Matematički fakultet

Seminarski rad iz Istraživanja podataka II

Analiza patogenosti Escherichia coli bakterije

Autori:
Anđelija Vasiljević
Mihailo Dedić

Profesor:
Prof. Dr. Nenad Mitić

16. januar 2024.

Sadržaj

1 Uvod

U današnjem dobu, sveprisutna digitalizacija bioloških podataka omogućava nam dublje razumevanje organizama i njihovih karakteristika. Ovaj rad istražuje bakteriju Escherichia coli, često poznatu kao E. coli. Iako većina slojeva E. coli nije štetna, određeni sojevi mogu predstavljati rizik po zdravlje ljudi. U skladu s tim, fokus ovog rada je na analizi patogenosti bakterije Escherichia coli. Radi boljeg razumevanja biološke osnove problema, preporučuje se čitanje narednog poglavlja - Biološka osnova, posebno ako čitalac nije upoznat s biološkim pojmovima.

Ovo istraživanje leži u proučavanju razlika u upotrebi kodona u kompletnoj sekvenci (genomu) Escherichia coli i patogenim ostrvima ove bakterije. Pojam "upotreba kodona"odnosi se na kombinaciju broja pojavljivanja svih kodona u genomu. Za svaki genom, patogeno ostrvo ili deo genoma bez patogenog ostrva urađeno je brojanje kodona putem metode pomerajućeg prozora. Konačna tabela sadržaće ove kombinacije broja pojavljivanja kodona i dodatno svaki red tabele će imati ciljni atribut koji označava da li kombinacija predstavlja genom, patogeno ostrvo ili deo genoma bez patogenog ostrva. Nakon prikupljanja podataka, cilj je konstruisati model klasifikacije koji će moći da predviđa ciljni atribut na osnovu kombinacije broja pojavljivanja svih kodona. Shodno tome, ovaj model bi trebalo da omogući da se razume kako su genomi prilagođeni različitim funkcijama i uslovima, kao što je patogenost u nasem slucaju.

2 Biološka osnova

Pre nego što se upustimo u analizu podataka, upoznaćemo se sa osnovnim biološkim pojmovima neophodnim za razumevanje ovog rada.

2.1 Genomi, nukleotidi i kodoni

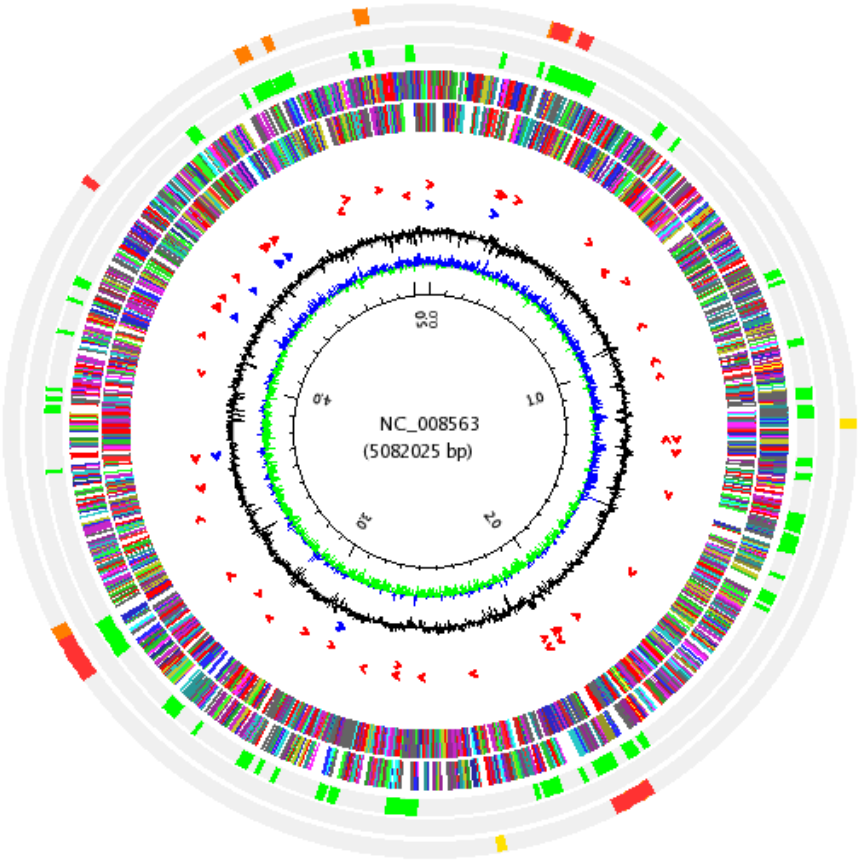
Genomi predstavljaju kompletnu genetsku informaciju jednog organizma. Sastoje se od nukleinskih kiselina Dezoksiribonukleinska kiselina (DNK) i Ribonukleinska kiselina (RNK). U slučaju većine živih bića (osim virusa) genomi su sačinjeni isključivo od DNK. Osnovne gradivne jedinice nukleinskih kiselina su nukleotidi. U slučaju DNK, postoje 4 tipa nukleotida - Adenin, Guanin, Timin, Citozin. Kodoni predstavljaju triplete nukleotida (tri susedna nukleotida).

