interpretaciju i donošenje odluka u medicinskim praksama.

Jedno od ključnih područja primjene strojnog učenja u medicini je dijagnostika bolesti. Algoritmi strojnog učenja mogu analizirati medicinske slike poput rendgenskih snimaka, MRI i CT skenova kako bi identificirali abnormalnosti i s pomoću toga pomogli u ranoj dijagnostici bolesti poput raka, neurodegenerativnih poremećaja i kardiovaskularnih oboljenja. Također, ML modeli mogu analizirati biološke podatke poput genetskih profila kako bi identificirali rizike za određene bolesti i pružili personalizirane preporuke za prevenciju i rano otkrivanje bolesti.

Osim dijagnostike, strojno učenje ima ključnu ulogu i u razvoju novih terapija i lijekova. Analizom velikih skupova podataka o kemijskim spojevima i biološkim procesima, ML algoritmi mogu identificirati potencijalne kandidate za lijekove, predvidjeti njihovu učinkovitost i sigurnost te tako ubrzati proces istraživanja i razvoja novih lijekova. Ova tehnologija omogućuje farmaceutskim tvrtkama da učinkovitije istraže različite kombinacije spojeva i ciljeva terapije, što može dovesti do bržeg otkrivanja novih lijekova i poboljšanja pristupa terapiji za pacijente.

Uzimajući u obzir sve ove faktore, jasno je da će strojno učenje i dalje imati značajnu ulogu u budućnosti medicinske skrbi. Nastavak istraživanja i razvoja u ovom području otvara vrata za inovativne pristupe dijagnostici, liječenju i prevenciji bolesti, pružajući nadu za poboljšanje zdravlja i kvalitete života pacijenata širom svijeta.

3. Klasifikacija leukemije

3.1. Priprema podataka

U ovom dijelu rada koristimo podatke o ekspresiji gena kako bismo razvili model za klasifikaciju leukemije tipa ALL i AML. Skup podataka sadrži podatke za učenje i testiranje te stvarne oznake za sve pacijente. Koristimo Jupyter bilježnicu i biblioteku **Pandas [7]** za obradu podataka. Prvi korak je učitavanje i priprema podataka za daljnju analizu i modeliranje.

Za početak, učitavamo skup podataka za učenje. Podaci se nalaze u datoteci data_set_ALL_AML_train.csv koja sadrži podatke o ekspresiji gena za 38 pacijenata. Učitali smo podatke koristeći pandas biblioteku te prikazali prvih nekoliko redaka kako bismo dobili uvid u strukturu podataka. Podaci sadrže stupce s opisima gena, akcesijskim brojevima, te vrijednostima ekspresije gena i odgovarajućim "call" vrijednostima koje označavaju prisutnost gena (A - Absent, P - Present, M - Marginal).

	Gene Description	Gene Accession Number	1	call	2	call.1	3	call.2	4	call.3	...
0	AFFX-BioB-5_at (endogenous control)	AFFX-BioB-5_at	-214	A	-139	A	-76	A	-135	A	...
1	AFFX-BioB-M_at (endogenous control)	AFFX-BioB-M_at	-153	A	-73	A	-49	A	-114	A	...
2	AFFX-BioB-3_at (endogenous control)	AFFX-BioB-3_at	-58	A	-1	A	-307	A	265	A	...
3	AFFX-BioC-5_at (endogenous control)	AFFX-BioC-5_at	88	A	283	A	309	A	12	A	...
4	AFFX-BioC-3_at (endogenous control)	AFFX-BioC-3_at	-295	A	-264	A	-376	A	-419	A	...

Slika 3.1: Prikaz prvih 5 redova tablice s podatcima odvojenim za učenje

Najprije pripremamo skup podataka s razinama ekspresije gena. Kako bismo pripremili podatke za analizu, prvo ćemo izdvojiti samo stupce koji sadrže vrijednosti ekspresije gena, izostavljajući sve stupce koji sadrže "call" vrijednosti te opise i akcesijske brojeve gena.

	0	1	2	3	4	5	6	7	8	9	...	7119	7120	7121	7122	7123	7124	7125	7126	7127	7128
1	-214	-153	-58	88	-295	-558	199	-176	252	206	...	185	511	-125	389	-37	793	329	36	191	-37
2	-139	-73	-1	283	-264	-400	-330	-168	101	74	...	169	837	-36	442	-17	782	295	11	76	-14
3	-76	-49	-307	309	-376	-650	33	-367	206	-215	...	315	1199	33	168	52	1138	777	41	228	-41
4	-135	-114	265	12	-419	-585	158	-253	49	31	...	240	835	218	174	-110	627	170	-50	126	-91
5	-106	-125	-76	168	-230	-284	4	-122	70	252	...	156	649	57	504	-26	250	314	14	56	-25

Slika 3.2: Prikaz prvih 5 redova tablice s podatcima odvojenim za učenje sa značajkama ekspresije gena

Zatim pripremamo skup podataka s prisutnošću gena te ćemo izdvojit stupce koji sadrže "call" vrijednosti te ćemo zamijeniti tekstualne oznake prisutnosti gena numeričkim vrijednostima: -1 za Absent (A), 0 za Marginal (M) i 1 za Present (P).

	0	1	2	3	4	5	6	7	8	9	...	7119	7120	7121	7122	7123	7124	7125	7126	7127	7128
1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	...	-1	-1	-1	1	-1	-1	-1	-1	-1	-1
2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	...	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	...	1	-1	-1	-1	-1	-1	1	-1	-1	-1
4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	...	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	...	1	-1	-1	1	-1	-1	-1	-1	-1	-1

Slika 4.3: Prikaz prvih 5 redova tablice s podatcima odvojenim za učenje sa značajkama prisutnosti gena

Testni skup podataka pripremamo na isti način kao i skup podataka za učenje. Učitavamo podatke iz datoteke data_set_ALL_AML_independent.csv koja sadrži podatke o ekspresiji gena za 34 pacijenata.

Pripremamo oznake za učenje i testiranje. Datoteka actual.csv sadrži oznake za sve pacijente u istraživanju. Oznake su u obliku tekstualnih vrijednosti koje označavaju tip leukemije: ALL ili AML. Zamijenimo te oznake numeričkim vrijednostima: 0 za ALL i 1 za AML te izdvojimo oznake za učenje i testiranje.

	patient	cancer
0	1	ALL
1	2	ALL
2	3	ALL
3	4	ALL
4	5	ALL

	patient	cancer
0	1	0
1	2	0
2	3	0
3	4	0
4	5	0

Slika 3.4: Prikaz tablica s oznakama klasa

Spajamo skupove podataka za učenje i testiranje te ih nasumično dijelimo u nove skupove za učenje i testiranje koristeći funkciju train_test_split iz biblioteke sklearn. Ovaj korak osigurava da naši skupovi podataka za učenje i testiranje budu uravnoteženi i reprezentativni.

Učitavanjem i transformacijom podataka osigurali smo da su podaci u prikladnom obliku za algoritme strojnog učenja. Sljedeći korak je primjena različitih algoritama za klasifikaciju kako bismo identificirali najbolje pristupe za točnu klasifikaciju tipa leukemije na temelju ekspresije gena .

3.2. Analiza algoritama i rezultati

Najprije, u kodu smo inicijalizirali model logističke regresije koristeći LogisticRegression() funkciju. Ova funkcija stvara objekt koji predstavlja model logističke regresije.

Kod 3.1: Model logističke regresije za klasifikaciju leukemije

```
modelLR = LogisticRegression()
```

Nakon inicijalizacije modela, koristili smo funkciju fit() kako bismo trenirali model na podacima za učenje i odgovarajućim oznakama. U ovom koraku, model se prilagođava podacima kako bi naučio odnos između ulaznih značajki (train_data) i oznaka (train_labels).

Kod 3.2: Funkcija prilagođavanja logističke regresije

```
modelLR.fit(train_data, train_labels)
```

Nakon što je model treniran, koristimo funkciju predict() kako bismo predvidjeli klasne oznake za testne podatke (test_data). Model koristi naučene parametre kako bi generirao predviđanja o kojoj klasi pripada svaki primjer testnog skupa.

