UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

Lucija Miličić

BIOINFORMATIČKA ANALIZA PODATAKA O DETEKCIJI I TRETMANU PREEKLAMPSIJE

master rad

Beograd, 2024.

Mentor:

prof. dr Jovana Kovačević, vanredni profesor
Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

prof. dr Mladen Nikolić, vanredni profesor
Univerzitet u Beogradu, Matematički fakultet

dr Vladimir Kovačević, naučni saradnik
Istraživačko-razvojni institut za veštačku inteligenciju Srbije

Datum odbrane: _____

Zahvaljujem se prof. dr Jovani Kovačević na ukazanom poverenju, sugestijama i podršci tokom izrade ovog rada, kao i na stečenom znanju iz oblasti bioinformatike. Hvala dr Vladimiru Kovačeviću za obezbeđene podatke, domensko znanje i korisne savete. Hvala prof. dr Mladenu Nikoliću na stečenom znanju iz oblasti mašinskog učenja.

Hvala porodici i prijateljima koji su verovali u mene.

Naslov master rada: Bioinformatička analiza podataka o detekciji i tretmanu preeklampsije

Rezime: Preeklampsija je oboljenje koje se može pojaviti tokom trudnoće i dovesti do ozbiljnih posledica, pa čak i smrti, zbog čega je rano otkrivanje ključno za dalji tretman i sprečavanje komplikacija. U otkrivanju ove bolesti od pomoći mogu biti specifični molekuli, metaboliti, koji učestvuju u svim procesima organizma. Promena koncentracije pojedinih metabolita može predstavljati indikator značajnih promena unutar organizma, na primer pojave nekog oboljenja, uključujući i preeklampsiju.

Osnovni cilj ovog istraživanja je identifikacija potencijalnih biomarkera među svim izmerenim metabolitima. Dakle, potrebno je identifikovati metabolite čija promena koncentracije može ukazivati na pojavu preeklampsije. Poslednja tehnološka dostignuća omogućila su merenje koncentracije čak do hiljadu različitih metabolita, što pruža mogućnost za primenu metoda mašinskog učenja u analizi ovih visokodimenzionih podataka. Osnovna ideja predstavlja treniranje modela nadgledanog učenja koji bi naučio veze između izmerenih koncentracija metabolita i informacije o tome da li oni odgovaraju pacijentkinji sa preeklampsijom. Od različitih metoda nadgledanog učenja najbolje se pokazao *XGBoost* klasifikator uz odgovarajući algoritam za izbor optimalnog podskupa atributa.

Rezultujući model korišćen je za identifikaciju najvažnijih atributa, koji predstavljaju potencijalne kandidate za biomarkere. Statističkim testovima je dodatno potvrđena razlika u koncentracijama ovih metabolita kod pacijentkinja sa preeklampsijom u odnosu na zdrave. Kod pacijentkinja sa preeklampsijom detektovana je povećana koncentracija pojedinih metabolita koji učestvuju u metabolizmu lipida, koji pripadaju grupama jedinjenja za koje je u literaturi potvrđena veza sa preeklampsijom. Dodatno, rezultati pokazuju razliku u koncentraciji u odnosu na zdrave trudnice i kod pojedinih metabolita za koje još uvek nisu poznata istraživanja koja bi ih povezala sa pojavom ove bolesti ili njenim simptomima.

Ključne reči: bioinformatika, mašinsko učenje, metabolomika, preeklampsija

Sadržaj

Glava 1

Uvod

U ovom poglavlju dat je pregled osnovnih pojmova iz molekularne biologije koji se koriste u ovom istraživanju, zatim je predstavljena metabolomika kao naučna disciplina i na kraju su navedene osnovne informacije o preeklampsiji, oboljenju koje predstavlja fokus ovog istraživanja.

1.1 Osnovni pojmovi

Dezoksiribonukleinska kiselina, skraćeno DNK, predstavlja molekul koji se sastoji od dva dugačka lanca nukleotida, jedinjenja koja sadrže azotne baze. U sastav DNK ulaze sledeće azotne baze: citozin, guanin, adenin i timin. Ribonukleinska kiselina, skraćeno RNK, predstavlja dugačak molekul koji se sastoji od jednog lanca nukleotida. U sastav RNK ulaze sledeće azotne baze: citozin, guanin, adenin i uracil. Određeni delovi DNK sekvence se nazivaju genima i predstavljaju zapise genetičke informacije. Dakle, DNK predstavlja strukturu koja čuva nasledne informacije svakog organizma. Celokupna DNK sekvenca nekog organizma, odnosno niz svih njegovih gena naziva se genom [15].

Za obavljanje svih životnih funkcija neophodni su proteini, jedinjenja koja nastaju na osnovu informacije zapisane u genima. U tom procesu RNK učestvuje u ulozi prenosioca informacije. Centralna dogma molekularne biologije definiše proces sinteze proteina kao protok genetičke informacije od DNK, preko RNK do proteina. Procesom koji se naziva transkripcija, informacija se prepisuje sa DNK i tako nastaje RNK, nakon čega se procesom translacije RNK prevodi u proteine [20].

Lanac ljudske DNK sadrži oko 3 milijarde nukleotida. Postupkom koji se naziva sekvenciranje određuje se tačan redosled nukleotida u ovom lancu. Jedno od naj-

značajnijih dostignuća na polju bioinformatike predstavlja *Human Genome Project*, odnosno projekat sekvenciranja kompletne ljudske DNK, koji je kao rezultat dao referentni ljudski genom. Ovo otkriće dovelo je do različitih istraživanja po pitanju razumevanja gena, kao što su otkrivanje koji geni upravljaju kojim procesom u organizmu, određivanje gena koji su očuvani kod različitih vrsta tokom evolucije ili utvrđivanje veza između određenih gena i pojave nekih oboljenja [2].

Disciplina koja se bavi ovakvom vrstom izučavanja gena naziva se genomika. Njen razvoj podstakao je primenu sličnog pristupa u izučavanju bioloških procesa u celosti, ali na različitim nivoima [8]. Tako dolazi do razvoja srodnih disciplina, prvo transkriptomike čiji je fokus informaciona RNK, zatim proteomike koja se bavi sistematičnim izučavanjem svih proteina i njihovih funkcija i na kraju metabolomike koja predstavlja sveobuhvatnu analizu metabolita i njihovih procesa. Ove naučne discipline su poznate pod zajedničkim nazivom "omike" i postale su sve zastupljenije u različitim naukama.

1.2 Metabolomika

Naučna disciplina koja se bavi izučavanjem metabolita naziva se metabolomika. Metaboliti se mogu definisati kao mali molekuli koji učestvuju u svim biohemijskim procesima ili nastaju kao njihovi produkti. Pripadaju raznovrsnim grupama jedinjenja, kao što su lipidi, peptidi, aminokiseline, ugljeni hidrati, vitamini, masne kiseline i drugi, što njihovu sveobuhvatnu analizu čini komplikovanom [5]. Metabolom, skup svih metabolita nekog organizma, oslikava njegovo celokupno stanje, odnosno i nasledne karakteristike, ali i one uzrokovane načinom života ili spoljnim faktorima. Promene u koncentracijama metabolita mogu ukazivati na promene u samom organizmu, kao što je pojava nekog oboljenja [4]. Upravo zbog toga metaboliti mogu biti značajni za razumevanje oboljenja, pronalazak potencijalnih biomarkera ili testiranje dejstva terapije [24].

Skorašnja tehnološka dostignuća omogućila su merenje i do hiljadu različitih metabolita, najčešće iz uzoraka krvi, što predstavlja prvi korak u metaboličkim istraživanjima [5]. Obrada ovih visokodimenzionih podataka zahteva posebno prilagođenu metodologiju koju pružaju različite tehnike mašinskog učenja.

1.3 Preeklampsija

Preeklampsija predstavlja ozbiljno oboljenje koje se može javiti tokom trudnoće u čak 3-5% slučajeva i predstavlja glavni uzrok smrtnosti kod majki i fetusa. Karakteriše se visokim krvnim pritiskom i proteinuriom u ranoj fazi, a kasnije može dovesti i do disfunkcije placente, oštećenja jetre i bubrega, hematoloških i drugih komplikacija [14]. Zbog ovih ozbiljnih posledica, od velike važnosti je rana detekcija i tretman preeklampsije.

Dijagnoza preeklampsije pre pojave simptoma još uvek nije poznata, a utvrđivanje rizika za njihovu pojavu se zasniva na tradicionalnim pregledima. Aktuelna istraživanja testiraju metode koje bi za ovaj problem koristile biomarkere [21].

Jedan od pristupa u otkrivanju potencijalnih biomarkera predstavlja metabolomička analiza. Cilj je pronalazak metabolita koji bi pomogli u ranoj detekciji ove bolesti, ali i utvrđivanje njihove veze sa drugim faktorima, poput godina starosti ili telesne mase. Dodatno, metabolomičke tehnike se mogu koristiti i za testiranje uticaja terapije. Iako još uvek nije poznat lek za ovo oboljenje, istraživanja su potvrdila pozitivan efekat aspirina na ublažavanje ovih simptoma [10].

Glava 2

Podaci

U ovom poglavu biće dat pregled podataka korišćenih u ovom istraživanju, njihova detaljna analiza i postupak pripreme podataka neophodan za dalji rad.

2.1 Analiza podataka

Skup podataka korišćen u ovom radu preuzet je iz istraživanja koje se bavi ispitivanjem uticaja aspirina za ublažavanje simptoma preeklampsije [10]. Prikupljeni podaci za 463 pacijentkinje i 968 metabolita strukuirani su u tri tabele:

1. `Pacijentkinje` - sadrži informacije o trimestru trudnoće u kom je uzet uzorak i gestacijskoj starosti ploda, godinama starosti pacijentkinje, njenoj telesnoj masi, etničkoj pripadnosti i slično. Pored ovih, za svaku pacijentkinju je dostupna informacija o tome da li ima ili nema preeklampsiju, kao i da li je pripadala grupi koja dobija aspirin ili placebo grupi. Na slici 2.1 prikazan je segment ove tabele.

2. `Metaboliti` - sadrži detaljne informacije o samim metabolitima, kao što su njihov zvanični hemijski naziv, klasifikacija, metabolički putevi u kojima učestvuju, kao i odgovarajući ključevi u različitim bazama podataka koje čuvaju informacije o hemijskim i biohemijskim jedinjenjima. Na slici 2.2 prikazan je segment ove tabele.