2.2 Patogenost, patogena ostrva

Patogena bakterija se odnosi na bakteriju koja može izazvati bolest ili infekciju kod organizama, uključujući ljude, životinje ili biljke. Kada kažemo da je bakterija "patogena", to znači da može uzrokovati štetne posledice (bolesti) kod svog domaćina. Kod E. coli, većina sojeva su bezopasni i prisutni u probavnom traktu ljudi i drugih organizama. Međutim, patogeni sojevi E. coli mogu izazvati razne bolesti.

Patogena ostrva su ključni delovi genoma koji su odgovorni za patogenost bakterije. U okviru ovog rada pojavljuju se sledeće vrste patogenih ostrva:

- PAI - patogena ostrva za koja je utvrđeno da se nalaze u genomu
- cPAI - kandidati za patogena ostrva
- nPAI - malo verovatna patogena ostrva



slika 2.1: E. coli NC_008563 sa PAIDB[2] baze podataka (na spoljašnjem krugu se nalaze patogena ostrva obojena crvenom (PAI), narandžastom (cPAI) i žutom (nPAI) bojom)

3 Priprema podataka

3.1 Prikupljanje podataka

Naš skup podataka obuhvata 95 genoma, preuzetih u FASTA formatu sa nacionalne baze podataka za bioinformatiku (NCBI)[1] Sjedinjenih Američkih Država. Službene oznake 95 kompletnih sekvenci su:

NC_017626, NC_017627, NC_013364, NC_013365, NC_013366, NC_013361, NC_013369, NC_013353, NC_013354, NC_018650, NC_018654, NC_018661, NC_018662, NC_018658, NC_018659, NC_018666, NC_008253, NC_011748, NC_017631, NC_008563, NC_009837, NC_009838, NC_020163, NC_010468, NC_012892, NC_012971, NC_013941, NC_017646, NC_004431, NC_017625, NC_017638, NC_011601, NC_009786, NC_009788, NC_009790, NC_009801, NC_011353, NC_011745, NC_002655, NC_007414, NC_017633, NC_012947, NC_009800, NC_011741, NC_011750, NC_017628, NC_007779, NC_010473, NC_012759, NC_020518, NC_000913, NC_016902, NC_016904, NC_017660, NC_011993, NC_022364, NC_017644, NC_017634, NC_017659, NC_017663, NC_022370, NC_012967, NC_017656, NC_017657, NC_011742, NC_011747, NC_011415, NC_011419, NC_013654, NC_013655, NC_010488, NC_010498, NC_002128, NC_002695, NC_013008, NC_013010, NC_017630, NC_017632, NC_011739, NC_011749, NC_011751, NC_017639, NC_017641, NC_017642, NC_017645, NC_007941, NC_007946, NC_017635, NC_017637, NC_017664, NC_017665, NC_017906, NC_017907, NC_017652, NC_017651.