Kod 3.3: Funkcija predikcije logističke regresije

```
predicted_labelsLR = modelLR.predict(test_data)
```

Testirali smo logističku regresiju koristeći podatke koji sadrže samo informacije o prisutnosti gena. Ovaj model postigao je visoku točnost od 91.18 % na testnom skupu podataka. Međutim, kada smo testirali model na skupu podataka za učenje, postigli smo savršenu točnost od 100.00%. Ova razlika u točnosti između testnog i skupa za učenje ukazuje na prenaučenost modela, tj. da model previše dobro "pamti" podatke za učenje i može imati problema s generalizacijom na nove podatke.

Zatim smo isprobali logističku regresiju koristeći podatke s razinama ekspresije gena. Ovaj model postigao je točnost od 97.06 % na testnom skupu podataka i 100.00 % na skupu podataka za učenje. Ovi rezultati sugeriraju da model dobro generalizira na nove podatke i da nije prenaučen.

Konačno, testirali smo logističku regresiju koristeći podatke s razinama ekspresije gena na slučajno podijeljenim skupovima za učenje i testiranje koristeći train_test_split funkciju iz sklearn.model_selection biblioteke u omjeru 7:3. Ovaj model postigao je izvrsnu točnost od 100.00 % kako na testnom, tako i na skupu podataka za učenje. To ukazuje na visoku robusnost modela i njegovu sposobnost da precizno klasificira podatke čak i na potpuno novim skupovima podataka.

	Model	Tip podataka	Testiranje na	Točnost
0	Logistička regresija	Prisutnost gena	Testni skup	91.18
1	Logistička regresija	Prisutnost gena	Trening skup	100.00
2	Logistička regresija	Razine ekspresije gena	Testni skup	97.06
3	Logistička regresija	Razine ekspresije gena	Trening skup	100.00
4	Logistička regresija	Razdvojeni podaci - razine ekspresije gena	Testni skup	100.00
5	Logistička regresija	Razdvojeni podaci - razine ekspresije gena	Trening skup	100.00

Slika 3.5: Prikaz rezultata modela logističke regresije dobivenih na različitim skupovima podataka

Zatim smo postupak ponovili ali s metodom stroja potpornih vektora.

Prvo, inicijalizirali smo SVM model koristeći svm.SVC() funkciju. Ova funkcija stvara objekt koji predstavlja SVM model.

Kod 3.4: Model stroja potpornih vektora za klasifikaciju leukemije

```
modelSVM = svm.SVC(kernel='linear')
```

Nakon inicijalizacije modela, koristili smo funkciju fit() kako bismo trenirali model na podacima za učenje i odgovarajućim oznakama.

Kod 3.5: Funkcija prilagođavanja stroja potpornih vektora

```
modelSVM.fit(train_data, train_labels)
```

Nakon što je model treniran, koristili smo funkciju predict() kako bismo predvidjeli klasne oznake za testne podatke (test_data).

Kod 3.6: Funkcija predikcije stroja potpornih vektora

```
predicted_labelsSVM = modelSVM.predict(test_data)
```

Testirali smo SVM koristeći podatke koji sadrže samo informacije o prisutnosti gena. Ovaj model postigao je visoku točnost od 91.18 % na testnom skupu podataka. Međutim, kada smo testirali model na skupu podataka za učenje, postigli smo savršenu točnost od 100.00 %. Ova razlika u točnosti između skupa za testiranje i učenje ukazuje na prenaučenost modela, tj. da model previše dobro "pamti" podatke za učenje i može imati problema s generalizacijom na nove podatke.

Zatim smo isprobali SVM koristeći podatke s razinama ekspresije gena. Ovaj model postigao je točnost od 97.06 % na testnom skupu podataka i 100.00 % na skupu podataka za učenje. Ovi rezultati sugeriraju da model dobro generalizira na nove podatke i da nije prenaučen.

Konačno, testirali smo SVM koristeći podatke s razinama ekspresije gena na slučajno podijeljenim skupovima za učenje i testiranje. Ovaj model postigao je izvrsnu točnost od 100.00 % kako na testnom, tako i na skupu podataka za učenje. To ukazuje na visoku robusnost modela i njegovu sposobnost da precizno klasificira podatke čak i na potpuno novim skupovima podataka.

	Model	Tip podataka	Testiranje na	Točnost
0	SVM	Prisutnost gena	Testni skup	91.18
1	SVM	Prisutnost gena	Trening skup	100.00
2	SVM	Razine ekspresije gena	Testni skup	97.06
3	SVM	Razine ekspresije gena	Trening skup	100.00
4	SVM	Razdvojeni podaci - razine ekspresije gena	Testni skup	100.00
5	SVM	Razdvojeni podaci - razine ekspresije gena	Trening skup	100.00

Slika 3.6: Prikaz rezultata modela SVM dobivenih na različitim skupovima podataka

Sada koristimo duboko učenje. Koristimo biblioteke TensorFlow i Keras za izgradnju neuronske mreže koja će naučiti složene uzorke u podacima i predviđati ciljnu varijablu. Prva funkcija, deep_learning, implementira osnovni model dubokog učenja. Ova funkcija sastoji se od tri sloja neuronske mreže: ulaznog sloja, jednog skrivenog sloja i izlaznog sloja. Svaki sloj sastoji se od neurona s aktivacijskom funkcijom "ReLU" (Rectified Linear Activation) za skriveni sloj i "sigmoid" funkcijom za izlazni sloj. Arhitektura modela se sastoji od sljedećih slojeva:

- Ulazni sloj: broj neurona odgovara broju atributa (značajki) u podacima.
- Prvi skriveni sloj: 64 neurona s aktivacijskom funkcijom "ReLU".
- Drugi skriveni sloj: 16 neurona s aktivacijskom funkcijom "ReLU".
- Izlazni sloj: 1 neuron s aktivacijskom funkcijom "sigmoid", koji daje vjerojatnost klasifikacije.

Model se trenira koristeći binarni gubitak (binary crossentropy loss) kao funkciju gubitka i Adam optimizator za optimizaciju. Konačno, model se evaluira na testnim podacima, pružajući informacije o gubitku i točnosti.

Kod 3.6: Funkcija modela dubokog učenja

```
def deep_learning(train_data, test_data, train_labels, test_labels):
input_layer_neurons = train_data.shape[1]
first_hidden_layer_neurons = 64
    second_hidden_layer_neurons = 16

    model = Sequential()
    model.add(Dense(first_hidden_layer_neurons,
input_dim=input_layer_neurons, activation='relu'))
model.add(Dense(second_hidden_layer_neurons, activation='relu'))
model.add(Dense(1, activation='sigmoid'))

    model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])

    model.fit(train_data, train_labels, epochs=100, batch_size=32, verbose=1)
predicted_labels = model.predict(test_data)
```

U svrhu pouzdanije evaluacije performansi modela dubokog učenja, provodimo eksperiment pokretanjem modela deset puta te izračunavamo aritmetičku sredinu točnosti dobivenih iz tih pokretanja. Ovaj pristup omogućuje nam dobivanje stabilnijih i pouzdanijih rezultata, smanjujući utjecaj slučajnih varijacija i pružajući bolji uvid u prosječne performanse modela. Model testiramo na testnog skupu prisutnosti gena te testnom skupu razine ekspresije gena.

