Primjena strojnog učenja u metagenomici
Tehnička dokumentacija
Verzija 1.0

Studentski tim: Luka Bulić
Lucia Crvelin
Niko Kaštelan
Mirta Krajinović
Lucija Topolko
Marko Žagar

Nastavnik: izv. prof. dr. sc. Mirjana Domazet-Lošo

Sadržaj

1. Uvod

1.1 Metagenomika

Metagenomika predstavlja područje bioloških znanosti koje istražuje genetsku raznolikost mikrobnih zajednica u različitim okolišima. Tradicionalna genomika fokusira se na proučavanje genoma pojedinih organizama, dok metagenomika omogućava analizu genetskog materijala iz cijelih zajednica mikroorganizama. Ovaj pristup otvara vrata dubljem razumijevanju strukture, funkcije i dinamike mikrobnih ekosustava, pridonoseći ključnim spoznajama o njihovoj ulozi u ekologiji, bolestima i drugim biološkim procesima. Metagenomika se temelji na sekvenciranju genoma iz uzoraka okoliša, kao što su tla, vode, zrak, ali i ljudski probavni trakt ili koža. Tehnološki napredak u sekvenciranju omogućava analizu velikih količina genetskih informacija, često bez potrebe za izolacijom i kultivacijom pojedinih mikrobnih vrsta. Jedan od ključnih izazova metagenomike jest analiza golemih i kompleksnih skupova podataka, često sastavljenih od stotina tisuća ili čak milijuna genoma. Stoga, metagenomika postavlja zahtjeve za razvojem sofisticiranih bioinformatičkih alata i metoda strojnog učenja. Ovaj spoj biologije, tehnologije i računalne znanosti otvara nova vrata za istraživanje mikrobnih ekosustava i njihovu primjenu u područjima poput medicinske dijagnostike, biotehnologije i očuvanja okoliša.

U ovom projektu, provodi se integracija metagenomike s područjem strojnog učenja. Proučavanje strojnog učenja u metagenomici predstavlja važnu komponentu ovog projekta, gdje se koriste napredne tehnologije za analizu metagenomskih podataka i razvoj modela koji omogućuju predviđanje i interpretaciju mikrobnih zajednica.

1.2 Strojno učenje

Strojno učenje predstavlja granu umjetne inteligencije koja se bavi razvojem algoritama i modela s mogućnosti učenja iz podataka, prilagodbi promjenama te donošenju odluka i predviđanja bez eksplicitnog programiranja. Ova disciplina omogućuje sustavima da poboljšavaju svoje performanse s iskustvom te da se nose s kompleksnim zadacima i problemima. Strojno učenje često se kategorizira prema nadziranom i nenadziranom pristupu. U nadziranom strojnom učenju, model se trenira na označenim podacima, gdje su ulazi povezani s odgovarajućim izlazima. Metode korištene u projektu pripadaju metodama nadziranog strojnog učenja. Ovaj pristup omogućuje modelima da generaliziraju i predviđaju nove, neviđene podatke na temelju prethodnog iskustva. Nadzirano strojno učenje dijeli se na metode klasifikacije i metode regresije. Metode klasifikacije fokusiraju se na podatke s diskretnim izlaznim vrijednostima, gdje model procjenjuje kojoj kategoriji ili klasi pripada određeni unos. S druge strane, metode regresije koriste se nad podacima s kontinuiranim, numeričkim izlaznim vrijednostima, pružajući modelima sposobnost predviđanja kvantitativnih varijabli. U kontekstu analize metagenomskih podataka, primjena ovih metoda omogućuje nam klasifikaciju zdravstvenih stanja organizma ili predviđanje specifičnih karakteristika mikrobnih ekosustava. Ovaj pristup nadziranog strojnog učenja omogućuje nam raznovrsne načine obrade podataka, prilagodljive različitim scenarijima analize, te istraživanje potencijalnih veza između bakterijskih vrsta i zdravstvenih uvjeta. U nastavku, istražujemo primjenu strojnog učenja u analizi metagenomskih podataka, istražujući metode klasifikacije s naglaskom na predviđanju zdravstvenih stanja organizma temeljem mikrobnih ekosustava.

Korišteni skupovi podataka obuhvaćaju informacije o pojedinim bakterijskim vrstama te raznolike demografske podatke, poput dobi, indeksa tjelesne mase i zemlje stanovanja, čineći ih ulaznim varijablama. Izlazna podatak o zdravstvenom stanju ispitanika. Cilj je anticipirati ovu izlaznu vrijednost putem izgradnje modela koji se temelji na metagenomskim informacijama o specifičnim bakterijskim vrstama.