Iz svakog genoma, korišćenjem podataka iz PAIDB[2] baze patogenih ostrva, izdvojena su sva patogena ostrva (PAI, cPAI, nPAI). Za svaku od ovih grupa formirana je tabela. Najvažniji atributi u ovim tabelama su početak i kraj patogenog ostrva u sekvenci, kao i njegova dužina. Na primer, za genom NC_017626 rezultat izgleda ovako:

- PAI ostrva: ovaj genom ne sadrži PAI ostrva, tako da tabela nije dodata
- cPAI ostrva:

	cPAI Number	Start	End	Size	No. of ORFs	G+C content
1	1	233779	275861	42083	39	50.64 %
2	2	2272389	2345178	72790	61	53.02 %
3	3	3249573	3277026	27454	33	35.86 %
4	4	3387125	3408657	21533	22	47.12 %
5	5	3429556	3495247	65692	62	48.4 %
6	6	4347120	4377124	30005	43	57.28 %
7	7	4822309	4852368	30060	32	49.55 %
8	8	4881858	4890349	8492	10	46.89 %
9	9	4918128	4927380	9253	14	49.62 %
10	10	5081722	5109850	28129	29	44.35 %
11	11	5118217	5152041	33825	53	52.82 %

slika 3.1

- nPAI ostrva:

	nPAI Number	Start	End	Size	No. of ORFs	G+C content
1	1	1316133	1327832	11700	11	51.45 %

slika 3.2

3.2 Brojanje kodona

Centralni deo pripreme podataka predstavlja brojanje kodona. Za nukleotide A, T, C i G formiramo sve moguće varijacije sa 3 elementa. Njih ima ukupno $4^3 = 64$. Za brojanje kodona je korišćen metod pomerajućeg prozora. Metod pomerajućeg prozora je tehnika u analizi sekvenci koja podrazumeva kretanje "prozora"fiksne veličine duž sekvence korak po korak. U kontekstu brojanja kodona u genomima, prozor obuhvata određeni broj nukleotida, u ovom slučaju tri. Prozor se pomera duž sekvence dok ne dođe do kraja.

Za genome je korišćena prva generacija sekvenciranja. Prilikom korišćenja ove metode moguće je da mašina ne prepozna odgovarajući nukleoid, što se obeležava jednim od narednih slova: Y, R, S, W, M, K, H, D, B, V, N. Ovo ponašanje podrazumeva da sekvencer nije siguran koji nukleoid se nalazi na tom mestu. Na primer, slovo Y nam govori da mašina nije sigurna da li je u pitanju citozin(C) ili timin(T). Kod kompletnih sekvenci gde se pojavi ovaj slučaj, brojanje kodona se nastavlja dalje bez uključivanja takvog kodona.

Za svaki genom, prolazimo kroz njegovu celokupnu sekvencu i za svaki ispravan kodon, uvećavamo njegov brojač u tabeli. Opisani postupak izvršavamo i za svako od patogenih ostrva.

Na kraju, želimo da utvrdimo i broj odgovarajućih kodona u genomima koji se nalaze van patogenih ostrva. Ovaj broj ćemo dobiti tako što od broja odgovarajućih kodona u kompletnom genomu, oduzmemo broj odgovarajućih kodona iz svih patogenih ostrva.

3.3 Formiranje konačne tabele za model

Nakon izvršenog brojanja kodona u različitim sekvencama, dolazimo i do završnog koraka - formiranje konačne tabele za ulaz u model.

Svaki red ove tabele predstavlja broj odgovarajućih kodona za neku od sekvenci. Koja je to vrsta sekvence, definisano je sa prvim atributom - Tip, koji ima sledeće vrednosti:

- P - PAI ostrvo
- C - cPAI ostrvo
- N - nPAI ostrvo
- K - kompletan genom
- R - kompletan genom bez PAI,cPAI i nPAI ostrva

Kao poslednju stvar dodajemo kolonu ID - službena oznaka tog genoma. Zajedno sa izbrojanim kodonima konačna tabela je završena.

Prikaz nekoliko redova iz tabele na slici **3.3**:

	Tip	ID	AAA	AAC	AAG	AAT	ACA
567	C	NC_017651	844	647	419	557	563
568	C	NC_017651	523	351	253	382	375
569	C	NC_017651	2531	1934	1346	2046	1667
570	C	NC_017651	679	420	341	523	310
571	C	NC_017651	513	395	253	321	365
572	N	NC_017651	197	138	111	198	104
573	N	NC_017651	305	246	202	219	205
574	K	NC_017626	124519	92605	71352	94964	67223
575	R	NC_017626	115714	86144	66327	87559	61586
576	K	NC_017627	2658	1811	1262	1980	1855
577	R	NC_017627	2243	1540	1096	1697	1615
578	K	NC_013364	127814	94230	72789	97125	69544

slika 3.3

3.4 Normalizacija vrednosti kodona

Kako bismo bolje identifikovali koji kodoni se često javljaju, a koji ređe, vršimo normalizaciju vrednosti odgovarajućih kodona. S obzirom na to da kompletne sekvence imaju po nekoliko miliona nukleotida, a patogena ostrva samo nekoliko hiljada, normalizaciju ćemo sprovesti deljenjem svake vrednosti kodona sa dužinom odgovarajuće sekvence.

Na primer, ako posmatramo vrednost određenog kodona u nekoj sekvenci, ova vrednost će biti podeljena dužinom te sekvence. Ovaj postupak nam omogućava da dobijemo normalizovanu vrednost koja odražava relativnu učestalost tog kodona u okviru date sekvence. Prikaz nekoliko redova iz konačne tabele na slici **3.4**:

	ID	Tip	AAA	AAC	AAG	AAT	ACA
567	NC_017651	C	0.024163994502977553	0.01852382043060009	0.011996106275767292	0.015947091158955564	0.016118873110398533
568	NC_017651	C	0.02687426134319922	0.018036072144288578	0.013000359693746468	0.01962900159292945	0.019269307846462157
569	NC_017651	C	0.023643599133099172	0.018066661684477993	0.012573798669755623	0.01911292130632987	0.015572453478813243
570	NC_017651	C	0.03178392547863128	0.019660160089875017	0.015962177596779478	0.024481580302391986	0.014511070542526799
571	NC_017651	C	0.025117508813160987	0.019339992166079123	0.012387387387387387	0.015716803760282023	0.017871132001566783
572	NC_017651	N	0.020774016661394074	0.014552356849098386	0.01170515659601392	0.020879468522619425	0.010966993567436465
573	NC_017651	N	0.021812200529214044	0.017592791246513622	0.014446113137381105	0.015661875134091396	0.014660659372094686
574	NC_017626	K	0.0237542057128446	0.01766604470031059	0.013611658349512026	0.01811606575152848	0.012823978434090801
575	NC_017626	R	0.02380469918265335	0.017721554923263307	0.01364479909680634	0.018012648907944972	0.012669479958024865
576	NC_017627	K	0.02345032025832407	0.015977626030031937	0.011134049723854393	0.017468635858345245	0.016365817937995165
577	NC_017627	R	0.022630506285691226	0.015537663700385415	0.011057973646508061	0.017121698246463667	0.01629436810137821
578	NC_013364	K	0.023796717120234918	0.017543967438932637	0.013552030626259873	0.018082965483458904	0.012947868742153576

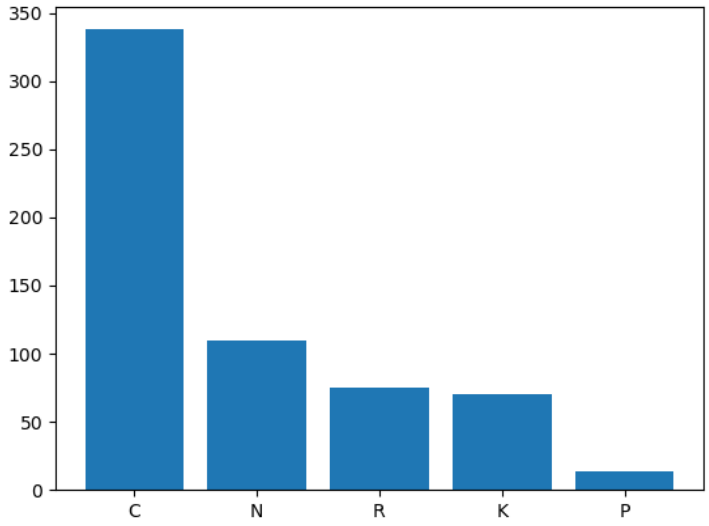
slika 3.4

3.5 Pretprocesiranje

Sada kada imamo normalizovanu tabelu, sledeći korak je sprovođenje pretprocesiranja kako bismo je pripremili za korišćenje u modelu. Na samom početku delimo skup na trening i test. Koristićemo 80 posto skupa za trening i 20 posto za test. Svaki od ovih skupova dalje je podeljen na ulazne atribute i ciljni (klasni) atribut y.

