Универзитет „Св. Кирил и Методиј“ во Скопје
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

Документација за проект по предметот

Вовед во науката за податоци

Тема:

АНАЛИЗА И КЛАСИФИКАЦИЈА НА ПОДАТОЦИ ОД ЧОВЕЧКИ МИКРОБИОМИ

Професор:

д-р Слободан Калајџиски

Изработиле:
Марко Ивановски 226089
Ана Папалазова 221006

Септември, 2025

Содржина

1. ВОВЕД

Во последните години анализата на човечкиот микробиом е една од најистражуваните теми, бидејќи бактериите кои живеат во нашето тело играат значајна улога во одржувањето на здравјето и во појавата на различни болести. Целта на овој проект е да се анализираат податоци од микробиоми на пациенти и здрави индивидуи и да се испита дали може да се изврши прецизна класификација на различни состојби како здрав пациент, дебел пациент, пациент со Кронова болест или пациент со Улцеративен колитис, со примена на техники за предобработка, визуелизација и моделирање на податоците.

2. КОРИСТЕНО МНОЖЕСТВО

Податоците кои се користат во овој проект се преземени од CAMDA натпреварот. Оригинално, податоците беа достапни во повеќе датотеки, но за потребите на проектот тие беа споени во една табела. Со тоа добивме множество со 613 редици и 2121 колони. Секоја редица во множеството претставува примерок од пациент, а колоните ги содржат метаподатоците и релативните застапености на различни бактериски видови. Првата колона претставува класна ознака (Healthy, Obese, CD – Chron's disease или UD – Ulcerative colitis).

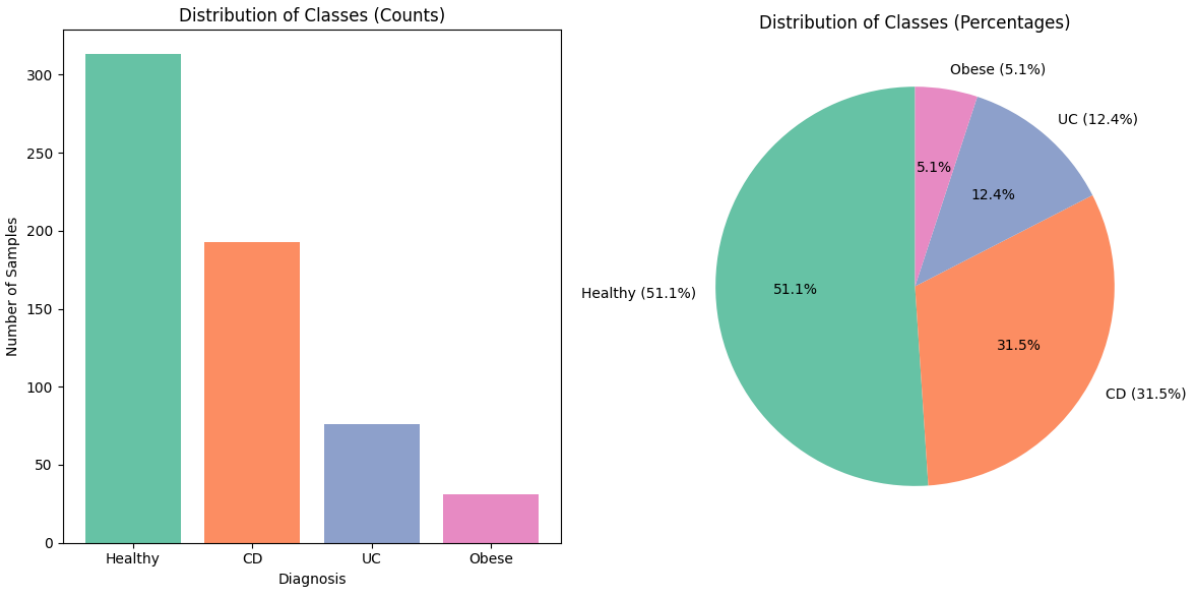
	Diagnosis	Project	SampleID	GMHI	hiPCA	Shannon_entropy	Butyrivibrio_crossotus	GGB3614_SGB4886	GGB1630_SGB2238
464	CD	HMP2	SRR5946620	-2.078329	9.012148	2.595824	0.000000	0.000000	0.000000
281	UC	PRJEB1220	ERR210422	3.839564	0.320413	5.191117	0.000000	0.000000	0.000000
251	CD	PRJNA389280	SRR5983386	-0.417530	2.505147	3.872293	0.000000	0.000000	0.000000
507	Healthy	HMP2	SRR5936195	1.894733	0.728368	5.154253	0.000000	0.000000	0.000000
438	Healthy	HMP2	SRR5935965	-0.787923	3.503061	4.595323	0.000000	0.000000	0.000000
109	Healthy	HMP2	SRR5947069	1.094875	10.700448	4.265028	0.000000	0.000000	0.000000
211	CD	PRJEB1220	ERR209701	-3.198056	3.850767	3.785529	0.000000	0.000000	0.000000
575	Healthy	HMP2	SRR5935989	2.939043	0.235928	5.776213	0.003340	0.000000	0.000000
34	UC	HMP2	SRR5946749	-2.240186	3.818899	2.838379	0.000000	0.000000	0.000000
309	Obese	PRJEB1220	ERR209160	-0.700031	2.839688	4.718414	0.000000	0.000000	0.000000

Слика 1. Примерок од користеното множество

3. ВИЗУЕЛИЗАЦИИ

3.1. Дистрибуција на дијагнози

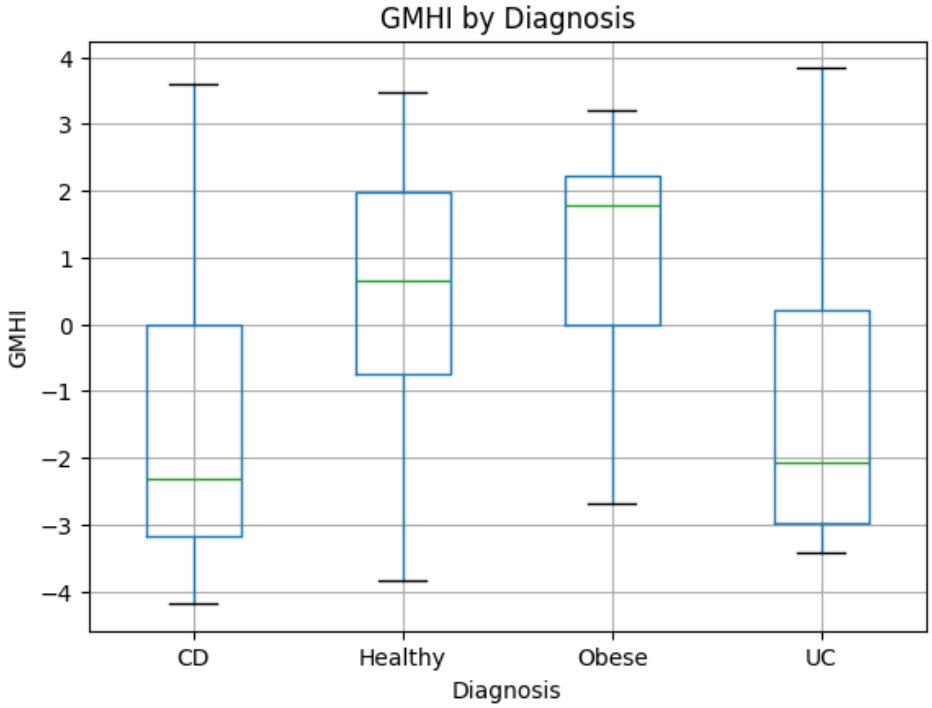
Пред да преминеме на специфични метрики, важно е да ја разбереме самата распределба на класите во множеството. За таа цел направивме анализа на класната распределба за да провериме дали множеството е избалансирано. Како што може да се види од графиконите, одредени класи се повеќе застапени, што е важно да се има предвид при тренирање на моделите.



Слика 2. Класна распределба на дијагнозите

3.2. Boxplot на GMHI по дијагнози

Следно се разгледа Gut Microbiome Health Index (GMHI) според различните дијагнози.



