

Psihološki efekti kovida

Projekat u okviru kursa Istraživanje podataka 1

autor: Jelisaveta Gavrilović 188/2020
profesor: Nenad Mitić
asistent: Marija Erić

SADRŽAJ

SADRŽAJ	2
UVOD	3
ANALIZA PODATAKA	4
VIZUELIZACIJA PODATAKA	5
MATRICA KORELACIJE	8
PREPROCESIRANJE PODATAKA	10
UKLANJANJE GREŠAKA	10
BRISANJE NEPOTREBNIH KOLONA	10
NEDOSTAJUĆE VREDNOSTI	10
KODIRANJE KATEGORIČKIH ATRIBUTA	11
DODATNE PRIPREME	12
KLASIFIKACIJA	14
STABLA ODLUČIVANJA	14
K-NAJBLIŽIH SUSEDА	20
POREĐENJE MODELA KLASIFIKACIJE	23
KLASTEROVANJE	26
K-SREDINA	27
HIJERARHIJSKO KLASTEROVANJE	31
POREĐENJE MODELA KLASTEROVANJA	37
PRAVILA PRIDRUŽIVANJA	38
ZAKLJUČAK	39
LITERATURA	40

UVOD

Koronavirusna bolest 2019 (engl. Coronavirus disease 2019), prepoznatljiva pod skraćenicom kovid 19 ili COVID-19, zarazna je bolest uzrokovana teškim akutnim respiratornim sindromom virus korona 2 (SARS-CoV-2). Bolest se od 2019. proširila na ceo svet što je dovelo do pandemije virusa korona 2019/20. Pandemija je ostavila snažan pečat na svakodnevni život ljudi već više od dve godine. Među mnogim posledicama pandemije nalaze se i psihološke, koje su doprinele masovnim pojavama anksioznosti i depresije.

Skup podataka je preuzet sa: <https://www.kaggle.com/datasets/hemanthhari/psychological-effects-of-covid>. Podaci su prikupljeni 2020. godine u cilju razumevanja mišljenja ljudi o "zaključavanju" i koliko je to uticalo na njihovu promenu načina života.

Cilj ovog projekta je kako su različiti aspekti života, kao što su radne navike i novo radno okruženje, produktivnost, društveni i porodični odnosi, opuštenost,... se promenili i u kojoj meri su se pojedinci prilagodili novim okolnostima, a najviše kako je to uticalo na emocionalnom nivou.

ANALIZA PODATAKA

Pre početka izgradnje modela i daljih istraživanja potrebno je upoznati se sa skupom podataka i razumeti podatke sa kojima radimo, odnosno potrebno je izvršiti detaljnu analizu podataka.

Kompletna lista atributa:

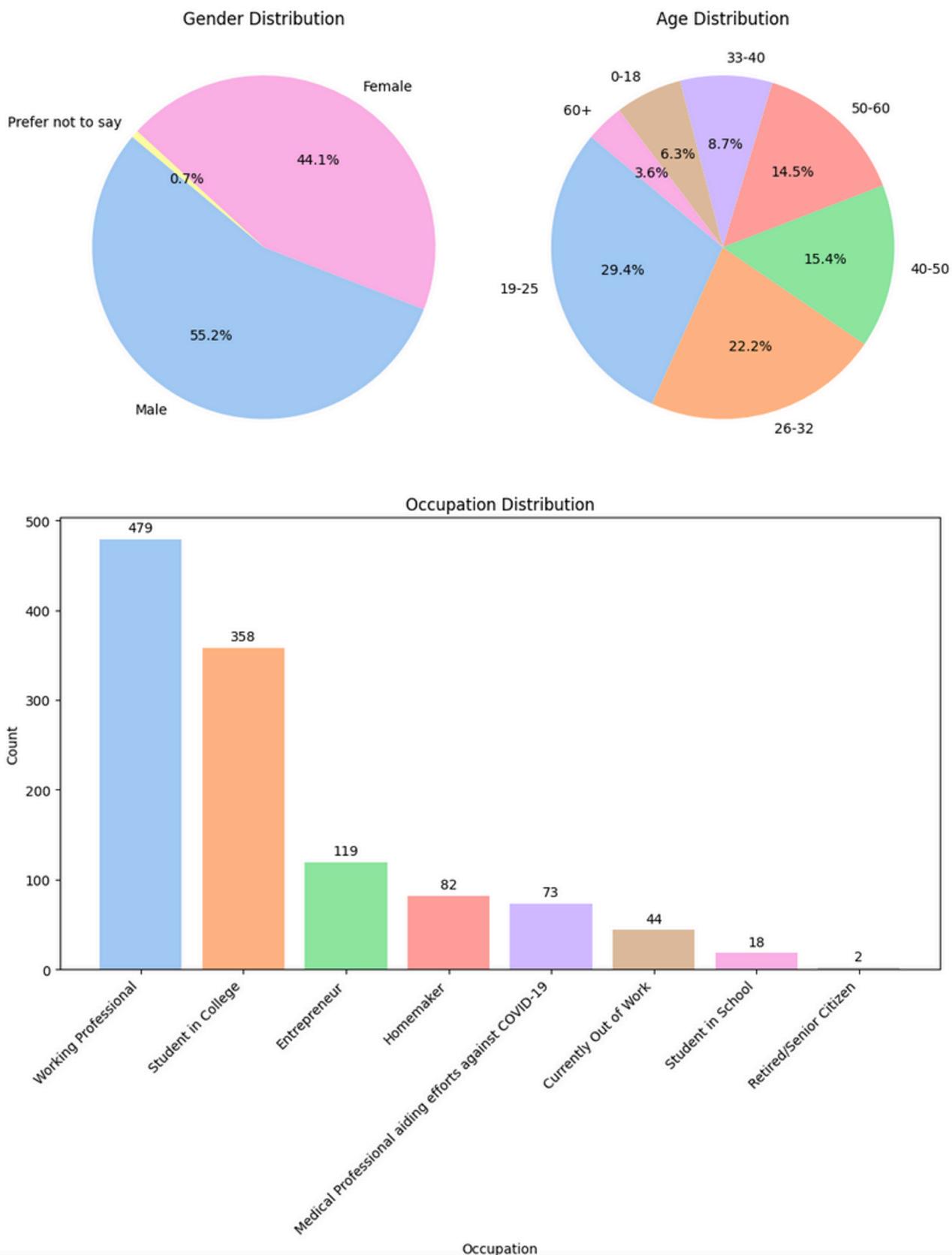
- **age** starosna grupa osobe
- **gender** pol
- **occupation** zanimanje / sektor u kojem osoba radi
- **line_of_work** vrsta posla kojim se bavi
- **time_bp** vreme provedeno na poslu pre pandemije
- **time_dp** vreme provedeno na poslu tokom pandemije
- **travel_time** vreme provedeno u putu do posla
- **easeof_online** ocena prelaska na online rad
- **home_env** ocena udobnosti kućnog okruženja
- **prod_inc** ocena povećanja produktivnosti
- **sleep_bal** ocena kvaliteta sna
- **new_skill** da li je osoba stekla nove veštine
- **fam_connect** ocena povezivanja sa porodicom
- **relaxed** ocena nivoa opuštenosti
- **self_time** ocena koliko osoba ima slobodnog vremena
- **like_hw** da li osoba voli raditi od kuće
- **dislike_hw** da li osoba ne voli raditi od kuće
- **prefer** da li osoba preferira rad od kuće ili u kancelariji
- **certaindays_hw** da li osoba preferira određene dane za rad od kuće
- **X, time_bp, travel_new, net_diff** - prilagođene kolone

	age	gender	occupation	line_of_work	time_bp	time_dp	travel_time	easeof_online	home_env	prod_inc	...	fam_connect	relaxed	self_time	like_hw	...
0	19-25	Male	Student in College	NaN	7	5	0.5	3	3	0.0	...	1.0	-0.5	-0.5	100	...
1	Dec-18	Male	Student in School	NaN	7	11	0.5	4	2	-0.5	...	1.0	1.0	1.0	1111	...
2	19-25	Male	Student in College	NaN	7	7	1.5	2	2	1.0	...	0.5	0.5	0.5	1100	...
3	19-25	Male	Student in College	NaN	7	7	1.5	3	1	0.0	...	0.0	-1.0	-0.5	100	...
4	19-25	Female	Student in College	NaN	7	7	1.5	2	2	0.0	...	0.0	0.5	0.0	1010	...