Prva provera koju smo izvršili bila je da li postoje nedostajuće vrednosti u našim podacima. Rezultat ove provere pokazao je da nemamo nedostajućih vrednosti ni u jednoj od kolona tabele.

Nakon toga, vršimo proveru balansiranosti klasa. Ona je prikazana na sledecoj slici **3.5**:

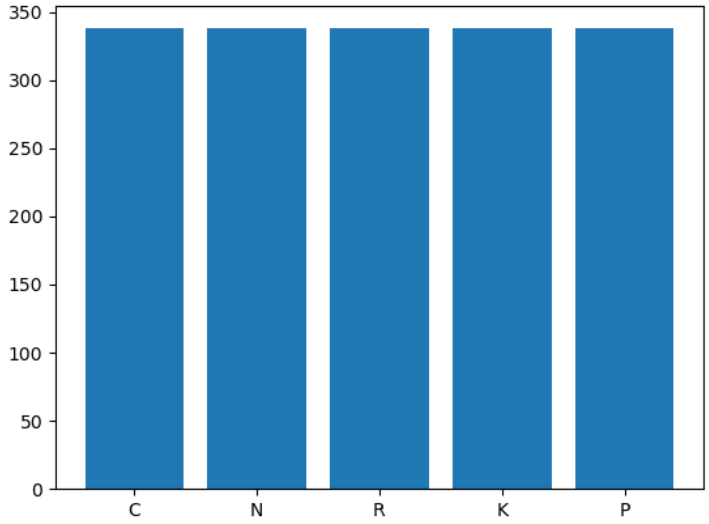


slika 3.5

Kao što se vidi sa slike 3.5, klase su veoma nebalansirane. Efikasnost različitih modela može značajno varirati u radu s nebalansiranim klasama. Generalno, modeli koji su osetljivi na nebalansiranost klasa često imaju problema u prepoznavanju manjinskih klasa. Mi ćemo pokušati da se suočimo sa tim na dva načina.

3.5.1 Tehnika naknadnog uzorkovanja (SMOTE)

Prva tehnika koju ćemo primeniti je jedna od tehnika naknadnog uzorkovanja (en. oversampling) poznata kao SMOTE(en. Synthetic Minority Oversampling Technique). SMOTE konstruiše dodatne instance klasa sa manjim brojem elemenata kako bi se smanjila nebalansiranost klasa. Iako ne treba koristiti ovu tehniku na test skupu, primenom na trening skupu nadamo se da će model bolje naučiti o klasama s manjim brojem elemenata. Na slici **3.6** možemo videti odnos broja elemenata u klasama posle primene SMOTE tehnike:



slika 3.6

3.5.2 Dodavanje težina klasama

Neki modeli podržavaju dodavanje težina klasama (en. class weights). Ovaj postupak izvodimo tako što određujemo klasu s najmanje elemenata i njoj dodeljujemo težinu 1. Sve ostale težine određujemo u odnosu na broj elemenata klase s najmanje elemenata. Mapu koja predstavlja težine dodeljene svim klasama možemo videti na slici **3.7**:

```
{'C': 0.04142011834319527,
 'N': 0.1272727272727726,
 'R': 0.18666666666666668,
 'K': 0.2,
 'P': 1}
```