	Eksperiment	Testni skup	Točnost
0	1	Razine ekspresije gena	0.9117
1	2	Razine ekspresije gena	0.9706
2	3	Razine ekspresije gena	0.9876
3	4	Razine ekspresije gena	0.8823
4	5	Razine ekspresije gena	0.8529
5	6	Razine ekspresije gena	0.8567
6	7	Razine ekspresije gena	0.8823
7	8	Razine ekspresije gena	0.9411
8	9	Razine ekspresije gena	0.9452
9	10	Razine ekspresije gena	0.9468
10	Aritmetička sredina	Razine ekspresije gena	0.9177

Slika 3.7: Prikaz rezultata dubokog učenja nakon deset pokretanja na skupu za testiranje razine ekspresije gena

	Eksperiment	Testni skup	Točnost
0	1	Prisutnosti gena	0.79410
1	2	Prisutnosti gena	0.91180
2	3	Prisutnosti gena	0.88230
3	4	Prisutnosti gena	0.85290
4	5	Prisutnosti gena	0.68235
5	6	Prisutnosti gena	0.85290
6	7	Prisutnosti gena	0.88230
7	8	Prisutnosti gena	0.97050
8	9	Prisutnosti gena	0.82350
9	10	Prisutnosti gena	0.91170
10	Aritmetička sredina	Prisutnosti gena	0.85640

Slika 3.8: Prikaz rezultata dubokog učenja nakon deset pokretanja na skupu za testiranje prisutnosti gena

Sada koristimo algoritam K-sredina (KMeans) za grupiranje podataka koji sadrže informacije o prisutnosti gena. Naš cilj je istražiti sposobnost algoritma da razdvoji pacijente s različitim vrstama leukemije (AML i ALL) koristeći te podatke. Za ovo grupiranje koristimo funkciju clustering, koja provodi postupak grupiranja i procjenjuje točnost dodijeljenih oznaka.

Funkcija clustering prvo definira broj grupa (klastera) na dva, budući da imamo dvije klase (AML i ALL). Zatim stvara KMeans objekt s dva klastera i primjenjuje ga na naše podatke. Nakon primjene, svakoj točki podataka dodijeljena je oznaka klastera. Tada se stvara rječnik koji povezuje oznake klastera s oznakama klasa (AML i ALL). Za svaku klasu, pronalazi se najčešće korištena oznaka klastera. Na temelju toga, stvara se novi skup oznaka. Točnost se zatim izračunava kao omjer broja točnih oznaka klastera i ukupnog broja oznaka.

	Eksperiment	Testni skup	Točnost
0	1	Prisutnosti gena	0.7105
1	2	Prisutnosti gena	0.7368
2	3	Prisutnosti gena	0.7105
3	4	Prisutnosti gena	0.6842
4	5	Prisutnosti gena	0.7631
5	6	Prisutnosti gena	0.6569
6	7	Prisutnosti gena	0.6842
7	8	Prisutnosti gena	0.6578
8	9	Prisutnosti gena	0.7368
9	10	Prisutnosti gena	0.8151
10	Aritmetička sredina	Prisutnosti gena	0.7156

Slika 3.9: Prikaz rezultata grupiranja nakon deset pokretanja na skupu za testiranje prisutnosti gena

	Eksperiment	Testni skup	Točnost
0	1	Razine ekspresije gena	0.8947
1	2	Razine ekspresije gena	0.9473
2	3	Razine ekspresije gena	0.8947
3	4	Razine ekspresije gena	0.8421
4	5	Razine ekspresije gena	0.8947
5	6	Razine ekspresije gena	0.8421
6	7	Razine ekspresije gena	0.8947
7	8	Razine ekspresije gena	0.9211
8	9	Razine ekspresije gena	0.8684
9	10	Razine ekspresije gena	0.7847
10	Aritmetička sredina	Razine ekspresije gena	0.8784

Slika 3.10: Prikaz rezultata grupiranja nakon deset pokretanja na skupu za testiranje prisutnosti gena

3.3. Usporedba rezultata

Analiziramo rezultate dobivene primjenom četiri različite metode za klasifikaciju leukemije.

1. **Logistička regresija i SVM:** Ove klasične metode klasifikacije pružaju visoku interpretabilnost rezultata i brzu implementaciju. Oba modela daju slične rezultate jer ako su podaci linearno odvojivi, oba modela će postaviti sličnu granicu odluke za klasifikaciju. Također rezultati su točniji kada je skup podataka veći.

2. **Duboko učenje:** Duboke neuronske mreže su moćne u ekstrakciji složenih značajki iz podataka kroz više slojeva transformacija. Osim toga, moguće je postići visoke

performanse modela s velikim skupovima podataka i visokom dimenzionalnošću. Rezultati dubokog učenja variraju jer težine u neuronskim mrežama se obično inicijaliziraju nasumično, što može dovesti do različitih početnih stanja i različitih rezultata nakon treniranja.

3. **Grupiranje:** Iako su rezultati grupiranja bili najlošiji u usporedbi s ostalim metodama, grupiranje ima svoje prednosti. Grupiranje omogućuje otkrivanje skrivenih uzoraka i struktura u podacima bez potrebe za označenim primjerima. Također, grupiranje može biti korisno za istraživanje podataka i inicijalno razumijevanje njihove strukture.

4. Klasifikacija raka dojke

4.1. Priprema podataka

Najprije pripremamo podatke za analizu izražaja gena. Koristeći skup podataka "Breast_GSE45827", pripremamo podatke za daljnje analize strojnih učenja. Svaki redak u skupu podataka predstavlja jedan uzorak i sadrži 54675 izražaja gena s klasifikacijom tipa raka dojke. U ovom skupu podataka, tipovi raka dojke su klasificirani u sljedeće kategorije:

1. **Normalno tkivo (normal)**: Ovo je referentna kategorija koja predstavlja zdravo tkivo dojke.

2. **Bazalni rak (basal)**: Ovaj tip raka dojke često se naziva "triple-negative" jer ne izražava receptore estrogena, progesterona i HER2. Obično je agresivniji i ima lošiju prognozu.

3. **HER2 pozitivan (HER)**: Karakteriziran visokim izražajem HER2 proteina, ovaj tip raka dojke može biti agresivan, ali je osjetljiv na terapije koje ciljaju HER2.

4. **Luminalni A (luminal_A)**: Ovaj tip raka dojke obično ima najbolju prognozu i visoko je pozitivan na estrogenske receptore (ER) i/ili progesteronske receptore (PR) te negativan na HER2.

5. **Luminalni B (luminal_B)**: Slično luminalnom A tipu, ali s višim proliferacijskim indeksima i/ili pozitivnim HER2 statusom, što ga čini nešto agresivnijim.

6. **Stanična linija (cell_line)**: Ova kategorija uključuje uzorke iz staničnih linija koje se koriste u istraživačkim laboratorijima za proučavanje raka dojke.

	samples	type	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_at	...
0	84	basal	9.850040	8.097927	6.424728	7.353027	3.029122	6.880079	4.963740	4.408328	...
1	85	basal	9.861357	8.212222	7.062593	7.685578	3.149468	7.542283	5.129607	4.584418	...
2	87	basal	10.103478	8.936137	5.735970	7.687822	3.125931	6.562369	4.813449	4.425195	...
3	90	basal	9.756875	7.357148	6.479183	6.986624	3.181638	7.802344	5.490982	4.567956	...
4	91	basal	9.408330	7.746404	6.693980	7.333426	3.169923	7.610457	5.372469	4.424426	...

Slika 4.1: Prikaz prvih 5 redova tablice sa podacima za klasifikaciju raka dojke

Kako bismo se fokusirali na izražaj gena, izdvajamo samo relevantne kolone, izostavljajući kolone "samples" i "type".

	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_at	1405_i_at	1431_at	...
0	9.850040	8.097927	6.424728	7.353027	3.029122	6.880079	4.963740	4.408328	8.870780	3.278896	...
1	9.861357	8.212222	7.062593	7.685578	3.149468	7.542283	5.129607	4.584418	7.767646	3.589636	...
2	10.103478	8.936137	5.735970	7.687822	3.125931	6.562369	4.813449	4.425195	9.417956	3.552253	...
3	9.756875	7.357148	6.479183	6.986624	3.181638	7.802344	5.490982	4.567956	9.022345	3.312473	...
4	9.408330	7.746404	6.693980	7.333426	3.169923	7.610457	5.372469	4.424426	9.400056	3.368243	...

