SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINARSKI RAD

KLASIFIKACIJA KARCINOMA NA TEMELJU PODATAKA O EKSPRESIJI GENA

Domagoj Sviličić

Voditelj: doc. dr. sc. Krešimir Križanović

Zagreb, siječanj, 2025.

Klasifikacija karcinoma na temelju podataka o ekspresiji gena

Domagoj Sviličić

Sažetak

Klasifikacija karcinoma na temelju podataka o ekspresiji gena je jedno od područja u kojem se primjenjuje strojno učenje. U ovom radu opisana je metoda klasifikacije karcinoma na temelju podataka o ekspresiji gena. Korišteni su podaci o ekspresiji gena iz dva skupa podataka: skup podataka o karcinomu dojke i skup podataka o karcinomu pluća. Korišteni su različiti algoritmi strojnog učenja za klasifikaciju karcinoma. Rezultati klasifikacije su uspoređeni i analizirani.

Ključne riječi: strojno učenje; bioinformatika; ekspresija gena; klasifikacija; karcinom

Sadržaj

1. Uvod

Strojno učenje je područje umjetne inteligencije koje se bavi razvojem algoritama i tehnika koji omogućuju računalima učenje iz podataka. Strojno učenje se koristi u različitim područjima, uključujući bioinformatiku. Bioinformatika je interdisciplinarno područje koje se bavi primjenom računalnih tehnika u biologiji. Jedno od područja u kojem se primjenjuje strojno učenje u bioinformatici je klasifikacija karcinoma na temelju podataka o ekspresiji gena. Razvojem tehnika analize ekspresije gena u posljednjim godinama omogućen je uvid u uzroke i liječenje različitih bolesti, uključujući karcinom. Ekspresija gena odnosi se na proces kojim se informacije kodirane u DNA pretvaraju u funkcionalne molekule poput proteina. Budući da su proteini odgovorni za niz ključnih funkcija poput rasta stanica, neispravno "uključivanje/isključivanje" pojedinih proteina može značajno promijeniti funkcionalnost stanice i time utjecati na nastanak i napredovanje karcinoma.

2. Priprema podataka

Podaci korišteni u svrhe ovog seminarskog rada preuzeti su s kaggle.com platforme. Skup podataka o karcinomu dojke sadrži podatke o ekspresiji gena za 801 pacijenta. Skup podataka o karcinomu dojke dostupan je na poveznici `https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida/`, dok je skup podataka o leukemiji dostupan na poveznici `https://www.kaggle.com/datasets/crawford/gene-expression`. Skup podataka o karcinomu dojke sadrži podatke o ekspresiji gena za 801 pacijenta.

Podaci za klasifikaciju leukemije su preuzeti iz istraživanja Golub et al. (1999) i nalaze se u obliku tri tablice. Prva tablica sadrži podatke prvih 38 pacijenata koje koristimo za treniranje modela (2.1.), druga tablica sadži podatke 34 pacijenata koje koristimo za testiranje i strukturom je potupuno analogna prvoj, dok treća tablica sadrži oznaku vrste leukemije za svakog pacijenta (2.2.). Naš skup podataka sadrži dvije vrste oznaka: ALL koja predstavlja akutnu limfoblastičnu leukemiju i AML koja predstavlja akutnu mijeloičnu leukemiju. To su ujedno i klase kojima pridružujemo primjere.

	Gene Description	Gene Accession Number	1	call	2	call.1	3	call.2	4	call.3	...	29	call.33	30	call.34	31	call.35	32	call.36	33	call.37
0	AFFX-BioB-5_at (endogenous control)	AFFX-BioB-5_at	-214	A	-139	A	-76	A	-135	A	...	15	A	-318	A	-32	A	-124	A	-135	A
1	AFFX-BioB-M_at (endogenous control)	AFFX-BioB-M_at	-153	A	-73	A	-49	A	-114	A	...	-114	A	-192	A	-49	A	-79	A	-186	A
2	AFFX-BioB-3_at (endogenous control)	AFFX-BioB-3_at	-58	A	-1	A	-307	A	265	A	...	2	A	-95	A	49	A	-37	A	-70	A
3	AFFX-BioC-5_at (endogenous control)	AFFX-BioC-5_at	88	A	283	A	309	A	12	A	...	193	A	312	A	230	P	330	A	337	A
4	AFFX-BioC-3_at (endogenous control)	AFFX-BioC-3_at	-295	A	-264	A	-376	A	-419	A	...	-51	A	-139	A	-367	A	-188	A	-407	A

Slika 2.1. Izgled tablice s podacima za treniranje

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

Slika 2.2. Izgled tablice actual.csv s oznakama

Podaci za klasifikaciju raka dojke preuzeti su iz repozitorija CuMiDa i nalaze se u obliku jedne tablice. Svaki redak u tablici sadrži 54675 vrijednosti ekspresije gena i oznaku klase za svaki uzorak. Oznake koje se pridjeljuju primjerima su: basal, HER, luminal_A, luminal_B, cell_line i normal. Prvih pet oznaka predstavljaju različite vrste raka dojke, dok oznaka normal označava da osoba ne boluje od nijedne vrste raka dojke. Struktura tablice prikazana je na slici 2.3.

	samples	type	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_at	...	AFFX-ThrX-M_at	AFFX-TrpnX-3_at	AFFX-TrpnX-5_at	AFFX-TrpnX-M_at
0	84	basal	9.850040	8.097927	6.424728	7.353027	3.029122	6.880079	4.963740	4.408328	...	4.901594	2.966657	3.508495	3.301999
1	85	basal	9.861357	8.212222	7.062593	7.685578	3.149468	7.542283	5.129607	4.584418	...	5.405839	2.934763	3.687666	3.064299
2	87	basal	10.103478	8.936137	5.735970	7.687822	3.125931	6.562369	4.813449	4.425195	...	5.184286	2.847684	3.550597	3.158535
3	90	basal	9.756875	7.357148	6.479183	6.986624	3.181638	7.802344	5.490982	4.567956	...	5.086569	3.031602	3.524981	3.272665
4	91	basal	9.408330	7.746404	6.693980	7.333426	3.169923	7.610457	5.372469	4.424426	...	5.235318	2.956232	3.445501	3.193947