U procesu izgradnje modela, važno je izbjeći prenaučenost. Prenaučenost označava situaciju u kojoj model postane prekomjerno prilagođen podacima za treniranje zbog čaga gubi sposobnost generalizacije na nepoznate ili nove podatke. To može rezultirati smanjenom točnošću modela na stvarnim svakodnevnim situacijama, jer je suviše usmjeren na specifičnosti podataka za treniranje. Rješenje za sprječavanje prenaučenosti jest podjela skupa podataka. Dvije ključne podskupine, podskup za treniranje i podskup za ispitivanje, omogućuju modelu učenje na jednom skupu podataka i evaluaciju na drugom, nepovezanom skupu. Time se omogućuje procjena točnosti modela na nepoznatim podacima te pridonosi njegovoj sposobnosti generalizacije. U situacijama kada količina dostupnih podataka nije dovoljna za pouzdano korištenje podskupova za treniranje i ispitivanje, primjenjuje se tehnika unakrsne validacije. Unakrsna validacija uključuje podjelu skupa podataka na više particija, gdje se u svakoj iteraciji jedna particija koristi

za testiranje, dok se preostale koriste za treniranje. Ova metoda pruža višestruke procjene performansi modela, smanjujući time utjecaj specifičnosti pojedinih skupova podataka i omogućavajući bolje procjene očekivane performanse na novim podacima. Tijekom izgradnje modela nad metagenomskim podacima, primijenili smo unakrsnu validaciju kako bismo osigurali pouzdanu evaluaciju i visoku točnost klasifikacije.

1.3 Programska podrška

U sklopu našeg istraživačkog projekta, pristupili smo analizi na dva različita načina: jedna grupa koristila je alat Weka, dok je druga grupa koristila programski jezik Python.

Weka, akronim od "Waikato Environment for Knowledge Analysis," predstavlja snažan otvoreni alat razvijen na Sveučilištu Waikato u Novom Zelandu. Ovaj alat namijenjen je radu s podacima i implementaciji raznolikih tehnika strojnog učenja. Weka pruža širok spektar funkcionalnosti, uključujući učitavanje, analizu, manipulaciju podacima te implementaciju različitih algoritama za klasifikaciju, regresiju, grupiranje i druge zadatke strojnog učenja. U našem projektu, Weka je poslužila za pretprocesiranje podataka, odabir značajki te evaluaciju modela pomoću evaluatorskih metoda poput Correlation i InfoGain, što je bio ključan korak u pripremi podataka za fazu klasifikacije.

S druge strane, Python je poslužio kao alat s bogatim bibliotekama, omogućujući učinkovito učitavanje, manipulaciju i čišćenje podataka. Korištenjem biblioteka poput NumPy i pandas, analizirana smo distribuciju vrijednosti atributa, istražena korelacija među podacima te su identificirane važne značajke. U domeni strojnog učenja, Python se istaknuo, posebice kroz korištenje scikit-learn biblioteke, jedne od najčešće upotrebljavanih biblioteka u tom području. Dodatno, model je optimiziran korištenjem XGBoost algoritma, popularnog algoritma pojačavanja u strojnom učenju.

Nakon izgradnje modela korištenjem algoritama kao što su slučajna šuma i potporni vektori, analizirali smo rezultate, obuhvaćajući različite metrike točnosti modela. Broj točno i pogrešno klasificiranih uzoraka, AUC vrijednosti te matrice zbunjenosti pružile su dublji uvid u performanse svakog pojedinog algoritma. Unakrsna validacija modela bila je ključna u određivanju optimalnih parametara za svaku metodu klasifikacije.

Usporedbom učinkovitosti i rezultata dobivenih kroz različite tehnologije, ovaj pristup pružio nam je cjelovitu sliku o skupu podataka i različitostima u rezultatima klasifikacije između korisnika Weke i Pythona. Ova diverzifikacija pristupa dodala je dubinu i širinu našem istraživanju, omogućujući nam uvid u analizu metagenomskih podataka.

2. Metode

Na točnost predviđanja bolesti utječe nekoliko čimbenika. To su: odabir značajki za treniranje modela, metoda klasifikacije, omjer skupa za treniranje i skupa za testiranje te specifične varijable unutar metoda klasifikacije. U sljedećim odlomcima ćemo ukratko objasniti metode koje smo koristili u našem radu.

2.1 Odabir značajki

Odabir značajki je metoda kojom se iz cijelog skupa značajki odabere određen broj najbitnijih značajki kako bi se model pojednostavio i davao bolje rezultate. Za odabir značajki potrebno je definirati broj značajki koje želimo zadržati te metodu za odabir značajki. Isprobali smo nekoliko metoda te usporedili rezultate koje smo dobili svakom metodom.

2.1.1 Korelacijska metoda

U korelacijskoj metodi se pomoću Pearsonovog korelacijskog koeficijenta mjeri linearna ovisnost između značajki. Prvo se izračunaju sve korelacije između značajki, a zatim se izbacuju značajke koje su prekorelirane kako bismo izbacili redundantnost iz sustava. Nakon toga se odabire određen broj najkoreliranijih značajki sa ciljnim stupcem. Nedostatak ove metode je loše prepoznavanje nelinearnih ovisnosti.