Slika 4.2: Prikaz prvih 5 redova uređene tablice s podacima za klasifikaciju raka dojke

Kako bismo mogli koristiti oznake klasa u algoritmima strojnih učenja, potrebno ih je transformirati u numeričke vrijednosti. Ovo omogućuje algoritmima da rade s klasama kao numeričkim vrijednostima.

	type
0	1
1	1
2	1
3	1
4	1

Slika 4.3: Prikaz tablice s oznakama klasa

Nakon što su podaci pripremljeni, dijelimo ih na skupove za testiranje i učenje. Skup za učenje podataka koristi se za učenje modela, dok se testni skup koristi za evaluaciju performansi modela.

4.2. Analiza algoritama i rezultati

Najprije ćemo primijeniti logističku regresiju na skup podataka "Breast_GSE45827" kako bismo klasificirali različite tipove raka dojke na temelju izražaja gena. Iako je osnovno korištena za binarne probleme, može se proširiti i na višeklasne probleme klasifikacije, što je slučaj na ovom skupu podataka. Nakon pripreme podataka, provest ćemo eksperimente s različitim veličinama

testnih skupova (30 %, 40 %, 50 %) i analizirati performanse modela.

	Veličina testnog skupa	Točnost na skupu za testiranje	Točnost na skupu za treniranje
0	30%	82.61	100
1	40%	86.89	100
2	50%	89.47	100

Slika 4.4: Prikaz točnosti na različitim skupovima za testiranje i učenje modela logističke regresije za klasifikaciju raka dojke

Nakon primjene logističke regresije, dalje ćemo istražiti stroj potpornih vektora (SVM) u rješavanju istog problema. SVM radi tako da konstruira hiperravnine u višedimenzijskom prostoru koji optimalno razdvajaju instance različitih klasa. U našem kontekstu, svaka instanca predstavlja uzorak izražaja gena, dok različite klase odgovaraju različitim tipovima raka dojke. SVM nastoji pronaći hiperravnine koje jasno razdvajaju te uzorke u prostoru izražaja gena prema njihovim klasama.

	Veličina testnog skupa	Točnost na skupu za testiranje	Točnost na skupu za treniranje
0	30%	84.78	100.0
1	40%	88.52	100.0
2	50%	90.79	100.0

Slika 4.5: Prikaz točnosti na različitim skupovima za testiranje i učenje modela SVM za klasifikaciju raka dojke

Zatim radimo model dubokog učenja. Duboko učenje može biti izuzetno učinkovito za višeklasnu klasifikaciju, uključujući i klasifikaciju tumora u različite kategorije. Duboke neuronske mreže imaju sposobnost modeliranja složenih i apstraktnih značajki iz podataka te otkrivanja dubokih veza između atributa i ciljnih klasa pružajući visoku preciznost i mogućnost otkrivanja suptilnih razlika između različitih tipova tumora.

Definirali smo model s pomoću biblioteke TensorFlow. Naš model ima tri sloja: ulazni sloj, prvi skriveni sloj sa 64 neurona i aktivacijskom funkcijom ReLU te drugi skriveni sloj s 32 neurona i

istom aktivacijskom funkcijom. Izlazni sloj ima 6 neurona jer imamo 6 klasa, a aktivacijska funkcija softmax daje vjerojatnost za svaku klasu.

Pri kompilaciji modela koristili smo funkciju gubitka sparse_categorical_crossentropy, optimizator adam i metriku točnosti. Trenirali smo model na podacima za učenje tijekom 50 epoha. Nakon toga smo evaluirali model na testnim podacima, dobivajući gubitak i točnost.

	Eksperiment	Točnost
0	1	0.8913
1	2	0.9347
2	3	0.8478
3	4	0.9130
4	5	0.7826
5	6	0.8913
6	7	0.8478
7	8	0.7608
8	9	0.8695
9	10	0.9130
10	Aritmetička sredina	0.8651

Slika 4.6: Prikaz rezultata dubokog učenja nakon deset pokretanja na skupu za testiranje

U dubokom učenju, smanjenje dimenzionalnosti značajki može biti korisno kako bi se poboljšao performans, interpretabilnost i efikasnost modela. Koristimo smanjenje značajki kako bismo poboljšali performanse našeg modela dubokog učenja. Velik broj značajki može dovesti do prenaučenosti, usporiti proces učenja i otežati interpretaciju rezultata. Smanjenjem dimenzionalnosti skupa podataka, olakšavamo proces učenja, smanjujemo računalne zahtjeve i poboljšavamo interpretabilnost modela.

Koristit ćemo tehniku analize glavnih komponenti (PCA) kako bismo identificirali najvažnije značajke za naš problem klasifikacije. Ove tehnike omogućuju nam da reduciramo dimenzionalnost skupa podataka tako da zadržimo što više informacija, ali smanjimo šum i nebitne varijable. Nakon smanjenja značajki, planiramo ponovno trenirati naš model dubokog učenja i evaluirati njegove performanse na smanjenom skupu podataka.

	Eksperiment	Točnost
0	1	0.6739
1	2	0.5869
2	3	0.7173
3	4	0.7174
4	5	0.6521
5	6	0.6956
6	7	0.6987
7	8	0.5868
8	9	0.6956
9	10	0.5869
10	Aritmetička sredina	0.6611

Slika 4.6: Prikaz rezultata dubokog učenja s odabirom značajki nakon deset pokretanja na skupu za testiranje

Nakon tog smo radili grupiranje podataka. Koristili smo KMeans algoritam za grupiranje naših podataka u šest klastera, što odgovara broju različitih klasa u skupu podataka: basal, HER, cell_line, luminal_A, luminal_B, i normal. Naš cilj je bio identificirati grupe podataka i vidjeti kako se te grupe poklapaju s unaprijed definiranim klasama.

Definirali smo KMeans objekt sa šest klastera (broj klasa u našem skupu podataka) i pokrenuli algoritam na treniranim podacima.

Kod 4.1: Funkcija KMeans

```
num_clusters = 6
kmeans = KMeans(n_clusters=num_clusters, n_init=10) kmeans.fit(train_data)
```

Nakon treniranja modela, svaki podatkovni primjer bio je dodijeljen jednom od klastera. Stvorili smo rječnik koji mapira svaku klasu na klastere kojima pripadaju podatkovni primjeri te smo identificirali najčešće klastere unutar svake klase. Dodijelili smo nove oznake podacima prema najčešćim klasterima za svaku klasu i izračunali točnost tako što smo usporedili te oznake s originalnim klasterima.

Kod 4.2: Oznake za grupiranje

```
labels = ["basal", "HER", "cell_line", "luminal_A", "luminal_B", "normal"]
groups_dict = {label: list() for label in labels}
```

Kod 4.3.: Ispis grupiranja

```
Clustering with all data:
Patient 1 belongs to cluster 3
Patient 2 belongs to cluster 3
Patient 3 belongs to cluster 3
Patient 4 belongs to cluster 3
Patient 5 belongs to cluster 3
Patient 6 belongs to cluster 3
Patient 7 belongs to cluster 3
Patient 8 belongs to cluster 3
Patient 9 belongs to cluster 3
Patient 10 belongs to cluster 3
Patient 11 belongs to cluster 3
Patient 12 belongs to cluster 4
Patient 13 belongs to cluster 4
Patient 14 belongs to cluster 4
Patient 15 belongs to cluster 4
Patient 16 belongs to cluster 4
Patient 17 belongs to cluster 4
Patient 18 belongs to cluster 4
Patient 19 belongs to cluster 4
Patient 20 belongs to cluster 4
Patient 21 belongs to cluster 4
Patient 22 belongs to cluster 4
Patient 23 belongs to cluster 4
Patient 24 belongs to cluster 4 ...
Patient 150 belongs to cluster 5
Patient 151 belongs to cluster 5
```

	Eksperiment	Točnost
0	1	0.7152
1	2	0.6225
2	3	0.7218
3	4	0.7086
4	5	0.6688
5	6	0.6821
6	7	0.7152
7	8	0.7086
8	9	0.7019
9	10	0.6754
10	Aritmetička sredina	0.6920

Slika 5.6: Prikaz rezultata grupiranja nakon deset pokretanja na skupu za testiranje prisutnosti gena

4.3. Usporedba rezultata

Sljedeći korak u našoj analizi podataka raka dojke jest usporedba performansi modela klasifikacije. Ova usporedba omogućuje nam dublji uvid u primjenjivost svake tehnike strojnog učenja u kontekstu našeg problema. Kroz detaljnu analizu rezultata na testnom skupu podataka, možemo istražiti kako se različiti modeli nose s klasifikacijom različitih klasa raka dojke, što nam pomaže u donošenju odluka o izboru optimalnog modela.