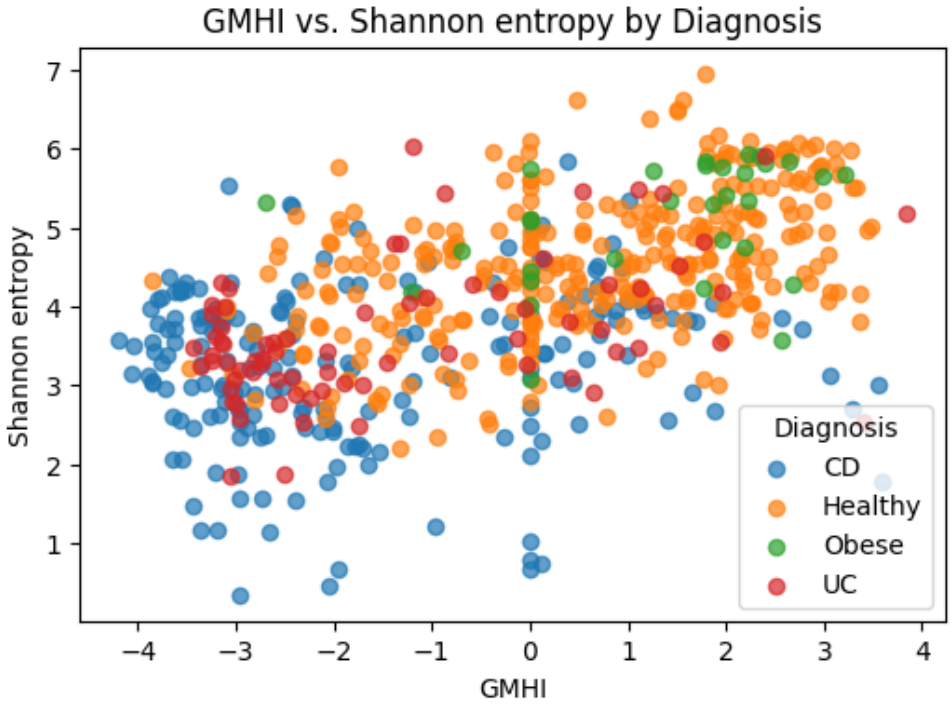
Слика 3. Boxplot на GMHI по дијагнози

Групите на пациенти со Кронова болест и Улцеративен колитис имаат значително пониски GMHI вредности од групите на здрави пациенти и дебели пациенти, што покажува дека GMHI може јасно да ги разликува воспалителните болести на цревата од поздравите профили на микробиоми. Групата на дебели пациенти има највисок просек

од сите, што укажува дека во ова множество, според GMHI метриката, дебелината не секогаш значи нездрава микробиома.

3.3. Scatter Plot на GMHI наспроти Shannon ентропија

За да ги комбинираме GMHI и Shannon ентропијата (мерка на разновидност), направивме scatter plot каде секоја точка е еден примерок обоен според дијагнозата.



Слика 4. Scatter plot на GMHI наспроти Shannon ентропија

Што покажуваат кластерите:

Точките за Кронова болест (сини) се групирани во долниот лев дел - низок GMHI и низок диверзитет. Тоа значи дека пациентите со Кронова болест обично имаат помалку видови бактерии и микробиом кој има слаба оценка на индексот за здравје.

Точките за Улцеративен колитис (црвени) се наоѓаат веднаш над или десно од кластерот за Кронова болест, но сепак најчесто во долниот лев квадрант, што повторно укажува на нездрав ниско-диверзитетен профил.

Здравите (портокалови) примероци се шират кон средниот и горниот десен агол - умерен до висок GMHI и висок диверзитет. Тие обично имаат и избалансиран микробиом и голем број различни видови.

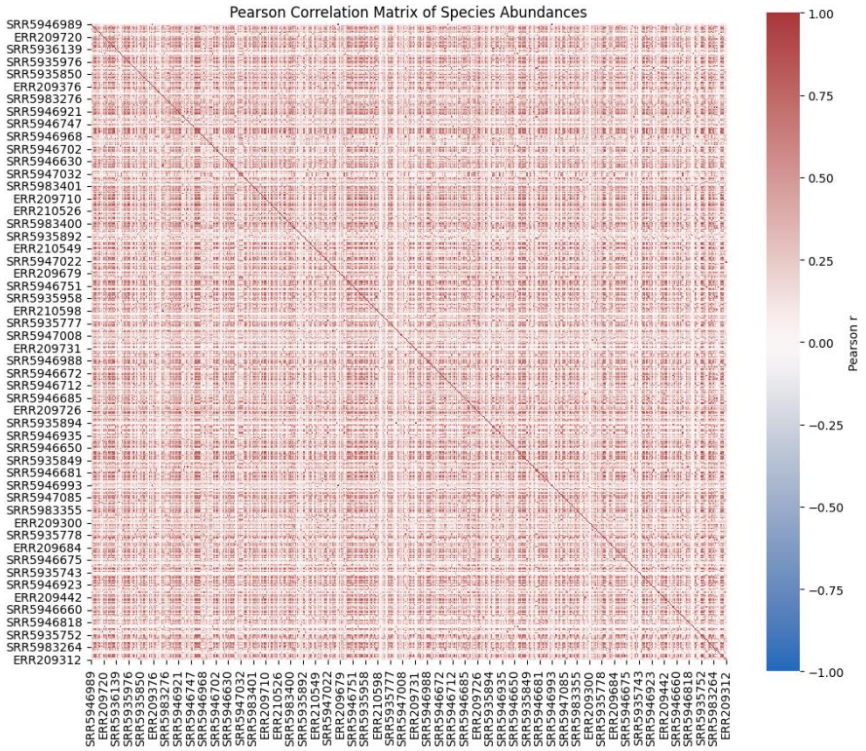
Дебелите (зелени) примероци најчесто го зафаќаат горниот десен агол покрај здравите, покажувајќи дека во оваа кохорта дебелите лица може да имаат и здрав GMHI резултат и висок диверзитет - понекогаш дури и да ја надминат здравата група.

3.4. Анализа на Pairwise корелација

За дополнителна анализа на множеството, извршена е пресметка на парните (pairwise) корелации помеѓу различните видови бактерии. Целта е да се откријат силно зависни и редундантни карактеристики кои можат да предизвикаат мултиколинеарност во моделите за машинско учење. Со отстранување на ваквите карактеристики, постигнуваме поедноставена репрезентација на податоците, побрза конвергенција при тренирањето и покорисна интерпретација на feature importance.

3.4.1. Пирсонова корелација

Пирсоновиот коефициент ја мери линеарната зависност меѓу две променливи. Вредностите се движат од -1 (совршена негативна врска) до +1 (совршена позитивна врска). Ќе користиме топлински мапи за да ги визуелизираме и анотираме најсилните позитивни/негативни парови.