2.1.2 Info gain metoda

Info gain metoda mjeri koliko pojedina značajka doprinosi smanjenju entropije u skupu podataka. Entropija predstavlja nesigurnost ili nepredvidivost u procjeni rezultata, a računa se kao negativna suma umnoška vjerojatnosti pojavljivanja određenih ishoda i logaritma istih. Što neka značajka više smanjuje entropiju to je ona važnija i veća je vjerojatnost da će ju metoda zadržati

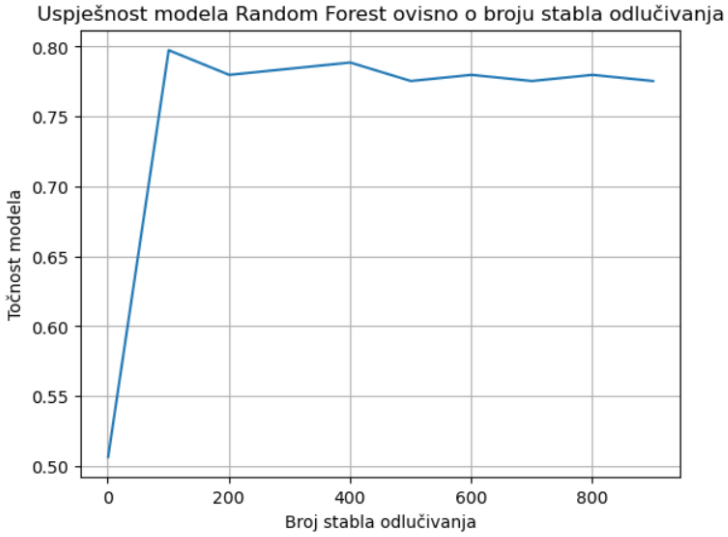
2.2 Metode klasifikacije

Metode klasifikacije ili klasifikatori su algoritmi koji služe za kategoriziranje podataka u određene klase. Kod nas klase predstavljaju bolesti, dok su podaci pojedini pacijenti. Proces klasifikacije smo izvodili u tri faze. Prva faza jest nasumična raspodjela podataka u skupinu za treniranje modela i skupinu za testiranje modela. Pokazalo se da najbolje rezultate model daje za testnu skupinu od 20% ukupnih pacijenata. Slijedi treniranje modela nad testnim skupom, gdje svaki od algoritama na svoj način izgradi algoritam kategorizacije. Na kraju smo proveli testiranje učinkovitosti modela te usporedili rezultate pojedinih klasifikatora.

2.2.1 Slučajna šuma

Slučajna šuma je algoritam koji se bazira na izgradnji unaprijed zadanog broja stabla odlučivanja. Algoritam slučajne šume se sastoji od četiri koraka kroz koje prolazi. Prvi korak je odabir podataka gdje se za svako stablo odabire bootstrap uzorak, odnosno uzorak nasumičnih redaka sa mogućim ponavljanjem. Potom se za svako stablo bira podskup značajki. Nasumičnim odabirom redaka i atributa u prva dva koraka se postiže raznolikost stabla što doprinosi generalizaciji modela. Nakon toga svako stablo rekurzivno stvara čvorove tako da bira značajku koja najbolje kategorizira podatke sve dok se ne izgradi stablo. Na svakom čvoru će se uzorak usmjeriti lijevo ili desno, ovisno o vrijednosti atributa koji se nalazi na tom čvoru, sve dok ne dođe do lista stabla koji predstavlja odluku stabla o kategoriji. Nakon što uzorak prođe kroz sva izgrađena stabla, uzima se kategorija koja se pojavila u većini stabla odlučivanja.

Prednost slučajne šume je što nasumičnost stabla smanjuje mogućnost prenaučenosti, a odabirom više stabla odlučivanja postižemo robusnost modela. Zbog toga se slučajna šuma smatra vrlo preciznim algoritmom.



Slika 1 - Utjecaj broja stabla odlučivanja na uspješnost modela

2.2.2 Stroj potpornih vektora

Stroj potpornih vektora (SVM) je model koji se može koristiti za regresiju i klasifikaciju. Osnovna ideja ovog modela jest smjestiti podatke u n-dimenzionalan prostor te zatim pronaći (n-1)-dimenzionalnu ravninu koja najbolje razdvaja podatke u ciljne kategorije. Iako je stroj potpornih vektora predviđen za binarnu klasifikaciju, pomoću njega je moguće raditi i višeklasnu klasifikaciju. To se može ostvariti na dva načina. Prvi način je One-to-One gdje se za svaki par ciljnih kategorija napravi zaseban SVM te se pronalazi granica koja dijeli podatke iz te dvije kategorije. Drugi način je One-to-Rest gdje se za svaku kategoriju gradi zaseban SVM gdje se pronalazi granica koja dijeli podatke iz te kategorije svih ostalih kategorija. Na kraju svaki SVM daje glas za jedan od dvije kategorije koje mu pripadaju te kategorija s najviše glasova postaje konačno predviđanje modela.
Prednost stroja potpornih vektora je efikasnost u visoko dimenzijalnim prostorima te mogućnost rada s nelinearnim podacima. Metoda One-to-Rest je zbog manjeg broja potrebnih SVM-a jest resursno manje zahtjevna te se većinski koristi u praksi.