slika 3.7

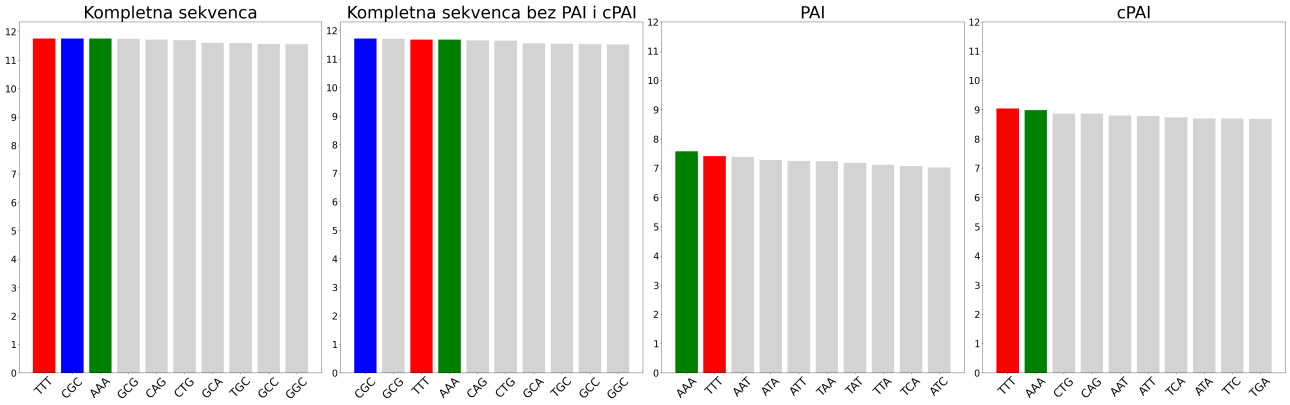
S obzirom na veoma mali skup podataka koje imamo nećemo primenjivati tehnike naknadnog odstranjivanja elemenata (en. undersampling). Takođe iz istog razloga nećemo koristiti tehnike smanjivanja broja atributa.

4 Analiza kodona

Pre nego što se posvetimo procesu klasifikacije, analiziraćemo dobijene podatke kako bismo stekli dublje razumevanje genomske strukture. Fokusiraćemo se na analizu kodona u određenim genomima bakterije E. coli kako bismo identifikovali potencijalno značajne obrasce.

4.1 Analiza E. coli NC_013364

Prvi korak u analizi je posmatranje primera E. coli NC_013364, koja sadrži jedno PAI i osam cPAI ostrva. Analiziraćemo tri najzastupljenija kodona u ovoj sekvenci. Najčešći kodoni identifikovani brojanjem su TTT, CGC i AAA. Za cPAI su sumirane vrednosti svih 8 sekvenci u jednu radi jednostavnijeg prikaza. Rezultati su predstavljeni na logaritamskoj skali kako bi se bolje vizualizovala distribucija. Grafikon prikazan na slici **4.1** ilustruje ove rezultate.

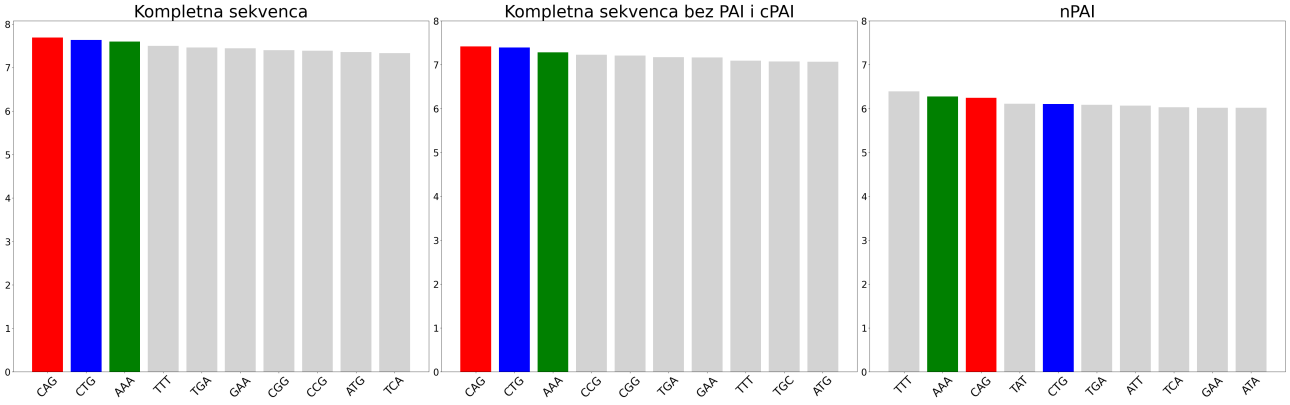


slika 4.1: Distribucija najzastupljenijih kodona u E. coli NC_013364

Na svakom pojedinačnom grafikonu prikazano je deset najčešćih kodona za određenu sekvencu. Zapažamo da i PAI i cPAI imaju kodone AAA i TTT među prvih nekoliko vrednosti. Kada iz kompletnih sekvenci izbacimo ostrva, primetimo da se prve tri vrednosti više ne poklapaju. Ovi zaključci pružaju uvid u to kako određeni kodoni u ostrvima mogu promeniti kompletnu sekvencu i doprineti patogenosti.