3. `Metaboliti kod pacijentkinja` - sadrži z-vrednosti koncentracija metabolita izmerenih iz krvi pacijentkinja koje su učestvovale u istraživanju. Na slici 2.3 prikazan je segment ove tabele.

sIDs	ga.w	ASA_tri_chr	weight	crl	smoking	trimester	pe	ptIDs	consss_batch	BErm	conception	ASA	prev.pe	age	Compliance	height	STUDY	sle	race
BAYL-09198	32.0	PLACEBO - 3rd Trimester	69.0	67.6	No	3	0.0	S2_28	batch-2023	batch_2023	Spontaneous	False	Nullip	31.3	0.0	174.0	ASPRE_LR	No	Black
BAYL-09199	12.1	PLACEBO - 1st Trimester	70.3	56.2	No	1	0.0	S2_38	batch-2023	batch_2023	Spontaneous	False	Nullip	31.3	0.0	164.0	ASPRE_LR	No	White
BAYL-09200	23.0	PLACEBO - 2nd Trimester	70.3	56.2	No	2	0.0	S2_38	batch-2023	batch_2023	Spontaneous	False	Nullip	31.3	0.0	164.0	ASPRE_LR	No	White
BAYL-09201	32.0	PLACEBO - 3rd Trimester	70.3	56.2	No	3	0.0	S2_38	batch-2023	batch_2023	Spontaneous	False	Nullip	31.3	0.0	164.0	ASPRE_LR	No	White
BAYL-09202	12.3	PLACEBO - 3rd Trimester	98,2	58,5	No	3	1.0	S3_13	batch-2023	batch_2023	Spontaneous	False	Nullip	24,4	0.0	156,0	ASPRE_HR	No	White

Slika 2.1: Segment tabele `Pacijentkinje`. Redovi tabele označavaju različite uzorke, među kojima neki odgovaraju istoj pacijentkinji, ali različitom trimestru trudnoće. Kolona *ptIDs* predstavlja identifikator pacijentkinje i na osnovu njega se mogu grupisati uzorci iz različitih trimestara iste pacijentkinje, ukoliko su dostupni. Preeklampsija je predstavljena kolonom *pe*, gde 1 ili 0 predstavlja da pacijentkinja ima ili nema preeklampsiju. Kolonom *ASA* označeno je da li je pacijentkinja dobijala terapiju aspirinom. Naziv kliničke studije iz koje potiče uzorak naveden je u koloni *STUDY*.

BIOCHEMICAL	PATHWAY_SORTORDER	SUPER_PATHWAY	SUB_PATHWAY	COMP_ID	PLATFORM	CHEMICAL_ID	PUBCHEM	CAS	KEGG	HMDB_ID
S-1-pyrroline-5-carboxylate	64.0	Amino Acid	Glutamate Metabolism	42370	Pos Early	35	11966181	2906-39-0	C04322	HMDB0001301
spermidine	553.0	Amino Acid	Polyamine Metabolism	485	Pos Early	50	1102	124-20-9	C00315	HMDB0001257
1-methylnicotinamide	4336.0	Cofactors and Vitamins	Nicotinate and Nicotinamide Metabolism	27665	Pos Early	55	457	1005-24-9	C02918	HMDB0000699
12,13-DiHOME	2048.0	Lipid	Fatty Acid, Dihydroxy	38395	Neg	62	10236635	263399-35-5	C14829	HMDB0004705
5-hydroxyindoleacetate	295.0	Amino Acid	Tryptophan Metabolism	437	Neg	71	1826	54-16-0	C05635	HMDB0000763

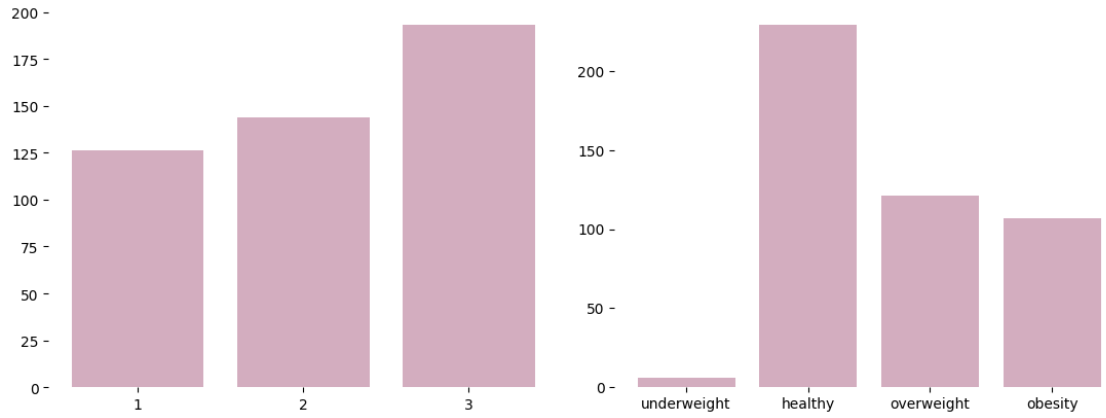
Slika 2.2: Segment tabele `Metaboliti`. Kolona *BIOCHEMICAL* sadrži zvanični naziv jedinjenja, a naredne tri kolone se odnose na metaboličke puteve u kojima navedeno jedinjenje učestvuje. Identifikator metabolita predstavljen je kolonom *COMP_ID*. Poslednjih pet kolona sadrži identifikatore tog jedinjenja u različitim bazama podataka hemijskih jedinjenja.

	42370	485	27665	38395	437	528	1417	1549	531	1414	...	53157	53266	53267
BAYL-08997	-1.259948	-0.610334	-1.201809	-0.954275	0.240999	-0.485987	0.598959	-0.286401	-1.043185	-0.660514	...	-1.094129	-0.419479	0.158619
BAYL-08998	0.170557	-0.811169	0.387317	-0.906589	-0.454249	-0.409053	0.612680	-0.315412	-1.095163	0.001887	...	-1.116107	0.000000	-0.424652
BAYL-08999	0.814204	-0.554787	-1.062940	4.929942	-0.000406	0.022588	3.050217	2.366699	1.132455	-0.142467	...	0.391963	0.284916	0.020418
BAYL-09000	0.038958	-0.603603	0.525122	0.974496	-0.542391	-0.494779	-0.347355	-0.392808	-0.713237	-0.359097	...	-0.119144	-0.807474	1.317694
BAYL-09001	-0.504466	-0.549375	-0.973446	-0.740805	0.000000	-0.049422	-0.787278	-0.598359	-0.781751	-0.515097	...	-0.351091	0.000000	-0.405493

Slika 2.3: Segment tabele `Metaboliti kod pacijentkinja`. Tabela sadrži z-vrednosti izmerenih metabolita za svaku pacijentkinju. Kolonama su predstavljeni metaboliti, označeni svojim identifikatorom iz tabele `Metaboliti`.

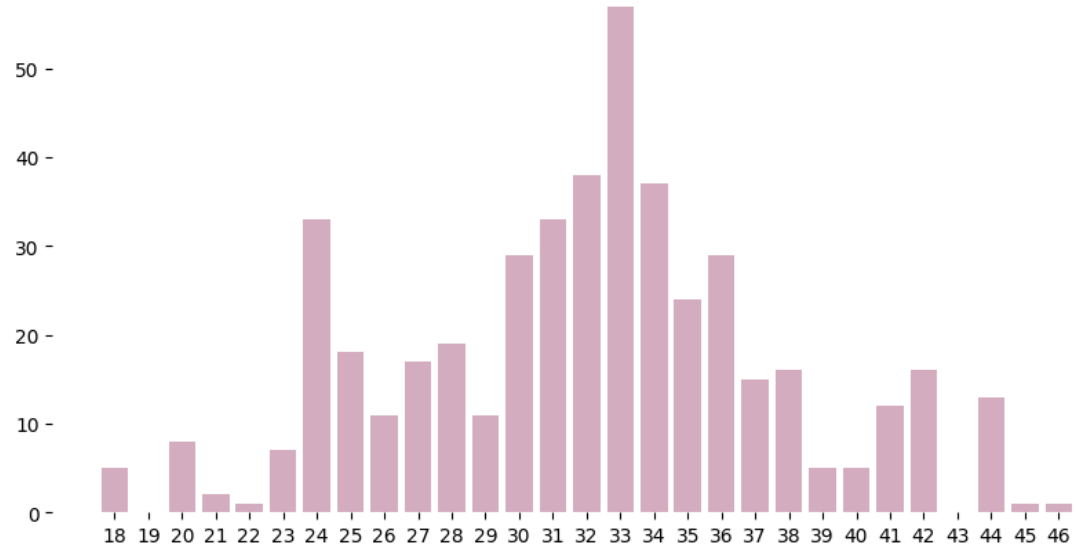
U korišćenom skupu podataka dostupna je informacija o telesnoj masi i visini svake pacijentkinje, pa je u skladu sa tim moguće izračunati BMI. Ovim podatkom proširene su informacije o pacijentkinjama. Na slici 2.4 prikazan je broj pacijentkinja

po trimestru trudnoće, kategorijama indeksa telesne mase i godinama starosti. Može se uočiti veći broj pacijentkinja sa visokim *BMI*, što je očekivano s obzirom na to da su u pitanju trudnice. Na dijagramu se može videti i da je većina pacijentkinja u tridesetim godinama.



(a)

(b)



(c)

Slika 2.4: Broj instanci prema a) trimestru, b) kategoriji *BMI*, c) godinama starosti. Najveći broj uzoraka je iz trećeg trimestra, veliki broj pacijentkinja ima visok *BMI* i pretežno su u tridesetim godinama.

Za prikazani broj instanci po trimestrima potrebno je naglasiti da su za pojedine pacijentkinje dostupni podaci iz više trimestara. Od ukupno 253 različite pacijentkinje, samo za 85 su poznati podaci iz sva tri trimestra, od kojih su samo dve sa preeklampsijom. Pretpostavlja se da se razlika u broju pacijentkinja po trimestri-

ma javlja zbog naknadnog priključivanja pojedinih pacijentkinja u kasnijem periodu trudnoće, tek nakon dobijene dijagnoze, odnosno pojave simptoma.

Dimenzije početnog skupa podataka prikazane su u tabeli 2.1. U nastavku će biti praćena promena ovih vrednosti tokom postupka pripreme podataka.

Broj atributa	Broj instanci	Preeklampsija	Kontrolna grupa
968	463	174	289

Tabela 2.1: Dimenzije izvornog skupa podataka

2.2 Priprema podataka

Kao što je pomenuto, tabela `Pacijentkinje` između ostalog sadrži informacije da li pacijentkinje uzimaju aspirin ili ne. S obzirom da se aspirin koristi za ublažavanje simptoma preeklampsije, mogao bi dovesti do promene koncentracija metabolita u odnosu na stvarne. Zbog toga su takve pacijentkinje isključene iz skupa podataka i u daljoj analizi su korišćeni samo podaci pacijentkinja koje nisu uzimale aspirin i koje čine placebo grupu. Dodatno, isključene su instance kod kojih nije dostupna koncentracija za više od 80% metabolita, a preostale nedostajuće z-vrednosti zamenjene su vrednošću 0.