1. Logistička regresija i SVM:

   Prednosti logističke regresije i SVM-a za klasifikaciju raka dojke je to što su relativno jednostavna za implementaciju i interpretaciju. Analizirajući naše rezultate, primijetili smo da oba modela postižu slične rezultate kada se primjenjuju na našim podacima. Također, primijetili smo da oba modela postižu bolje rezultate na manjim skupovima za učenje. Manji skupovi za učenje često sadrže manje raznolike primjere, što može dovesti do stabilnijih rezultata i manje prenaučenosti.

2. Duboko učenje:

   Duboke neuronske mreže su sklone prenaučenosti, što znači da mogu "zapamtiti" uzorke iz skupa za učenje podataka i loše generalizirati na novim, neviđenim podacima. Osim toga, kompleksnost modela može otežati interpretaciju rezultata i otkrivanje uzročno-posljedičnih veza. U našoj analizi, primijetili smo da rezultati dubokog učenja variraju zbog nasumičnosti prisutne u samom procesu treniranja i evaluacije modela. Korištenjem PCA metode za smanjenje dimenzionalnosti podataka, uočili smo da se dobivaju lošiji rezultati. Ovo sugerira da PCA metoda može izgubiti neke relevantne informacije prisutne u originalnom skupu podataka, što može dovesti do gubitka detaljnosti i smanjenja sposobnosti modela da uoči složene uzorke.

3. Grupiranje:

   Analizirajući rezultate korištenja algoritma grupiranja, primijetili smo da smo postigli najlošije rezultate u usporedbi s ostalim modelima. Različiti faktori mogu utjecati na ovakve rezultate, uključujući varijabilnost u rezultatima zbog slučajnosti u njegovom radu. Rezultati sugeriraju da algoritam grupiranja možda nije optimalan izbor za klasifikaciju raka dojke na ovom skupu podataka. Potrebno je daljnje istraživanje i optimizacija kako bi se poboljšala njegova učinkovitost i primjenjivost.

5. Klasifikacija preživljavanja raka dojke

5.1. Priprema podataka

U ovom dijelu rada smo proveli predviđanje preživljavanja raka dojke koristeći modele strojnog učenja s kliničkim podacima i profilima ekspresije gena. Koristili smo podatke "Breast Cancer Gene Expression Profiles (METABRIC)".

Prvih 31 stupaca podataka sadržavalo je kliničke informacije, uključujući status smrti. Ostali stupci sadržavali su informacije vezane uz gene, koje uključuju ekspresiju gena i informacije o mutacijama gena.

Najprije smo razdvojili podatke u tri skupa:

1. **Klinički skup podataka**: Ovaj skup sadrži kliničke informacije o pacijentima, uključujući status smrti, dob pacijenta, tip raka dojke, histološki stupanj neoplazme, status HER2, primijenjena kemoterapija i druge relevantne podatke.
2. **Skup podataka o ekspresiji gena**: Ovaj skup sadrži informacije o ekspresiji gena, koje su važne za analizu veza između gena i preživljavanja pacijenata.
3. **Skup podataka o mutaciji gena**: Iako nećemo koristiti ovaj skup u daljnjim koracima projekta, on sadrži informacije o mutacijama gena koje su zabilježene u skupu podataka.

	patient_id	age_at_diagnosis	type_of_breast_surgery	cancer_type	cancer_type_detailed	cellularity	chemotherapy	pam50_+_claudin-low_subtype	...
0	0.0	75.65	1.0	0.0	1.0	0.0	0.0	6.0	...
1	2.0	43.19	0.0	0.0	0.0	0.0	0.0	2.0	...
2	5.0	48.87	1.0	0.0	1.0	0.0	1.0	3.0	...
3	6.0	47.68	1.0	0.0	1.0	2.0	1.0	3.0	...
4	8.0	76.97	1.0	0.0	1.0	0.0	1.0	3.0	...

Slika 5.1: Prikaz prvih 5 redova kliničkog skupa podataka

	brca1	brca2	palb2	pten	tp53	atm	cdh1	chek2	nbn	nf1	stk11	bard1	mlh1	msh2	msh6	...
0	-1.3990	-0.5738	-1.6217	1.4524	0.3504	1.1517	0.0348	0.1266	-0.8361	-0.8578	-0.4294	-1.1201	-0.4844	-0.7483	-1.6660	...
1	-1.3800	0.2777	-1.2154	0.5296	-0.0136	-0.2659	1.3594	0.7961	0.5419	-2.6059	0.5120	0.4390	1.2266	0.7612	0.1821	...
2	0.0670	-0.8426	0.2114	-0.3326	0.5141	-0.0803	1.1398	0.4187	-0.4030	-1.1305	0.2362	-0.1721	-1.7910	3.0955	0.6608	...
3	0.6744	-0.5428	-1.6592	0.6369	1.6708	-0.8880	1.2491	-1.1889	-0.4174	-0.6165	1.0078	-0.4010	-1.3905	4.8798	0.0615	...
4	1.2932	-0.9039	-0.7219	0.2168	0.3484	0.3897	0.9131	0.9356	0.7675	-0.2940	-0.2961	0.6320	-0.3582	0.3032	0.8747	...

Slika 5.2: Prikaz prvih 5 redova skupa podataka o ekspresiji gena

Najprije smo stvorili dodatan skup podataka koji smo dobili spajanjem skupa s kliničkim podacima i skupa s razinama ekspresije gena s ciljem poboljšanja efikasnosti i dubljeg razumijevanja bioloških mehanizama koji mogu utjecati na preživljavanje pacijenata s rakom dojke.

Kod 5.1.: Spajanje skupa podataka s kliničkim podacima i skupa s razinama ekspresije gena

```
X_overall = np.concatenate((X_clinical, X_gene_expression), axis=1)
```

Zatim smo podijelili svaki skup podataka na skup za učenje i skup za testiranje koristeći funkciju train_test_split iz biblioteke scikit-learn. Ova podjela nam je omogućila da razdvojimo podatke potrebne za učenje modela od podataka potrebnih za evaluaciju modela.

