Primjena strojnog učenja u metagenomici
Projektna dokumentacija

Verzija 2.0

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

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

Sadržaj

Prijedlog i plan projekta

Na koji način koristiti predložak?

Tekst pisan u italic formi opisuje koje informacije je potrebno uključiti u pojedino poglavlje Prijedloga.
Za upis vlastitog teksta, potrebno je pritisnuti <ENTER> nakon italic teksta.
Tekst upisan u <trokutastim zagradama> treba zamijeniti s onim što se navodi.

Svi članovi tima trebaju pročitati plan i suglasiti se s njime, a to potvrđuju svojim potpisom na kraju dokumenta.

1. Puni naziv projekta

[Navesti puni naziv projekta iz kojeg će biti vidljiva problematika/tematika kojom se projekt bavi.]

Primjena strojnog učenja u metagenomici

2. Opis problema/teme projekta

[Objasniti problem, odnosno temu projekta, objasniti ukratko tip, uvjete i kontekst projekta.]

Projekt istražuje inovativne pristupe primjene strojnog učenja za analizu metagenomskih podataka s glavnim ciljem razumijevanja složenosti mikrobnih ekosustava u ljudskom organizmu. Naglasak je stavljen na otkrivanje povezanosti između različitih bakterijskih vrsta i različitih zdravstvenih stanja. S obzirom na svoj eksplorativni karakter, ovaj projekt ima potencijal transformirati način na koji sagledavamo utjecaj mikroorganizama na ljudsko zdravlje.

U okviru istraživanja, primijenjene su sofisticirane metode klasifikacije strojnog učenja za analizu metagenomskih podataka, s ciljem izgradnje modela kojim se može predvidjeti zdravstveno stanje. Osim analize podataka o prisutnim bakterijskim vrstama u mikrobiomu, pažnja je usmjerena i prema demografskim čimbenicima poput dobi, indeksa tjelesne mase i države stanovanja, čime se obogaćuje analiza. Odabir ključnih značajki predstavlja bitan segment ovog pristupa, s ciljem poboljšanja preciznosti modela te otkrivanja potencijalnih biomarkera povezanih s različitim zdravstvenim uvjetima.

U kontekstu projekta, posebno je važno razumijevanje tehnika strojnog učenja i analize podataka. Implementacija projekta zahtijeva upotrebu sofisticiranih algoritama i alata prilagođenih specifičnostima metagenomskih podataka. S obzirom na multidisciplinarni pristup, uključujući biološke, računalne i medicinske aspekte, projekt se postavlja kao važan korak prema sveobuhvatnom razumijevanju mikrobioma ljudskog organizma.

Rezultati dobiveni u okviru istraživačkog projekta uspoređeni su s rezultatima postojećih radova, dodatno potvrđujući relevantnost dobivenih spoznaja. Očekuje se da će ova istraživanja pridonijeti ne samo znanstvenom razumijevanju mikrobioma, već i razvoju novih strategija u dijagnostici i individualiziranim terapeutskim pristupima.

3. Cilj projekta

[Navesti predviđeni cilj ili ciljeve projekta. Definiranje ciljeva omogućuje određivanje pravca u kojem će se kretati izvođenje projekta. Navesti predviđeno trajanje projekta.]

Predviđeno trajanje projekta je 12 tjedana.

Predviđeni ciljevi projekta su:

- Istraživanje raznolikih pristupa strojnog učenja za analizu metagenomskih podataka
- Dobiti uvid u različite metode strojnog učenja koje su primjenjive na analizu metagenomskih podataka, istražujući širok spektar pristupa i njihovu primjenjivost
- izgradnja modela korištenjem različitih metoda klasifikacije za predviđanje zdravstvenog stanja
- Implementirati modele temeljene na različitim metodama klasifikacije kako bi se anticipiralo zdravstveno stanje organizma
- Usporedba točnosti modela za različite bolesti

- Provesti usporedbe točnosti izgrađenih modela za različite bolesti, s ciljem prepoznavanja najučinkovitijih pristupa
- Identifikacija ključnih bakterijskih vrsta
- Identificirati bakterijske vrste čija prisutnost u mikrobiomu pokazuje korelaciju s prisutnošću određenih bolesti, naglašavajući važnost tih vrsta kao potencijalnih pokazatelja zdravstvenog stanja
- Istraživanje konteksta bolesti i njihova povezanost s bakterijskim vrstama
- Provesti istraživanje o kontekstu bolesti, analizirajući njihovu kompleksnu povezanost s bakterijskim vrstama
- Istražiti razlike utjecaja konzumiranja probiotičkih dodataka i probiotičkih namirnica