Slika 2.3. Izgled tablice s podacima za klasifikaciju raka dojke

Za podatke o leukemiji karakteristično je kako je za svakog pojedinog pacijenta oznaka ekspresije gena iskazana numerički i kategorički oznakom A, P ili M. A označava da gen nije prisutan (engl. Absent), P označava da je gen prisutan (engl. Present) i M označava da ne možemo sa sigurnšću tvrditi da li je gen prisutan ili ne (engl. Marginal). Znakovi su zamijenjeni diskretnim numeričkim vrijednostima -1, 1, 0. Pomoću pandas biblioteke izvučeni su relevatni stupci i spareni s odgovarajućim oznakama kako bi se formirali trening i test skupovi. U nastavku će se da dotični skup paraleleno koristiti i numeričke i kategoričke oznake kako bi se prikazala razlika u rezultatima klasifikacije.

3. Rezultati i rasprava

3.1. Klasifikacija raka dojke

3.1.1. Logistička regresija

Logistička regresija je tehnika strojnog učenja koja se koristi za modeliranje binarnih ili višeklasnih izlaza. U ovom radu, korištena je za višeklasnu klasifikaciju šest vrsta raka dojke. Logistička regresija je jednostavna za implementaciju i interpretaciju, ali može biti ograničena u slučajevima kada su podaci kompleksni ili kada postoji više nelinearnih odnosa među značajkama. Unatoč tome, logistička regresija je pokazala solidne performanse, pružajući pouzdane rezultate uz relativno nisku složenost modela. Dobiveni rezultati pokazuju visoku točnost modela, što sugerira učinkovitost ove metode u detekciji različitih tipova raka dojke. Eksperiment je provedenih na nekoliko različith omjera podjele podataka na trening i test skupove, a rezultatni ispis je prikazan na slici 3.1.

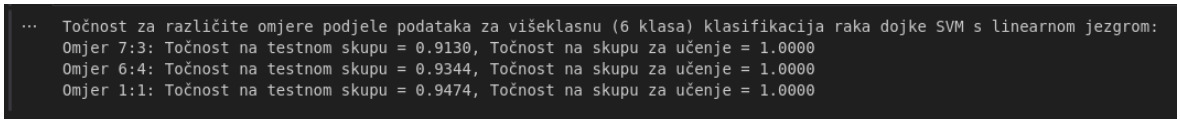
```
···  Točnost za različite omjere podjele podataka za višeklasnu (6 klasa) klasifikacija raka dojke logističkom regresijom:
     Omjer 7:3: Točnost na testnom skupu = 0.8913, Točnost na skupu za učenje = 1.0000
     Omjer 6:4: Točnost na testnom skupu = 0.9344, Točnost na skupu za učenje = 1.0000
     Omjer 1:1: Točnost na testnom skupu = 0.9342, Točnost na skupu za učenje = 1.0000
```

Slika 3.1. Rezultati klasifikacije logističkom regresijom uz korištenje podataka o karcinomu dojke

3.1.2. Stroj potpornih vektora (SVM) s linearnom jezgrom

Podrška vektorima (SVM) s linearnom jezgrom je algoritam koji traži hiper-ravninu koja najbolje razdvaja klase u višedimenzionalnom prostoru. Glavna prednost SVM-a s linearnom jezgrom leži u njegovoj jednostavnosti i učinkovitosti kod linearno razdvojivih skupova podataka. Za linearno neodvojive podatke moguće je koristiti i druge jezgre poput radijalne ili polinomijalne. U ovom istraživanju, SVM je primijenjen za klasifikaciju

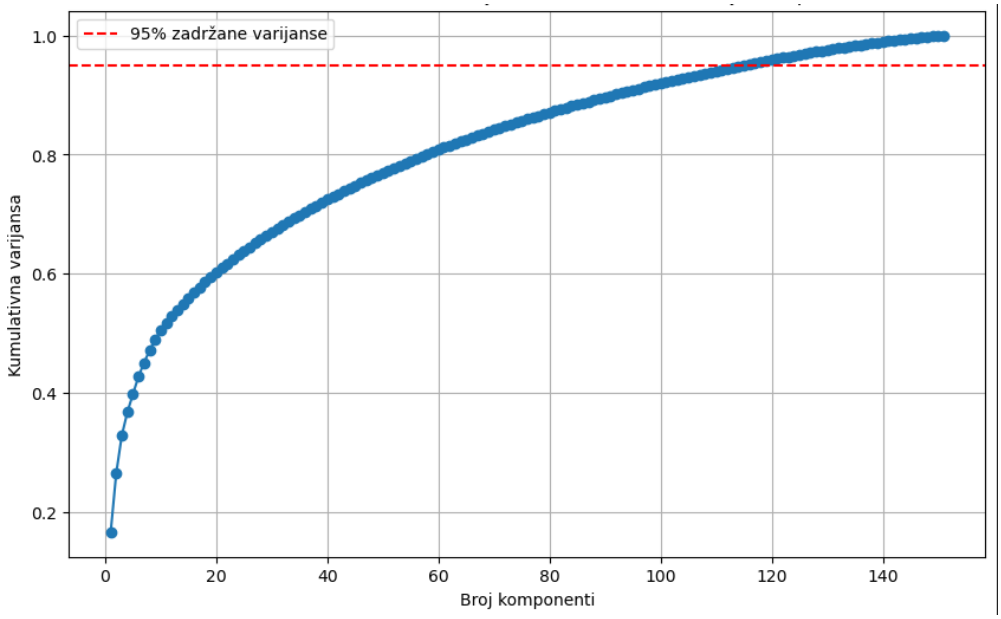
šest vrsta raka dojke, pokazujući visoku točnost. SVM model koristi Lagrangovu dualnost za maksimizaciju margine između klasa, čime osigurava robusnost i generalizaciju modela. Ispitivanja na različitim omjerima podjele podataka (npr. 7:3, 6:4) pokazala su dosljedno visoke performanse, s točnostima koje se kreću iznad 90% na testnim skupovima.