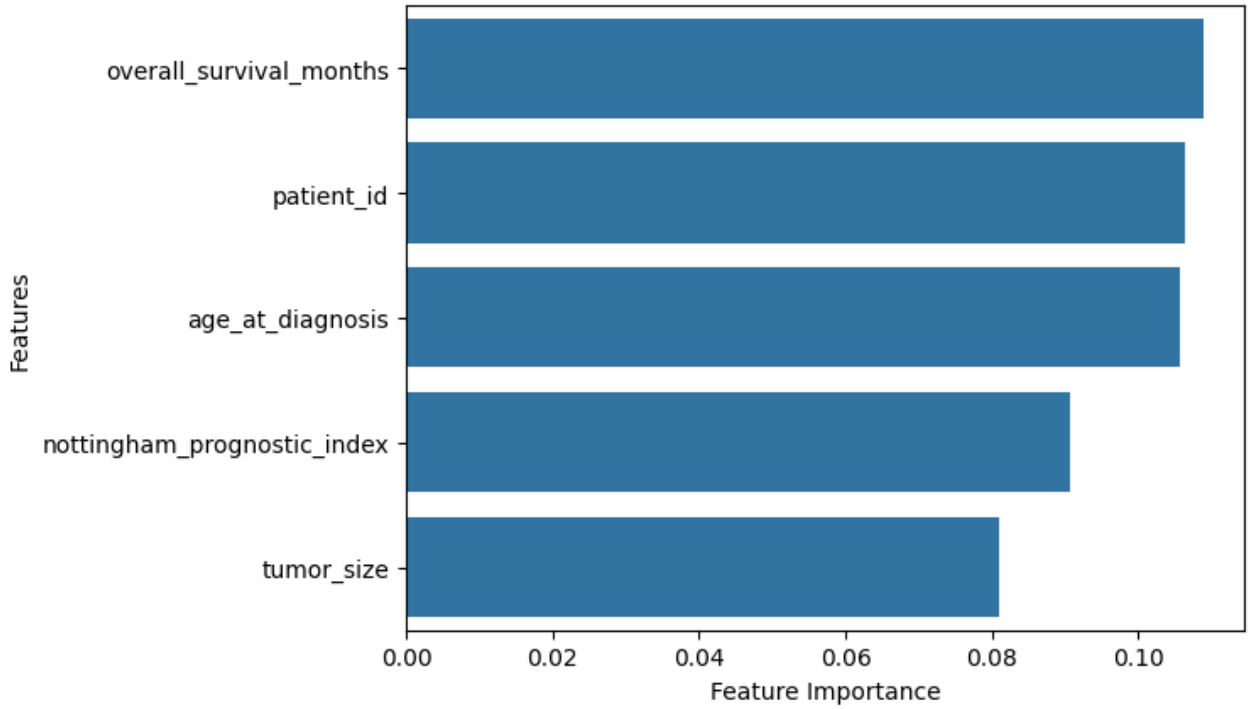
Nakon podjele, stvorili smo tri nova dodatna skupa podataka koje smo dobili reduciranjem dimenzionalnosti tri skupa podataka kako bismo pripremili podatke za daljnju analizu i modeliranje. Koristili smo UMAP (Uniform Manifold Approximation and Projection) kao tehniku za smanjenje dimenzionalnosti. Želimo istražiti kako redukcija skupa podataka utječe na naše mogućnosti predviđanja preživljavanja pacijenata s rakom dojke. Ova analiza pomoći će nam procijeniti koristi ili eventualne nedostatke redukcije dimenzionalnosti u poboljšanju kvalitete i točnosti naših modela.

Kod 5.2.: UMAP reduktor za cjelokupni skup podataka

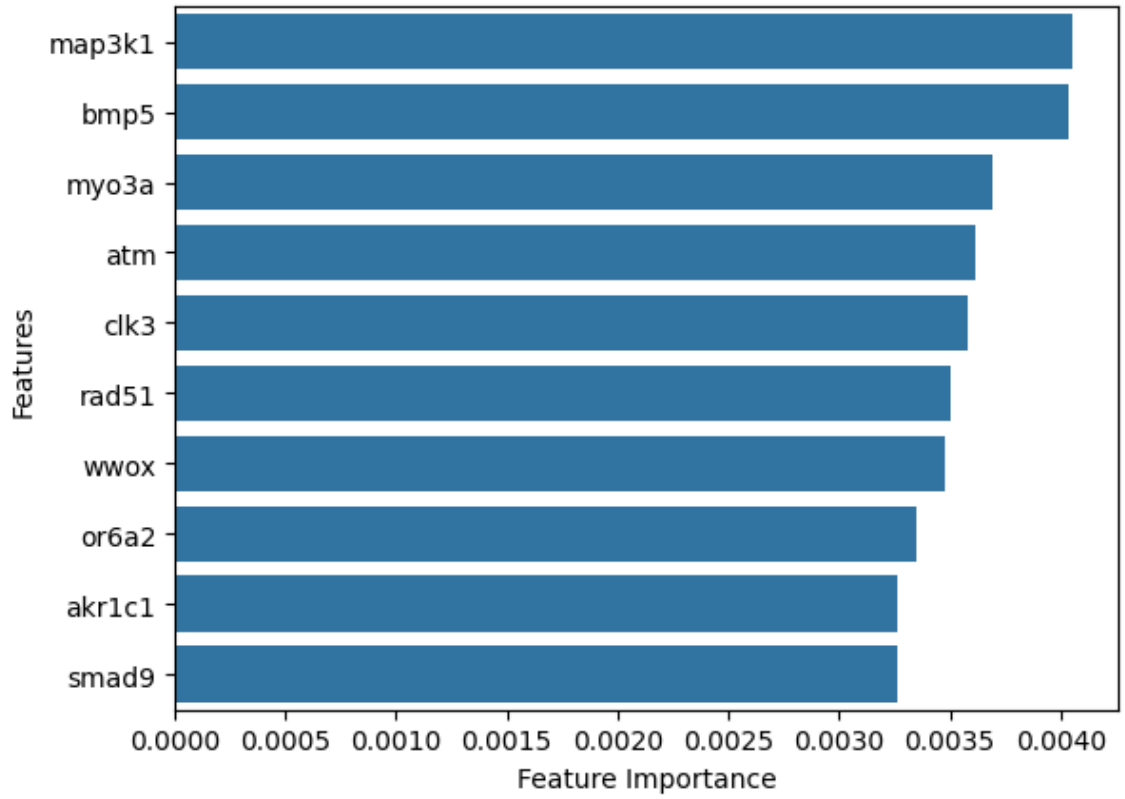
```
reducer_whole = UMAP(n_neighbors=100, n_components=40, n_epochs=1000,
                     min_dist=0.5, local_connectivity=2, random_state=42,)
X_train_reduced = reducer_whole.fit_transform(X_train, y_train_gene_expression)
```

Nakon smanjenja dimenzionalnosti podataka, sljedeći korak bio je identificirati najvažnije značajke svakog skupa podataka. Identifikacija ključnih značajki omogućuje nam bolje

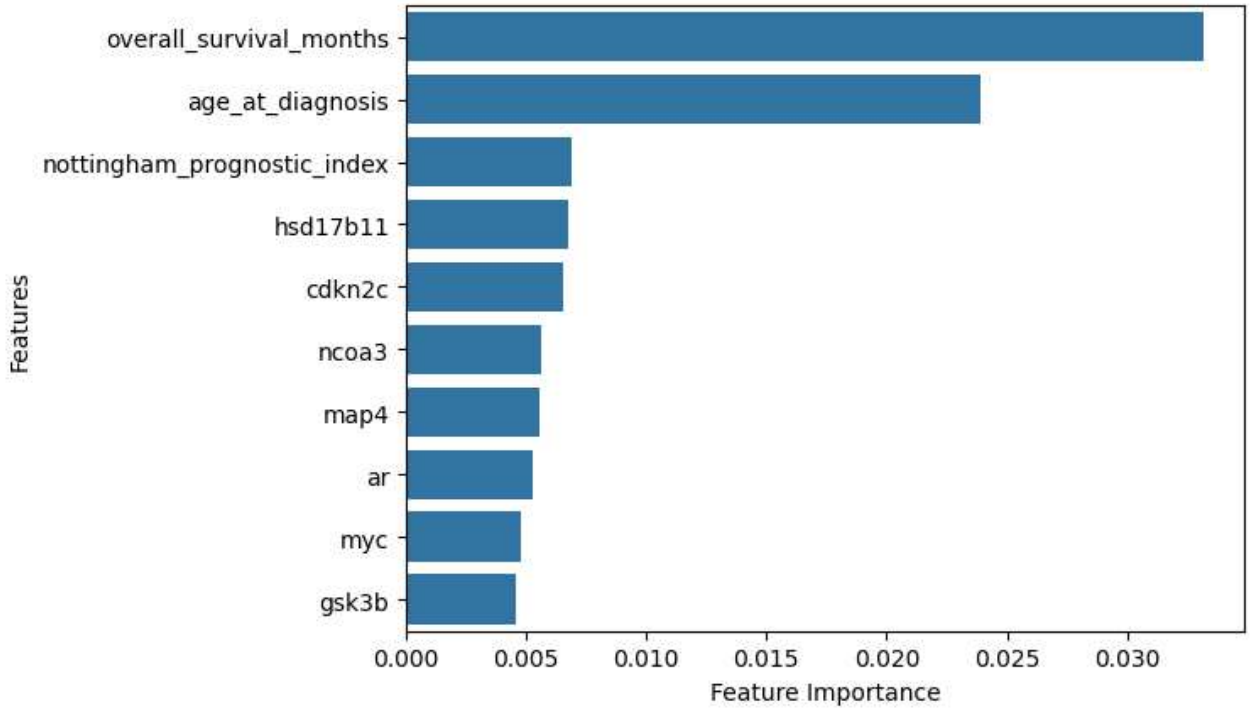
razumijevanje bioloških i kliničkih faktora koji utječu na preživljavanje pacijenata, što može poslužiti kao osnova za daljnja istraživanja.