4. Voditelj studentskog tima

<Ime i prezime studenta>
Lucija Topolko

5. Rezultat(i)

[Navesti što će se isporučiti na kraju projekta, voditi računa da osim rezultata u vidu nekog proizvoda ovdje treba navesti i svu dokumentaciju.]

Fokus našeg istraživanja bio je ustanoviti vezu između ljudskog mikrobioma i oboljenja od određenih bolesti (specifično ciroze jetre, pretilosti, upalnih bolesti crijeva i raka debelog crijeva), koristeći skupove podataka iz referentnog rada *„Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights“* i implementirajući algoritme strojnog učenja u programskim alatima *Jupyter Notebook* i *Weka.* Promatrane značajke za bolesti bile su prvenstveno brojnost bakterijskih vrsta u čovjekovom mikrobiomu, a dodatno su razmatrane i demografske značajke.

Za analizu podataka korišteno je 6 algoritama strojnog učenja: algoritam slučajne šume, algoritam logističke regresije, *BayesNet* algoritam, J48 algoritam, SVM algoritam te naknadno *XGBoost* algoritam. Kao mjere uspješnosti modela primarno su korištene točnost pri klasifikaciji i matrica zabune, a dodatno i ROC krivulja.

Proveden je proces odabira značajki nad algoritmima slučajne šume, J48 i *XGBoost,* koji su se pokazali najuspješnijima u klasificiranju bolesti. Pri odabiru značajki korištene su metoda *InfoGain* za algoritam slučajne šume i *XGBoost.* U ovoj fazi u obzir su uzete samo značajke koje se odnose na brojnost bakterijskih vrsta. Zatim su provedeni dodatni postupci u pokušaju poboljšanja modela: metoda umjetnog povećavanja nad podatcima u kategoriji raka debelog crijeva u svrhu uravnoteženja skupa podataka korištena je kako bi pospješila klasifikaciju algoritma slučajne šume i *XGBoost*, te uključivanje određenih demografskih značajki i metoda penaliziranja lažno negativnih rezultata za podatke u kategoriji raka debelog crijeva za algoritam J48. Točnost navedenih algoritama u spomenutim faza prikazana je u sljedećoj tablici:

	Slučajna šuma + *InfoGain*	***XGBoost* + *InfoGain***	**J48**	**Slučajna šuma + uravnoteženi podaci**	***XGBoost* + uravnoteženi podaci**	**J48 + demografski podaci + penalizacija**
Ciroza jetre	88.89%	88.89%	96.96%	83.33%	96.15%	95.81%
Pretilost	75.00%	81.58%	90.17%	78.79%	85.71%	96.68%
Upalne bolesti crijeva	84.00%	85.71%	95.17%	80.00%	72.13%	97.19%
Rak debelog crijeva	0.00%	27.27%	10.42%	11.11%	93.55%	99.43%

Daljnjim proučavanjem značajki bakterijskih vrsta koje su se pokazale najutjecajnijima na ishod klasifikacije, ustanovljeno je da postoji stanovita razina međusobnog preklapanja kako između rezultata različitih primijenjenih algoritama, tako i s drugim znanstvenim istraživanjima koja povezuju stanje ljudskog mikrobioma s pojavama navedenih bolesti. Dodatna potvrda točnosti naših rezultata bila je i znatno preklapanje između zaključaka naše analize i analize provedene u referentnom radu *„Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights“* čije smo skupove podataka koristili.

Sam kraj projekta bio je posvećen istraživanju vanjskih utjecaja na mikrobiom (a posljedično i eventualnu pojavu bolesti), s fokusom na utjecaj različitih režima prehrane. Proučavanjem ove literature stečen je dublji uvid u medicinski aspekt problematike projekta te su pronađeni odgovori na neka pitanja koja su se nametnula tijekom analize skupova podataka. Ova faza ujedno je bila i zaključna faza našeg projekta.