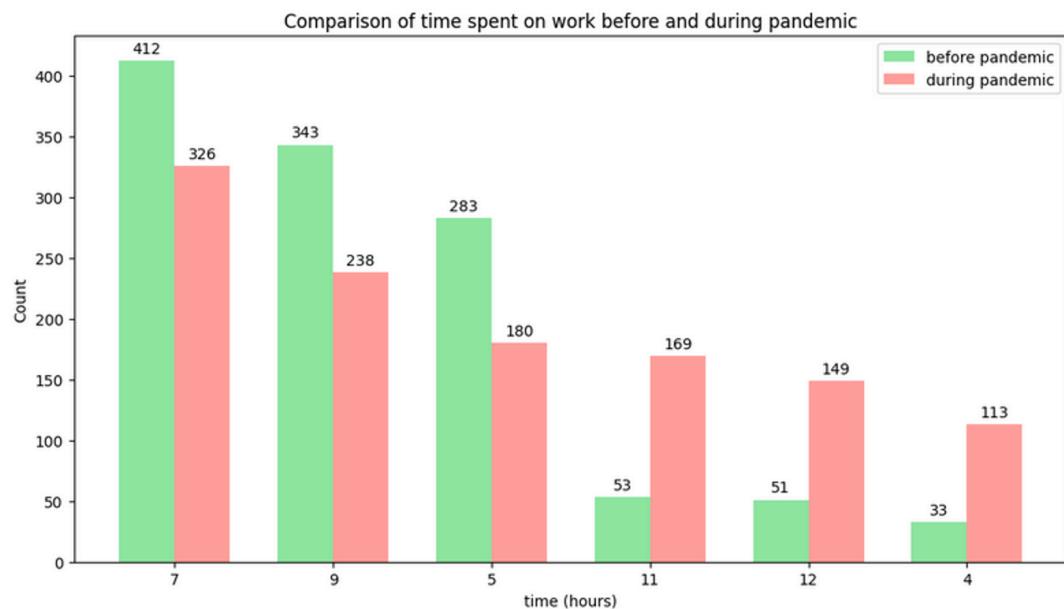
5 rows x 22 columns

VIZUELIZACIJA PODATAKA

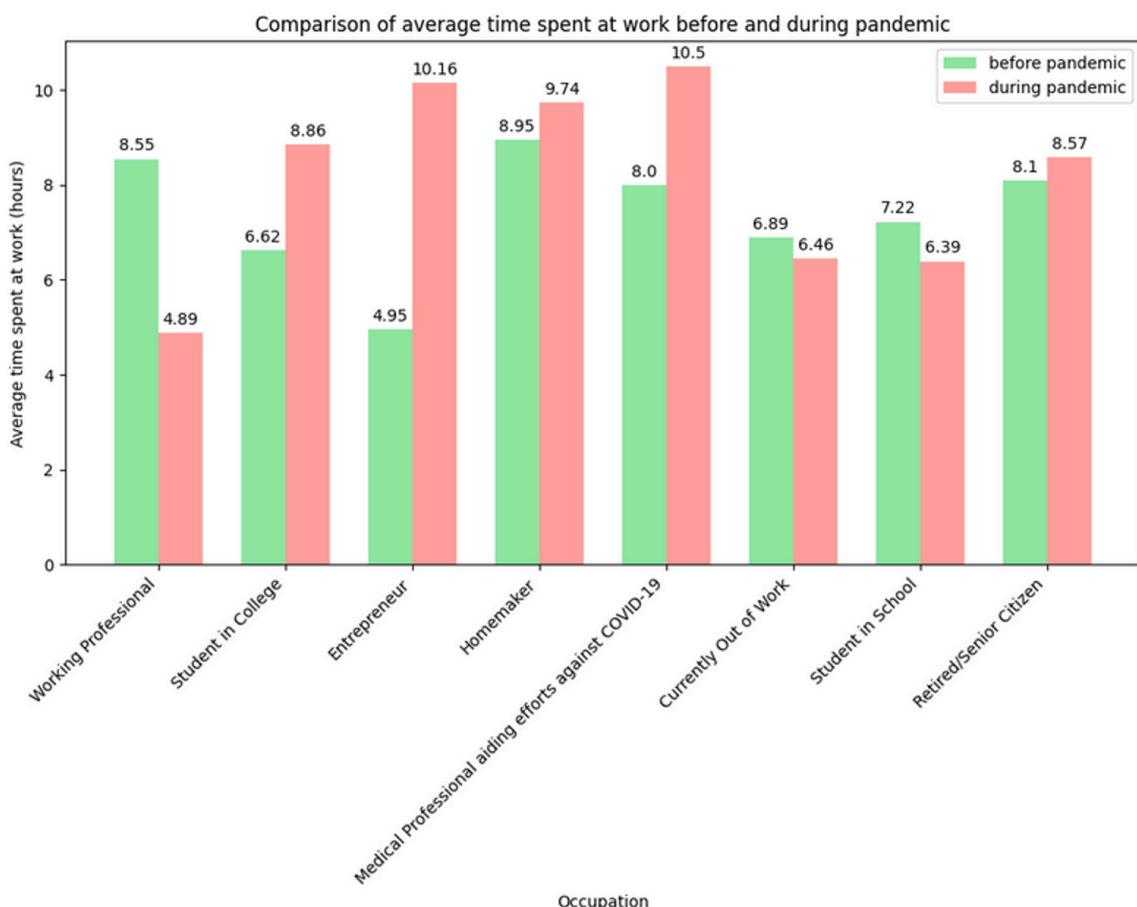
Vizualno predstaviti podatke nam omogućava lakše upoznavanje i razumevanje podataka, njihovih raspodela, lakše identifikovanje obrazaca, korelacija i eventualnih anomalija.