Slika 3.2. Rezultati klasifikacije SVM-om uz korištenje podataka o karcinomu dojke

3.1.3. Duboko učenje s PCA

Duboko učenje s analizom glavnih komponenti (PCA) kombinira sposobnost neuronskih mreža da modeliraju složene nelinearne odnose s PCA-ovom tehnikom za smanjenje dimenzionalnosti. PCA transformira originalne podatke u novi skup ortogonalnih komponenti koje zadržavaju maksimalnu varijancu podataka, smanjujući dimenzionalnost i time olakšavajući obradu podataka. U ovom radu, PCA je smanjio broj značajki na 117 komponenti (pri čemu je zadržao 95% originalne varijabilnosti), nakon čega je slijedilo treniranje neuronske mreže. Neuronska mreža je sadržavala slojeve s 64 i 32 čvora, s ReLU aktivacijskom funkcijom, te završni sloj s 6 izlaznih čvorova za klasifikaciju. Takav pristup omogućio je smanjenje overfittinga i poboljšanje točnosti, osobito kod velikih i složenih skupova podataka. Grafički prikaz smanjenja broja značajki prikazan je na grafikonu 3.3., a rezultati klasifikacije za tako treniranu neuronsku mrežu kretali su se u prosjeku 64.52% za deset pokretanja nad testnim skupom.



Slika 3.3. PCA - kumulativna varijanca u zavisnosti o broju komponenti

3.1.4. Biblioteka Lazy Predict

Lazy Predict je biblioteka dizajnirana za brzo testiranje raznih modela strojnog učenja bez potrebe za ručnim podešavanjem parametara. Ona automatski trenira i evaluira niz različitih klasifikatora te pruža detaljan izvještaj o njihovim performansama. U ovom radu, Lazy Predict je korišten kako bi se ubrzao proces odabira najboljeg modela za klasifikaciju raka dojke. Prednost ove biblioteke leži u njenoj sposobnosti da brzo identificira najučinkovitije modele za određeni skup podataka, smanjujući vrijeme potrebno za ručno testiranje više modela. To omogućuje istraživačima da se fokusiraju na dublju analizu i interpretaciju rezultata najboljih modela. Rezultatni ispis nakon korištenja biblioteke prikazan je slici 3.13.

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score
LogisticRegression	0.96	0.97	None	0.96
LinearDiscriminantAnalysis	0.92	0.94	None	0.92
BernoulliNB	0.92	0.93	None	0.92
SGDClassifier	0.92	0.93	None	0.92
RidgeClassifierCV	0.95	0.93	None	0.95
RidgeClassifier	0.95	0.93	None	0.95
ExtraTreesClassifier	0.93	0.93	None	0.93
Perceptron	0.89	0.91	None	0.90
LinearSVC	0.88	0.90	None	0.89
PassiveAggressiveClassifier	0.88	0.90	None	0.89
NearestCentroid	0.88	0.88	None	0.88
CalibratedClassifierCV	0.87	0.87	None	0.87
RandomForestClassifier	0.92	0.87	None	0.91
XGBClassifier	0.88	0.84	None	0.87
LGBMClassifier	0.88	0.80	None	0.87
BaggingClassifier	0.87	0.79	None	0.85
GaussianNB	0.86	0.77	None	0.83
SVC	0.80	0.73	None	0.77
KNeighborsClassifier	0.80	0.72	None	0.78
DecisionTreeClassifier	0.78	0.71	None	0.76
ExtraTreeClassifier	0.68	0.64	None	0.64
AdaBoostClassifier	0.42	0.36	None	0.32
QuadraticDiscriminantAnalysis	0.17	0.24	None	0.17
LabelSpreading	0.22	0.17	None	0.08
LabelPropagation	0.22	0.17	None	0.08
DummyClassifier	0.24	0.17	None	0.09

Slika 3.4. Ispis Lazy predict biblioteke za klasifikaciju raka dojke

Lazypredict dodatno omogućuje i prikaz vremena trajanja treniranja i testiranja modela, što je prikazano na slici 3.5.

Model	Time Taken
LogisticRegression	1.75
LinearDiscriminantAnalysis	2.05
BernoulliNB	0.94
SGDClassifier	1.15
RidgeClassifierCV	0.84
RidgeClassifier	0.77
ExtraTreesClassifier	0.93
Perceptron	1.11
LinearSVC	2.87
PassiveAggressiveClassifier	2.37
NearestCentroid	0.80
CalibratedClassifierCV	5.36
RandomForestClassifier	1.06
XGBClassifier	36.63
LGBMClassifier	27.56
BaggingClassifier	6.25
GaussianNB	1.01
SVC	1.17
KNeighborsClassifier	0.86
DecisionTreeClassifier	1.70
ExtraTreeClassifier	0.68
AdaBoostClassifier	17.66
QuadraticDiscriminantAnalysis	1.10
LabelSpreading	0.91
LabelPropagation	0.87
DummyClassifier	0.67

Slika 3.5. Ispis vremenskih trajanja Lazy predict biblioteke za klasifikaciju raka dojke

Primjetno je da su XGBClassifier, LGMClassifier, AdaBoostClassifier modeli bili nasporiji u izvedbi. Ovi modeli su vjerojatno bili najsporiji zbog kompleksnosti i veličine ansambla. Spomenuti boosting algoritmi treniraju niz modela jedan za drugim, gdje svaki novi model pokušava ispraviti greške prethodnih. Taj proces može biti vrlo računalno intenzivan, pogotovo kada se radi o velikim datasetovima i velikom broju iteracija. Također, konfiguracije modela i hiperparametri poput dubine stabala odluke, broja stabala i stope učenja mogu značajno utjecati na vrijeme treniranja, a bitno je naglasiti kako naš podskup sadrži i poprilično velik broj značajki.

- **Najbolje Performanse**: LogisticRegression postiže najbolje rezultate sa točnošću od 0.96 i vremenskim trajanjem treniranja od 1.75 sekundi.
- **Brzi Modeli**: BernoulliNB i RidgeClassifier su među najbržima sa treniranjem ispod 1 sekunde, a imaju solidne performanse.