6. Slični projekti

[Navesti projekte koji su povezani s dotičnim projektom.]

1. Tušek, R. Primjena strojnog učenja u metagenomici, završni rad, Fakultet elektrotehnike i računarstva, 2018
2. Pasolli, E., Truong, D. T., Malik, F., Waldron, L., Segata, N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput Biol. 2016 Jul 11;12(7):e1004977. doi: 10.1371/journal.pcbi.1004977

7. Resursi

[Navesti ljudske i ostale resurse potrebne za uspješno dovršenje projekta. Popuniti tablicu raspoloživih članova tima s podacima važnim za projekt. Mogu se navesti znanja i vještine člana koje mogu biti od koristi za projekt, na primjer znanja Java-e, XML-a, iskustvo u radu s MS Projectom, sudjelovanje u sličnim projektima ili bilo kakva korisna informacija. Ako projekt koristi i druge resurse napraviti posebnu tablicu za njih. U kolonu Napomene treba upisati sve termine kad dotični član tima neće biti raspoloživ za rad na projektu (putovanja, odmori, odsustva).]

Tablica ljudskih resursa

Ime i prezime	E-mail adresa
Luka Bulić	luka.bulic@fer.hr
Lucia Crvelin	lucia.crvelin@fer.hr
Niko Kaštelan	niko.kastelan2@fer.hr
Mirta Krajinović	mirta.krajinovic@fer.hr
Lucija Topolko	lucija.topolko@fer.hr
Marko Žagar	marko.zagar@fer.hr

8. Glavni rizici

[Navesti glavne zapreke za ostvarenje uspjeha projekta, te posljedice ukoliko projekt ne uspije.]

Budući da je jedan od glavnih ciljeva projekta izrada što točnijeg modela na temelju podataka koje smo odabrali, jedan od najvećih rizika je reprezentativnost skupa podataka. Kako bismo dobili što opširniji skup podataka, nastojali smo uzeti podatke iz više istraživanja koja proučavaju različite bolesti, no usprkos tome, i dalje su postojali problemi poput prenaučenosti (npr. logično visoka korelacija između indeksa tjelesne mase i pretilosti) te nekonzistentnih ili nejasnih atributa. Najviše takvih problema su stvarali demografski atributi. Naime, ako se u podatkovnom skupu nalaze podatci o nekoj bolesti iz samo jedne države, onda atribut države postaje beskoristan te čak može i smanjiti kvalitetu modela zbog pojave prenaučenosti. Model će u takvom slučaju pronaći potpunu korelaciju između države i bolesti te stoga postaje neupotrebljiv ako poželimo dodati nove podatke povezane s tom državom ili bolesti.

Također, skupovi za određene bolesti su puno veći od nekih drugih bolesti što predstavlja mogući problem kod izrade modela koji radi na principu vjerojatnosti i točnosti u postotcima. Ako se ne poduzmu mjere kako bi se to ispravilo, model će konstantno zanemarivati manje skupove i raspoređivat će ih u one veće kako bi povećao točnost, dok će točnost predviđanja manjih skupova biti jako blizu nule.

9. Smanjivanje rizika

[Navesti korake koji će se poduzeti kako bi se što je moguće više umanjio svaki od prethodno navedenih rizika.]

Većina navedenih rizika koji se odnose na skup podataka može se smanjiti pretprocesiranjem i filtriranjem podataka prije kreiranja modela, a to postaje moguće nakon analize i upoznavanja s podatcima. Odabir značajki je metoda s kojom smo uklanjali nepotrebne značajke kao i značajke koje dovode do prenaučenosti. To su ponajprije bile demografske značajke, no po potrebi smo uklonili i značajke poput indeksa tjelesne mase ili mjesta uzimanja uzorka. Tako smo primjerice uklonili indeks tjelesne mase pri promatranju pretilosti, jer za to postoji očita korelacija, no pri promatranju drugih bolesti kod kojih ne postoji očita korelacija, kao što je dijabetes, ostavili smo kod odabira značajki i indeks tjelesne mase za kreiranje modela.

Problem različitih veličina skupova podataka čest je problem u radu s podatcima te alati s kojima smo radili projekt nude jednostavne načine kako bi se taj problem što bolje riješio. Pritom smo, kao dvije glavne metode, koristili uvećavanje slabo zastupljenog skupa i smanjivanje bolje zastupljenog skupa čime smo dobili puno preciznije rezultate za manje skupove koji su sada imali sličnu točnost kao i veći skupovi.