U korišćenom skupu podataka, nazivi pojedinih metabolita nisu poznati, zbog čega su kolone koje im ogovaraju takođe isključene iz daljeg istraživanja. Iako je njihova koncentracija izmerena kod određenog broja pacijentkinja, bilo kakvo otkriće vezano za njih ne bi imalo značaj. Tabela 2.2 prikazuje dimenzije skupa nakon navedenih izmena.

	Broj atributa	Broj instanci	Preeklampsija	Kontrolna grupa
izvorni skup podataka	968	463	174	289
nakon izmena opisanih u 2.2	798	312	97	215

Tabela 2.2: Dimenzije skupa podataka. Prikazane su dimenzije izvornog skupa podataka i promenjene vrednosti nakon isključivanja instanci koje odgovaraju aspirin grupi i metabolita za koje nisu dostupni podaci.

2.3 Efekat serija

Podaci koji potiču iz laboratorije gotovo nikada nisu tehnički homogeni. Različito vreme izvođenja eksperimenata ili različita serija reagenasa koji se koriste mogu uticati na podatke i stvoriti privid podele skupa podataka na različite grupe, odnosno serije (eng. *batch*) [7]. Ovaj efekat poznat je pod nazivom *efekat serija* (eng. *batch effect*) i neophodno ga je ukloniti iz podataka, kako ne bi uticao na rezultate i doveo do pogrešnih zaključaka.

Podaci u korišćenom skupu podataka potiču iz 4 različita klinička istraživanja, pa je potrebno utvrditi da li se u skupu javlja efekat serija. U te svrhe, izvršena je UMAP projekcija podataka, kako bi bilo moguće vizuelizovati instance visoke dimenzije. Algoritam UMAP (eng. *Uniform Manifold Approximation and Projection*), predstavlja jedan od algoritama koji se koriste za redukciju dimenzionalnosti podataka, kojim je moguće modelirati visokodimenzione podatke sa neodređenom topološkom strukturom [12].

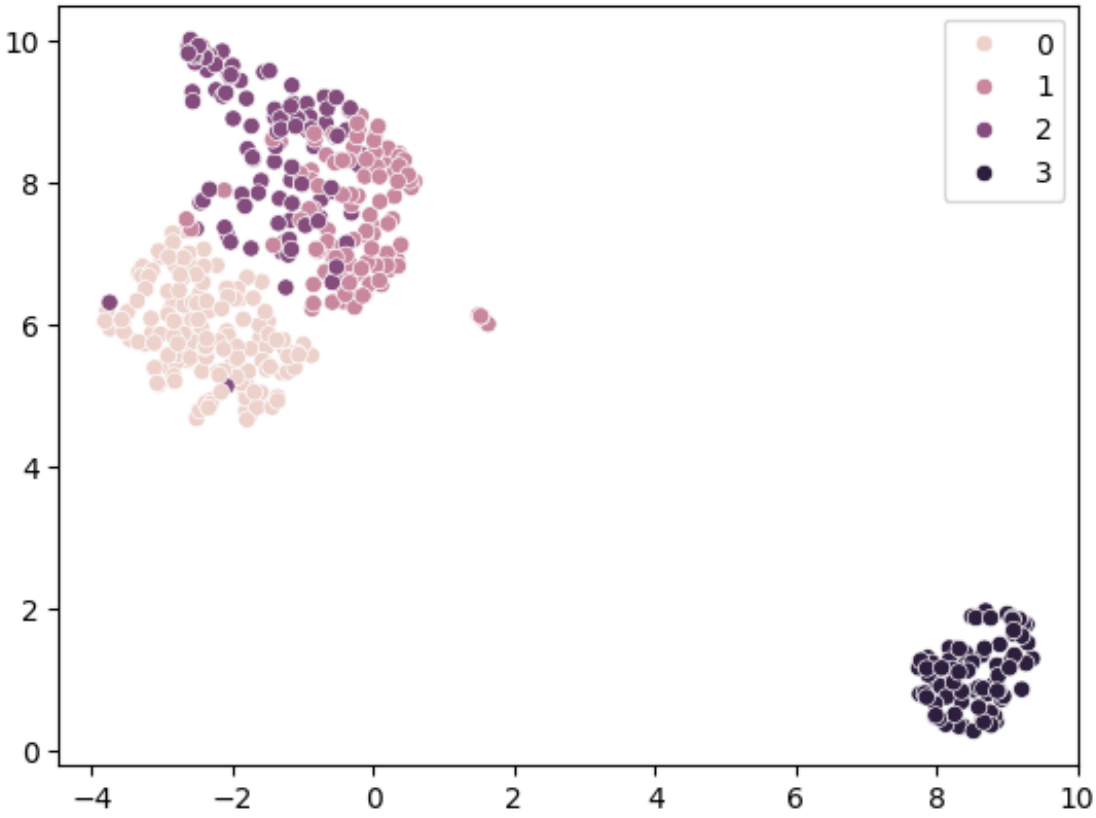
Na ovaj način, podaci su predstavljeni tačkastim dijagramom (eng. *scatter plot*) i označeni različitim bojama na osnovu studije iz koje potiču. Na slici 2.5 mogu se jasno uočiti klasteri među podacima, veštački nastali usled prisustva efekta serija.

Jedan od najčešće korišćenih alata za uklanjanje efekta serija, poznat pod nazivom *ComBat*, pokazao je odlične performanse na različitim bioinformatičkim skupovima podataka, posebno kada se radi o skupovima sa malim brojem instanci [1]. Iz tog razloga je u ovom radu korišćena verzija ovog alata implementirana u programskom jeziku *Python*.

Za primenu ovog alata se, pored podataka koje je potrebno korigovati, prosleđuju i oznake unapred poznatih serija. Informacija iz koje studije potiče koji podatak dostupna je zajedno sa ostalim informacijama o pacijentkinjama.

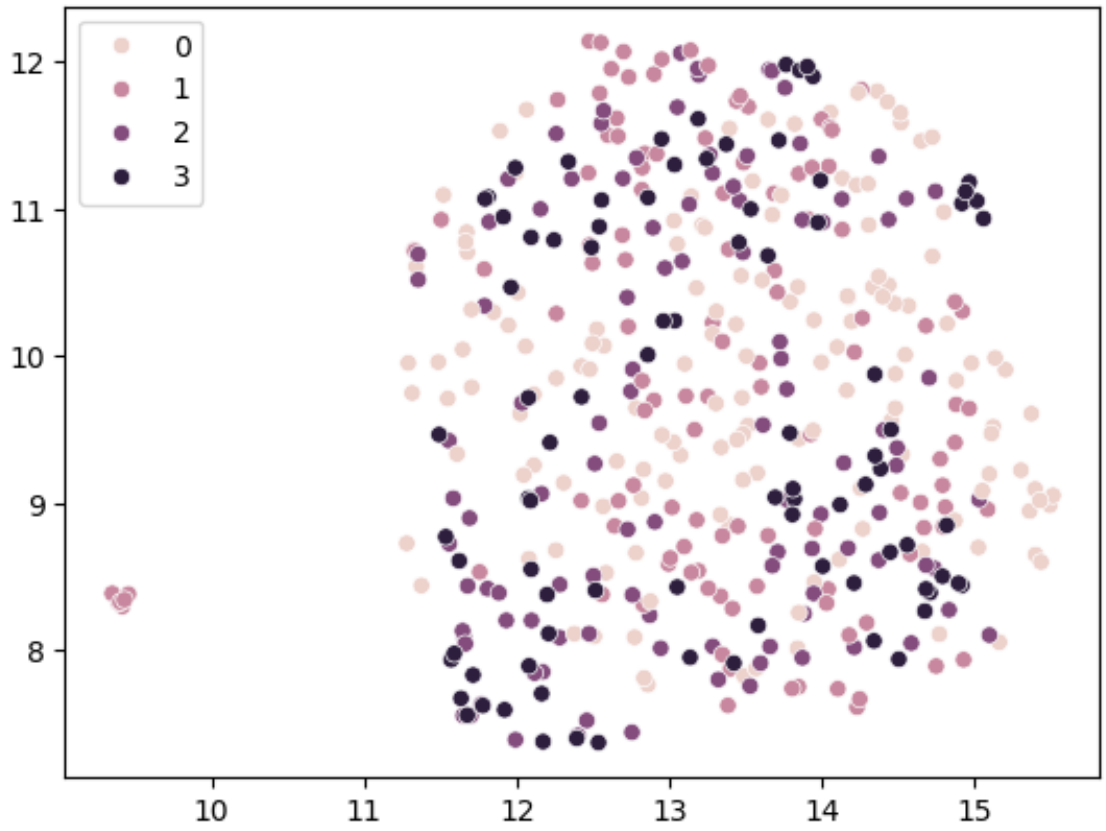
Slika 2.6 predstavlja UMAP projekciju podataka nakon korekcije efekta serija. Na ovoj slici se može primetiti da instance više nisu grupisane u klastere prema studijama iz kojih su dobijeni.

Još jedno opažanje na slikama 2.5 i 2.6 odnosi se na odudarajuće podatke. Može se uočiti grupa od nekoliko instanci udaljenih od ostalih podataka. Njihovi indeksi su locirani na projekciji nakon uklanjanja efekta serija korišćenjem *IQR* (*Interquartile range*) metode za detekciju elemenata van granica. Upoređivanjem vrednosti koncentracija metabolita koji odgovaraju izdvojenim instancama sa prosečnim vrednostima preostalih uočene su značajne razlike za pojedine metabolite. Utvrđeno da



Slika 2.5: UMAP projekcija podataka pre uklanjanja efekta serija. Različitim bojama prikazani su uzorci iz različitih kliničkih istraživanja. Pre primene alata *ComBat* mogu se uočiti prisutni klasteri kao posledica efekta serija.

izdvojene instance predstavljaju odudarajuće podatke, zbog čega su uklonjene iz dalje analize. Nove dimenzije skupa prikazane su u tabeli 2.3.



Slika 2.6: UMAP projekcija podataka nakon uklanjanja efekta serija. Različitim bojama prikazani su uzorci iz različitih kliničkih istraživanja. Nakon primene alata *ComBat* nisu prisutni klasteri.

	Broj atributa	Broj instanci	Preeklampsija	Kontrolna grupa
izvorni skup podataka	968	463	174	289
nakon izmena opisanih u 2.2	798	312	97	215
nakon izmena opisanih u 2.3	798	305	97	208

Tabela 2.3: Dimenzije skupa podataka. Prikazane su dimenzije izvornog skupa podataka, zatim promenjene vrednosti nakon isključivanja instanci koje odgovaraju aspirin grupi i metabolita za koje nisu dostupni podaci i na kraju nakon isključivanja instanci za koje je utvđeno da predstavljaju odudarajuće podatke.