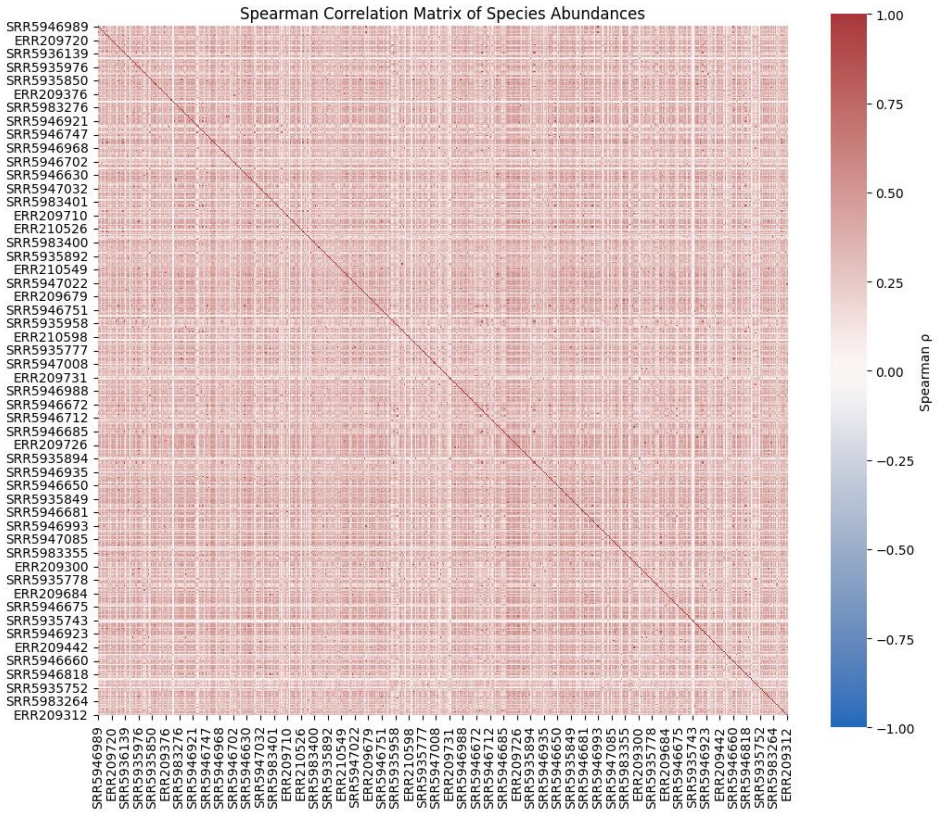


Слика 5. Heat мапа на Пирсонови коефициенти

Карактеристиките со |r| близу 1 се речиси линеарно редундантни. Со забележување на тие парови, можеме да избереме еден што ќе го отфрлиме и со тоа да ја намалиме мултиколинеарноста, што им помага на многу класификатори посигурно да конвергираат и позначајно да ги толкуваат важностите на карактеристиките.

3.4.2. Сперманова корелација

Спермановиот коефициент, за разлика од Пирсоновиот, ја мери монотоната зависност меѓу променливите (односно дали една варијабла расте или опаѓа заедно со другата, без разлика дали врската е линеарна). Овој метод е поотпорен на отстапувања и подобро ја фаќа нелинеарната структура на податоците.



Слика 6. Heat мапа на Сперманови коефициенти

Од сликата може да се забележи дека повторно се издвојуваат групи на силно поврзани карактеристики.

3.4.3. Идентификација и елиминација на силно-корелирани карактеристики

Со цел да го поедноставиме множеството, ги филтриравме сите парови карактеристики кои имаат апсолутна корелација поголема од 0.85, и кај Пирсон, и кај Сперман. На тој начин, добивме 287 видови бактерии кои се сметаат за редундантни и беа отстранети од понатамошната анализа. Оваа редукција е клучна бидејќи се елиминира мултиколинеарноста, се подобрува стабилноста на моделите и се овозможува полесна интерпретација на биомаркерите кои остануваат. По ова чистење, добиваме прочистен датасет со значително намален број на карактеристики, кој ќе се користи во следните чекори на анализа и моделирање.

4. ПРЕТПРОЦЕСИРАЊЕ

За да обезбедиме сигурни и репрезентативни резултати од моделите, потребно е да се извршат неколку клучни чекори на претпроцесирање. Во оваа фаза ги подготвуваме податоците од така што моделите ќе можат правилно да ги научат структурите и зависностите.

4.1. Log-трансформација на abundances

Бидејќи голем број видови бактерии имаат нула или многу ниска застапеност, а малку видови доминираат со екстремно високи вредности, директното користење на овие вредности може да доведе до проблеми при тренирање. Затоа применивме логаритамска трансформација со `log1p`, која ги компресира екстремно високите вредности, нулите ги остава непроменети и создава порамномерна дистрибуција. Овој чекор е клучен за да се избегне доминација на аномалии и да се добијат значајни метрики за варијанса и растојание меѓу примероците.

4.2. Енкодирање на класите

Класите (дијагнози) првично се зачувани како текстуални вредности, но поголемиот дел од алгоритмите за машинско учење бараат нумерички вредности, па поради тоа со помош на LabelEncoder, секоја класа ја мапиравме на нумеричка вредност:

- Кронова болест → 0
- Здрав → 1
- Дебел → 2
- Улцеративен колитис → 3

Ова едноставно енкодирање овозможува коректно тренирање на моделите, стратифицирано делење на податоците и избегнување на грешки или некоректно толкување на текстуални вредности.

4.3. Делење на множеството

За да се провери колку добро генерализираат моделите, податоците се поделени на 80% кои се за тренирање и 20% кои ќе се користат за тестирање. Делењето е направено стратифицирано според класите, за да се задржи истата распределба на дијагнози во двете подгрупи. На овој начин, тест сетот останува репрезентативен и независен за финална евалуација.

4.4. Балансирање на класите

Во множеството постои нерамномерна распределба на примероци меѓу класите. Ова може да доведе моделот да ги фаворизира мнозинските класи (на пр. Здрав) и да ги игнорира поретките (на пр. Дебел или Кронова болест). За да го ублажиме овој проблем, применивме тежини на класите кои се пресметани обратно пропорционално на застапеноста на секоја класа. Со ова грешките на малубројните класи се казнуваат посилно, се обезбедува поправедна одлука и се подобрува чувствителноста на моделите за сите групи пациенти.

4.5. Стандардирање на карактеристиките

Поради тоа што вредностите на застапеност на бактериите и индексите на разновидност (GMHI, Shannon) имаат различни размери и единици, потребно е стандардизирање. Применивме StandardScaler кој од секоја карактеристика ја одзема нејзината средна вредност и ја дели со стандардната девијација. Овој чекор спречува карактеристиките со големи вредности да доминираат, овозможува фер оптимизација

кај алгоритми базирани на растојание или градиент и обезбедува побрза и постабилна конвергенција на моделите.

5. ТРЕНИРАЊЕ И ЕВАЛУАЦИЈА НА МОДЕЛИ

Следно ќе трениираме и евалуираме различни модели за класификација, со цел да ја предвидиме дијагнозата на пациентите врз основа на микробиомските податоци. Се користат пет модели со различна комплексност за да се тестира кој дава најдобри резултати:

- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)
- Gaussian Naive Bayes
- XGBoost

5.1. Заеднички пристап на моделите

За сите модели се применува унифициран pipeline со цел фер споредба и минимизирање на варијациите:

- Податоците се балансираат со SMOTE за да се обезбеди рамномерна застапеност на сите класи.
- Се применува стандардизација (StandardScaler) за да се изедначат скалите на карактеристиките.
- Се користи Stratified k-fold cross-validation (k=5) за проценка на стабилноста и генерализацијата на моделите.
- Се врши hyperparameter tuning со GridSearchCV за да се најдат оптималните вредности.
- Финалната проценка се врши на тест сет (20% од податоците) кој е изолиран уште на почетокот.

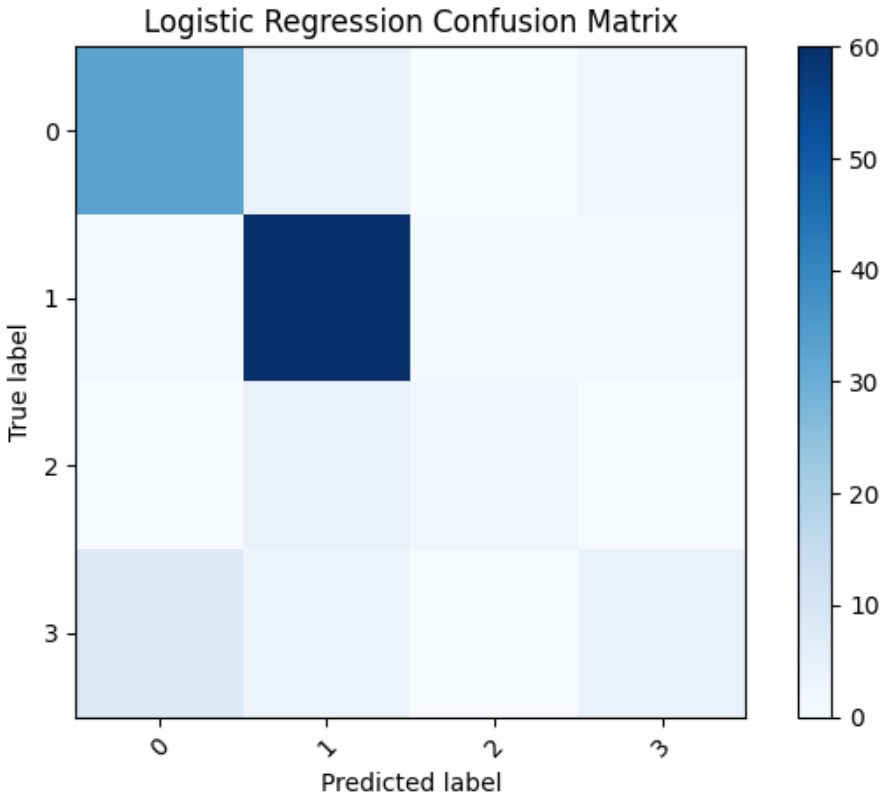
5.2. Модели

5.2.1. Logistic Regression

Логистичката регресија претставува едноставен, но истовремено и моќен модел за мултикласна класификација. Поради својата линеарна структура, моделот е лесно интерпретабилен и овозможува брзо тренирање, што го прави одличен референтен избор за споредба со посложените алгоритми.