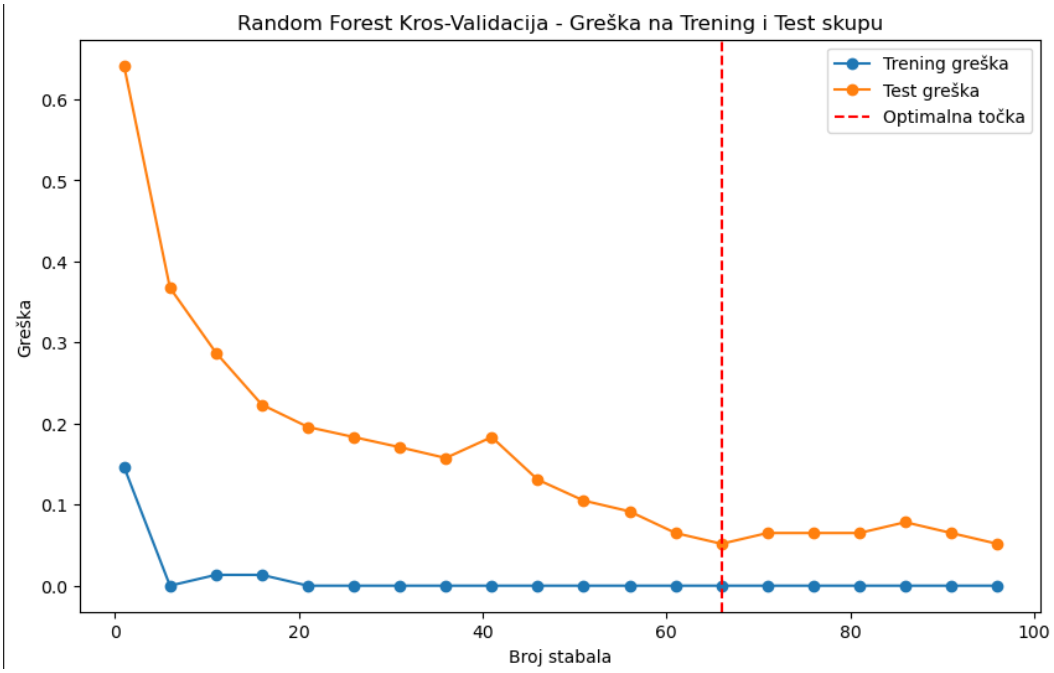
- **Varijabilnost u Performansama**: KNeighborsClassifier i SVC pokazuju niže performanse u usporedbi s drugim modelima, ali njihova točnost je još uvijek respektabilna.

- **Najsloženiji Modeli**: XGBClassifier i LGBMClassifier imaju znatno duže vrijeme treniranja, ali njihove performanse nisu najbolje, što sugerira da dodatna kompleksnost nije nužno korisna za naš skup podataka.

- **Izuzetno Niske Performanse**: Modeli poput QuadraticDiscriminantAnalysis i LabelSpreading pokazuju znatno niže performanse i očito nisu najprikladniji izbor za ovaj zadatak.

3.1.5. Random Forest

Random Forest je ansambl metoda koja koristi mnoštvo stabala odlučivanja za donošenje konačne odluke putem većinskog glasanja. Svako stablo u šumi trenira se na slučajnom podskupu podataka, što smanjuje varijansu i povećava robusnost modela. Random Forests je osobito učinkovit u rješavanju problema klasifikacije gdje su podaci visoko dimenzionalni i sadrže puno buke. U kontekstu raka dojke, ova tehnika je pokazala značajnu točnost, zahvaljujući sposobnosti da uhvati složene odnose između značajki bez overfittinga. To čini Random Forests jednim od preferiranih izbora kada je potrebno osigurati stabilne i pouzdane predikcije. Jedan od hiperparametara koji se može prilagoditi je borj stabala u šumi. U našem je slučaju metodom kros validacije odabrano 66 stabala uz točnost 0d 0.91%. Na slici 3.6. prikazan je grafikon koji prikazuje točnost modela u ovisnosti o broju stabala u šumi. Optimalan broj stabala odabran je tako da model radi najmanju moguću pogrešku na skupu za testiranje

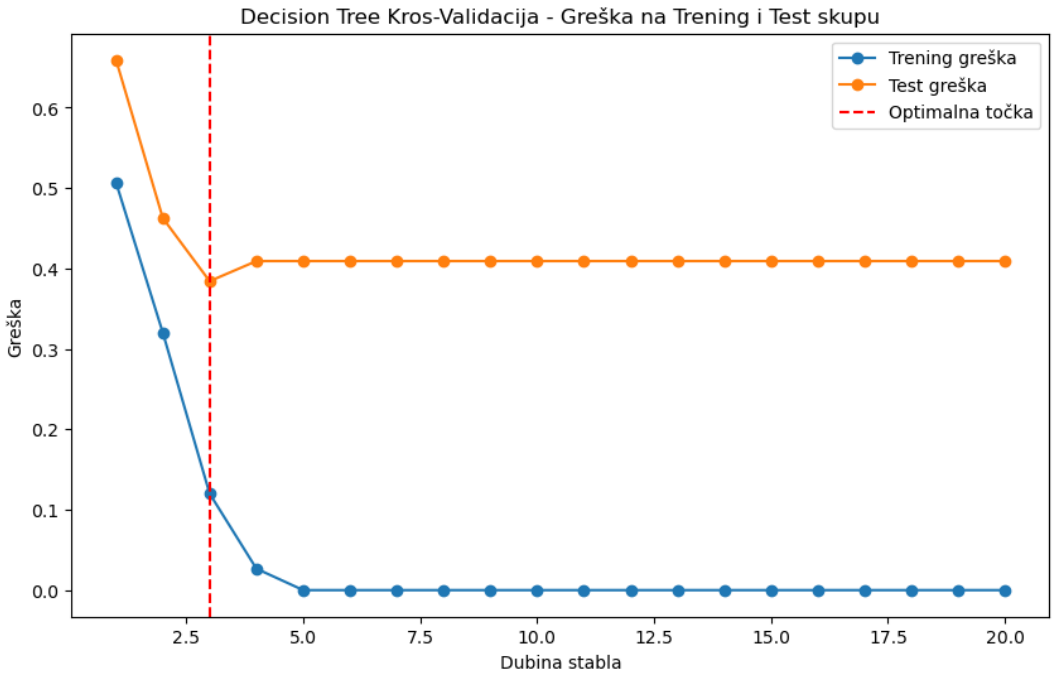


Slika 3.6. odabir optimalnog broja stabala metodom kros validacije

Točnost ove metode odgovara rezultatima dobivenim korištenjem Lazy Predict biblioteke. Manje razlike su moguće zbog interne implementacije modela unutar biblioteke.