10. Glavne faze projekta

[Navesti glavne faze projekta, te ukratko objašnjenje po kojem načelu je projekt podijeljen na te faze- vremenska organizacija, smanjenje rizika, raspoloživost resursa i/ili nešto drugo.]

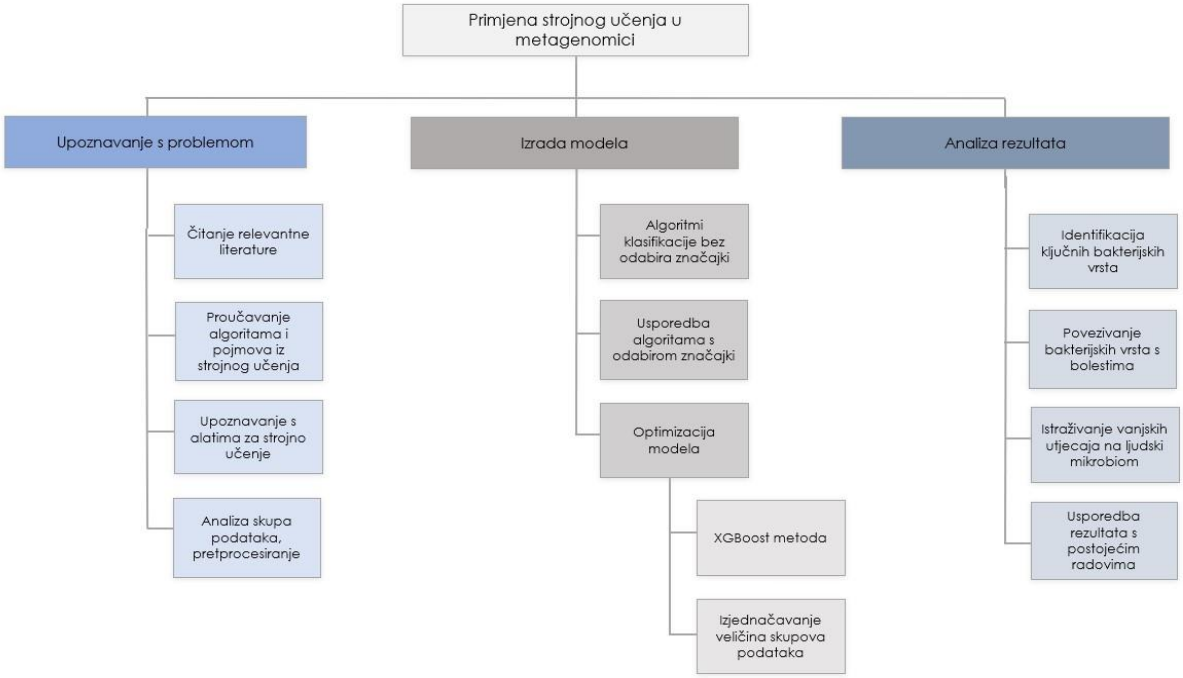
1. Faza: proučavanje literature, upoznavanje s algoritmima strojnog učenja, pretprocesuiranje podataka
2. Faza: usporedba algoritama klasifikacije bez odabira značajki
3. Faza: usporedba značajki (bakterijske vrste, demografski podaci) dobivenih za različite bolesti
4. Faza: uravnoteživanje skupova podataka, usporedba različitih metoda za odabir značajki
5. Faza: usporedba konačnih rezultata s rezultatima postojećih medicinskih istraživanja, istraživanje vanjskih utjecaja na ljudski mikrobiom

Navedene faze projekta proizlaze jedna iz druge, logičkim slijedom od upoznavanja s problematikom projekta, ovladavanja potrebnom tehnologijom, primjene stečenog znanja na produciranje konkretnih rezultata, do pronalaženja korelacije dobivenih rezultata s rezultatima drugih znanstvenih istraživanja.

Faze projekta nadograđuju se jedna na drugu, na način da se u svakoj fazi stremilo ka korištenju rezultata i stečenih znanja proizašlih iz prethodne faze kako bi se došlo do što relevantnijih zaključaka i eventualno proširio opseg istraživanja.