Slika 5.3.: Prikaz najvažnijih značajki u kliničkom skupu podataka



Slika 5.4.: Prikaz najvažnijih značajki u skupu podataka ekspresije gena



Slika 5.5.: Prikaz najvažnijih značajki u spojenom skupu podataka

5.2. Analiza algoritama i rezultati

Sljedeći korak u našem istraživanju bio je razvoj i evaluacija modela strojnog učenja za predviđanje preživljavanja kod pacijenata s rakom dojke. Koristili smo Random Forest klasifikator s parametrima n_estimators=50 (broj stabala u ansamblu), max_features="sqrt" (broj značajki za razmatranje pri odabiru najboljeg rasporeda), te random_state=44 (postavljen za reproducibilnost rezultata). Implementacija Random forest modela omogućuje nam dublje razumijevanje povezanosti između značajki i ciljne varijable te optimizaciju predviđanja preživljavanja pacijenata s rakom dojke na temelju kliničkih i genetskih podataka.

Kod 5.3.: Model Random Forest

```
rf_model = RandomForestClassifier(n_estimators=50, max_features="sqrt",
random_state=44)
rf_model.fit(X_train, y_train)
```

Zatim smo model isprobali na različitim skupovima podataka, uključujući kliničke podatke, podatke o izražaju gena, kombinirane podatke te reducirane podatke svakog skupa podataka.

	Model	Tip podataka	Testiranje na	Točnost
0	Random forest	Klinički podaci	Testni skup	0.7794
1	Random forest	Klinički podaci-reducirani skup	Testni skup	0.7205
2	Random forest	Razine ekspresije gena	Testni skup	0.6281
3	Random forest	Razine ekspresije gena-reducirani skup	Testni skup	0.5357
4	Random forest	Spojeni podaci	Testni skup	0.6701
5	Random forest	Spojeni podaci-reducirani skup	Testni skup	0.5672

Slika 5.6.: Prikaz rezultata modela Random Forest dobivenih na različitim skupovima podataka

Model koji je treniran na kliničkim podacima postigao je najvišu točnost na testnom skupu od 77.94 %. To pokazuje da klinički podaci sami po sebi imaju visoku informativnost za predviđanje preživljavanja pacijenata s rakom dojke. Smanjenje dimenzionalnosti kliničkih podataka rezultiralo je nešto nižom točnošću na testnom skupu od 72.05 %, što upućuje na potrebu za dodatnom optimizacijom redukcije kako bi se očuvala ili poboljšala prediktivna moć modela.

Model temeljen na podacima o razinama ekspresije gena pokazao je točnost od 62.81 % na testnom skupu. Ovi rezultati sugeriraju da podaci o razinama ekspresije gena imaju nešto manju prediktivnu moć u usporedbi s kliničkim podacima. Nakon smanjenja dimenzionalnosti podataka o razinama ekspresije gena, točnost modela na testnom skupu pala je na 53.57 %. Ovo može ukazivati na preveliku redukciju dimenzionalnosti.

Kombinacija kliničkih podataka i podataka o razinama ekspresije gena rezultirala je točnošću od 67.01 % na testnom skupu, što pokazuje da integracija različitih vrsta podataka može poboljšati performanse modela. Međutim, nakon smanjenja dimenzionalnosti ovih spojenih podataka, točnost na testnom skupu iznosi 56.72 %.

Idući korak bio je primijeniti duboko učenje za predviđanje preživljavanja pacijenata s rakom dojke. Duboko učenje pruža nam mogućnost da modeliramo kompleksne veze između različitih značajki kliničkih podataka i ekspresije gena radi boljeg predviđanja ishoda.

Koristili smo biblioteku TensorFlow za implementaciju modela neuronske mreže. Naš model sadrži tri sloja: ulazni sloj s brojem neurona jednakim broju atributa (značajki), prvi skriveni sloj s 64 neurona i aktivacijskom funkcijom ReLU, te drugi skriveni sloj s 16 neurona također s aktivacijskom funkcijom ReLU. Izlazni sloj ima jedan neuron s sigmoidnom aktivacijskom funkcijom, što je pogodno za binarnu klasifikaciju. Koristili smo binarnu križnu entropiju kao funkciju gubitka i optimizator Adam za učenje modela. Model je treniran kroz 100 epoha s veličinom batch-a od 32 primjera. Nakon treniranja, evaluirali smo točnost modela na testnom skupu podataka kako bismo procijenili njegovu performansu.

Ovaj pristup omogućuje nam da ispitamo kako duboko učenje može doprinijeti poboljšanju prediktivne moći naših modela u odnosu na klasični model Random forest, pružajući detaljne uvide u važnost i utjecaj različitih skupova podataka na rezultate predviđanja preživljavanja pacijenata.

	Eksperiment	Testni skup	Točnost
0	1	Klinički podaci	0.6785
1	2	Klinički podaci	0.7100
2	3	Klinički podaci	0.6554
3	4	Klinički podaci	0.7521
4	5	Klinički podaci	0.6978
5	6	Klinički podaci	0.7037
6	7	Klinički podaci	0.6978
7	8	Klinički podaci	0.6862
8	9	Klinički podaci	0.7252
9	10	Klinički podaci	0.6983
10	Aritmetička sredina	Klinički podaci	0.7005

Slika 5.7.: Prikaz rezultata dubokog učenja nakon deset pokretanja na testnom skupu klinički podaci

	Eksperiment	Testni skup	Točnost
0	1	Razine ekspresije gena	0.6386
1	2	Razine ekspresije gena	0.6428
2	3	Razine ekspresije gena	0.6260
3	4	Razine ekspresije gena	0.6127
4	5	Razine ekspresije gena	0.6049
5	6	Razine ekspresije gena	0.6172
6	7	Razine ekspresije gena	0.6587
7	8	Razine ekspresije gena	0.6232
8	9	Razine ekspresije gena	0.6354
9	10	Razine ekspresije gena	0.6521
10	Aritmetička sredina	Razine ekspresije gena	0.6311

Slika 5.9.: Prikaz rezultata dubokog učenja nakon deset pokretanja na testnom skupu razine ekspresije gena

	Eksperiment	Testni skup	Točnost
0	1	Spojeni podaci	0.7205
1	2	Spojeni podaci	0.7016
2	3	Spojeni podaci	0.6974
3	4	Spojeni podaci	0.7205
4	5	Spojeni podaci	0.7226
5	6	Spojeni podaci	0.7345
6	7	Spojeni podaci	0.7569
7	8	Spojeni podaci	0.7362
8	9	Spojeni podaci	0.7252
9	10	Spojeni podaci	0.7129
10	Aritmetička sredina	Spojeni podaci	0.7228

Slika 5.8.: Prikaz rezultata dubokog učenja nakon deset pokretanja na testnom skupu spojenih podataka

Ovi rezultati prikazuju performanse modela dubokog učenja na različitim skupovima podataka. Variranje točnosti između različitih eksperimenata ukazuje na osjetljivost modela.

Modeli temeljeni samo na kliničkim podacima pokazali su varijabilne performanse s točnostima koje variraju od 65.54 % do 75.21 %. Aritmetička sredina točnosti iznosi 70.05 %. To ukazuje da klinički podaci sami po sebi imaju značajnu prediktivnu moć, ali i oni su podložni varijacijama u performansama.

Modeli temeljeni samo na podacima o razinama ekspresije gena pokazali su umjerene performanse s točnostima koje variraju od 60.49 % do 65.87 %. Aritmetička sredina točnosti iznosi 63.11 %. Ovo ukazuje da su podaci o razinama ekspresije gena sami po sebi korisni, ali imaju ograničenu prediktivnu moć.