Promena radnog vremena

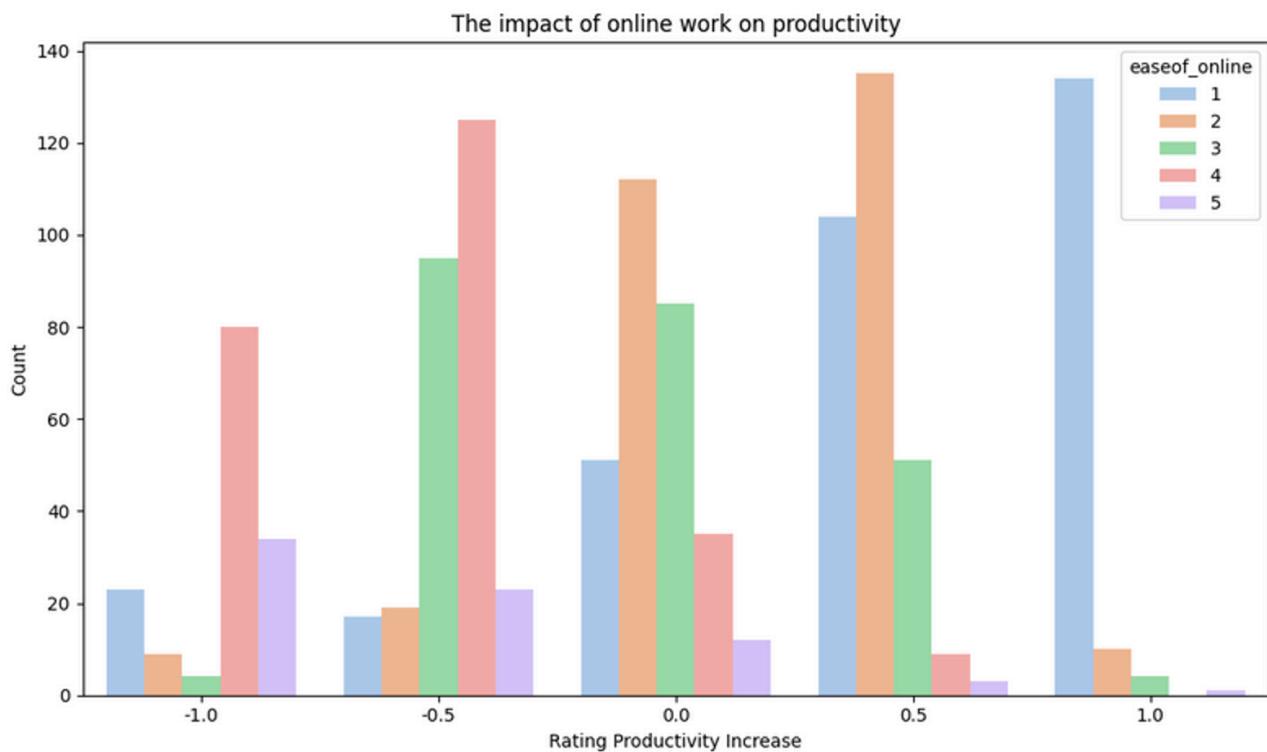


Sa grafika možemo da zaključimo da se tokom korone smanjio broj osoba koji su radili 5, 7 i 9 sati pre korone i da se broj ljudi koji rade 11 i 12 sati dnevno tri puta povećao.



Na prvom mestu po povećanju radnog vremena su preduzetnici, koji su duplo povećali prosečan broj sati provedenog na poslu. Na drugom mestu su medicinski radnici (povećanje 2.5h) i na trećem studenti (povećanje 2.2h). Pandemija se negativno odrazila na mnoge zaposlene čije se radno vreme smanjilo skoro duplo.

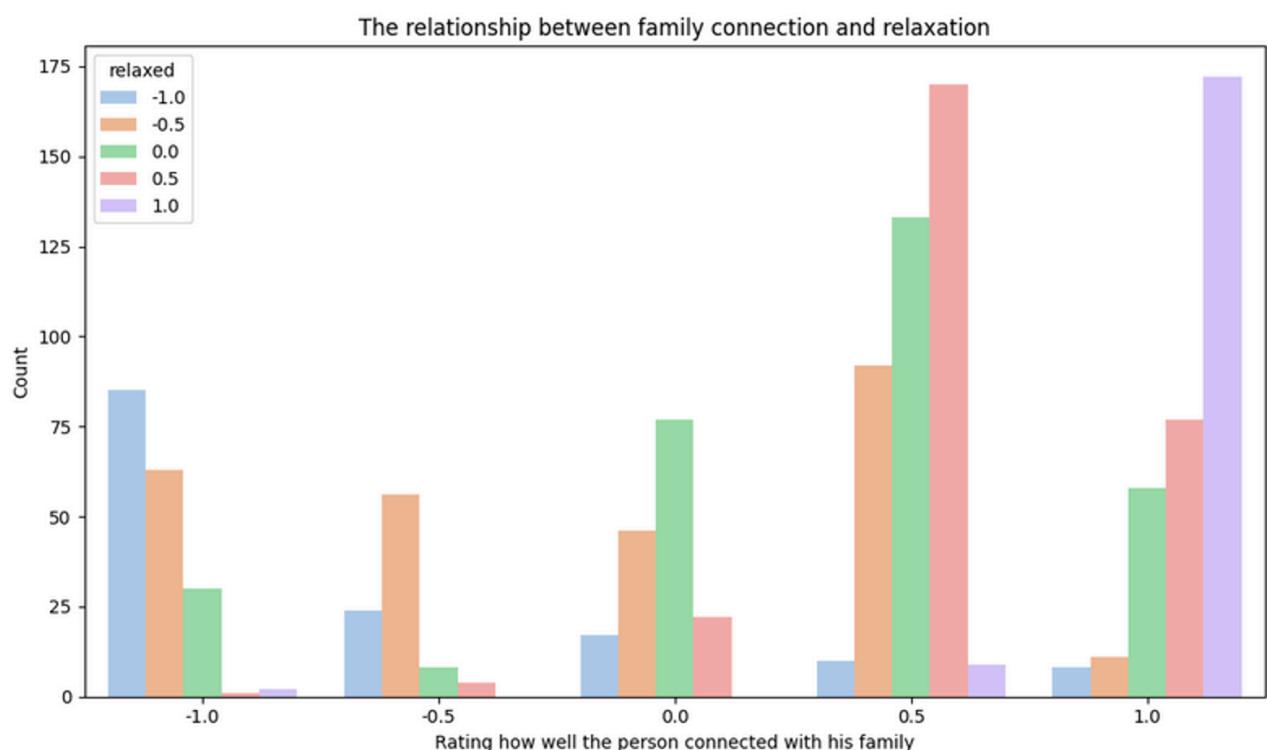
Produktivnost



1 - lako prilagođavanje, 5 - teško prilagođavanje na online režim rad

Produktivnije osobe su bile one koje su se lakše prilagodile online načinu rada.