11. Struktura raspodijeljenog posla (engl. *Work Breakdown Structure* - WBS)

[Nacrtati WBS s navedenim aktivnostima projekta.]



12. Kontrolne točke projekta (engl. *milestones*)

[Općenito, kontrolna točka projekta je događaj ili rezultat neke aktivnosti koji ukazuje na to je li projekt u skladu sa zadanim rokovima ili kasni. Ta informacija se upisuje u kolonu o statusu projekta. Ako projekt kasni moraju se poduzeti akcije da se rokovi dostignu. Za svaku kontrolnu točku treba odrediti točan datum. Po potrebi se mogu dodavati ili oduzimati redovi tablice.]

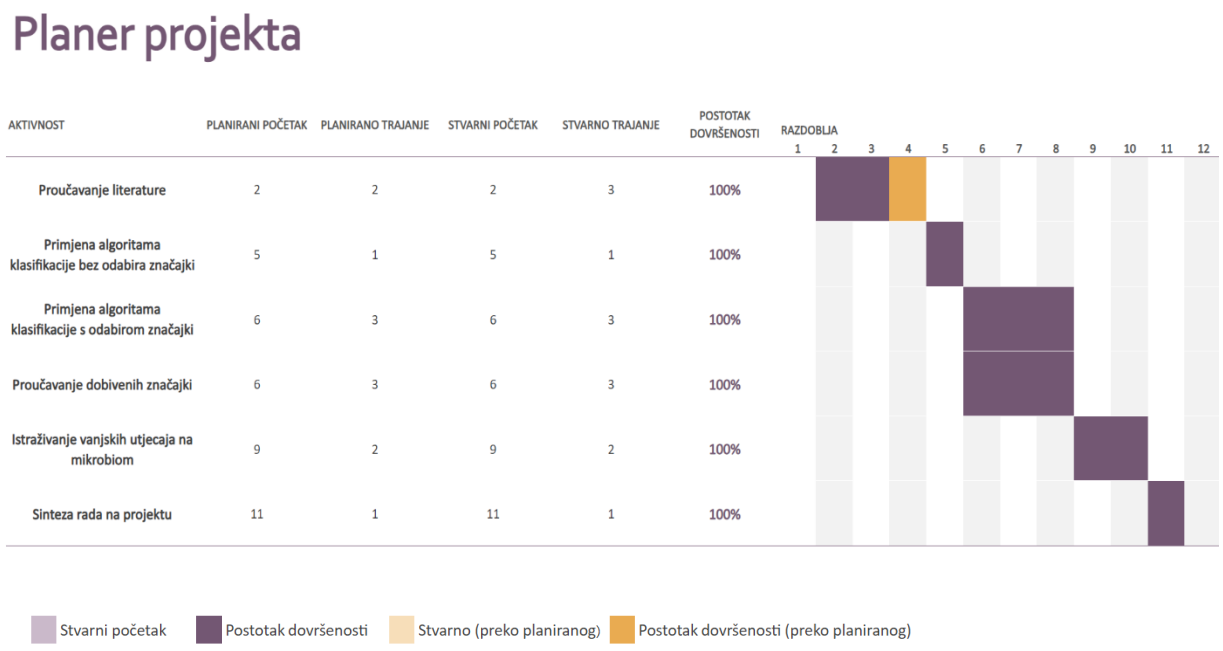
Tablica kontrolnih točki projekta

Kontrolne točke	Planirani datum	Realizirani datum	Status projekta
Analiza nepoznatih pojmova i algoritama iz referentnog rada	18.10.2023.	18.10.2023.	Projekt u skladu sa zadanim rokovima
Proučavanje programskog paketa Weka i Python knjižnica za strojno učenje	25.10.2023.	25.10.2023.	Projekt u skladu sa zadanim rokovima
Primjena algoritama klasifikacije u strojnom učenju bez odabira značajki	8.11.1023	8.11.2023.	Projekt u skladu sa zadanim rokovima
Primjena algoritma slučajna šuma s odabirom značajki, kontekst dobivenih rezultata	15.11.2023.	15.11.2023.	Projekt u skladu sa zadanim rokovima
Primjena algoritma XGBoost s odabirom značajki	6.12.2023.	6.12.2023.	Projekt u skladu sa zadanim rokovima

Istraživanje vanjskih utjecaja na mikrobiom	13.12.2023.	13.12.2023.	Projekt u skladu sa zadanim rokovima
Završna prezentacija	17.1.2023.	17.1.2023.	Projekt završen u skladu sa zadanim rokovima

13. Gantogram

[Izraditi Gantogram pomoću programa MS Project, Open Workbench, Microsoft Excel - pri, i sl. Pohraniti prikaz Gantograma (screenshot) i postaviti ga unutar ovog poglavlja kao ubačenu sliku.]