Glava 3

Metode

U ovom poglavlju biće prikazan izbor odgovarajućih metoda mašinskog učenja u skladu sa konkretnim ciljem istraživanja i prisutnim ograničenjima, zatim detalji implementacije samog modela i na kraju njegova evaluacija.

Analiza metabolomičkih podataka predstavljenih numeričkim atributima u formi tabele je problem pogodan za primenu nekih od tehnika mašinskog učenja. U odnosu na konkretan problem, ali i cilj istraživanja, neophodno je izabrati odgovarajuće metode. Ograničenja sa kojima se susrećemo u bioinformatičkim problemima najčešće se odnose na veličinu skupa podataka. Izmereni broj metabolita, odnosno broj atributa skupa, vrlo često je veći od broja instanci [11], što uz prisutan šum i veliki broj nedostajućih vrednosti otežava procesiranje ovakvih podataka i može lako dovesti do preprilagođavanja. Skup podataka koji je korišćen u ovom istraživanju sadrži 305 instanci sa 798 atributa, od kojih 97 instanci odgovara pacijentkinjama sa preeklampsijom (Tabela 2.3).

Osnovni cilj ovog istraživanja je identifikacija potencijalnih biomarkera među svim izmerenim metabolitima. Dakle, treba utvrditi koji od atributa u skupu imaju najjaču vezu sa kolonom koja sadrži informaciju o tome da li pacijentkinja ima preeklampsiju ili ne. Umesto analize pojedinačnih metabolita ili njihovih kombinacija, ideja je treniranje modela nadgledanog učenja na ovom skupu podataka, koji bi naučio te veze.

Rezultujući model se ne bi koristio kao alat za dalju predikciju oboljenja, već kao alat za utvrđivanje metabolita koji najviše doprinose odluci. To podrazumeva izbor tehnike koja ima svojstvo interpretabilnosti. Aktuelna istraživanja razvijaju interpretabilne metode koje daju dobre rezultate na visokodimenzionim skupovima sa manjim brojem instanci [18], a kako njihova implementacija još uvek nije dostupna,

potrebno je prevazići ograničenja na drugi način.

3.1 Redukcija skupa atributa

Prvi korak implementacije modela nadgledanog učenja podrazumeva redukciju dimenzionalnosti skupa podataka. Umesto primene algoritama poput PCA, koji kreiraju nove atribute u manje dimenzionom prostoru, ideja je zadržati originalne atribute, ali tako da najbolje moguće opisuju skup. Algoritmi za odabir atributa korišćeni u ovom radu su *mRMR* i *SelectKBest* i u nastavku će biti ukratko opisani.

Algoritam *mRMR* [9] (eng. *Minimum redundancy maximum relevance*) pronalazi podskup atributa zadate veličine, koji imaju najveću korelisanost sa ciljnom promenljivom, a najmanju međusobnu korelisanost. Prvi kriterijum, odnosno relevantnost odabranih atributa, određuje se F-testom, a drugi kriterijum koji se odnosi na najmanju moguću redundantnost među njima, određuje se izračunavanjem Pirsonovog koeficijenta korelacije[1] [17].

SelectKBest [19] predstavlja skup algoritama biblioteke *scikit learn* koji prema zadatom kriterijumu biraju k najboljih atributa. Odabrana metrika određuje konkretan algoritam koji se koristi, a u ovom radu su testirane *f_ classif* koja koristi ANOVA statistiku i *mutual_ info_ classif* koja koristi meru zajedničke informacije.

Dva pomenuta pristupa na sličan način implementiraju rangiranje atributa prema izračunatim statistikama. Razlika je u tome što *mRMR*, kao pohlepni algoritam, iterativno dodaje jedan po jedan atribut, izostavljajući one za koje se ispostavi da su redundantni u odnosu na već dodate, dok *SelectKBest* samo odabere prvih k. Iz tog razloga, kreirana su dva modela koja se razlikuju po metodi za odabir atributa.

Hiperparametar koji figuriše kod obe metode odnosi se na veličinu skupa atributa koje treba zadržati. Kod oba pristupa određen je na isti način i to korišćenjem tehnike unakrsne validacije. Dodatno, statistika koju koristi *SelectKBest* tretirana je takođe kao hiperparametar i testirane su obe navedene funkcije (*f_ classif* i *mutual_ info_ classif*). Metoda za izbor atributa ulančana je sa narednim koracima i najbolja kombinacija svih hiperparametara izabrana je za kompletan model.

Ograničenje koje se javlja kod korišćenja *mRMR* je interfejs koji nije kompatibilan sa *scikit-learn* bibliotekom, što otežava ulančavanje ove metode sa klasifikatorom, kao i primenu algoritma pretrage optimalne konfiguracije. Za prevazilaženje ovog

[1] Ove statistike se odnose na slučaj neprekidnih tipova atributa, dok se za diskretne koristi mera zajedničke informacije (eng. *mutual information*).

problema iskorišćena je klasa *BaseEstimator* spomenute biblioteke koja omogućava proširivanje klase *mRMR* implementiranjem metoda *fit* i *transform*.

Algoritam *mRMR* ima mogućnost kreiranja više alternativnih rešenja, odnosno može se zadati vrednost za željeni broj rešenja. To može biti korisno, ako je na primer poznato da je neki metabolit teže izolovati, tada se može odabrati rešenje u kojem je on izostavljen. Ova opcija je implementirana tako da se razmatraju samo relevantni atributi, ali potencijalno različitim redosledom, što može rezultovati različitim konačnim skupom atributa. Sama implementacija ne garantuje da će kreirana rešenja zaista biti različita. Testiranje različitih vrednosti ovog parametra pokazalo je da se dobijeni podskupovi razlikuju najčešće samo u jednom atributu, stoga je korišćena podrazumevana vrednost, odnosno kreirano je samo jedno rešenje. Ostali parametri su takođe korišćeni sa svojim podrazumevanim vrednostima.

3.2 Izbor klasifikatora

Nakon izbora optimalnog podskupa atributa, cilj je detektovati potencijalne biomarkere među njima. Kao što je već spomenuto, taj korak će biti izvršen uz pomoć metoda mašinskog učenja. Kako je poznata vrednost ciljne promenljive koja se odnosi na to da li pacijentkinja ima ili nema preeklampsiju, problem koji se rešava predstavlja binarnu klasifikaciju. To dovodi do problema izbora odgovarajuće metode klasifikacije za konkretan problem.

Model	Tačnost	*AUC*	*F1*	Preciznost	Odziv
Logistička regresija	0.65	0.63	0.39	0.47	0.35
SVM (linearni)	0.66	0.64	0.42	0.48	0.38
SVM (rbf)	0.66	0.7	0.39	0.46	0.34
Slučajne šume	0.7	**0.81**	0.35	0.57	0.26
XGBoost	**0.73**	**0.81**	**0.55**	**0.59**	**0.52**

Tabela 3.1: Poređenje različitih tehnika klasifikacije. Prikazane su ocene kvaliteta za svaku od testiranih tehnika. Podebljane su maksimalne vrednosti po kolonama i sve odgovaraju *XGBoost* klasifikatoru.

Na samom početku istraživanja testirane su neke od osnovnih tehnika binarne klasifikacije i poređenje dobijenih mera kvaliteta prikazano je u tabeli 3.1. U skladu sa očekivanjem, logistička regresija i metod potpornih vektora (eng. *Support vector*

machine, SVM) pokazali su slabe performanse na konkretnom problemu, dok je bolje rezultate pokazao je pristup korišćenjem ansambla. O ovom pristupu će biti više reči u nastavku, s obzirom da se pokazao kao najpogodniji za ovaj problem.

3.3 Obučavanje modela

Ansambl metoda predstavlja jednu od često korišćenih metoda mašinskog učenja. Najjednostavniji način kreiranja ansambla je prosta agregacija (eng. *bagging*), odnosno grupisanje većeg broja nezavisnih klasifikatora. Metod slučajnih šuma (eng. *Random Forest*) zasniva se na ovom pristupu i jedan je od metoda testiranih na problemu preeklampsije. Pojačavanje (eng. *Boosting*) predstavlja drugačiji pristup za konstrukciju ansambla, kod kog se naredni klasifikator dodaje tako da popravi performanse dotadašnjeg. Ovaj pristup je pokazao izuzetne performanse u različitim domenima primene, konkretno model gradijentnog pojačavanja pod nazivom *XGBoost* (eng. *eXtreme Gradient Boosting*) [3]. Iako metod slučajnih šuma nije imao posebno loše performanse, *XGBoost* se pokazao kao bolji kandidat i on će biti detaljnije razmatran u nastavku.

XGBoost je biblioteka koja, kao što je već pomenuto, za konstrukciju ansambla koristi gradijentno pojačavanje. Princip pojačavanja podrazumeva građenje ansambla dodavanjem jednog po jednog modela, kako bi se pojačale dotadašnje performanse, za razliku od proste agregacije (npr. metod slučajnih šuma) kod koje se to radi paralelno. Gradijentno pojačavanje implementira ovaj princip po uzoru na algoritam gradijentnog spusta. Implementacija se zasniva na minimizaciji funkcije greške i to u pravcu lokalnog poboljšanja. Iterativnim algoritmom dodaje se novo stablo koje u tom trenutku najviše smanjuje funkciju greške, a kako bi ispitivanje svih mogućih struktura stabala bilo gotovo nemoguće, započinje se od jednog lista i pohlepnim pristupom dodaju nove grane [3].

Dodatna prednost ovog izbora je interpretabilnost. Obučeni model svakom od atributa pridružuje važnost (*feature importance*) koja se računa na osnovu zadate funkcije. Podrazumevana funkcija je dobit (eng. *Gain*) i opisuje koliko u proseku svaki od atributa doprinosi poboljšanju modela, odnosno smanjenju funkcije greške. Na osnovu ovog parametra, atributi se mogu rangirati, odnosno mogu se uočiti oni koji najviše doprinose poboljšanju predviđanja. Na konkretnom primeru identifikacije potencijalnih biomarkera ideja je obučiti ovakav model da pravi predviđanja vezana za preeklampsiju i rezultujuće važnosti atributa tumačiti kao značaj odgovarajućih

metabolita.

Prva opcija koju je potrebno podesiti određuje koja funkcija greške će se izračunavati tokom obučavanja modela. Kako je problem koji se rešava binarne prirode, jer je potrebno predvideti da li pacijentkinja ima ili nema preeklampsiju, odabrana je opcija *binary:logistic*. Na osnovu ove vrednosti se za funkciju greške postavlja sigmoidna funkcija, kao kod logističke regresije. Vrednosti ove funkcije se mogu tumačiti kao verovatnoća pripadnosti pozitivnoj klasi, odnosno klasi sa preeklampsijom.