2.2.3 Logistička regresija

Logistička regresija transformira linearne kombinacije ulaznih značajki pomoću logističke funkcije tako da izlaz modela bude u rasponu od 0 do 1 i predstavlja vjerojatnost pripadnosti jednoj od klasa. Tijekom treninga, model optimizira koeficijente koji se nalaze uz značajke kako bi minimizirao grešku između stvarnih kategorija i predviđenih vjerojatnosti. Višeklasna klasifikacija se izvodi metodom One-to-Rest koja je analogna onoj iz stroja potpornih vektora.

2.2.4 XGBoost

XGBoost ili eXtreme Gradient Boosting je ansambl metoda za klasifikaciju i regresiju što znači da koristi nekoliko slabijih modela te tako stvara snažan i robustan model. Konkretno, XGBoost koristi stabla odlučivanja kao osnovne modele, a gradient boosting za poboljšanje njihove preciznosti. Nakon izgradnje stabla odlučivanja algoritam koristi optimizacijske tehnike gradijentnog spusta što znači da kroz iteracije prilagođava težine svakog stabla te na taj način ispravlja greške prethodnih iteracija.

Prednost XGBoosta je njegova sposobnost rada sa velikim i visokodimenzionalnim podacima. Osim toga XGBoost ima ugrađenu sposobnost regularizacije što sprječava prenaučenost modela.

3. Podaci

Podatke koje smo koristili u analizi smo preuzeli s repozitorija istraživanja (Pasolli et al., 2016.) čije smo rezultate htjeli replicirati i potvrditi. Podaci su sakupljeni iz osam različitih studija te formirani u dva skupa podataka. U prvom skupu su prikazani udjeli pojedinih bakterija u mikrobiomu, dok su u drugom skupu prikazane prisutnosti taksonomskih markera. U našem projektu, usredotočili smo se na prvi navedeni skup podataka.

Budući da su podaci kombinirani iz 8 različitih izvora, susreli smo se s različitim nedosljednostima koje smo trebali razriješiti prije analize samih podataka. Prvo smo grupirali iste ili slične bolesti u jedinstvene grupe. Na primjer, pretili pacijenti su u nekim redovima bili označeni kao “obese”, dok su u drugima bili označeni kao “overweight”. Zatim smo izbacili neke od bolesti za koje nismo imali dovoljno velike uzorke pacijenata da bismo mogli provesti kvalitetnu analizu nad njima. Potom je bilo potrebno na neki način izmijeniti polja ili retke s nedostajućim vrijednostima kako bi algoritmi klasifikacije radili. Imali smo nekoliko pristupa: zamjenjivanje sa srednjom vrijednošću u numeričkim stupcima, brisanje cijelih redaka sa nedostajućim vrijednostima te zamjenjivanje sa vrijednostima koje su izvan originalne domene vrijednosti (na primjer -1 za numeričke stupce). Posljednji pristup je davao najbolje rezultate.

Nakon što smo riješili probleme s podacima, bilo je potrebno dalje oblikovati podatke kako bismo postigli što bolje rezultate. Trebalo je uravnotežiti broj zdravih i bolesnih pacijenata, kao i brojeve pacijenata s različitim bolestima. Osim toga, neki od demografskih atributa bili su prekorelirani s određenom bolesti koju je trebalo previdjeti. Neki primjeri takvih atributa su “diabetic” koji je prekoreliran s dijabetesom i “bmi” koji je prekoreliran s pretilosti. U višeklasnoj klasifikaciji bilo je potrebno ukloniti takve stupce kako ne bismo dobili jako precizne, ali nerelevantne rezultate. U binarnoj klasifikaciji nije bilo potrebno ukloniti sve takve stupce, već samo one koji su prekorelirani s bolesti koju pokušavamo predvidjeti. Na primjer, "bmi” nećemo ukloniti u binarnoj klasifikaciji gdje predviđamo je li pacijent obolio od dijabetesa. Iako je “bmi” jedan od atributa koji izravno utječe na pretilost, nije nam poznata izravna veza s dijabetesom.