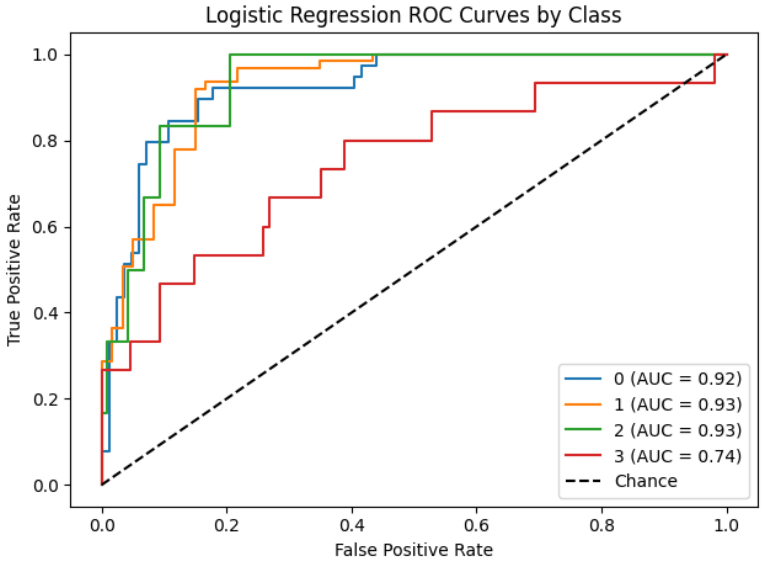
Во нашиот експеримент, најдобрите хиперпараметри се добиени со вредности C=80 и penalty='l2'. При евалуацијата, моделот постигна точност од 0.81, со precision 0.72**,** recall 0.60**,** F1 0.63 и ROC-AUC 0.88**.** Овие резултати укажуваат на добра вкупна прецизност, но и на потешкотии при правилното препознавање на помалку застапените класи, особено Улцеративниот колитис и Дебелината.

Анализата на Confusion matrix дополнително го потврдува ова.



Слика 7. Confusion matrix за Logistic Regression

Најчестите грешки се јавуваат кај Кроновата болест (0) и Улцеративниот колитис (3), кои често се предвидуваат како Здрави (1). Класата Дебел (2) е најтешка за точна предикција, што се должи на нејзината мала застапеност во податоците.



Слика 8. ROC за Logistic Regression

Од друга страна, ROC кривите покажуваат подлабока слика за способноста на моделот. Иако Дебел е слабо предвиден во Confusion Matrix, неговата висока AUC вредност укажува дека Логистичката регресија добро ги рангира овие примероци, но

дека тековниот праг на одлучување не е оптимален. Спротивно на тоа, кривата за Улцеративниот колитис е блиску до линијата на случајност, што сигнализира послаба одвоеност и ограничена способност на моделот да ги разликува овие пациенти од останатите групи.

Покрај предвидувачките резултати, Logistic Regression нуди и вредни биолошки увиди преку анализа на Feature Importance. Најзначајните карактеристики се специфични бактериски видови со најголеми коефициенти, што ни кажува кои микроби се најсилно поврзани со дадената дијагноза.

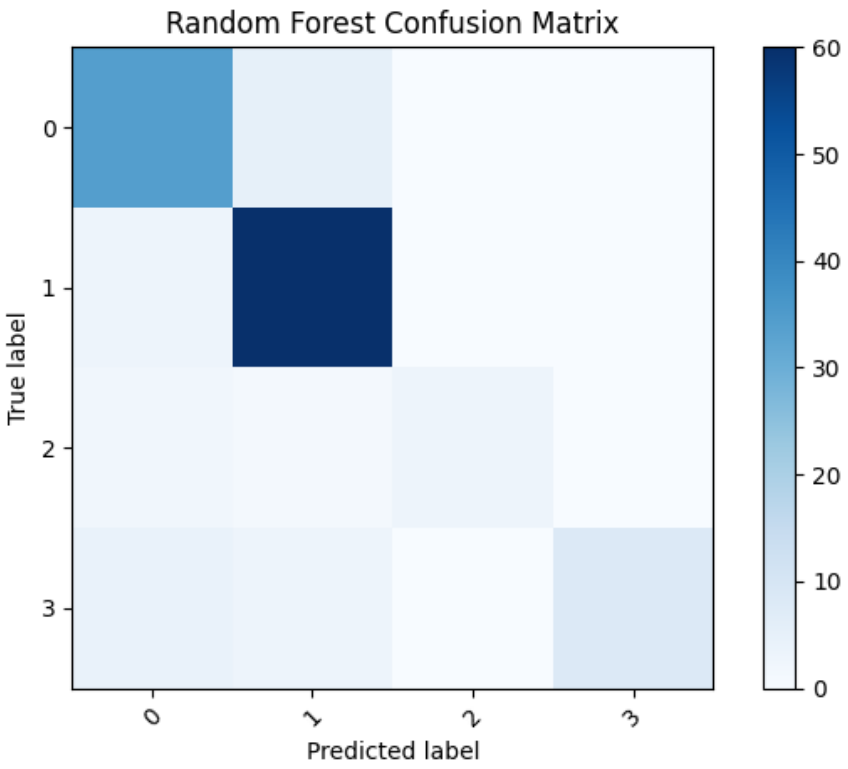
Слика 9. Feature importances за Logistic Regression

5.2.2. Random Forest

Random Forest е метод базиран на голем број одлучувачки дрва, каде секое дрво учи на различна подгрупа од податоците. Овој пристап му овозможува на моделот да фати нелинеарни врски и сложени интеракции помеѓу карактеристиките, што е особено корисно кај микробиомски податоци со голем број поврзани таксони.

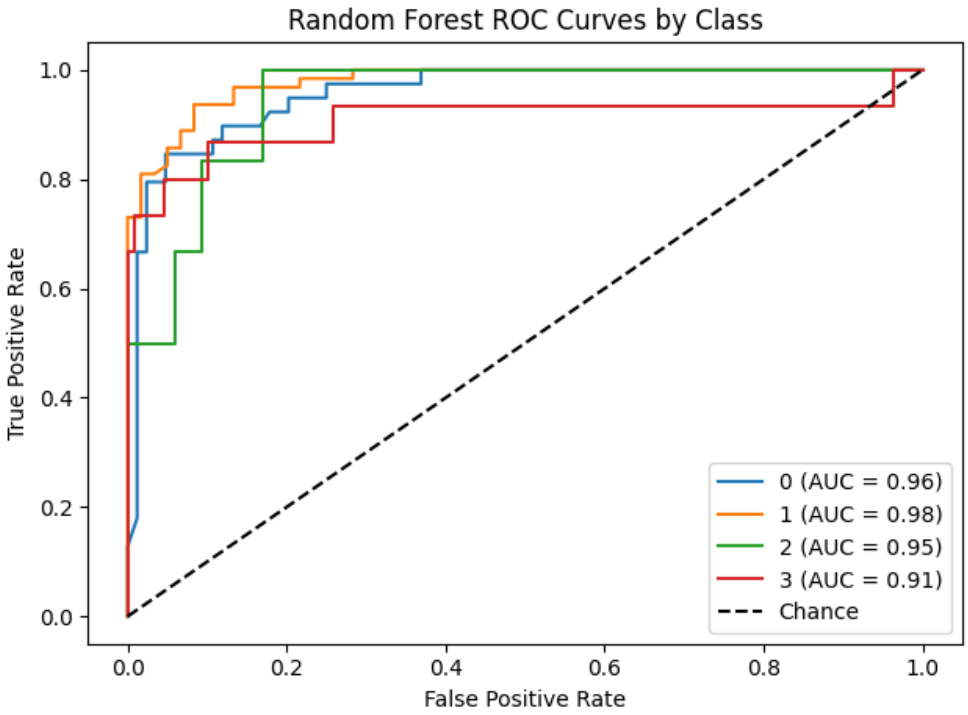
Моделот беше обучен со избрани параметри кои овозможуваат балансирање помеѓу прецизност и избегнување на преголемо вклопување. Најдобрите хиперпараметри кои ги добивме се: n_estimators=500, max_depth=None, min_samples_leaf=1. Постигнатите резултати се точност 0.85, precision 0.92, recall 0.71, F1 0.78 и ROC-AUC 0.95, која е највисока вредност меѓу повеќето тестирани модели.