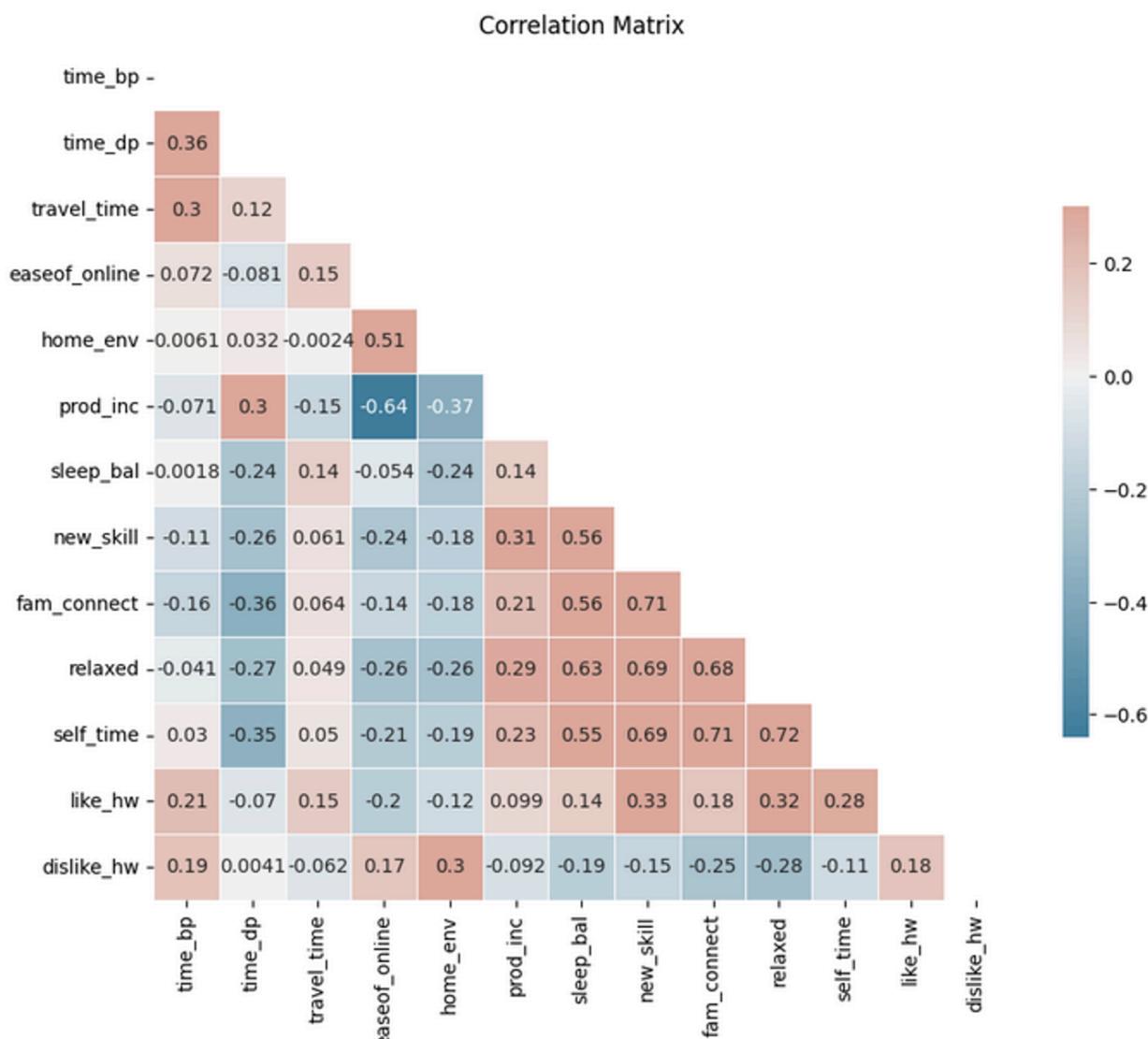
Porodični odnosi i opuštenost



MATRICA KORELACIJE

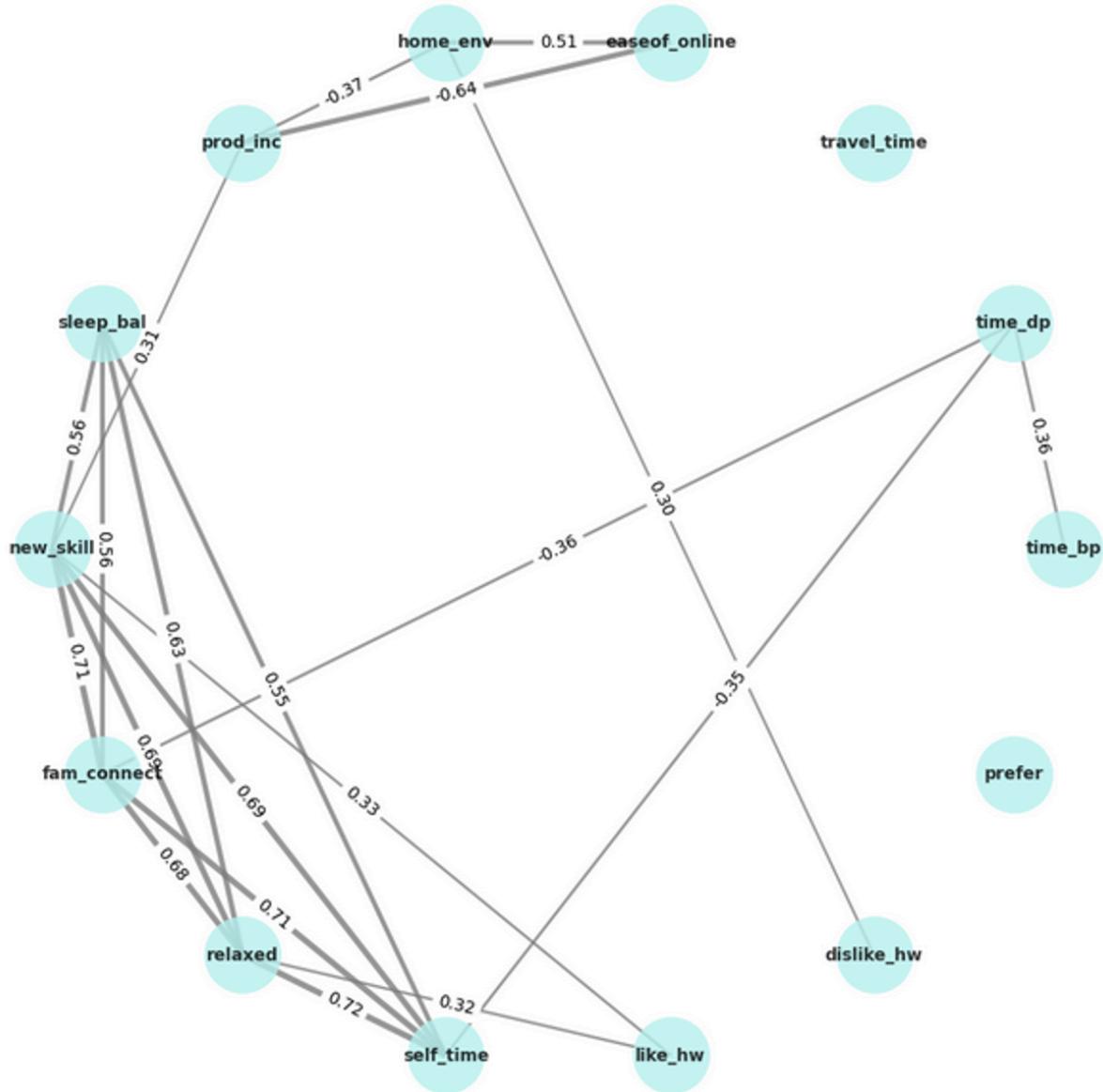
Korelacija je mera koja meri zavisnost između dva numerička atributa, odnosno kako vrednosti jednog atributa utiču na vrednosti drugog atributa. Što je po apsolutnoj vrednosti koeficijenta korelacije veći, to je jača veza između ova dva atributa.

Matrica korelacijske upravo prikazuje korelacije između svaka dva numerička atributa iz skupa podataka.



Da bi nam bilo lakše razumeti koji atributi zavisi od koga, formiraćemo graf sa pragom 0.3 po apsolutnoj vrednosti.

Graph based on Correlation Matrix



Na osnovu grafa možemo da zaključimo da od svakog čvora (sem travel_time i prefer) postoji put - direktna ili preko nekog drugog čvora grana do bar jednog od čvorova: prod_inc, sleep_bal, new_skill, fam_connect, relaxed ili self_time. To nam ukazuje da su ovi atributi imali dosta uticaja na ostale attribute, odnosno ova stanja su imala dosta uticaja na samu osobu tokom pandemije.

PREPROCESIRANJE PODATAKA

Preprocesiranje podataka je ključna faza u analizi podataka koja uključuje pripremu i obradu podataka - uklanjanje grešaka, rukovanje nedostajućim vrednostima, normalizaciju,... Cilj je osigurati kvalitet i pouzdanost podataka, što poboljšava performanse analitičkih modela i omogućava donošenje boljih odluka.

UKLANJANJE GREŠAKA

```
1 data["age"].unique()  
array(['19-25', 'Dec-18', '33-40', '60+', '26-32', '40-50', '50-60'],  
      dtype=object)
```

Primećujemo da se u koloni 'age' nalazi vrednost 'Dec-18' koju je potrebno zameniti sa '0-18'.

BRISANJE NEPOTREBNIH KOLONA

U bazi podataka postoje takozvane prilagođene kolone koje se mogu zanemariti.

X, net_diff - Custom Column (= Unnamed: 19)

time_bp.1 - Custom Column

travel_new - Custom Column

Zbog toga ćemo obrisati ove kolone.

NEDOSTAJUĆE VREDNOSTI

```
age          0  
gender       0  
occupation   0  
line_of_work 696  
time_bp       0  
time_dp       0  
travel_time   0  
easeof_online 0  
home_env      0  
prod_inc      0  
sleep_bal     0  
new_skill     0  
fam_connect   0  
relaxed       0  
self_time     0  
like_hw       0  
dislike_hw    0  
prefer        0  
certaindays_hw 0  
dtype: int64
```

Kolona 'line_of_work' sadrži informacije o oblasti u kojoj rade ispitanici koji su označeni kao 'Working Professional'. S obzirom da ova kolona sadrži nedostajuće vrednosti i da nama nije trenutno bitna konkretna oblast, ovu kolonu ćemo takođe izbaciti.