4. Rezultati

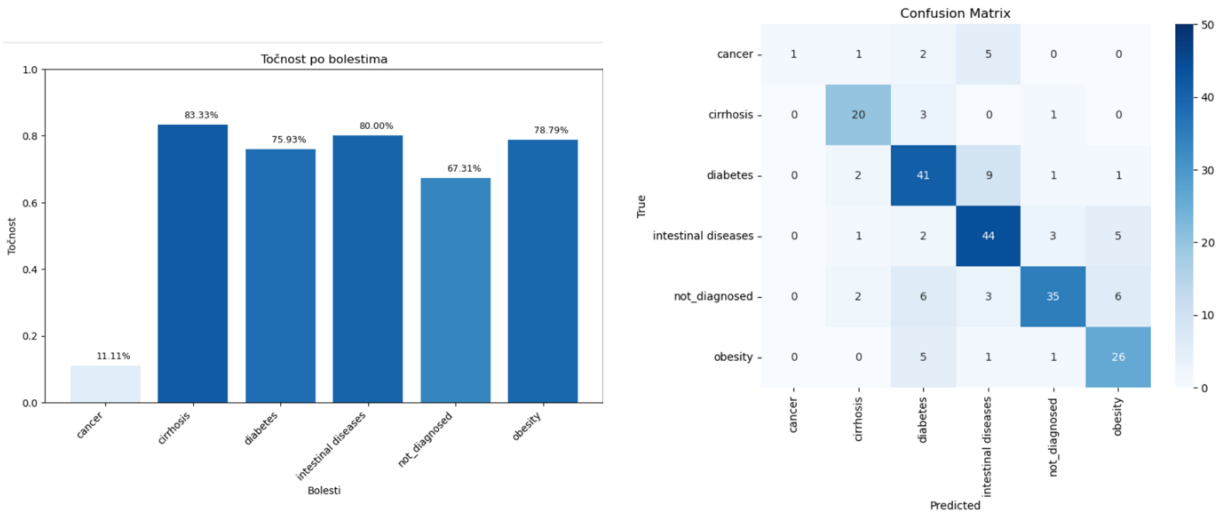
4.1 Mjere kvalitete modela

Naš je projektni tim bio podijeljen u dvije skupine. Grupa 1 rješavala je problem koristeći programski jezik Python te njemu pridružene biblioteke za strojno učenje. Grupa 2 rješavala je zadani problem u programskom okruženju WEKA koji se služi programskim jezikom Java.

Obje skupine dobile su neovisne rezultate koristeći različita programska okruženja. Osnovna mjera kvalitete modela bila je točnost u klasifikaciji. Točnost se odnosila na postotak točno svrstanih entiteta u skupu entiteta odabranom za taj model. Detaljniji prikaz točnosti modela dobiven je iz matrice zabune (*confusion matrix*) prema kojoj je uočeno koji su entiteti krivo klasificirani te koja im je pogrešno dodijeljena klasa. Ova nam je analiza dala uvid u osjetljivost i specifičnost naših klasifikatora, tj. sposobnost da se točno klasificiraju bolesni i zdravi pojedinci.

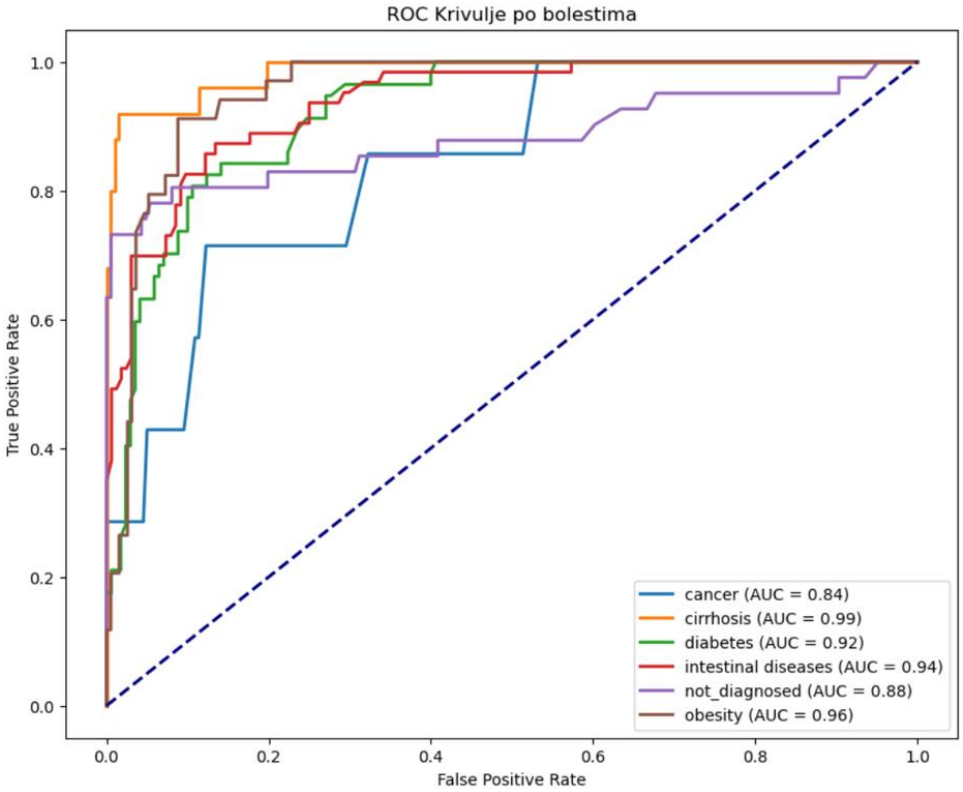
ROC (*receiver operating characteristic*) krivulja grafički prikazuje stopu prirasta stvarno pozitivnih rezultata (y-os) na prirast lažno pozitivnih rezulrata (x-os). Općenito, strmiji prirast pri malim vrijednostima na x-osi te ranije postizanje platoa krivulje označava bolju točnost modela. Numerički, ROC krivulja se može izraziti površinom ispod krivulje (*area under curve, AUC*) koja je također veća što je veća točnost modela.