Odabir drugih hiperparametara je, kao što je spomenuto kod selekcije atributa, izvršen primenom unakrsne validacije za izbor optimalne konfiguracije. Ulančane metode za smanjenje dimenzionalnosti, standardizaciju podataka i klasifikaciju, predstavljaju kompletan model za koji je potrebno odabrati optimalne hiperparametre. Za klasifikator su testirane različite vrednosti za maksimalan broj stabala i za njihovu maksimalnu dubinu. Loš izbor ovih parametara, na primer preduboka stabla ili preveliki broj estimatora, lako mogu dovesti do preprilagođavanja modela, posebno imajući u vidu mali broj instanci u skupu. Pretraga najbolje kombinacije je izvršena korišćenjem algoritma *GridSearchCV*.

Pri izvršavanju ove pretrage, pod najboljim hiperparametrima podrazumevaju se oni sa kojima model daje najbolje rezultate na osnovu neke zadate metrike. Pri prvom pokušaju, za ovu metriku je odabrana tačnost predviđanja. Iako je tačnost klasifikacije često korišćena kao mera kvaliteta modela, visoke vrednosti ne garantuju zaista dobar model. Na konkretnom slučaju, druge metrike poput preciznosti i odziva, pokazale su da dobijeni model značajno greši u predviđanjima. Dok su instance koje odgovaraju zdravim trudnicama uglavnom ispravno prepoznate, pacijentkinje sa preeklampsijom su često označene kao zdrave. Zamenom metrike na osnovu koje se evaluiraju hiperparametri dobijen je model sa nešto boljim performansama. Umesto tačnosti, korišćene su *F1* mera (odnos preciznosti i odziva) i *AUC score* (površ ispod *ROC* krive).

3.4 Evaluacija modela

Kako bi se na osnovu konačnog modela mogla vršiti identifikacija značajnih atributa, on mora da pokaže određeni kvalitet. U nedostatku velike količine podataka, nije moguće vršiti evaluaciju podelom na podskupove za obučavanje, validaciju i ocenu kvaliteta. Tokom procesa implementacije, pokazalo se da bi način podele skupa doveo do značajnih razlika u ocenama. Iako je izvršena stratifikacija, pri slučajnom

izboru skupa za testiranje, jedan isti model davao je različite vrednosti izračunatih metrika. Iz tog razloga potrebno je promeniti pristup kako bi se dobila objektivna i relevantna ocena kvaliteta.

Za izbor hiperparametara korišćen je pristup sa primenom unakrsne validacije, konkretno primenom algoritma *GridSearchCV* sa 5 slojeva (eng. *folds*). Ovaj pristup omogućava dobijanje nepristrasne ocene za svaku od testiranih konfiguracija, na osnovu čega se na kraju bira najbolja.

Evaluacija konačnog modela izvršena je primenom ugnežđene unakrsne validacije. Funkciji za računanje zadatih metrika metodom unakrsne validacije, u ovom slučaju se ne prosleđuje jedan konkretan model, već instanca klase *GridSearchCV* sa svim hiperparametrima koje je potrebno testirati. Ponovo je izvršena stratifikovana podela na 5 slojeva. Implementacija klase *GridSearchCV* je takva da se metodi *fit* i *transform* odnose na konfiguraciju sa najboljim performansama u tom ciklusu. Dakle, na kraju svakog ciklusa se najbolji model obuči na podskupu koji je u tom krugu namenjen za trening i za njega izračunaju zadate metrike. Prosečne vrednosti metrika koje su izračunate na pojedinačnim slojevima aproksimiraju stvarnu ocenu kvaliteta krajnjeg modela koji se dobija obučavanjem na celom skupu podataka.

Isti princip primenjen je za oba kreirana modela koji se razlikuju po načinu izbora atributa u prvom koraku. Dobijena vrednost za veličinu optimalnog skupa atributa je 40, a za maksimalnu dubinu stabla 10 u oba slučaja. Kao bolja funkcija za rangiranje atributa kod *SelectKBest* selektora, pokazala se ANOVA.

Ocene kvaliteta dobijenih modela prikazane su u tabeli 3.2. Kako u literaturi nisu pronađene slične tehnike za istraživanje preeklampsije sa kojima bi dobijeni rezultati mogli biti upoređeni, za detekciju značajnih metabolita biće korišćena dobijena dva modela, kao klasifikatori koji su pokazali najbolji rezultat na korišćenom skupu podataka.

Model	Tačnost	*AUC*	*F1*	Preciznost	Odziv
mRMR selektor	0.74	0.81	0.56	0.63	0.52
SelectKBest selektor	0.7	0.76	0.47	0.54	0.43

Tabela 3.2: Ocene rezultujućih modela. Prikazane su ocene kvaliteta za dva dobijena modela koji se razlikuju u načinu izbora atributa (selektoru atributa). Model sa *mRMR* selektorom daje nešto bolje performanse.

Glava 4

Rezultati

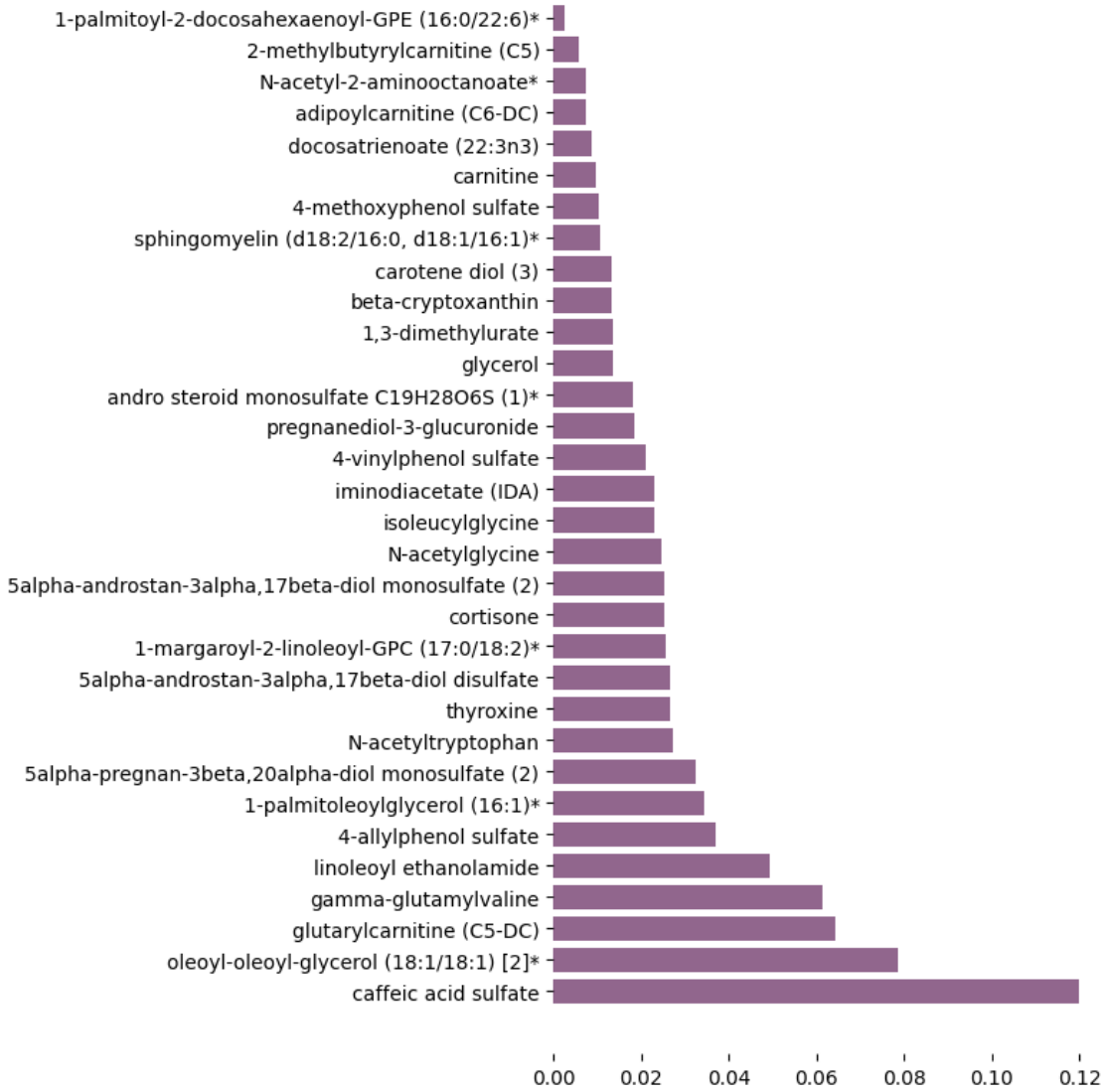
U ovom poglavlju su prikazani rezultati statističkih testova, a zatim je dat pregled najznačajnijih metabolita i diskusija dobijenih rezultata u poređenju sa literaturom. Za detekciju metabolita koji najviše doprinose odluci klasifikacije implementirana je funkcija za mapiranje važnosti atributa na identifikator metabolita kojem odgovara. Ove vrednosti dostupne su kao svojstvo obučenog *XGBoost* klasifikatora pod nazivom *feature_importances_*.

4.1 Statistički testovi

Razlika u raspodeli koncentracija većine izdvojenih metabolita kod pacijentkinja sa preeklampsijom u odnosu na zdrave trudnice potvrđena je statističkim testovima. Korišćeni testovi su *Mann-Whitney* test i T-test implementirani u okviru biblioteke *SciPy*. Oba testa se koriste za utvrđivanje da li postoji statistički značajna razlika između dve grupe nezavisnih uzoraka. Razlika je u tome što se *Mann-Whitney* test može primeniti i u slučaju kad podaci ne dolaze iz normalne raspodele. Nulta hipoteza za oba metoda je da uzorci iz dve grupe potiču iz iste raspodele. Za p-vrednost veću od zadatog praga ova hipoteza se odbacuje i zaključuje se da među njihovim raspodelama postoji značajna razlika. Zadati prag za p-vrednost je 0.05.