Confusion Matrix покажува дека Random Forest многу добро ги класифицира примероците од Кроновата болест и Здрави особи. Сепак, и понатаму постојат грешки кај Улцеративниот колитис и Дебели особи, кои често се предвидуваат како Здрави. Ова укажува дека, иако моделот е силен, проблемите со нерамнотежа во класите сè уште се одразуваат.



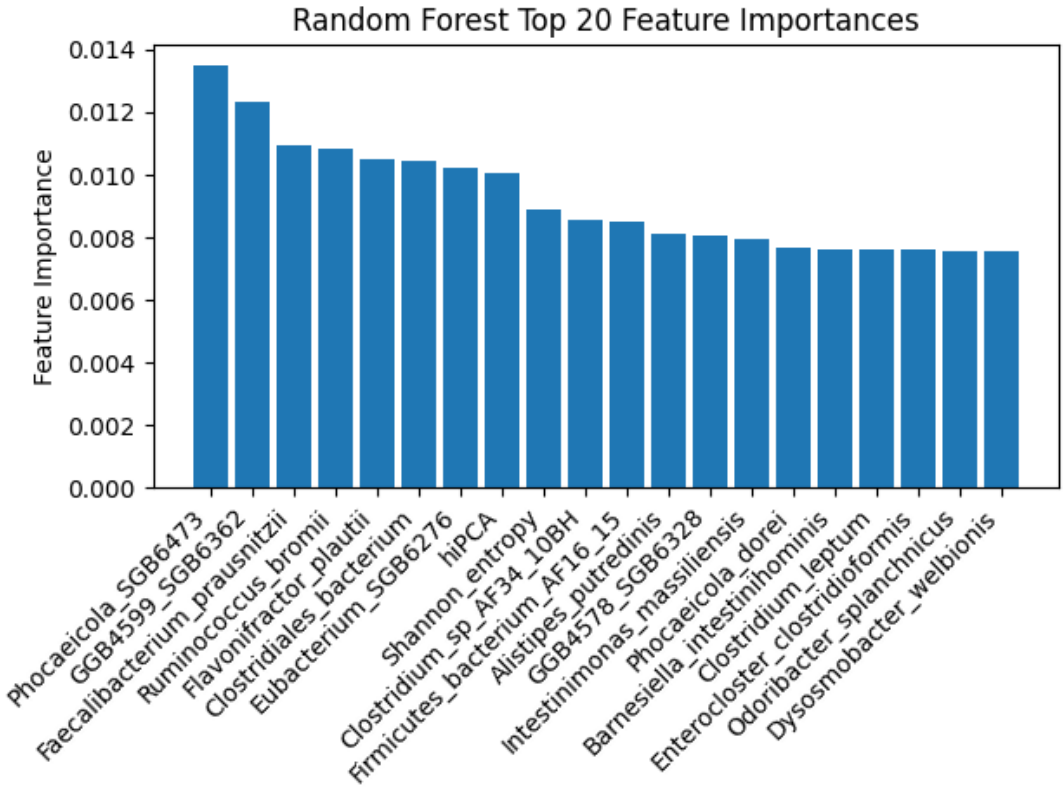
Слика 10. Confusion matrix за Random Forest

Во однос на ROC кривите, сите класи постигнуваат висока сепарабилност, КБ (AUC=0.96), Здрав (AUC=0.98), Дебел (AUC=0.95) и УК (AUC=0.91). Кривите близу 1.0 покажуваат дека резултатите од веројатноста на моделот исклучително ефикасно ги рангираат вистинските позитивни резултати пред лажните позитивни резултати.



Слика 11. ROC за Random Forest

Дополнително, прикажани се и Feature Importances, кои ги истакнуваат најрелевантните бактериски таксони за класификацијата.



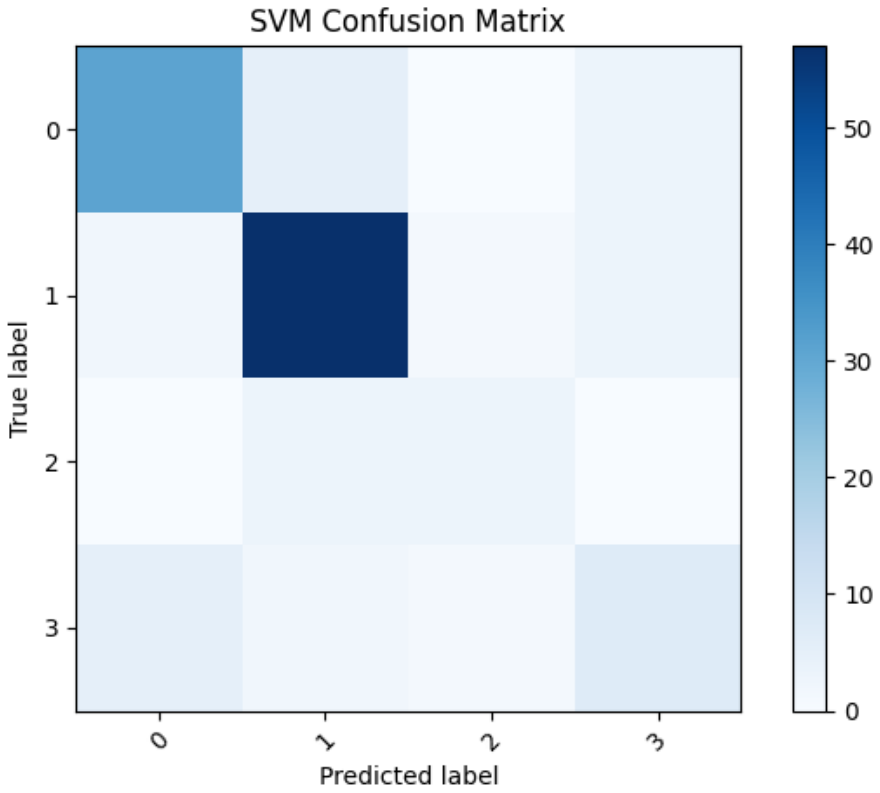
Слика 12. Feature importances за Random Forest

5.2.3. Support Vector Machine

SVM е модел кој гради оптимални граници за раздвојување на класите во високо-димензионален простор. Благодарение на kernel функциите, може да фати и нелинеарни врски меѓу карактеристиките, што често се застапени кај микробиомски податоци.

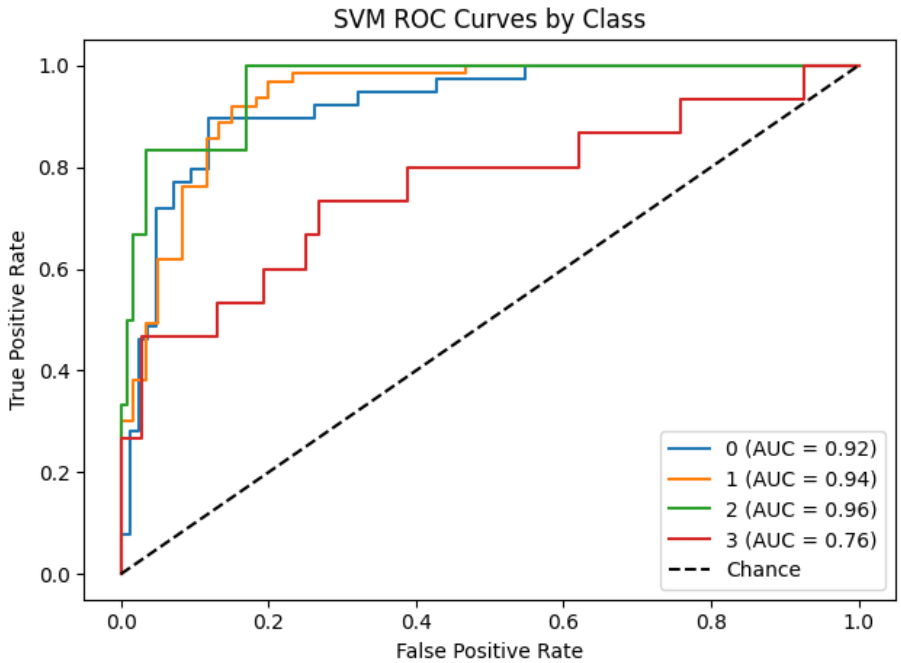
Моделот беше обучен со избрани параметри, од кои како најдобри ги добивме: C=40, gamma='scale', kernel='linear'. Во анализата SVM постигна accuracy 0.80, precision 0.70, recall 0.67, F1 0.68 и ROC-AUC 0.89. Иако овие резултати се солидни, се забележува пониска чувствителност во споредба со други модели, што укажува на потешкотии во целосното препознавање на помалку застапените класи.