Grupa 1 klasifikacijski je problem riješila koristeći algoritam slučajne šume. Korištena je InfoGain metoda za provjeru doprinosa točnosti predikcije pojedinih značajki, a broj značajki ograničen je na 150. Broj nedijagnosticiranih pacijenata smanjen je na 250. Riječ je bila o višeklasnom klasifikatoru koji je u obzir uzimao klase “cancer”, “obesity”, “intestinal diseases”, “cirrhosis”, “diabetes” i “not_diagnosed”. Koristeći navedene postavke dobiven je model s prikazanim točnostima za svaku od pojedinih klasa – Slika 2.



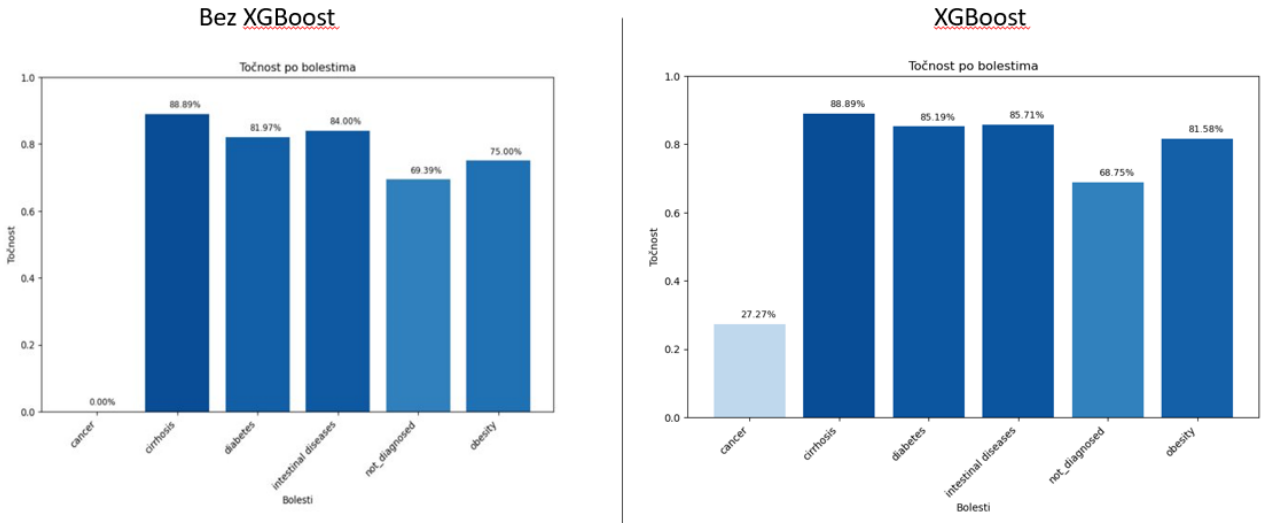
Slika 2 - Točnosti po klasama i matrica zabune

Za pojedinačne je bolesti također izvedena i ROC krivulja te su izračunate pojedinačne AUC vrijednosti za svaku klasu – Slika 3.



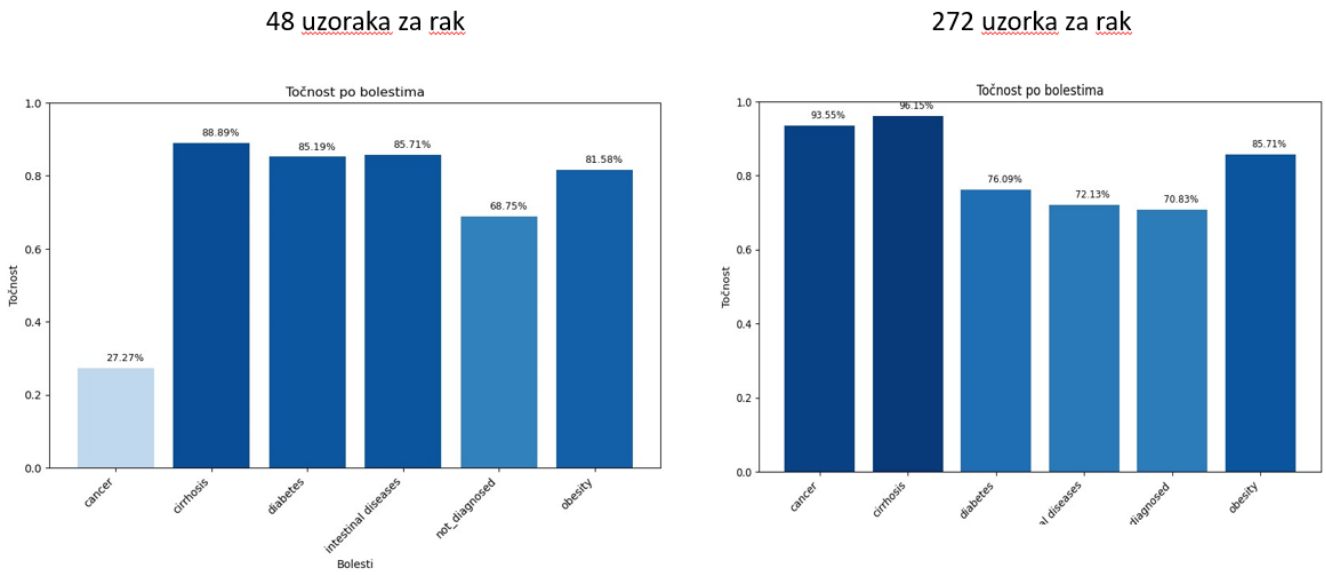
Slika 3 – ROC krivulje i AUC za svaku pojedinačnu klasu

U svrhu unaprjeđenja modela primijenjena je na isti skup značajki i entiteta i XGBoost metoda. Korištenjem ove metode znatno su poboljšane točnosti klasifikacije za sve kategorije – Slika 4.