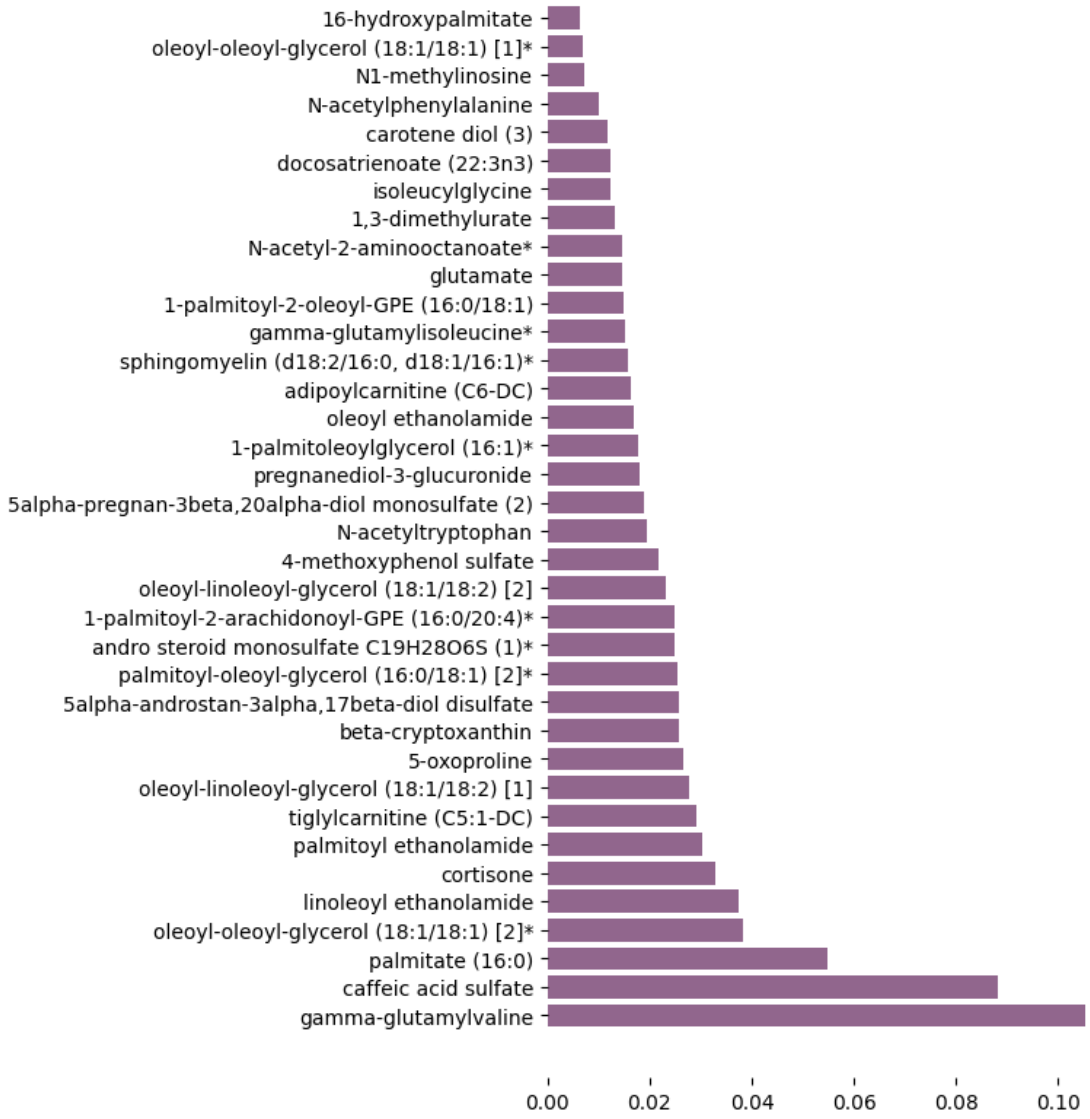
Za primenu T-testa potrebno je da vrednosti koncentracija metabolita budu iz normalne raspodele. Takvi metaboliti izdvojeni su korišćenjem *Shapiro* testa. Nulta hipoteza ovog testa je da uzorci ne dolaze iz normalne raspodele, pa su izdvojeni oni kod kojih je dobijena p-vrednost veća od zadatog praga 0.05.

Metaboliti za koje je potvrđena statistička značajnost prikazani su na slikama 4.1 i 4.2 sortirani prema važnosti atributa. Od 40 odabranih metabolita, kod oba



Slika 4.1: Statistički značajni metaboliti modela sa *mRMR* selektorom. Od 40 metabolita koje je ovaj selektor odabrao, za 32 je potvrđena statistička značajnost. Za metabolite je prikazana važnost (*feature importance*) u odnosu na *XGBoost* klasifikator i prema ovoj vrednosti su sortirani.

pristupa za selekciju atributa, pokazalo se da je statistički značajno 32 kod *mRMR* pristupa i 36 kod *SelectKBest*. Među ovim metabolitima nalazi se čak 20 zajedničkih, iako se pristup kojim su odabrani u određenoj meri razlikuje. Među 5 najbolje rangiranih nalaze se čak 4 ista, *caffeic acid sulfate*, *oleoyl-oleoyl-glycerol*, *gamma-glutamylvaline* i *linoleoyl ethanolamide*. Dva pomenuta pristupa za odabir atributa se ipak ne razlikuju u potpunosti, već su jednim delom slični (po načinu izbora



Slika 4.2: Statistički značajni metaboliti modela sa *SelectKBest* selektorom. Od 40 metabolita koje je ovaj selektor odabrao, za 36 je potvrđena statistička značajnost. Za metabolite je prikazana važnost (*feature importance*) u odnosu na *XGBoost* klasifikator i prema ovoj vrednosti su sortirani.

atributa koji su u korelaciji sa ciljnom promenljivom), što je opisano u prethodnom poglavlju. Iz tog razloga su detaljnije istraženi metaboliti koji su bolje rangirani, umesto zajedničkih metabolita sa lošijim rangom.

Svi prikazani metaboliti kod kojih je potvrđena statistička značajnost mogu biti potencijalni kandidati za biomarkere. Dalja analiza biće predstavljena na primeru deset najznačajnijih metabolita prema modelu sa *mRMR* selektorom koji je pokazao

nešto bolje performanse (Tabela 3.2). Bilo bi korisno ispitati i uticaj preostalih metabolita, što prevazilazi okvire ovog istraživanja.

4.2 Potencijalni kandidati za biomarkere

Skup podataka izmerenih metabolita, pored hemijskog naziva sadrži i reference za baze podataka u kojima se može pronaći više informacija o ovim jedinjenjima. Za informacije o klasi jedinjenja kojoj pripadaju i procesima u kojima učestvuju metaboliti, korišćena je Baza podataka ljudskog metaboloma (eng. *Human Metabolome DataBase, HMDB*) [23]. Tabela 4.1 sadrži informacije dobijene iz pomenute baze podataka [23] za 10 najznačajnijih metabolita prema modelu sa *mRMR* selektorom, rangiranih prema značajnosti.

Grupisanjem podataka po trimestru, za svaku od grupa posebno (preeklampsija i kontrolna grupa), kreirani su grafici promene koncentracije metabolita kroz trimestre prikazani na slici 4.3. Na x osi su predstavljeni trimestri, a na y osi prosek za z-vrednosti koncentracija odgovarajućeg metabolita. Na ovim graficima se može uočiti razlika u koncentracijama metabolita između grupe sa preeklampsijom i kontrolne grupe. U nastavku će biti diskutovani prikazani rezultati, pri čemu će metaboliti biti predstavljeni svojim rangom iz tabele 4.1.

Kod pacijentkinja sa preeklampsijom, detektovana je povišena koncentracija pojedinih lipida (metaboliti 2 i 7) i jedinjenja koja učestvuju u njihovom metabolizmu (metaboliti 3 i 5) u odnosu na zdrave trudnice. Nedavna istraživanja otkrila su da disfunkcija metabolizma lipida može početi u ranoj fazi trudnoće koja će kasnije razviti preeklampsiju, što znači da se mogu koristiti za predviđanje nastanka preeklampsije [6, 13]. Detaljnije je objašnjeno da su masne kiseline neophodne za ispravan napredak trudnoće, te da je povećana koncentracija masti (metaboliti 2 i 7) i jedinjenja koja učestvuju u njihovom razlaganju, kao što su acilkarnitini (metabolit 3), normalna i kod zdrave trudnoće, ali da je dodatno povećanje prisutno kod trudnica koje imaju preeklampsiju [13], što je u skladu sa rezultatima dobijenim u ovom radu. Pored navedenih, potvrđena je i povišena koncentracija steroidnih hormona iz grupe progestogena (metabolit 8) [10]. Ova klinička istraživanja potvrđuju mogućnost korišćenja pomenutih grupa jedinjenja, kojima pripadaju i navedeni metaboliti, za ranu detekciju preeklampsije.

Metabolit označen kao najznačajniji pri klasifikaciji, sulfat kofeinske kiseline (metabolit 1), detektovan je u smanjenoj koncentraciji u poređenju sa zdravim trudni-

Rang	Naziv metabolita	Kratak opis
1	*caffeic acid sulfate*	Organsko jedninjenje, nema informacija u literaturi.
2	*oleoyl-oleoyl-glycerol*	Diglicerid, lipid, sadrži dva lanca oleinske kiseline (nezasićena masna kiselina).
3	*glutarylcarnitine*	Spada u grupu acilkarnitina, koji su zaduženi za transport organskih i masnih kiselina u procesu njihove razgradnje. Detektovan u tkivu placente.
4	*gamma-glutamylvaline*	Dipeptid, sastoji se od dve aminokiseline, glutamina i valina.
5	*linoleoyl ethanolamide*	Spada u grupu masnih amida (etanolamid masne kiseline), nema informacija u literaturi.
6	*4-allylphenol sulfate*	Organsko jedinjenje iz klase fenilsulfata, nema informacija u literaturi.
7	*1-palmitoleoylglycerol*	Monoglicerid, lipid, sadrži jedan lanac palmitinske kiseline (zasićena masna kiselina)
8	*5alpha-pregnan-3beta,20alpha-diol monosulfate*	Sterol iz grupe steroidnih hormona povezanih sa progesteronom (progestogeni).
9	*N-acetyltryptophan*	N-acilovana aminokiselina triptofan, u većoj koncentraciji predstavlja toksin koji može dovesti do oštećenja bubrega ili kardiovaskularnih bolesti.
10	*thyroxine*	Hormon tiroidne žlezde, učestvuje u različitim procesima, reguliše metabolizam proteina, masti i ugljenih hidrata.

Tabela 4.1: Kandidati za biomarkere rangirani prema značajnosti. Prikazane su osnovne informacije za 10 najznačajnijih metabolita prema modelu sa *mRMR* selektorom. Informacije preuzete iz *Human Metabolome DataBase, HMDB* [23].

cama. O njemu u literaturi nema mnogo informacija i nisu poznata istraživanja o metaboličkim procesima u kojima učestvuje, pa ni o njegovoj povezanosti sa preeklampsijom.