4.2 Analiza E. coli NC_013366 (Plazmid)

Drugi primer koji analiziramo odnosi se na E. coli NC_013366, koja sadrži 2 nPAI ostrva. Kao i prethodno, vrednosti dve nPAI sekvence su sumirane u jednu radi pojednostavljenja analize. Rezultati analize najčešćih kodona u plazmidu prikazani su na slici **4.2**.



slika 4.2: Distribucija najzastupljenijih kodona u E. coli NC_013366 (Plazmid)

Iako se radi o plazmidu, važno je napomenuti da plazmidi mogu nositi gene koji čine bakterije patogenim. E. coli plazmid u ovom primeru sadrži samo dva nPAI ostrva, što strukturu konacne sekvence bez tih ostrva se ne menja mnogo.

Analiza kodona može da pruži bitan uvid u genomske karakteristike bakterija, posebno u vezi sa ostrvima koja često nose genetske informacije. Ovi rezultati služe kao osnova za dalje istraživanje i razumevanje patogenosti bakterija E. coli.

5 Modeli i rezultati

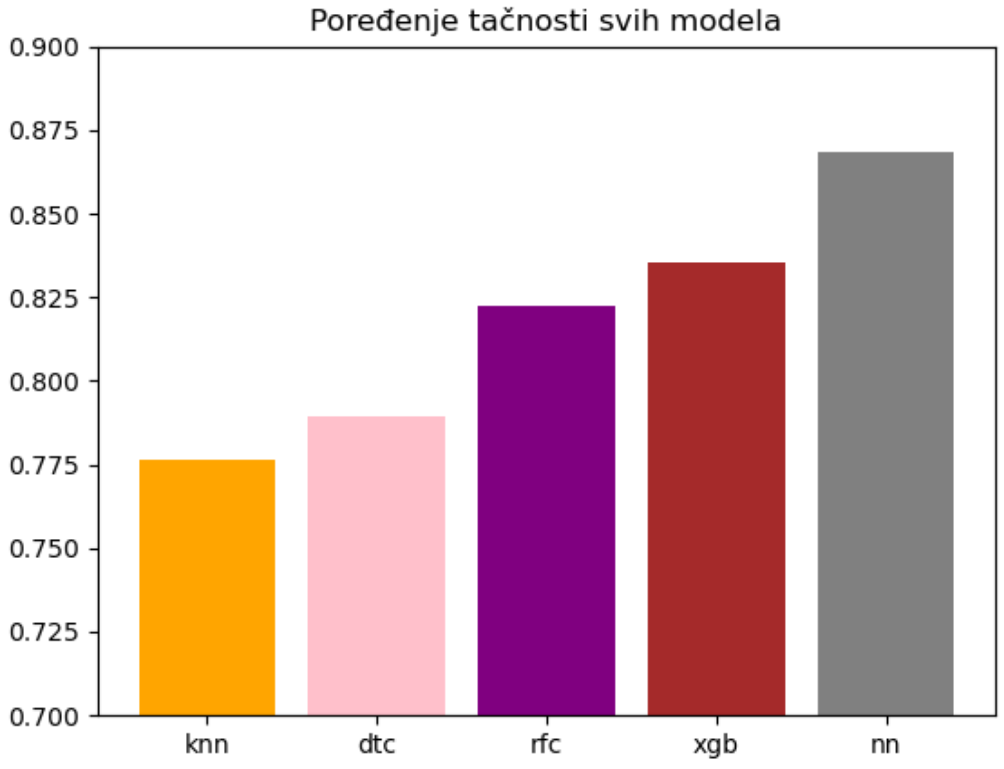
U okviru ovog rada primenjivaćemo modele klasifikacije na naš skup podataka.
Svaki od njih ćemo ubaciti u model matričnog pretraživanja (en. Grid Search(GS)). GS na osnovu matrice parametara i odgovarajućeg estimatora koji predstavlja jedan od naših modela vraća najbolji model na osnovu nekog kriterijuma određivanja najboljeg za odgovarajući validacioni skup.