KODIRANJE KATEGORIČKIH ATRIBUTA

Većina algoritama i modela zahteva numeričke podatke kao ulaz, što znači da ćemo morati kodirati naše nenumeričke kategorije atributе.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1175 entries, 0 to 1174
Data columns (total 19 columns):
 #   Column            Non-Null Count  Dtype  
--- 
 0   age               1175 non-null    object  
 1   gender             1175 non-null    object  
 2   occupation         1175 non-null    object  
 3   time_bp            1175 non-null    int64  
 4   time_dp            1175 non-null    int64  
 5   travel_time        1175 non-null    float64 
 6   easeof_online      1175 non-null    int64  
 7   home_env           1175 non-null    int64  
 8   prod_inc           1175 non-null    float64 
 9   sleep_bal          1175 non-null    float64 
 10  new_skill          1175 non-null    float64 
 11  fam_connect        1175 non-null    float64 
 12  relaxed             1175 non-null    float64 
 13  self_time          1175 non-null    float64 
 14  like_hw             1175 non-null    int64  
 15  dislike_hw          1175 non-null    int64  
 16  prefer              1175 non-null    object  
 17  certaindays_hw     1175 non-null    object  
 18  lifestyle_change    1175 non-null    category 
dtypes: category(1), float64(7), int64(6), object(5)
memory usage: 166.6+ KB
```

Pregled kategoričkih atributa i pripadajućih numeričkih vrednosti koje su im dodijeljene nakon label encoding-a:

Atribut	Kategoričke vrednosti	Numeričke vrednosti
age	0-18	0
	19-25	1
	26-32	2
	33-40	3
	40-50	4
	50-60	5
	60+	6
gender	Female	0
	Male	1
	Prefer not to say	2

Atribut	Kategoričke vrednosti	Numeričke vrednosti
occupation	Currently Out of Work	0
	Entrepreneur	1
	Homemaker	2
	Medical Professional aiding efforts against COVID-19	3
	Retired/Senior Citizen	4
	Student in College	5
	Student in School	6
prefer	Working Professional	7
	Complete Physical Attendance	0
certaindays_hw	Work/study from home	1
	Maybe	0
	No	1
	Yes	2

DODATNE PRIPREME

Ovim smo završili preprocesiranje koje je zajedničko i za klasifikaciju i klasterovanje. Sada prelazimo na specifične korake preprocesiranja za oba pristupa.

PREPROCESIRANJE ZA KLASIFIKACIJU

Podela na ulazne i ciljne attribute

Cilj istraživanja je razumeti kako su različite promene načina života uticale na psihološko stanje pojedinca.

Za merenje stepena promene načina života tokom pandemije COVID-19 konstruisaćemo ciljnu promenljivu 'lifestyle_change' na osnovu atributa: 'prod_inc', 'sleep_bal', 'new_skill', 'fam_connect', 'relaxed' i 'self_time', a zatim ih iz baze podataka izbaciti. Izbor baš ovih atributa je iz više razloga:

1. Matrica korelaciјe:

Na osnovu analize matrice korelacija zaključili smo da su ovi atributi imali najviše uticaja na ostale attribute koji se odnose na promenu načina života, što ukazuje na njihov potencijalni značaj.

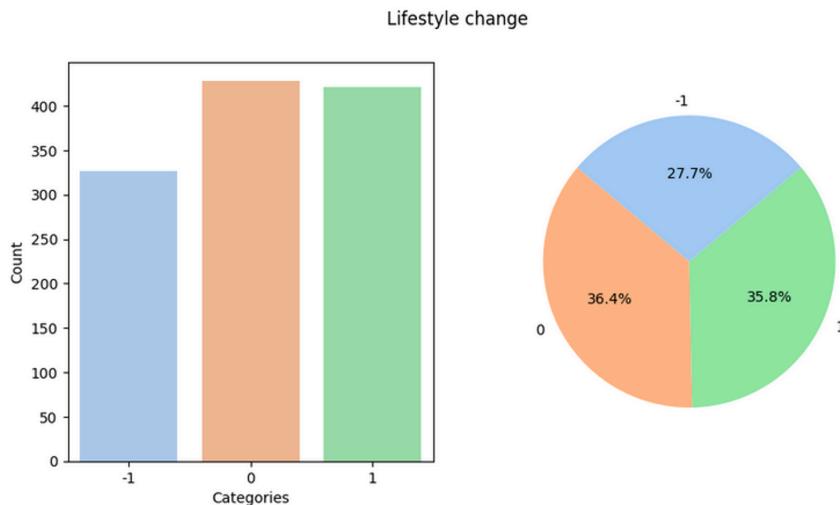
2. Direktan uticaj na psihološko stanje:

Svaki od odabranih atributa direktno se odnosi na aspekte koji mogu značajno uticati na psihološko stanje pojedinca. Na primer, produktivnost može biti povezano sa osećajem uspeha, opuštenost često povezujemo sa nivoom stresa, režim spavanja ima duboku vezu sa kvalitetom sna i smanjenjem umora, porodica često ima dosta veliki uticaj na nas same,...

Na osnovu ocena koje opisuju stepen promene konkretnog aspekta, svaku osobu ćemo klasifikovati u jednu od tri kategorije na osnovu srednje vrednosti pojedinačnih ocena:

- **negativna promena (-1)**: ako je srednja vrednost ocena manja od -0.25, smatraćemo da je osoba doživela negativnu promenu u svom načinu života
- **neutralna promena (0)**: ako je srednja ocena između -0.25 i 0.25, smatraćemo da se način života ove osobe nije mnogo promenio