Според Confusion Matrix, најчестите грешки се јавуваат кај Дебели пациенти и пациенти со Улцеративен колитис, кои често се предвидуваат како Здрави. Кроновата болест се идентификува подобро, но сепак со неколку погрешни предвидувања.



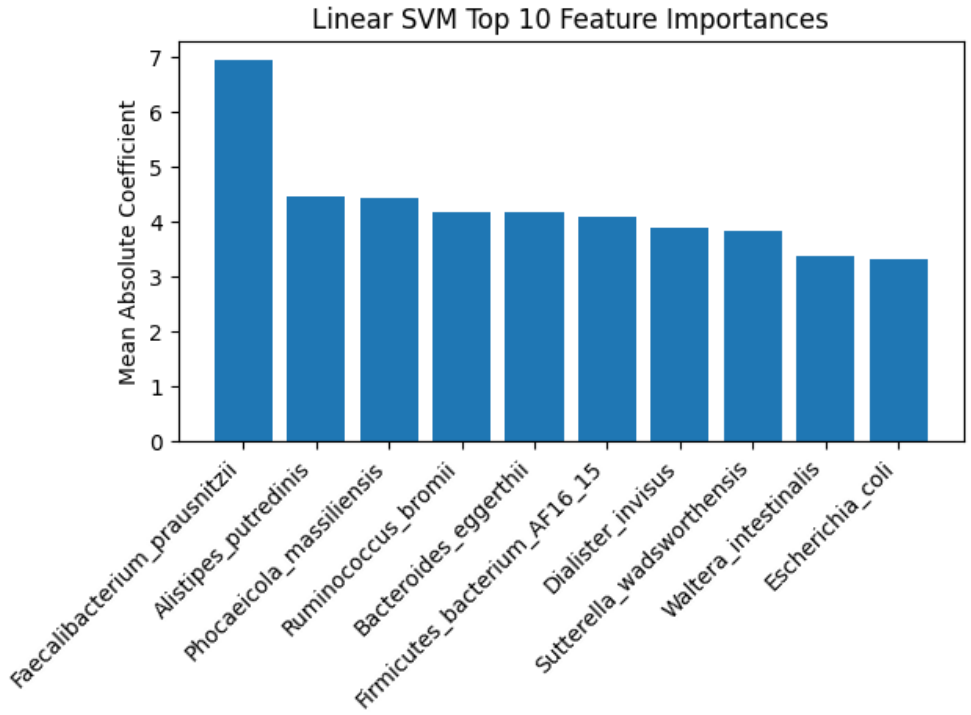
Слика 13. Confusion matrix за Support Vector Machine

ROC кривите потврдуваат дека SVM е стабилен модел, со висока AUC за Здрави, Дебели и пациенти со Кронова болест, но нешто пониска за Улцеративен колитис. Ова укажува дека, иако SVM има добро разделување за доминантните класи, сепак не успева целосно да ја искористи информацијата кај помалите групи.



Слика 14. ROC за Support Vector Machine

Дополнително, прикажани се и Feature Importances, кои ги истакнуваат најрелевантните бактериски таксони за класификацијата. Графиконот е сличен со оној од Logistic Regression, бидејќи и двата модели се линеарни и зависат од тежините на карактеристиките. Најзначајни бактерии повторно се Faecalibacterium prausnitzii, Alistipes putredinis и Phocaeicola massiliensis, што укажува дека овие видови претставуваат стабилни биомаркери за разликување на здрави од болни пациенти.



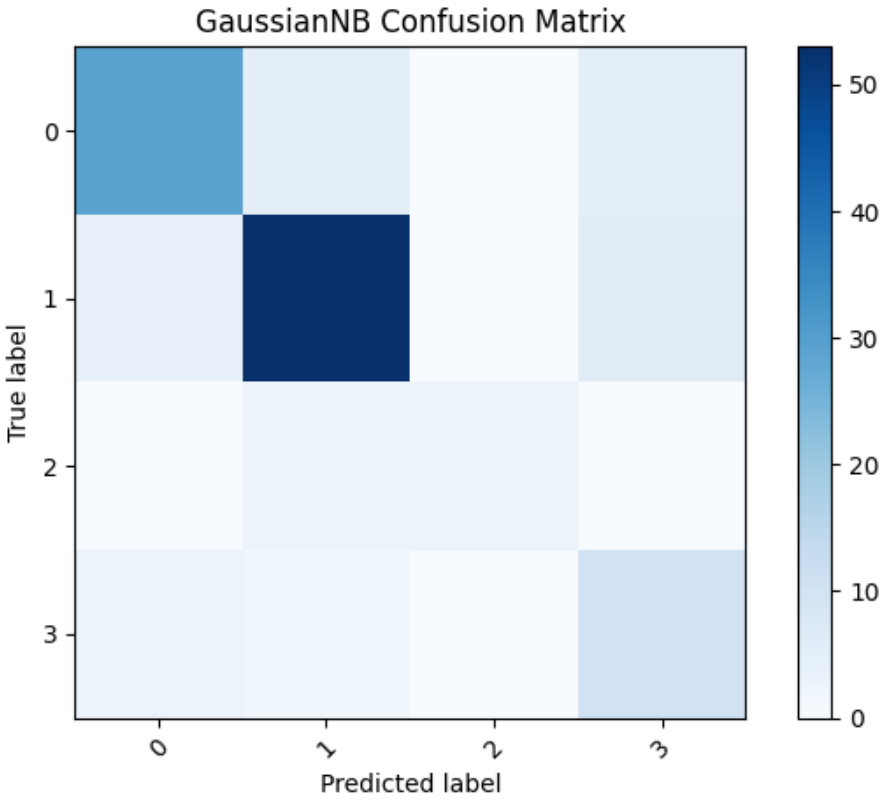
Слика 15. Feature importances за Support Vector Machine

5.2.4. Gaussian Naive Bayes

Gaussian Naive Bayes е едноставен веројатносен модел кој претпоставува дека карактеристиките следат нормална распределба и дека се независни една од друга. Иако оваа претпоставка е често нереална кај микробиомските податоци, таа овозможува моделот да биде брз, скалабилен и помалку подложен на пренатренираност на мали сетови.