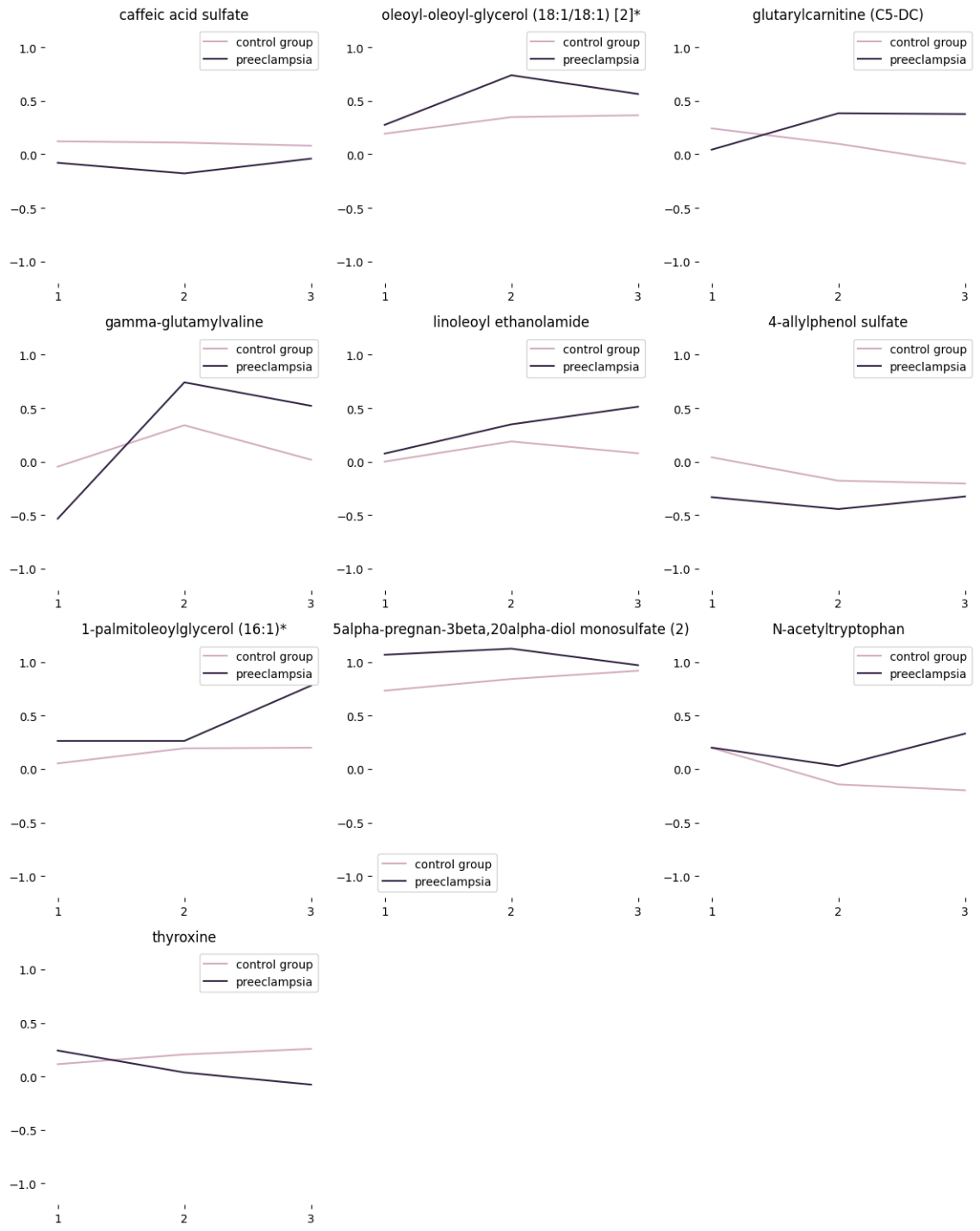
Smanjena koncentracija tiroksina (metabolit 10) detektovana je u drugom i trećem trimestru. Postoje istraživanja koja se bave vezom nivoa ovog hormona i preeklampsije, međutim rezultati nisu usaglašeni. Pojedine studije otkrile su značajno nižu koncentraciju tiroksina kod pacijentkinja sa preeklampsijom u odnosu na zdra-

ve trudnice, dok druga istraživanja tvrde da ta razlika nije statistički značajna [16]. Nije potvrđeno da se može koristiti pri ranom otkrivanju preeklampsije. Za preostale metabolite (4, 6 i 9) nisu poznata istraživanja njihove povezanosti sa preeklampsijom.

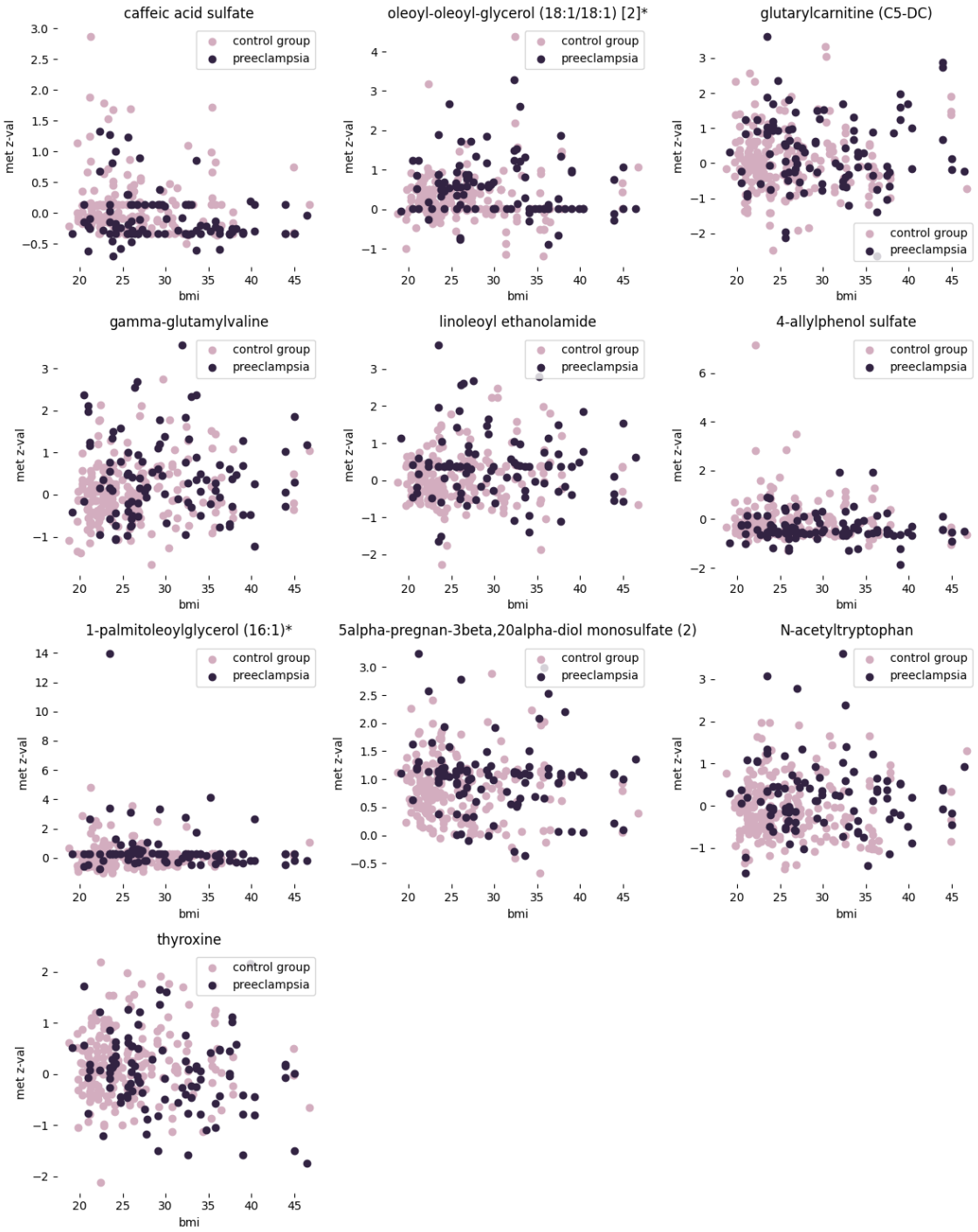
Dalje analize metabolita usmerene su na utvrđivanje povezanosti koncentracije detektovanih metabolita sa drugim faktorima kao što su indeks telesne mase (*BMI*) i godine starosti.

Izračunavanjem Pirsonovog koeficijenta korelacije između koncentracije metabolita i indeksa telesne mase dobijene su vrednosti između -0.2 i 0.2, iz čega se može zaključiti da ove vrednosti nisu korelisane. Slika 4.4 prikazuje zavisnost z-vrednosti koncentracije metabolita od vrednosti *BMI* za obe grupe, sa preeklampsijom i bez. Ni na jednom od prikazanih grafika nije prisutna pravilnost, ne može se zaključiti na primer da je koncentracija nekog od metabolita pretežno veća za veću vrednost *BMI* i slično. Dodatno, podelom podataka po trimestru u kom su metaboliti mereni, dolazi se do istog zaključka da nema korelacije između ove dve vrednosti. Ovakav rezultat nije u skladu sa nekim prethodnim istraživanjima [22].

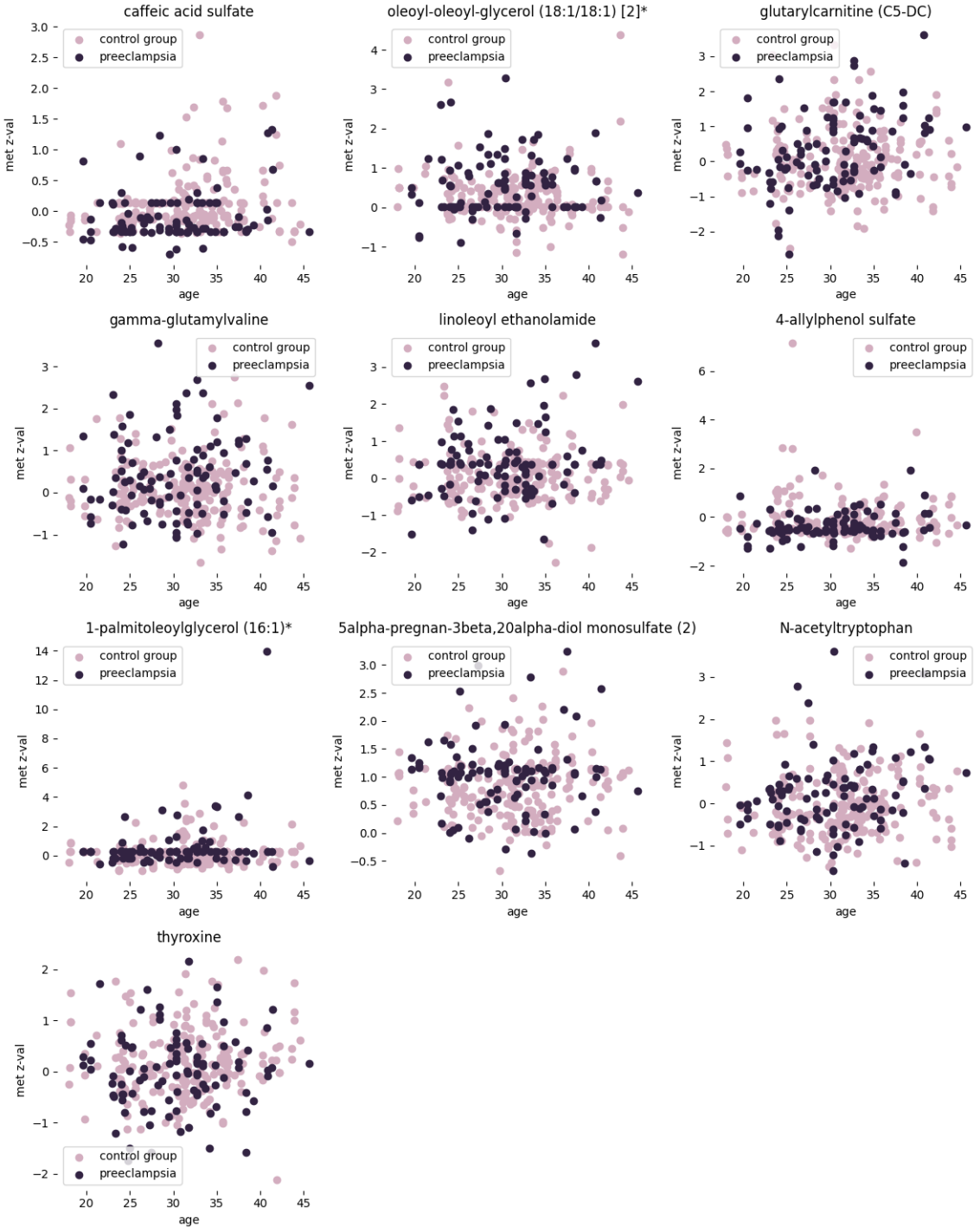
Na isti način ispitana je korelacija godina starosti sa koncentracijama pomenutih metabolita. Dolazi se do zaključka da nema korelacije među njima, zbog dobijenih vrednosti koje su ponovo bliske nuli. Na slici 4.5 prikazani su grafici zavisnosti koncentracije metabolita i godina starosti, na osnovu kojih se može izvesti sličan zaključak kao za vrednost *BMI*.