3.1.6. Decision Tree

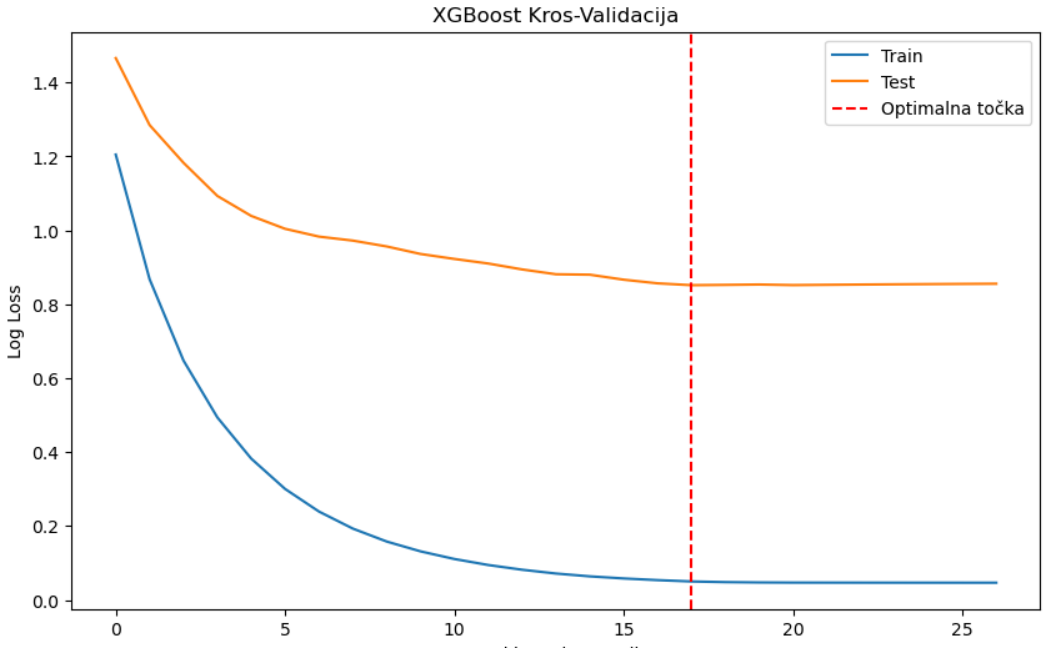
Decision Tree je metoda koja koristi strukturu stabla kako bi donijela odluke na temelju značajki podataka. Svaki čvor stabla predstavlja odluku na određenoj značajki, dok svaki listeni čvor predstavlja ishod ili klasu. Decision Tree je intuitivan i jednostavan za interpretaciju, što ga čini privlačnim za mnoge primjene. Međutim, zbog sklonosti prekomjernoj prilagodbi (overfittingu), često se kombinira s drugim tehnikama poput prerezivanja stabla ili korištenja ansambl metoda. U ovom radu, Decision Tree je ostvario točnost od 0.71% na testnom skupu, što je niže u usporedbi s drugim metodama. Ova metoda može biti korisna za osnovnu analizu podataka, ali može biti poboljšana korištenjem složenijih tehnika. Kros validacijom moguće je odrediti optimalnu dubinu stabla odluke što je prikazano na slici 3.7.



Slika 3.7. odabir optimalne dubine stabla metodom kros validacije

3.1.7. XGBoost

XGBoost (Extreme Gradient Boosting) je optimizirana implementacija algoritma gradijentnog pojačavanja dizajnirana za brzu izvedbu i visoku točnost. Koristi tehniku gradijentnog pojačavanja za stvaranje skupa modela slabih učitelja (najčešće stabala odlučivanja), koji se treniraju uzastopno, gdje svaki sljedeći model pokušava ispraviti greške prethodnog. XGBoost koristi regularizaciju kako bi spriječio prekomjernu prilagodbu, što ga čini izuzetno učinkovitim za velike i kompleksne skupove podataka. U ovom istraživanju, XGBoost je ostvario točnost od 0.88% na testnom skupu. Grafički prikaz određivanja optimalnog broja boosting rundi pomoću kros validacije prikazan je na slici 3.8.



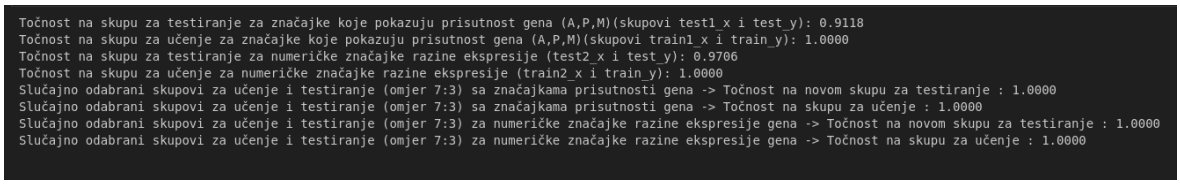
Slika 3.8. Odabir optimalnog broja boosting rundi metodom kros validacije

Interesantno je primjetiti kako krivulje pogreške na trening i test skupu imaju karakterističan monoton padajući oblik, a optimalan broj rundi ustvari predstavlja trenutak kada krivulja pogreške na testnom skupu postiže konvergenciju. Implementacija algoritma je dovoljno napredna da niti u jednoj rundi ne dolazi do povećanja pogreške na testnom skupu, što sugerira da je model dovoljno robusan i da ne dolazi do prenaučenosti.