Slika 4 – Usporedba točnosti prije i nakon primjene XGBoost metode

Unatoč primijenjenim metodama, točnost klasifikacije za kategoriju "cancer" ostala je izrazito niža od ostalih. Procijenjeno je da je uzrok tome mali broj instanci s oznakom "cancer" te je provedeno "umjetno povećavanje" broja instanci koje pripadaju toj kategoriji. Navedenom se tehnikom konačno dobio i visok rezultat točnosti za predviđanje kategorije cancer – Slika 5.



Slika 5 – Usporedba točnosti prije i nakon povećanja broja instanci "cancer" kategorije

Grupa 2 primijenila je algoritam šume J48 u sklopu WEKA programa. U ovom je slučaju pristup rješavanju problema bio kroz seriju binarnih klasifikatora za pojedinu bolest. U početku su ostavljene kao značajke podatci o svim bakterijama te su dobivene navedene točnosti, osjetljivosti i specifičnosti - Tablica 1.

Tablica 1 - Točnosti i osjetljivosti binarnih klasifikatora za pojedinačne bolesti

Značajke	Klase	Točnost	Osjetljivost	Specifičnost
Sve bakterijske populacije (S.B.P.)	"n", "ulcerative_colitis"	**95.1743%**	**53.3333%**	**97.8692%**
S.B.P.	"n", "obesity"	**90.1726%**	**24.2424%**	**96.5066%**
S.B.P.	"n", "cirrhosis"	**96.9613%**	**45.7627%**	**99.9026%**
S.B.P.	"n", "cancer"	**97.2407%**	**10.4167%**	**99.2697%**

U sljedećem koraku bile su uključene i druge značajke osim bakterija kako bi se poboljšala točnost binarnih klasifikatora. Navedene su značajke bile "spol", "država" i "bodysite", tj. mjesto uzimanja uzorka. Uključenjem navedenih značajki očekivane su bolje vrijednosti mjera kvaliteta modela. Konačno, za klasifikator kategorije "cancer" poduzeta je dodatna mjera jačeg penaliziranja lažno negativnih rezultata (100 puta) kako bi se konkretno utjecalo na osjetljivost modela. Za sve su klasifikatore ponovno određene mjere kvalitete – Tablica 2.

Tablica 2 - Točnosti i osjetljivosti nakon dodatka značajki i prilagodbe modela za "cancer"

Značajke	Klase	Točnost	Osjetljivost	Specifičnost
S.B.P., spol, država, "bodysite"	"n", "ulcerative_colitis"	**97.1850%**	**73.3333%**	**98.7161%**
S.B.P., spol, država, "bodysite"	"n", "obesity"	**96.6799%**	**77.2727%**	**98.5444%**
S.B.P., spol, država, "bodysite"	"n", "cirrhosis"	**95.8108%**	**65.3061%**	**97.9740%**
S.B.P., spol, država, "bodysite"	"n", "cancer"	**99.4291%**	**81.2500%**	**99.8539%**

Koristeći dva različita pristupa (višeklasna klasifikacija vs. binarni klasifikatori) dobiveni su prediktivni modeli strojnog učenja kojima se na temelju bakterijskog metagenoma mogu predvidjeti asocijacije između pojedinih bakterija i pojedinih bolesti.

4.2 Bakterijski rodovi i vrste u binarnoj klasifikaciji

Svaku od skupina bolesti kojima smo se bavili podvrgli smo binarnoj klasifikaciji izbacivši pritom sve demografske podatke. Zadržali smo samo atribute koji su se odnosili na bakterije u mikrobiomu, čime smo stvorili skup od 3302 relevantna atributa. Cilj nam je bio identificirati ključne vrste bakterija koje igraju značajnu ulogu u prepoznavanju određenih bolesti.

Kao rezultat za pretilost dobili smo da su za prepoznavanje pretilih osoba najvažnije bakterije iz porodice Veillonellaceae i Ruminococcaceae. Istraživanje je otkrilo značajan porast razine bakterija porodice Veillonellaceae u mikrobiomu pretile osobe u usporedbi s mikrobiomom zdravih osoba, dok je prisutnost bakterija porodice Ruminococcaceae značajno smanjena (Duan et al., 2021.). Osim toga, identificirana je povezanost između pretilosti i vrste Bacteroides ovatus. Iako istraživanje ne pruža izravnu vezu između pretilosti i ove vrste, naznačuje da njezina akumulacija može biti povezana s povećanim rizikom od razvoja dijabetesa tipa 2 kod pretilih osoba (Li et al., 2022.).