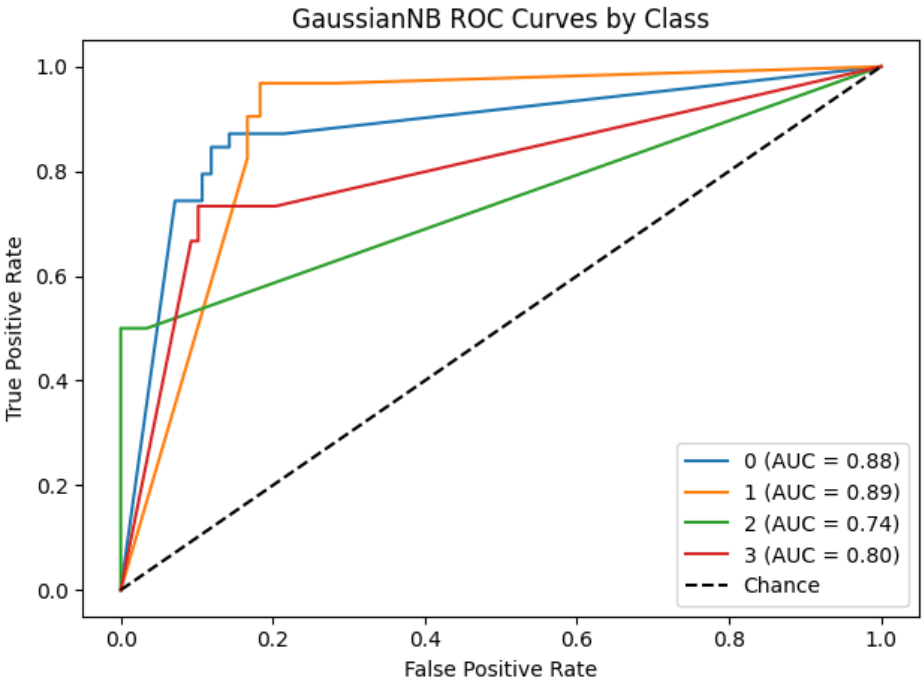
Кај овој модел како најдобар хиперпараметар се покажа var_smoothing=1e-09, што овозможи постабилни проценки на веројатностите. Моделот постигна accuracy 0.77, precision 0.78, recall 0.69, F1 0.71 и ROC-AUC 0.83. За ваков едноставен пристап, овие резултати се изненадувачки конкурентни и покажуваат дека Naive Bayes може да биде добар базичен модел.

Confusion Matrix открива дека моделот добро идентификува многу примероци со Кронова болест и Здрави примероци, но сепак погрешно означува значителен број на КБ како Здрави и обратно. Кластите Дебел и Улцеративен колитис често се предвидени како Здрави. Ова ја истакнува слабоста на GaussianNB во справувањето со помали групи.



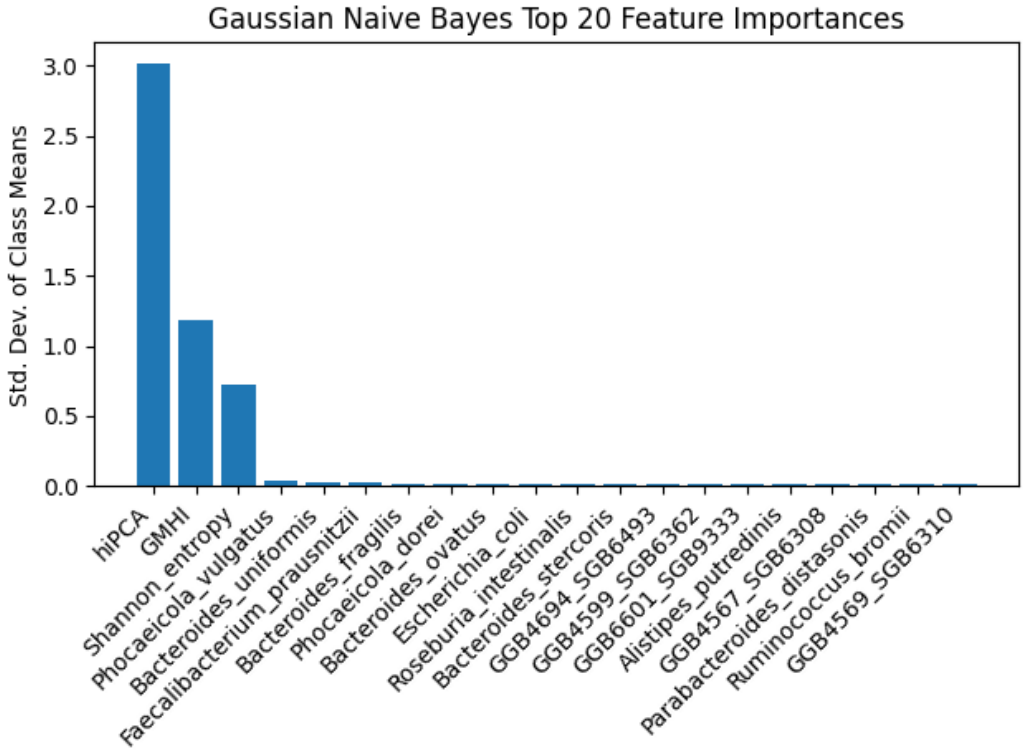
Слика 16. Confusion matrix за Gaussian Naive Bayes

ROC кривите покажуваат дека Здрави (AUC ≈ 0.89) и Кронова болест(≈ 0.88) имаат добра разделба, додека Улцеративен колитис (≈ 0.82) и Дебелина (≈ 0.75) значително заостануваат. Тоа значи дека Naive Bayes е подобар за доминантните класи, но има ограничена моќ кај малцинските.



Слика 17. ROC за Gaussian Naive Bayes

Дополнително, со анализа на најважните карактеристики, моделот ги истакнува карактеристиките со најголема варијација меѓу класите, што овозможува корисен биолошки увид.



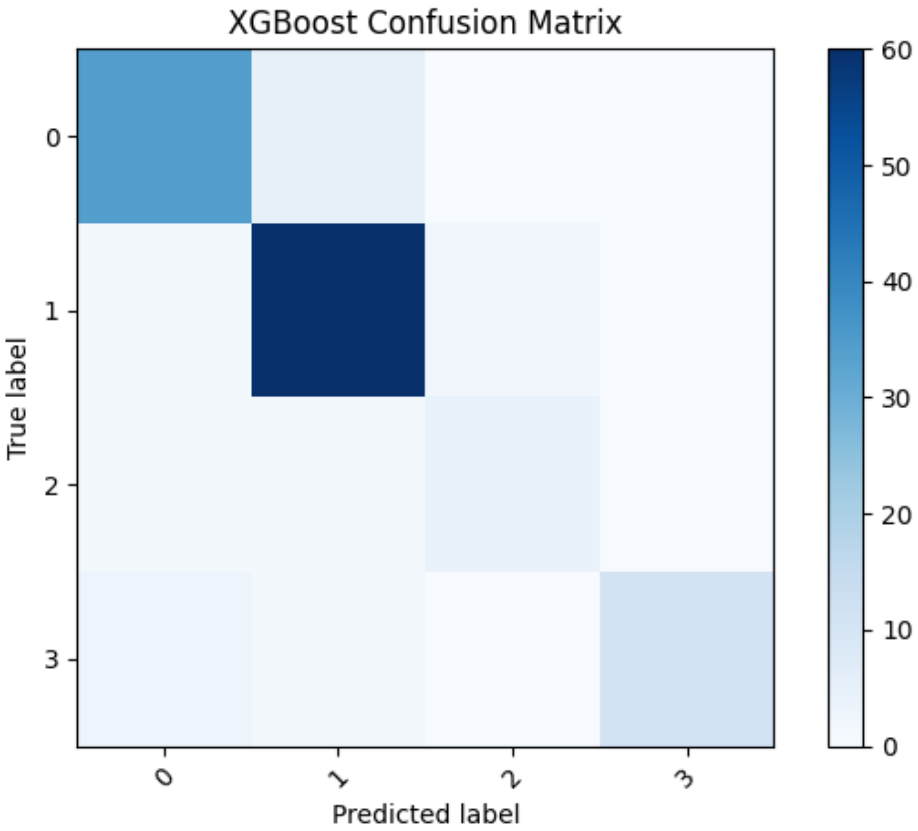
Слика 18. Feature importances за Gaussian Naive Bayes

5.2.5. XGBoost

XGBoost е градиент-бустинг алгоритам кој гради дрва последователно, каде секое ново дрво ги корегира грешките од претходните. Тој се смета за еден од најмоќните модели за табеларни податоци, особено кога има голем број корелирани карактеристики, како што е случајот со микробиомот.