Integracija kliničkih podataka s podacima o razinama ekspresije gena poboljšala je performanse modela, pri čemu je postignuta točnost od 69.74 % do 75.69 %. Aritmetička sredina točnosti za ove eksperimente iznosi 72.28 %. Ovi rezultati sugeriraju da kombinacija različitih vrsta podataka može značajno doprinijeti poboljšanju prediktivne moći modela.

5.3. Usporedba rezultata

1. Random Forest:

   Iz rezultata je evidentno da klinički podaci imaju najvišu prediktivnu moć za model Random Forest, s točnošću od 77.94 % na testnom skupu. Redukcija dimenzionalnosti kliničkih podataka rezultira blagim smanjenjem točnosti, dok je utjecaj redukcije dimenzionalnosti znatno izraženiji kod podataka o ekspresiji gena i spojenih podataka. Spojeni podaci, iako sadrže dodatne informacije, ne uspijevaju nadmašiti performanse samih kliničkih podataka. Spojeni podaci uključuju više značajki nego samo klinički podaci, što povećava dimenzionalnost skupa podataka. Visoka dimenzionalnost može otežati učenje modela i povećati rizik od prekomjernog prilagođavanja. Random Forest model može imati problema s obradom velikog broja značajki, posebno ako mnoge od tih značajki nisu informativne. Iako su Random Forest modeli snažni i mogu dobro raditi s različitim vrstama podataka, njihov kapacitet za hvatanje vrlo složenih nelinearnih odnosa je ograničeniji u usporedbi s neuronskim mrežama. Ovi rezultati upućuju na to da, iako je integracija različitih vrsta podataka često poželjna, u ovom slučaju klinički podaci sami pružaju najpouzdanije informacije za predviđanje preživljavanja pacijenata s rakom dojke. Također, smanjenje dimenzionalnosti, iako korisno u nekim scenarijima, treba provoditi pažljivo kako bi se izbjegao gubitak ključnih informacija.

2. Duboko učenje:

   Korištenjem neuronskih mreža postižemo bolju točnost za kombinirane kliničke podatke i podatke o ekspresiji gena. Sljedeće, možemo vidjeti da klinički podaci imaju lošiju izvedbu od kombiniranih podataka, ali bolju od podataka o ekspresiji gena. Klinički podaci imaju najinformativnije značajke što smo vidjeli pronalaskom najvažnijih značajki. Međutim i dalje neke od najvažnijih značajki su povezane s podacima o ekspresiji gena stoga kombinacija ova dva skupa podataka radi bolje nego svaki od njih pojedinačno. Klinički podaci i podaci o ekspresiji gena pružaju različite vrste informacija. Klinički podaci uključuju demografske informacije, detalje o dijagnozi, tretmane i rezultate pacijenata, dok podaci o ekspresiji gena pružaju molekularni profil tumora. Zbog tog kombinacijom ovih

dviju vrsta podataka dobivamo sveobuhvatniju sliku pacijenta, što može poboljšati točnost modela predviđanja. Veća raznolikost značajki može pomoći modelu da prepozna složenije obrasce koji bi mogli ostati neotkriveni ako se koriste samo klinički podaci ili samo podaci o ekspresiji gena.

6. Zaključak

U ovom radu, cilj nam je bio primijeniti različite tehnike strojnog učenja nad podacima ekspresije gena pacijenata oboljelih od leukemije i raka dojke. Cilj nam je bio ispravno klasificirati vrstu leukemije i raka dojke te usporediti rezultate dobivene različitim modelima kako bismo identificirali najefikasniji pristup za svaku vrstu klasifikacije.

Primijetili smo da su postignuti bolji rezultati za klasifikaciju leukemije u usporedbi s klasifikacijom raka dojke. Ovo može biti zbog manjeg broja primjera u podacima koji se odnose na klasifikaciju leukemije, kao i manje dimenzionalnosti podataka. Ograničena dimenzionalnost može olakšati modelima pronalaženje relevantnih uzoraka i značajki te poboljšati klasifikacijske rezultate. Također tehnike strojnog učenja mogu biti preciznije u problemima s dvije klase.

Naša iskustva s nenadziranim učenjem nisu bila zadovoljavajuća u usporedbi s metodama nadziranog učenja. Ipak, nenadzirano učenje može biti korisno kada podaci nemaju prethodno definirane ciljne varijable ili oznake.

Unatoč tome što nismo postigli bolje rezultate s dubokim učenjem u usporedbi s logističkom regresijom i SVM-om, primijetili smo da duboko učenje ima veći potencijal za davanje boljih rezultata uz pravilnu prilagodbu parametara. Klasifikacija raka dojke te procjenjivanje preživljavanja zahtijeva dublje razumijevanje veza između ulaznih podataka i ciljne varijable. Duboko učenje, s mogućnošću učenja složenih reprezentacija podataka, može pružiti veću fleksibilnost u modeliranju takvih složenih veza i potencijalno poboljšati rezultate.

Preporučujemo daljnje istraživanje i optimizaciju parametara za duboko učenje, posebno za klasifikaciju raka dojke. Takvo istraživanje može pružiti dublje uvide i potencijalno poboljšati rezultate klasifikacije.

Literatura

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Col ler, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield,E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression. 1999.

[2] Jan Šnajder. Stroj potpornih vektora. https://www.fer.unizg.hr/_download/repository/SU-202008-StrojPotpornihVektora%5B1%5D.pdf 2022

[3] Jan Šnajder. Strojno učenje. https://www.fer.unizg.hr/_download/repository/UI-2020-10StrojnoUcenje.pdf 2020

[4] Scikit-learn. Machine Learning in Python. Pristupljeno 12.5.2024. iz https://scikit-learn.org/stable/documentation.html

[5] DeepAI. Artificial Intelligence. Pristupljeno 12.5.2.24. iz https://deepai.org/

[6] Python (2024.) Pristupljeno 1.4.2024. iz https://www.python.org/

[7] Pandas (2024.) Pristupljeno 1.4.2024. iz https://pandas.pydata.org/

Dijagnostika bolesti na temelju podataka o ekspresiji gena

Sažetak

Analiza ekspresije gena ključna je za razumijevanje molekularnih mehanizama koji stoje iza različitih tipova karcinoma. Ona omogućava identifikaciju specifičnih gena čija je aktivnost povezana s nastankom i razvojem bolesti, što je od presudne važnosti za dijagnozu, prognozu i terapiju. U ovom radu, istražili smo primjenu različitih tehnika nadziranog i nenadziranog strojnog učenja na skupovima podataka za klasifikaciju leukemije i raka dojke koristeći podatke dobivene analizom ekspresije gena. Kroz analizu, implementaciju i evaluaciju algoritama logističke regresije, stroja potpornih vektora (SVM), dubokog učenja i grupiranja, istražili smo njihov doprinos u klasifikaciji karcinoma koristeći programske biblioteke sklearn i keras. Uspješno smo klasificirali različite oblike karcinoma te su naši rezultati bili u skladu s očekivanim. Ovaj rad pruža uvid u primjenu strojnog učenja u medicini te otvara nova područja za daljnja istraživanja.

Ključne riječi: ekspresija gena, logistička regresija, stroj potpornih vektora, duboko učenje, grupiranje

Disease diagnosis based on gene expression data

Abstract

Gene expression analysis is crucial for understanding the molecular mechanisms underlying various types of cancer. It enables the identification of specific genes whose activity is associated with the onset and development of the disease, which is essential for diagnosis, prognosis, and therapy. In this study, we explored the application of different supervised and unsupervised machine learning techniques on datasets for the classification of leukaemia and breast cancer using data obtained from gene expression analysis. Through the analysis, implementation, and evaluation of algorithms such as logistic regression, support vector machines (SVM), deep learning, and clustering, we investigated their contribution to cancer classification using the sklearn and keras libraries. We successfully classified different forms of cancer, and our results were consistent with expectations. This paper provides insight into the application of machine learning in medicine and opens new areas for further research.

Keywords: gene expression, logistic regression, support vector machine, deep learning, clustering