Kod binarne klasifikacije ciroze jetre također smo primijetili važnost porodice Veillonellaceae u klasifikaciji. I kod ove bolesti dokazan je porast razine bakterija ove porodice u mikrobiomu oboljelih (Chen et al., 2016.). Također, povećana je i razina bakterija vrste Megasphaera micronuciformis i roda Streptococcus, posebice vrste Streptococcus anginosus (Schwnger et al., 2019.).

Bakterija Eggerthella lenta pokazala se kao najvažnija u klasifikaciji dijabetesa tipa 2. Istraživanja pokazuju kako je porast ove vrste karakterističan za bolesti poput upale slijepog crijeva, dijabetesa i raka krvi (Jiang et al., 2021.). S druge strane, količina bakterija porodice Ruminococcaceae smanjena je, posebice vrsta Ruminococcus torques (Chu et al., 2022.), (Chen et al., 2022.).

Rak debelog crijeva naš je algoritam prepoznao prvenstveno na temelju vrsta Peptostreptococcus stomatis, Parvimonas micra i Fusobacterium nucleatum. Pretjerana zastupljenost ovih bakterija u mikrobiomu nađena je u više od 66% pacijenata oboljelih od raka debelog crijeva (Osman et al., 2021.).

U našem skupu podataka nalaze se dvije bolesti crijeva – Chronova bolest i ulcerozni kolitis. Kao vrsta važna za klasifikaciju istaknula se Subdoligranulum variabile. Istraživanje na djeci s Chronovom bolešću pokazalo je da je ova vrsta jače zastupljena kod osoba s tom bolešću (Quince et al., 2015.). Kod klasifikacije ulceroznog kolitisa istaknula se vrsta Odoribacter splanchnicus za koju je pokazano da ima pozitivne učinke i poboljšanje stanja oboljelih osoba (Lima et al., 2022.).

5. Literatura

Pasolli, E., Truong, D.T., Malik, F., Waldron, L., Segata, N. (2016.). Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput Biol, 2016 Jul 11;12(7):e1004977, https://doi.org/10.1371/journal.pcbi.1004977

Duan, M., Wang, Y., Zhang, Q., Zou, R., Guo, M., Zheng, H. (2021). Characteristics of gut microbiota in people with obesity. PLoS ONE, 16(8), https://doi.org/10.1371/journal.pone.0255446

Li, Y., Yang, Y., Wang, J., et al. (2022). Bacteroides ovatus-mediated CD27− MAIT cell activation is associated with obesity-related T2D progression. Cellular & Molecular Immunology, 19(7), 791–804, https://doi.org/10.1038/s41423-022-00871-4

Chen, Y., Ji, F., Guo, J. et al. Dysbiosis of small intestinal microbiota in liver cirrhosis and its association with etiology. Sci Rep 6, 34055 (2016), https://doi.org/10.1038/srep34055

Schwenger, K. J. P., Clermont-Dejean, N., Allard, J. P. (2019). The role of the gut microbiome in chronic liver disease: the clinical evidence revised. JHEP Reports, 1(3), 214-226. ISSN 2589-5559, https://doi.org/10.1016/j.jhepr.2019.04.004

Jiang, S., E, J., Wang, D., Zou, Y., Liu, X., Xiao, H., Wen, Y., Chen, Z. (2021). Eggerthella lenta bacteremia successfully treated with ceftizoxime: case report and review of the literature. Eur J Med Res 26, 111, https://doi.org/10.1186/s40001-021-00582-y

Chu, N., Juliana CN., Chan, Elaine Chow. (2022). A diet high in FODMAPs as a novel dietary strategy in diabetes? Clinical Nutrition, 41(10), 2103-2112. https://doi.org/10.1016/j.clnu.2022.07.036

Chen, W., Zhang, M., Guo, Y., Wang, Z., Liu, Q., Yan, R., Wang, Y., Wu, Q., Yuan, K., Sun, W. (2021) The Profile and Function of Gut Microbiota in Diabetic Nephropathy, Diabetes, Metabolic Syndrome and Obesity, 14:, 4283-4296, https://doi.org/10.2147/DMSO.S320169

Osman, M.A., Neoh, Hm., Ab Mutalib, NS. et al. (2021). Parvimonas micra, Peptostreptococcus stomatis, Fusobacterium nucleatum and Akkermansia muciniphila as a four-bacteria biomarker panel of colorectal cancer. Sci Rep 11, 2925. https://doi.org/10.1038/s41598-021-82465-0

Quince, C., Ijaz, U.Z., Loman, N. et al. (2015.) Extensive Modulation of the Fecal Metagenome in Children With Crohn's Disease During Exclusive Enteral Nutrition, American Journal of Gastroenterology 110(12):p 1718-1729, https://doi.org/10.1038/ajg.2015.357

Lima, S.F., Gogokhia, L., Viladomiu, M., Chou, L., Putzel, G. et al. (2022.) Transferable Immunoglobulin A–Coated Odoribacter splanchnicus in Responders to Fecal Microbiota Transplantation for Ulcerative Colitis Limits Colonic Inflammation, Gastroenterology, 162(1):p 166-178, https://doi.org/10.1053/j.gastro.2021.09.061