- **pozitivna promena (1)**: ako je srednja ocena veća od 0.25, smatraćemo da je osoba doživela pozitivnu promenu u svom životu.

Balansiranost klasa



S obzirom na ove vrednosti, možemo primetiti da podaci nisu savršeno balansirani, ali razlike nisu drastične. Za sada nećemo primeniti nikakve tehnike balansiranja. Ukoliko primetimo da neravnoteža negativno utiče na performanse modela, tada ćemo primeniti neku tehniku za balansiranje klasa.

Podela na trening i test skup i standardizacija

Za klasifikaciju nam je potrebno da podelimo skup na trening i test skup. Koristićemo 80% podataka za trening i preostalih 20% za testiranje. Na oba skupa primenićemo standardizaciju kako bismo sve atribute sveli na istu skalu vrednosti.

PREPROCESIRANJE ZA KLASTEROVANJE

Kod klasterovanja nema potrebe podeliti podatke na trening i test skup, tako da ćemo samo ove podatke samo standardizovati.

Nakon ovih transformacija naši podaci su spremni za klasifikaciju i klasterovanje.

KLASIFIKACIJA

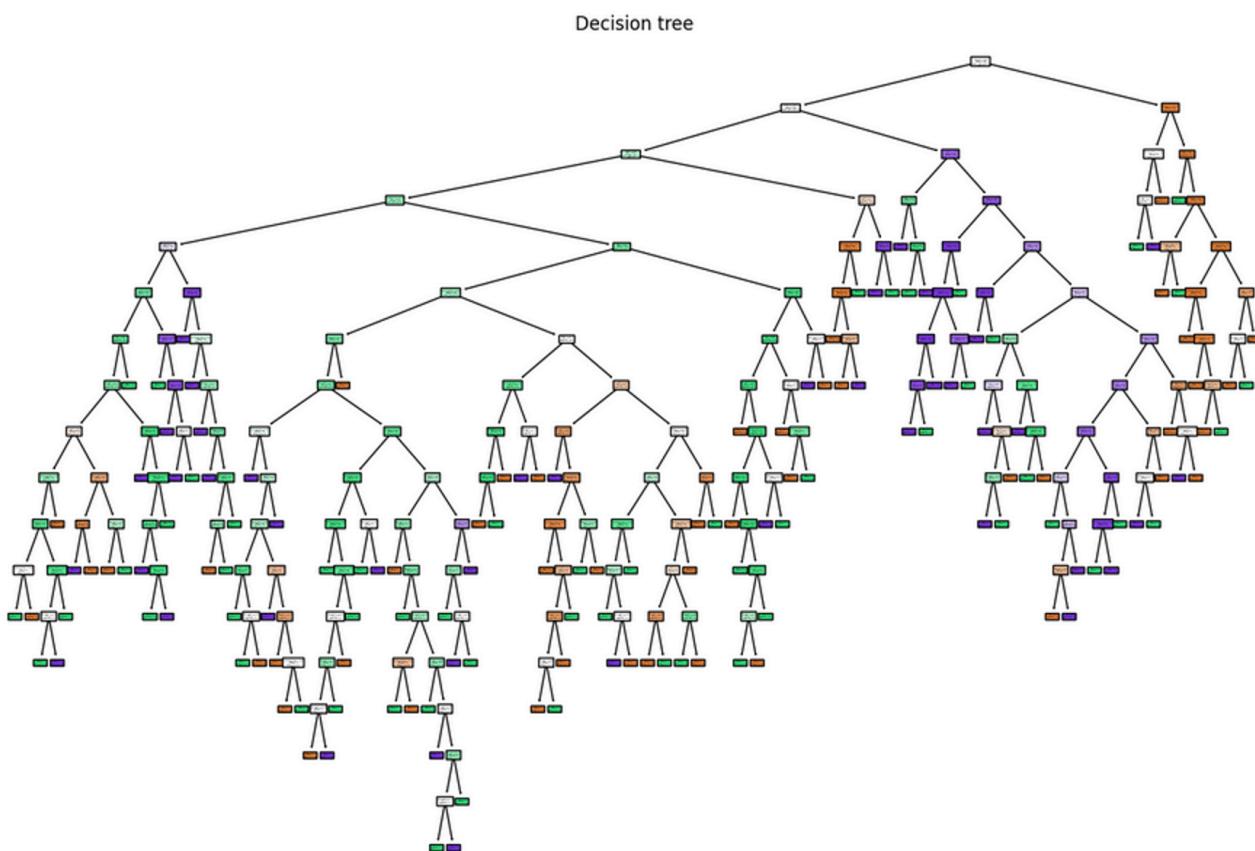
Osnovna ideja klasifikacije je razvijanje modela koji može "naučiti" kako da razvrsta podatke na osnovu uzoraka i karakteristika. Model se trenira na osnovu postojećih podataka sa poznatim klasama, a zatim se koristi za klasifikaciju novih, nepoznatih podataka. Važnost klasifikacije se ogleda u sposobnosti donošenja informisanih odluka i prognoziranja budućih događaja na osnovu analizu podataka.

STABLA ODLUČIVANJA

Decision Trees

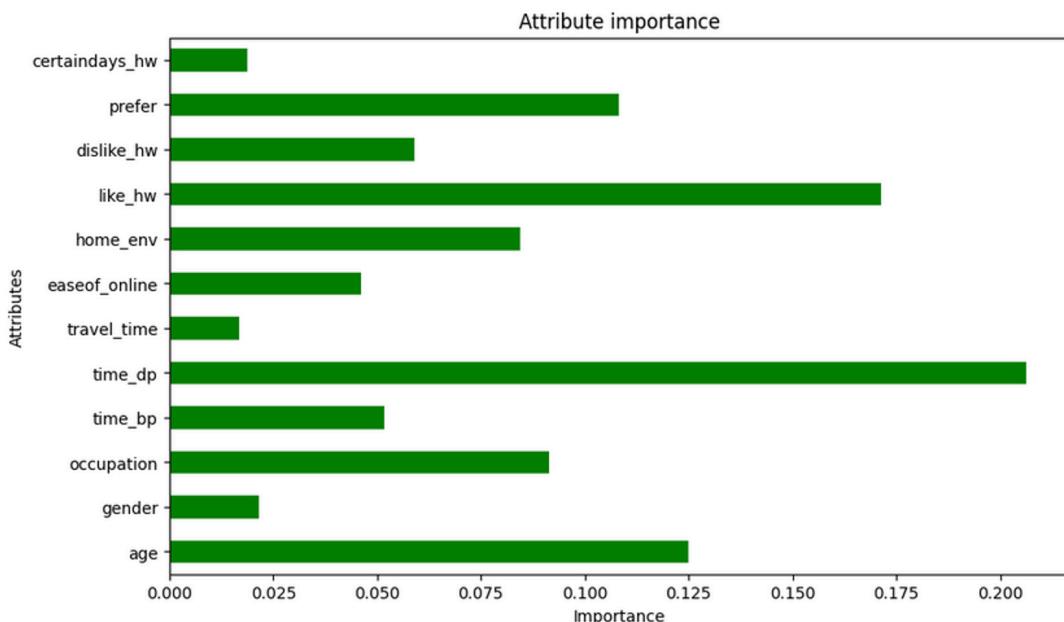
Stablo odlučivanja je model koji se koristi za donošenje odluka na osnovu hijerarhijskih pravila i karakteristika podataka. Glavne komponente su čvorovi i grane. Čvorovi predstavljaju tačke u stablu gde se postavljaju pitanja, a grane povezuju čvorove i označavaju odgovore na posavljena pitanja ili primenjena pravila. Nakon izgradnje stabla, podaci se klasifikuju putujući od korena do listova, gde svaki list predstavlja konačnu klasu podatka.

Koristeći `DecisionTreeClassifier` sa podrazumevanim vrednostima dobijamo stablo koje izgleda ovako:



Broj čvorova: 289, dubina: 17

Analizom važnosti atributa, možemo bolje razumeti koji atributi su bili ključni za oblikovanje rezultata.



Najvažniji atributi za predviđanje ciljne promenljive su "time_dp" i "like_hw", dok su najmanje uticaja imali "travel_time" i "gender".

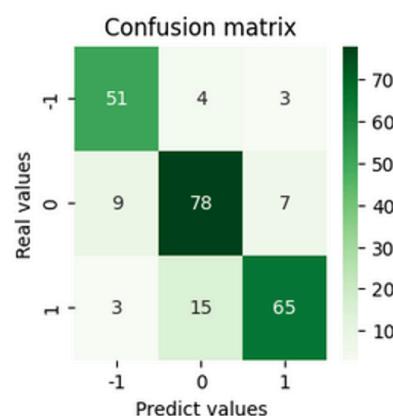
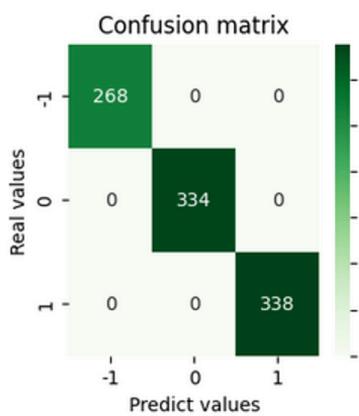
Performanse modela:

Train data:

	precision	recall	f1-score	support
-1	1.00	1.00	1.00	268
0	1.00	1.00	1.00	334
1	1.00	1.00	1.00	338
accuracy			1.00	940
macro avg	1.00	1.00	1.00	940
weighted avg	1.00	1.00	1.00	940

Test data:

	precision	recall	f1-score	support
-1	0.81	0.88	0.84	58
0	0.80	0.83	0.82	94
1	0.87	0.78	0.82	83
accuracy			0.83	235
macro avg	0.83	0.83	0.83	235
weighted avg	0.83	0.83	0.83	235



Na osnovu analize rezultata možemo zaključiti da je model odlučivanja veoma dobro istreniran na trening skupu, gde su sve metrike bile savršene. Međutim, na test skupu, primećujemo da model dosta greši, posebno u klasifikaciji instance klase 0 i 1. Ovakvi rezultati ukazuju na to da se naš model preprilagodio trening podacima.

Mere optimizacije:

- Podešavanje hiper-parametara:

Hiper-parametri igraju ključnu ulogu u oblikovanju performansi i ponašanja modela. Ovi parametri omogućavaju kontrolu nad složenošću stabla, njegovom dubinom, brojem listova,... Oni se ne uče tokom treniranja, već se podešavaju pri inicijalizaciji modela. GridSearchCV je metoda koja istražuje različite kombinacije hiper-parametara na osnovu unapred definisanog skupa vrednosti.

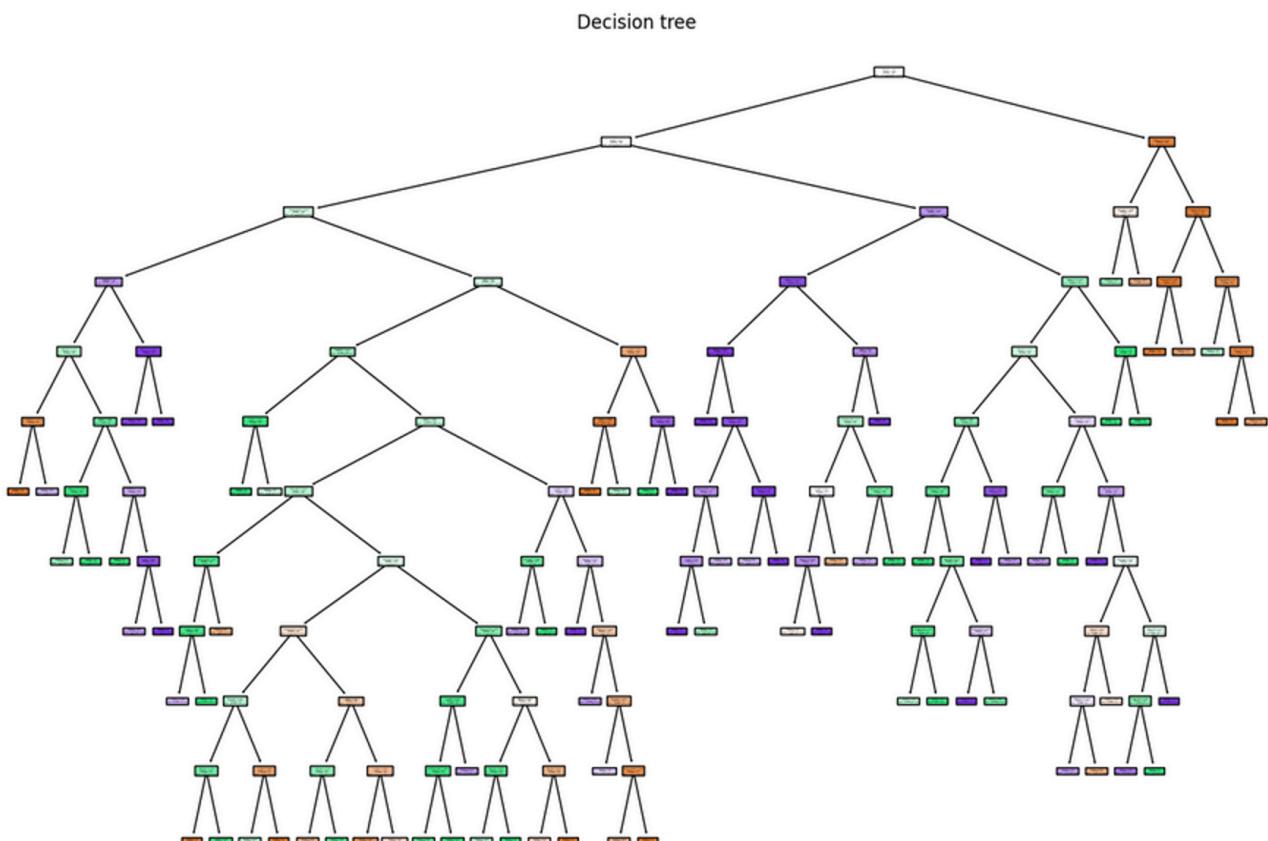
Skup vrednosti:

```
1 params = {
2     'criterion': ['gini', 'entropy'],
3     'max_depth': [7, 9, 11, 13, 15],
4     'min_samples_leaf': [3, 4, 5, 6],
5     'class_weight': [None, 'balanced']
6 }
```

Nakon treniranja i isprobavanja različitih kombinacija, model sa najboljim performansama ima hiper-parametre:

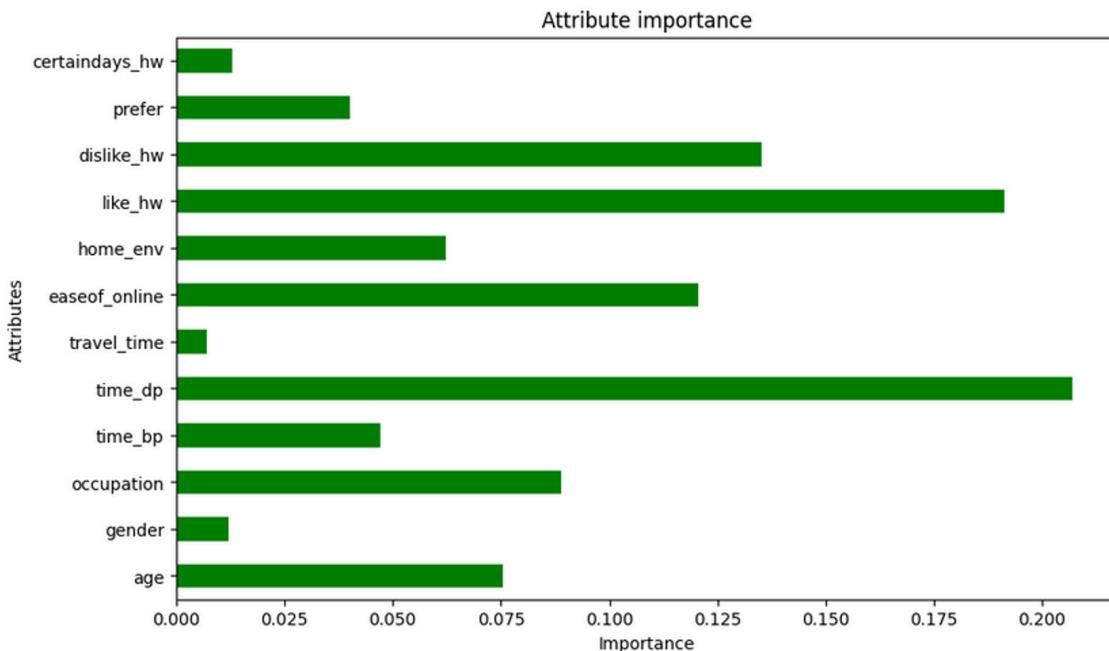
```
1 estimator_dtc.best_params_
{'class_weight': 'balanced',
 'criterion': 'entropy',
 'max_depth': 11,
 'min_samples_leaf': 3}
```

i stablo ima oblik:



Broj čvorova: 153, dubina: 11

Važnost atributa:



I u ovom modelu "time_dp" i "like_hw" su se istaknuli kao najvažniji atributi, dok "travel_time" i "gender" i dalje imaju najmanju važnost. Veću važnost imaju i atributi "dislike_hw", "easeof_online" prilikom predviđanja ciljne promenljive.

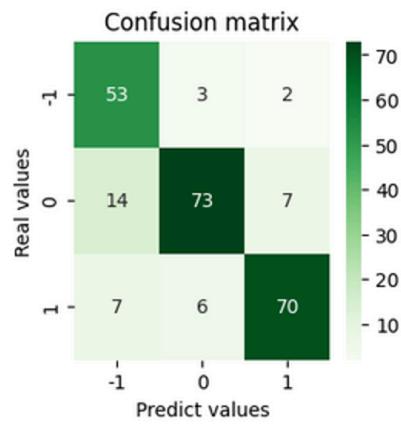
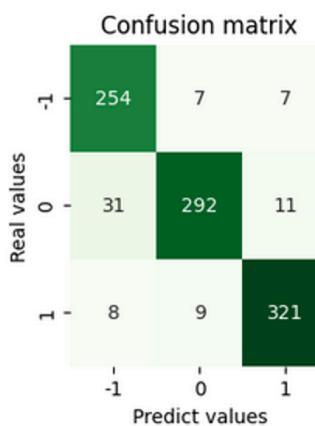
Performanse modela:

Train data:

	precision	recall	f1-score	support
-1	0.87	0.95	0.91	268
0	0.95	0.87	0.91	334
1	0.95	0.95	0.95	338
accuracy			0.92	
macro avg	0.92	0.92	0.92	940
weighted avg	0.92	0.92	0.92	940

Test data:

	precision	recall	f1-score	support
-1	0.72	0.91	0.80	58
0	0.89	0.78	0.83	94
1	0.89	0.84	0.86	83
accuracy			0.83	235
macro avg	0.83	0.84	0.83	235
weighted avg	0.85	0.83	0.84	235



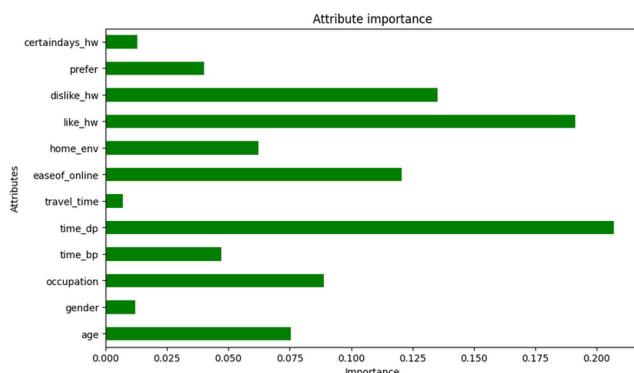
Preciznost, odziv i F1-ocene su visoke za svaku klasu, što ukazuje na to da je model sposoban da (donekle) kvalitetno klasificiše instance na skupu za obuku. Međutim, kada je model testiran na neviđenim podacima (test skup), performanse su nešto niže. S obzirom na razlike u performansama između skupa za obuku i test skupa i ovaj model pokazuje blagu tendenciju za preprilagođavanje.

- Slučajne šume:

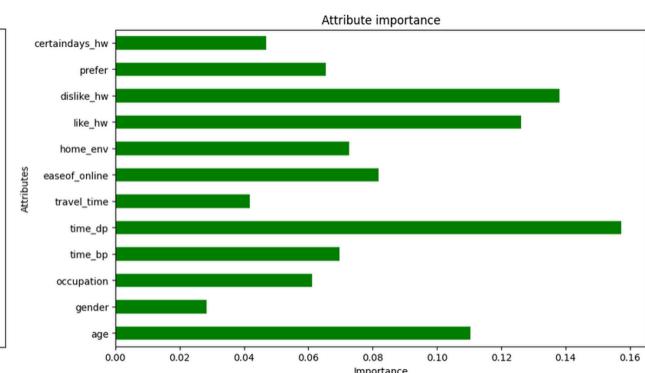
Slučajna šuma je ansambl algoritama koji se bazira na ideji da više modela može zajedno da donese bolje odluke od pojedinačnih modela. Svako stablo odlučivanja se trenira na različitom podskupu podataka, a zatim se kombinuju rezultati kako bi se donela konačna odluka.