Model ćemo trenirati sa nekoliko vrsta podataka. Prvo ćemo trenirati model sa običnim trening podacima, zatim nastavljamo sa trening podacima nad kojima je primenjena SMOTE tehnika, i ako je moguće, trening podacima kojima su dodeljene težine klasama.

Modeli koji su korišćeni su:

1. K najbližih suseda (KNN)
2. Drvo odlučivanja (DTC)
3. Nasumična šuma (RFC)
4. Ekstremno gradijentno pojačavanje (XGBoost)
5. Neuronske mreže

Za svaki model ćemo uzeti njegovu najbolju tačnost na test skupu koju je ostvario u jednom od njegovih trening skupova (običan,SMOTE i težine). Poređenje tačnosti modela možemo videti na slici **6.1**.



slika 5.1

Kada imamo neuravnotežene klase, gledanje samo tačnosti (accuracy) može biti obmanjujuće, jer model može postići visoku tačnost jednostavno predviđajući dominantnu klasu, zanemarujući manjinske klase. U takvim situacijama, korisno je analizirati i druge metrike kako biste dobili bolji uvid u stvarne performanse modela. Zato na **Tabela 1**, **Tabela 2** i **Tabela 3** možemo videti preciznost, odziv i f1 meru svih modela na test skupu za svaki od trening skupova redom.

Model	Preciznost	Odziv	F1 mera
K najbližih suseda	76.52%	77.63%	77.06%
Drvo odlučivanja	76.9%	78.95%	77.68%
Nasumična šuma	82.16%	82.23%	81.64%
Ekstremno gradijentno pojačavanje	80.93%	81.58%	81.2%
Neuronske mreže	**87.96%**	**86.84%**	**86.73%**

Tabela 1: Upoređivanje modela

Model	Preciznost	Odziv	F1 mera
K najbližih suseda	81.1%	75%	77.35%
Drvo odlučivanja	79.54%	76.97%	77.52%
Nasumična šuma	82.37%	81.58%	81.7%
Ekstremno gradijentno pojačavanje	83.84%	83.55%	83.63%
Neuronske mreže	**84.74%**	**84.21%**	**84.15%**

Tabela 2: Upoređivanje modela sa SMOTE

Model	Preciznost	Odziv	F1 mera
K najbližih suseda	-	-	-
Drvo odlučivanja	79.4%	75.66%	76.6%
Nasumična šuma	81.47%	82.24%	**82.47%**
Ekstremno gradijentno pojačavanje	-	-	-
Neuronske mreže	**82.22%**	**82.89%**	82.41%

Tabela 3: Upoređivanje modela sa težinama

Na osnovu analize rezultata prikazanih u tabelama, možemo izvući nekoliko ključnih zaključaka o performansama različitih modela:

Neuronske mreže ističu se kao najefikasniji model u svim analiziranim uslovima. Njihova veća preciznost, odziv i F1 mera sugerišu na sposobnost modela da bolje generalizuje i prepoznaje obrasce u podacima. Ovaj model se izdvaja kada nisu primenjene tehnike poput SMOTE i klasnih težina. Nasumična šuma pokazuju povećanje u F1 meri kada se primene težine klasa, dok ekstremno gradijentno pojačavanje pokazuje poboljšanje pri korišćenju SMOTE tehnike. Drvo odlučivanja i K najbližih suseda daju identične rezultate za sve tri metode.

6 Zaključak

Ovaj rad predstavlja uvid u genetske karakteristike bakterije Escherichia coli, sa posebnim naglaskom na analizi kodona u kontekstu patogenih ostrva. Kroz primenu tehnike 3-grama, fokusirali smo se na analizu kombinacija tri uzastopna nukleotida, prateći njihovu frekvenciju kroz genomsku sekvencu. Rezultati analize ukazuju na uticaj patogenih ostrva na distribuciju određenih kodona. Sa tehnikom pomerajućeg prozora omogućilo nam je da istreniramo modele što pokazuje da ova metoda može da donese dobre rezultate. Potencijalna poboljšanja koja se mogu razmotriti su pronalaženje boljih modela klasifikacije ili eksperimentisanje s drugim n-gramima.

Literatura

[1] NCBI baza podataka

[2] Pathogenicity Island Database (PAI DB)