3.2. Klasifikacija leukemije

3.2.1. Logistička regresija

Logistička regresija korištena je i za klasifikaciju podataka o leukemiji. Analiza rezultata pokazuje da logistička regresija može biti učinkovita u klasifikaciji različitih podtipova leukemije, uz zadržavanje visoke točnosti modela. Model je primjenjen koristeći zasebno numeričke i kategoričke značajke, a osim inicijalne podjele u trening i test skupove kakva je dana dvjema početnim tablicama, korištena je i podjela u omjeru 7:3. Rezultati su prikazani na slici 3.9.

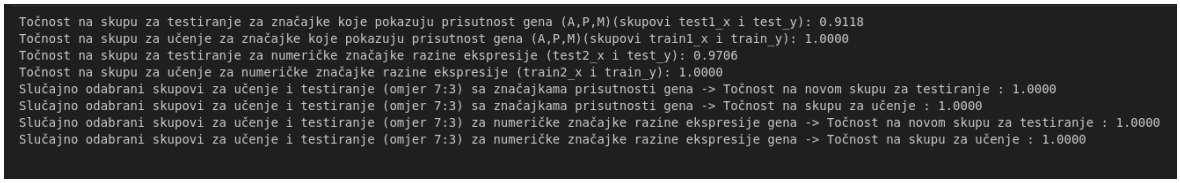


Slika 3.9. Ispis modela logističke regresije za klasifikaciju leukemije

Numeričke značajke su očekivano dale bolje rezultate, što je posljedica veće količine informacija koje one nose. Kategoričke značajke su u ovom slučaju bile manje informativne, što je rezultiralo nižom točnošću modela. Preraspodjelom podataka u trening i test skupove u omjeru 7:3, povećan je broj primjera u trening skupu, što je rezultiralo većom precinznosti na testnom skupu.

3.2.2. Stroj potpornih vektora (SVM) s linearnom jezgrom

Za klasifikaciju leukemije, SVM s linearnom jezgrom pokazao je značajnu točnost, što ukazuje na njegovu prikladnost za ovakve vrste podataka gdje je potrebna jasna separacija među klasama. Rezultatni ispis dan je na slici 3.10.



Slika 3.10. Ispis modela SVM-a za klasifikaciju leukemije

Rezultati su identični onima dobivenim korištenjem logističke regresije, što sugerira da su ove dvije metode jednako prikladne za klasifikaciju leukemije.

3.2.3. Biblioteka Lazy Predict

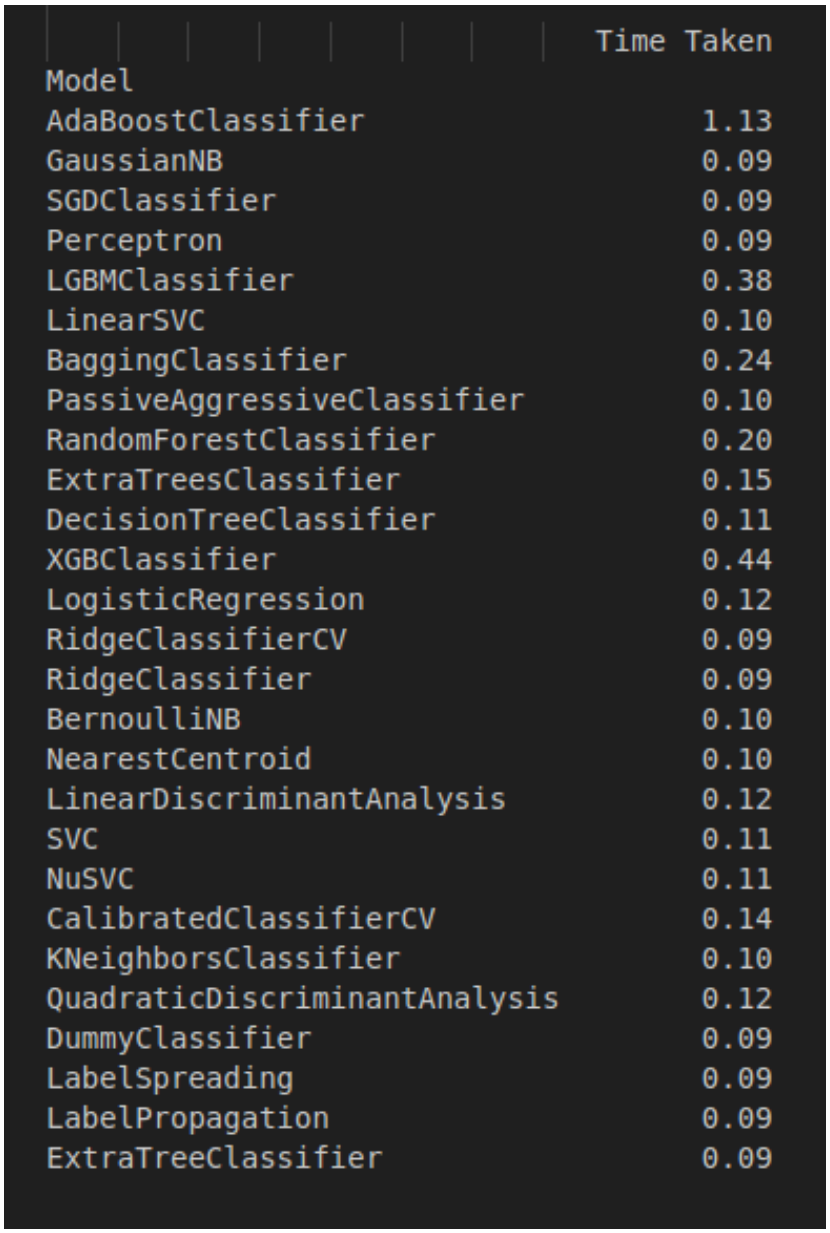
I kod podataka za leukemiju korištena je Lazy Predict biblioteka te su rezultati za numeričku verziju značajki prikazani na slici 3.11.