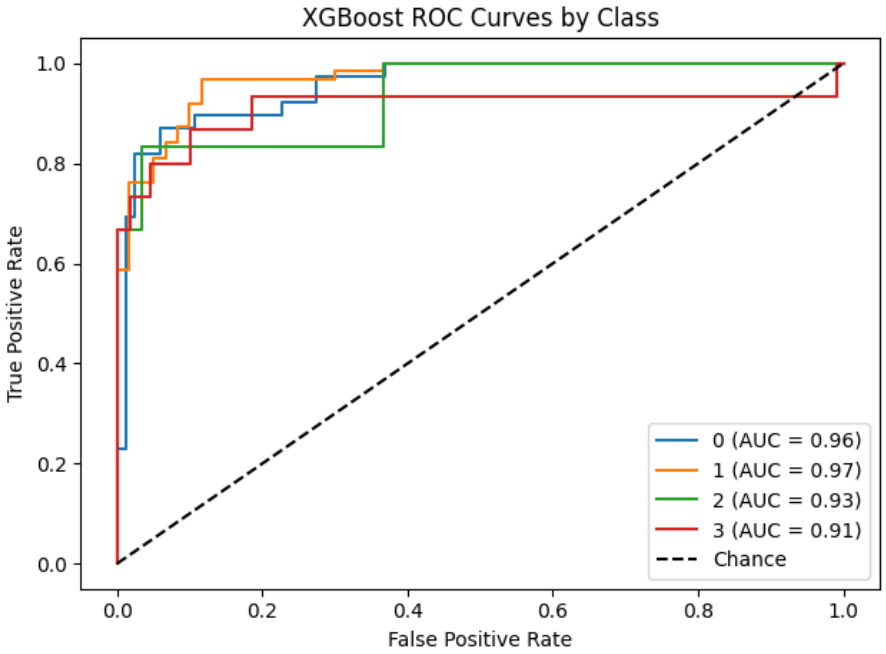
Како најдобри хиперпараметри се покажаа: clf_learning_rate=0.5, clf_max_depth=3, clf_n_estimators=100. Во нашите тестирања, XGBoost постигна најдобри резултати: accuracy 0.86, precision 0.90, recall 0.77, F1 0.82 и ROC-AUC 0.94. Овие бројки јасно ја истакнуваат неговата предност во однос на сите други тестирани модели.

Confusion Matrix покажува дека XGBoost сигурно ги идентификува примероците со Кронова болест и Здравите примероци и постигнува значајно подобрување во класификацијата на Дебели примероци. Иако кај Улцеративниот колитис и понатаму има одреден број погрешни предвидувања, грешките се значително намалени во споредба со Logistic Regression и Naive Bayes.



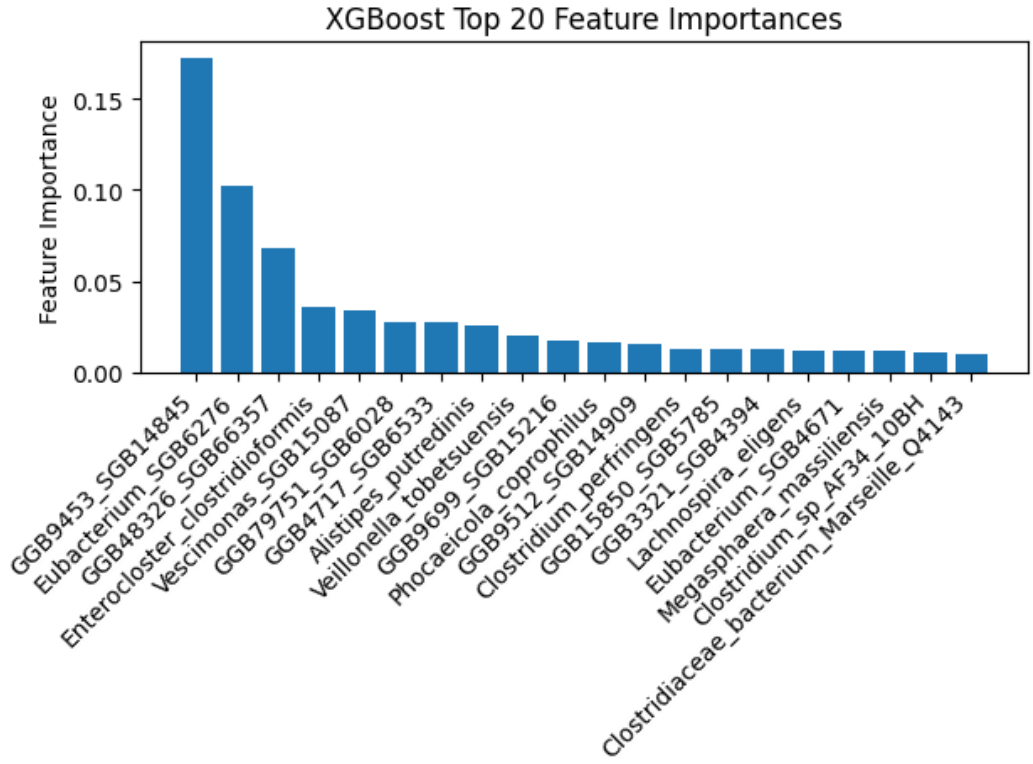
Слика 19. Confusion matrix за XGBoost

ROC кривите се скоро идеални: Здрави (AUC ≈ 0.97), Кронова болест (AUC≈ 0.96) и Дебели(AUC≈ 0.93) се одлично раздвоени, а дури и Улцеративен колитис (AUC≈ 0.91) има многу добра разделба. Стрмниот почеток на кривите укажува дека XGBoost е исклучително ефикасен во рангирањето на вистинските позитивни случаи пред лажните.



Слика 20. ROC за XGBoost

Анализата на Feature Importance дополнително потврдува дека моделот идентификува клучни бактериски видови како главни предиктори, што ја зголемува неговата биолошка применливост.



Слика 21. Feature importances за XGBoost

5.3. Споредба на модели

Од споредбата на моделите, XGBoost и Random Forest се покажаа како најуспешни за ова множество. XGBoost постигна највисоки вредности на accuracy и F1-score, што го прави најбалансиран модел во смисла на прецизност и чувствителност. Random Forest, пак, обезбеди највисока прецизност и многу добра ROC-AUC, со што овозможува и биолошка интерпретабилност преку анализа на feature importance.

Во споредба, Logistic Regression, SVM и Gaussian Naive Bayes имаат пониски резултати, но сепак служат како корисни референтни модели. Особено GaussianNB постигна подобри recall и precision од Logistic Regression иако е едноставен модел, што укажува на неговата вредност за побрза анализа или помали множества.

	model	accuracy	precision	recall_macro	f1_macro	roc_auc_ovr
0	LogisticRegression	0.804900	0.717200	0.599600	0.629600	0.880900
1	RandomForest	0.853700	0.915100	0.714400	0.775200	0.947300
2	SVM	0.796700	0.701200	0.666600	0.681900	0.894900
3	GaussianNB	0.772400	0.780800	0.687900	0.709200	0.827400
4	XGBoost	0.861800	0.898600	0.766300	0.816100	0.943400

Слика 22. Споредба на перформансите на сите тестирани модели

6. ЗАКЛУЧОК

Овој проект покажа дека ensemble методите (како XGBoost и Random Forest) се особено соодветни за анализа на микробиомски податоци поради нивната способност да ги доловат сложените интеракции меѓу таксоните. Резултатите укажуваат дека машинското учење може да идентификува биомаркери и да помогне во разликување на здрави и болни состојби базирани на микробиомски профили.

Во иднина, точноста и интерпретабилноста на моделите може да се подобрат преку интегрирање на биолошки информации на ниво на метаболички патеки (pathway-level features), примена на длабоки невронски мрежи за откривање на посложени и нелинеарни врски, како и со комбинирање на различни типови податоци – како што се метагеномски профили, клинички фактори и информации за исхрана.

7. РЕФЕРЕНЦИ

1. CAMDA Datasets. *Critical Assessment of Massive Data Analysis (CAMDA) – CAMDA Play datasets*. Достапно на: https://bipress.boku.ac.at/camda-play/elementor-4538/?redirect_to=https%3A%2F%2Fbipress.boku.ac.at%2Fcamda-play%2Fcamda-play%2Fdatasets%2F
2. de Winter, J. C. F., Gosling, S. D., & Potter, J. (2016). „Споредба на Пирсонов и Спирманов коефициент на корелација низ различни дистрибуции и големини на примероци: водич со симулации и емпириски податоци“. *Psychological Methods*, 21(3), 273–290. Достапно на: https://journals.lww.com/anesthesia-analgesia/fulltext/2018/05000/correlation_coefficients__appropriate_use_and.50.aspx
3. Mukaka, M. M. (2012). „Водич за соодветна употреба на коефициентот на корелација во медицински истражувања“. *Malawi Medical Journal*, 24(3), 69–71. Достапно на: https://pmc.ncbi.nlm.nih.gov/articles/PMC3576830/