Slika 4.3: Promena koncentracije metabolita kroz trimestre. Na svakom dijagramu x osom su predstavljeni trimestri, a y osom prosek po grupama za z-vrednost koncentracije odgovarajućeg metabolita. Različitim bojama označene su preeklampsija i kontrolna grupa. Za svaki od prikazanih metabolita može se primetiti razlika u ovim vrednostima između te dve grupe.



Slika 4.4: Zavisnost koncentracije metabolita od indeksa telesne mase. Na x osi predstavljena je vrednost *BMI*, a na y osi z-vrednost koncentracije metabolita za svaku pacijentkinju. Različitim bojama označene su vrednosti iz kontrolne, odnosno grupe sa preeklampsijom. Ni na jednom od prikazanih grafika ne može se uočiti trend ili pravilnost. Nema korelacije između ovih vrednosti.



Slika 4.5: Zavisnost koncentracije metabolita od godina starosti. Na x osi predstavljene su godine starosti, a na y osi z-vrednost koncentracije metabolita za svaku pacijentkinju. Različitim bojama označene su vrednosti iz kontrolne, odnosno grupe sa preeklampsijom. Ni na jednom od prikazanih grafika ne može se uočiti trend ili pravilnost. Nema korelacije između ovih vrednosti.

Glava 5

Zaključak

Za preeklampsiju kao oboljenje koje može dovesti do ozbiljnih posledica od velike važnosti su istraživanja koja bi omogućila njenu ranu detekciju. Jedan od pravaca ovih istraživanja zasniva se na metabolomičkoj analizi sa ciljem da se pronađu kandidati za potencijalne biomarkere. U ovom radu prikazana je jedna takva analiza metabolita prikupljenih iz više kliničkih istraživanja preeklampsije.

Početni skup je sadržao preko 900 metabolita i njegova dimenzija je redukovana primenom dva različita algoritma za odabir atributa, *mRMR* i *SelectKBest*. Pomoću obe metode, kao optimalna dimenzija skupa, dobijena je vrednost 40. Na ovim podacima je uz svaki selektor atributa, obučen *XGBoost* klasifikator koji na osnovu koncentracija izmerenih metabolita daje odgovor na pitanje da li pacijentkinja ima preeklampsiju. Kao bolji se pokazao model sa *mRMR* pristupom sa tačnošću od 0.74 i AUC od 0.8.

Na osnovu važnosti atributa (eng. *feature importances*) *XGBoost* klasifikatora, odabrani metaboliti su rangirani da bi se odredili oni koji najviše doprinose. Među njima se nalazi nekoliko metabolita koji pripadaju grupi lipida ili učestvuju njihovim metaboličkim putevima. Za ove grupe jedinjenja je u literaturi potvrđena povezanost sa preeklampsijom i istraživanja potencijalnih biomarkera za preeklamspiju predlažu pojedine kandidate iz te grupe.

Za ostale kandidate nema mnogo istraživanja o njihovoj povezanosti sa preeklampsijom. Dodatna istraživanja potrebna su za utvrđivanje procesa u kojima učestvuju ovi metaboliti, koji faktori mogu dovesti do njihove promene u koncentraciji i na koji način to može biti povezano sa preeklampsijom ili njenim simptomima. Bilo bi značajno uporediti rezultate dobijene u ovom radu sa rezultatima odgovarajuće kliničke studije.

Prikupljanje novih podataka iz drugih studija i proširivanje postojećeg skupa podataka, omogućilo bi dodatne metaboličke analize preeklampsije. Sličan postupak navedenom u ovom radu mogao bi se primeniti za svaki trimestar posebno, za šta u ovom istraživanju nije bilo dovoljno podataka. Najveći značaj imala bi analiza metabolita u prvom trimestru, odnosno pre same pojave simptoma karakterističnih za ovu bolest. Još jedan pravac daljeg razvoja ovog istraživanja mogao bi biti usmeren ka analizi uticaja mogućih terapija na predložene kandidate za potencijalne biomarkere.

Bibliografija

[1] A. Behdenna, M. Colange, and J. Haziza. pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *BMC Bioinformatics*, 24:459, 2023.

[2] David R. Bentley. The human genome project—an overview. *Medicinal Research Reviews*, 20(3):189–196, 2000.

[3] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[4] Katja Dettmer and Bruce D Hammock. Metabolomics–a new exciting field within the „omics" sciences. *Environmental Health Perspectives*, 112(7):A396–A397, 2004.

[5] Warwick B. Dunn, David I. Broadhurst, Helen J. Atherton, Royston Goodacre, and Julian L. Griffin. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem. Soc. Rev.*, 40:387–426, 2011.

[6] B. Fatemeh and Nobakht M. Gh. Application of metabolomics to preeclampsia diagnosis. *Systems Biology in Reproductive Medicine*, 64(5):324–339, 2018. PMID: 29965778.

[7] Wilson Goh, Bin Wen, Wei Wang, and Limsoon Wong. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology*, 35:498–507, 2017.

[8] Jaime Gosálvez and José A. Horcajadas. Introduction: Human genome projects: The omics starting point. In José A. Horcajadas and Jaime Gosálvez, editors, *Reproductomics*, pages xvii–xxix. Academic Press, 2018.

[9] Bo Li and Benjamin Haibe-Kains. UMAP, Uniform Manifold Approximation and Projection for Dimension Reduction, 2018. on-line at: `https://pypi.org/project/pymrmre/`.

[10] X. Li, A. Milosavljevic, S. H. Elsea, C. C. Wang, F. Scaglia, A. Syngelaki, K. H. Nicolaides, and L. C. Poon. Effective Aspirin Treatment of Women at Risk for Preeclampsia Delays the Metabolic Clock of Gestation. *Hypertension (Dallas, Tex. : 1979)*, page 1398–1410, 2021.

[11] Ulf W. Liebal, An N. T. Phan, Malvika Sudhakar, Karthik Raman, and Lars M. Blank. Machine learning applications for mass spectrometry-based metabolomics. *Metabolites*, 10(6), 2020.

[12] Leland McInnes. UMAP, Uniform Manifold Approximation and Projection for Dimension Reduction, 2018. on-line at: `https://umap-learn.readthedocs.io/en/latest/`.

[13] Yao Mengxin, Xiao Yue, Yang Zhuoqiao, Ge Wenxin, Liang Fei, Teng Haoyue, Gu Yingjie, and Yin Jieyun. Identification of biomarkers for preeclampsia based on metabolomics. *Clinical Epidemiology*, 14:337–360, 2022.

[14] Ben W J Mol, Claire T Roberts, Shakila Thangaratinam, Laura A Magee, Christianne J M de Groot, and G Justus Hofmeyr. Pre-eclampsia. *The Lancet*, 387(10022):999–1011, 2016.

[15] David W. Mount. *Bioinformatics : sequence and genome analysis*. Cold Spring Harbor Laboratory Press, 2004.

[16] Nineetha Muraleedharan and Jessy Sumangala1 Janardhanan. Thyroid hormone status in preeclampsia patients: A case–control study. *Muller Journal of Medical Sciences and Research*, 8(2):68–73, 2017.

[17] M. Radovic, M. Ghalwash, and N. Filipovic. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, 18(9), 2017.

[18] Camilo Ruiz, Hongyu Ren, Kexin Huang, and Jure Leskovec. High dimensional, tabular deep learning with an auxiliary knowledge graph. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 26348–26371. Curran Associates, Inc., 2023.

[19] scikit learn. SelectKBest, Feature selection. on-line at: `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html`.

[20] Laurentin Táriba and Hernán Eduardo. *Central Dogma of Molecular Biology*, pages 19–36. Springer Nature Switzerland, Cham, 2023.

[21] M.A. Telang, S.P. Bhutkar, and R.R. Hirwani. Analysis of patents on preeclampsia detection and diagnosis: A perspective. *Placenta*, 34(1):2–8, 2013.

[22] Scott W. Walsh. Obesity: a risk factor for preeclampsia. *Trends in Endocrinology & Metabolism*, 18:365–370, 2007.

[23] D. S. Wishart, A. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B. L. Lee, M. Berjanskii, R. Mah, M. Yamamoto, J. Jovel, C. Torres-Calzada, M. Hiebert-Giesbrecht, V. W. Lui, D. Varshavi, D. Varshavi, D. Allen, and V. ... Gautam. The Humane Metabolome Database. on-line at: `https://hmdb.ca/`.

[24] Aihua Zhang, Hui Sun, Ping Wang, Ying Han, and Xijun Wang. Modern analytical techniques in metabolomics analysis. *Analyst*, 137:293–300, 2012.

Biografija autora

Lucija Miličić (*Beograd, 11. avgust 1999.*) upisala je osnovne studije na Matematičkom fakultetu, modul Informatika, 2018. godine i završila ih 2022. godine sa prosekom 8.74. Iste godine upisuje master studije, takođe na Matematičkom fakultetu, gde od 2023. godine i radi kao saradnik u nastavi.