Rezultati za numeričke značajke razine ekspresije gena:

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score
AdaBoostClassifier	1.00	1.00	1.00	1.00
GaussianNB	1.00	1.00	1.00	1.00
SGDClassifier	1.00	1.00	1.00	1.00
Perceptron	1.00	1.00	1.00	1.00
LGBMClassifier	1.00	1.00	1.00	1.00
LinearSVC	0.95	0.96	0.96	0.95
BaggingClassifier	0.95	0.96	0.96	0.95
PassiveAggressiveClassifier	0.95	0.96	0.96	0.95
RandomForestClassifier	0.95	0.94	0.94	0.95
ExtraTreesClassifier	0.95	0.94	0.94	0.95
DecisionTreeClassifier	0.95	0.94	0.94	0.95
XGBClassifier	0.95	0.94	0.94	0.95
LogisticRegression	0.95	0.94	0.94	0.95
RidgeClassifierCV	0.95	0.94	0.94	0.95
RidgeClassifier	0.95	0.94	0.94	0.95
BernoulliNB	0.82	0.79	0.79	0.81
NearestCentroid	0.82	0.79	0.79	0.81
LinearDiscriminantAnalysis	0.82	0.78	0.78	0.80
SVC	0.77	0.72	0.72	0.75
NuSVC	0.77	0.72	0.72	0.75
CalibratedClassifierCV	0.73	0.67	0.67	0.68
KNeighborsClassifier	0.68	0.61	0.61	0.61
QuadraticDiscriminantAnalysis	0.50	0.56	0.56	0.45
DummyClassifier	0.59	0.55	0.55	0.57
LabelSpreading	0.59	0.50	0.50	0.44
LabelPropagation	0.59	0.50	0.50	0.44
ExtraTreeClassifier	0.50	0.49	0.49	0.50

Slika 3.11. Ispis Lazy Predict za klasifikaciju leukemije (numeričke značajke)

Vremena izvođenja za istu numeričku verziju značajki prikazana su na slici 3.12.



Model	Time Taken
AdaBoostClassifier	1.13
GaussianNB	0.09
SGDClassifier	0.09
Perceptron	0.09
LGBMClassifier	0.38
LinearSVC	0.10
BaggingClassifier	0.24
PassiveAggressiveClassifier	0.10
RandomForestClassifier	0.20
ExtraTreesClassifier	0.15
DecisionTreeClassifier	0.11
XGBClassifier	0.44
LogisticRegression	0.12
RidgeClassifierCV	0.09
RidgeClassifier	0.09
BernoulliNB	0.10
NearestCentroid	0.10
LinearDiscriminantAnalysis	0.12
SVC	0.11
NuSVC	0.11
CalibratedClassifierCV	0.14
KNeighborsClassifier	0.10
QuadraticDiscriminantAnalysis	0.12
DummyClassifier	0.09
LabelSpreading	0.09
LabelPropagation	0.09
ExtraTreeClassifier	0.09

Slika 3.12. Vremena izvođenja Lazy Predict za klasifikaciju leukemije (numeričke značajke)

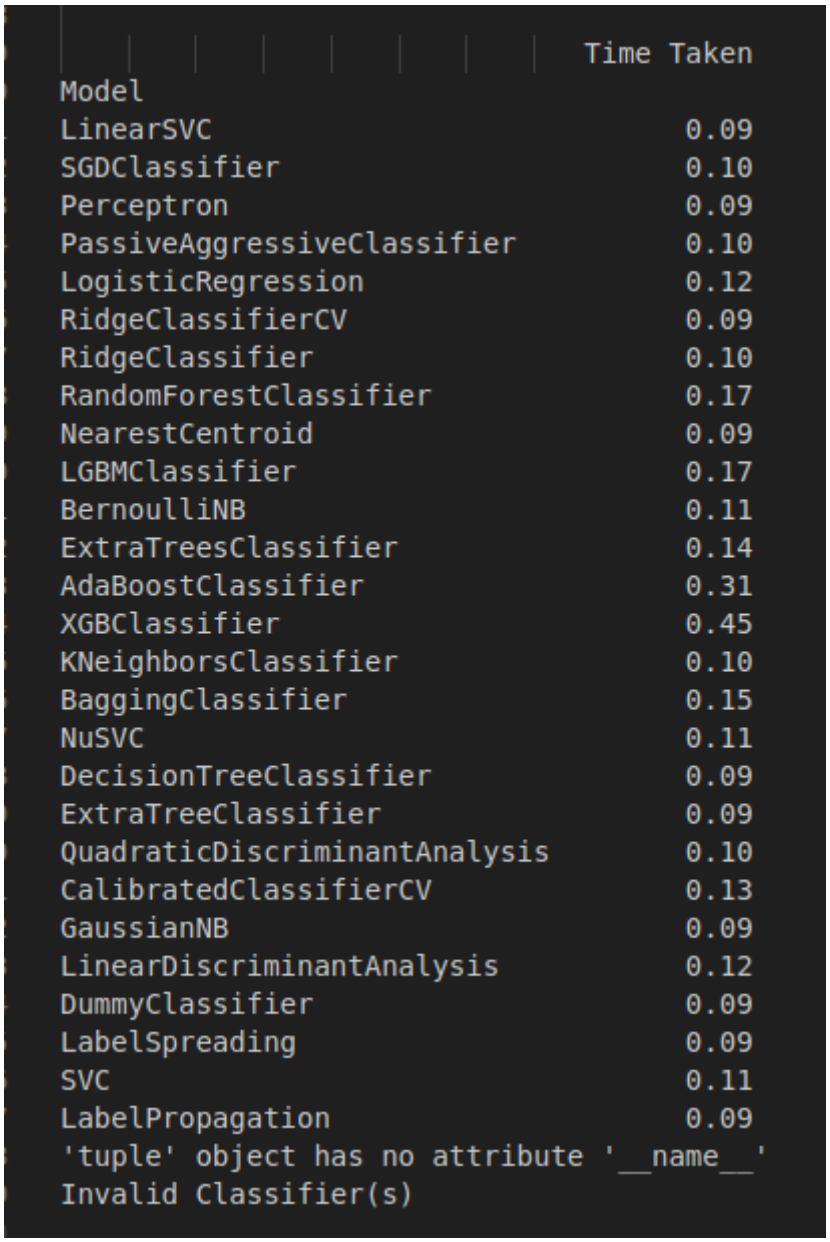
Na slici 3.13. prikazan je ispis Lazy Predict biblioteke za kategoričku verziju značajki. (oznake A, P, M označavaju odsutnost, prisutnost i marginalnost gena)