14. Zapisnici sastanaka

[Ovdje za svaki održani sastanak navesti: datum, vrijeme i mjesto održavanja sastanaka, popis nazočnih, glavne zaključke sastanka.]

Datum	Vrijeme	Mjesto	Prisutni	Zaključak
10.10.2023.	9:30	D-259, FER	izv. prof. dr. sc. Mirjana Domazet-Lošo, Luka Bulić, Lucia Crvelin, Niko Kaštelan, Mirta Krajinović, Lucija Topolko, Marko Žagar	Upoznavanje članova i odabir teme projekta. Proučiti referentni rad i nepoznate pojmove.

18.10.2023.	13:00	D-259, FER	izv. prof. dr. sc. Mirjana Domazet-Lošo, Luka Bulić, Lucia Crvelin, Niko Kaštelan, Mirta Krajinović, Lucija Topolko, Marko Žagar	Analiza nepoznatih pojmova i algoritama iz referentnog rada. Informirati se o načinu implementacije algoritama klasifikacije.3
25.10.2023.	13:00	D-262*, FER	izv. prof. dr. sc. Mirjana Domazet-Lošo, Luka Bulić, Lucia Crvelin, Niko Kaštelan, Mirta Krajinović, Lucija Topolko, Marko Žagar	Prezentiranje dobivenih rezultata. Usporediti rezultate dobivene različitim algoritmima klasifikacije bez odabira značajki.
8.11.1023	13:00	D-259, FER	izv. prof. dr. sc. Mirjana Domazet-Lošo, Lucia Crvelin, Niko Kaštelan, Mirta Krajinović, Lucija Topolko, Marko Žagar	Prezentiranje dobivenih rezultata. Pronaći najznačajnije karakteristike podataka pri primjeni algoritma slučajna šuma i proučiti kontekst.
15.11.2023.	13:15	D-259, FER	izv. prof. dr. sc. Mirjana Domazet-Lošo, Luka Bulić, Niko Kaštelan, Mirta Krajinović, Lucija Topolko, Marko Žagar	Prezentiranje dobivenih rezultata. Usporediti rezultate dobivene algoritmom slučajna šuma s rezultatima algoritma XGBoost.

6.12.2023.	13:00	D-259, FER	izv. prof. dr. sc. Mirjana Domazet-Lošo, Lucia Crvelin, Niko Kaštelan, Mirta Krajinović, Lucija Topolko, Marko Žagar	Prezentiranje konačnih rezultata. Istražiti utjecaj vanjskih utjecaja na ljudski mikrobiom.
13.12.2023.	13:30	D-259, FER	izv. prof. dr. sc. Mirjana Domazet-Lošo, Luka Bulić, Lucia Crvelin, Niko Kaštelan, Mirta Krajinović, Lucija Topolko, Marko Žagar	Prezentiranje istraženih vanjskih utjecaja na mikrobiom (unos probiotika, post). Napraviti završnu prezentaciju.
17.1.2023.	13:15	D-259, FER	izv. prof. dr. sc. Mirjana Domazet-Lošo, Luka Bulić, Lucia Crvelin, Niko Kaštelan, Mirta Krajinović, Lucija Topolko, Marko Žagar	Završna prezentacija

Suglasan s dokumentom (potpisuju članovi tima):

Luka Bulić Datum: _____ Potpis: _____

Lucia Crvelin Datum: _____ Potpis: _____

Niko Kaštelan Datum: _____ Potpis: _____

Mirta Krajinović Datum: _____ Potpis: _____

Lucija Topolko Datum: _____ Potpis: _____

Marko Žagar Datum: _____ Potpis: _____

Odobrio(potpisuje nastavnik):

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

Datum: _____ Potpis: _____