Rezultati za značajke prisutnosti gena:

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score
LinearSVC	1.00	1.00	1.00	1.00
SGDClassifier	1.00	1.00	1.00	1.00
Perceptron	1.00	1.00	1.00	1.00
PassiveAggressiveClassifier	1.00	1.00	1.00	1.00
LogisticRegression	0.95	0.94	0.94	0.95
RidgeClassifierCV	0.95	0.94	0.94	0.95
RidgeClassifier	0.95	0.94	0.94	0.95
RandomForestClassifier	0.95	0.94	0.94	0.95
NearestCentroid	0.95	0.94	0.94	0.95
LGBMClassifier	0.95	0.94	0.94	0.95
BernoulliNB	0.95	0.94	0.94	0.95
ExtraTreesClassifier	0.95	0.94	0.94	0.95
AdaBoostClassifier	0.91	0.89	0.89	0.91
XGBClassifier	0.91	0.89	0.89	0.91
KNeighborsClassifier	0.91	0.89	0.89	0.91
BaggingClassifier	0.86	0.83	0.83	0.86
NuSVC	0.82	0.78	0.78	0.80
DecisionTreeClassifier	0.77	0.76	0.76	0.77
ExtraTreeClassifier	0.64	0.66	0.66	0.64
QuadraticDiscriminantAnalysis	0.50	0.56	0.56	0.45
CalibratedClassifierCV	0.64	0.56	0.56	0.53
GaussianNB	0.64	0.56	0.56	0.53
LinearDiscriminantAnalysis	0.64	0.56	0.56	0.53
DummyClassifier	0.59	0.55	0.55	0.57
LabelSpreading	0.59	0.50	0.50	0.44
SVC	0.59	0.50	0.50	0.44
LabelPropagation	0.59	0.50	0.50	0.44

Slika 3.13. Vremena izvođenja Lazy Predict za klasifikaciju leukemije (kategoričke značajke)

Također je prikazano i vrijeme izvođenja za kategoričku verziju značajki na slici 3.14.



Model	Time Taken
LinearSVC	0.09
SGDClassifier	0.10
Perceptron	0.09
PassiveAggressiveClassifier	0.10
LogisticRegression	0.12
RidgeClassifierCV	0.09
RidgeClassifier	0.10
RandomForestClassifier	0.17
NearestCentroid	0.09
LGBMClassifier	0.17
BernoulliNB	0.11
ExtraTreesClassifier	0.14
AdaBoostClassifier	0.31
XGBClassifier	0.45
KNeighborsClassifier	0.10
BaggingClassifier	0.15
NuSVC	0.11
DecisionTreeClassifier	0.09
ExtraTreeClassifier	0.09
QuadraticDiscriminantAnalysis	0.10
CalibratedClassifierCV	0.13
GaussianNB	0.09
LinearDiscriminantAnalysis	0.12
DummyClassifier	0.09
LabelSpreading	0.09
SVC	0.11
LabelPropagation	0.09

'tuple' object has no attribute '__name__'
Invalid Classifier(s)

Slika 3.14. Vremena izvođenja Lazy Predict za klasifikaciju leukemije (kategoričke značajke)

Uvidom u rezultate za kategoričke značajke, tj. podatke sa značajkama prisutnosti gena uočavam kako su LinearSVC, SGDClassifier, Perceptron i PassiveAggressiveClassifier modeli koji imaju točnost, uravnoteženu točnost, ROC AUC i F1 Score od 1.00. To znači da su ti modeli ispravno klasificirali sve instance u skupu podataka.

3.2.4. Random Forests

Random Forests tehnika je također primijenjena na podatke o leukemiji, pokazujući visoku točnost i mogućnost generalizacije. Ova metoda pomaže u sprječavanju prekomjerne prilagodbe modela i poboljšava ukupnu stabilnost klasifikacije. Točnosti na tes-

tnom skupu i za numeričke i kategoričke značajke bila je identična i iznosila je 95.45% što odgovara rezultatima dobivenim korištenjem Lazy Predict biblioteke.

3.2.5. Decision Tree

Kao i kod raka dojke, Decision Tree je korišten za klasifikaciju leukemije. Točnost na prvom skupu (kategoričke značajke prisutnosti gena) iznosi 77.27% dok na drugom skupu (numeričke značajke razine ekspresije gena) iznosi 95.45% što ukazuje na uvjerljivo veću informativnost numeričkih značajki kod modela ove vrste. Dobiveni rezultati odgovaraju onima dobivenim korištenjem Lazy Predict biblioteke.

3.2.6. XGBoost

XGBoost je pokazao dobre rezultate u klasifikaciji leukemije, nudeći brzu obradu i visoku točnost. Točnost na prvom skupu (kategoričke značajke prisutnosti gena) iznosi 90.91% dok na drugom skupu (numeričke značajke razine ekspresije gena) iznosi 95.45% što ukazuje na veću informativnost numeričkih značajki kod modela ove vrste. Dobiveni rezultati ponovo odgovaraju onima dobivenim korištenjem Lazy Predict biblioteke.

4. Zaključak

Modeli strojnog učenja pokazali su se pouzdanim alatima u klasifikaciji raka dojke i leukemije, što ukazuje na njihovu učinkovitost u analizi bioloških podataka. Logistička regresija, SVM s linearnom jezgrom, Random Forests, Decision Tree i XGBoost su metode koje su najdetaljnije obrađene za klasifikaciju različitih vrsta raka u svrhu izrade ovog seminarskog rada, pružajući stabilne i pouzdane rezultate. Korištenje Lazy Predict biblioteke omogućilo je brzo testiranje različitih modela i odabir najboljeg modela za određeni skup podataka. Rezultati istraživanja ukazuju na važnost odabira odgovarajuće metode strojnog učenja za analizu bioloških podataka, s obzirom na složenost i raznolikost podataka. U budućnosti, moguće je proširiti istraživanje na druge vrste raka i bolesti, te primijeniti naprednije tehnike strojnog učenja kako bi se poboljšala točnost i generalizacija modela.