Kao i kod pojedinačnih stabala odlučivanja, slučajne šume, `RandomForestClassifier` moguće je pokrenuti sa podrazumevanim vrednostima, ali i podešavati hiper-parametre.

Važnosti atributa:



slučajne šume sa podrazumevanim vrednostima.



slučajne šume sa hiper-parametrima:
`{'criterion': 'entropy', 'n_estimators': 100}`

Najvažniji atribut u oba modela je "time_dp". Razlika je u drugom važnom atributu, gde je u prvom modelu to "easeof_online", a u drugom modelu to "dislike_hw". Takođe, srednje važni atributi variraju, naglašavajući razlike u uticaju atributa kao što su "age", "home_env", "occupation" i "prefer" između ta dva modela.

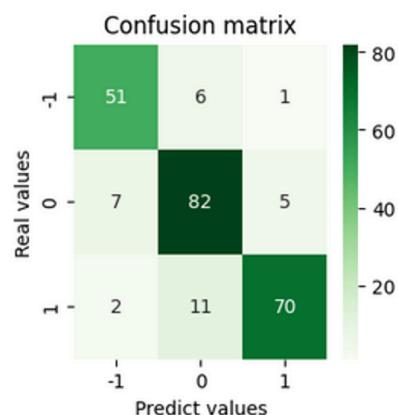
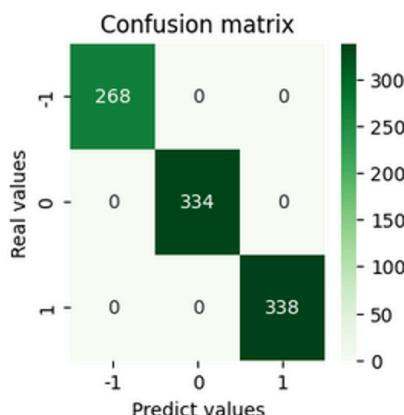
Preformanse modela sa podrazumevanim vrednostima:

Train data:

	precision	recall	f1-score	support
-1	1.00	1.00	1.00	268
0	1.00	1.00	1.00	334
1	1.00	1.00	1.00	338
accuracy			1.00	940
macro avg	1.00	1.00	1.00	940
weighted avg	1.00	1.00	1.00	940

Test data:

	precision	recall	f1-score	support
-1	0.85	0.88	0.86	58
0	0.83	0.87	0.85	94
1	0.92	0.84	0.88	83
accuracy			0.86	235
macro avg	0.87	0.87	0.86	235
weighted avg	0.87	0.86	0.86	235



Ovaj model je takođe sklon preprilagođavanju. Na skupu za trening, model postiže 100% preciznost, odziv i F1-ocjene za sve tri klase (-1, 0, 1). Kada je model testiran na test podacima, primećujemo blagi pad u performansama, iako su i dalje solidne.

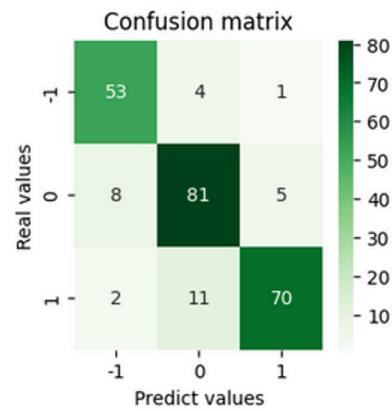
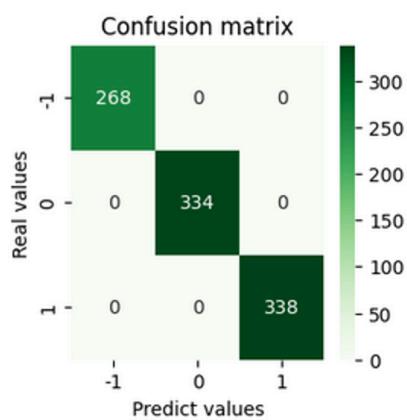
Performanse modela sa podešenim hiper-parametrima:

Train data:

	precision	recall	f1-score	support
-1	1.00	1.00	1.00	268
0	1.00	1.00	1.00	334
1	1.00	1.00	1.00	338
accuracy			1.00	940
macro avg	1.00	1.00	1.00	940
weighted avg	1.00	1.00	1.00	940

Test data:

	precision	recall	f1-score	support
-1	0.84	0.91	0.88	58
0	0.84	0.86	0.85	94
1	0.92	0.84	0.88	83
accuracy			0.87	235
macro avg	0.87	0.87	0.87	235
weighted avg	0.87	0.87	0.87	235



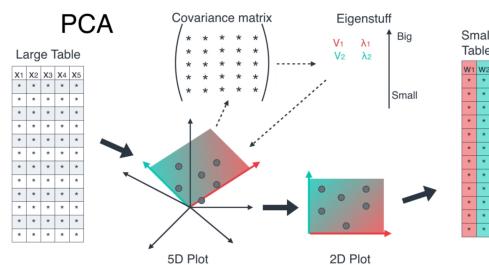
Ovaj model nije doneo nikakva poboljšanja u odnosu na prethodnog modela.

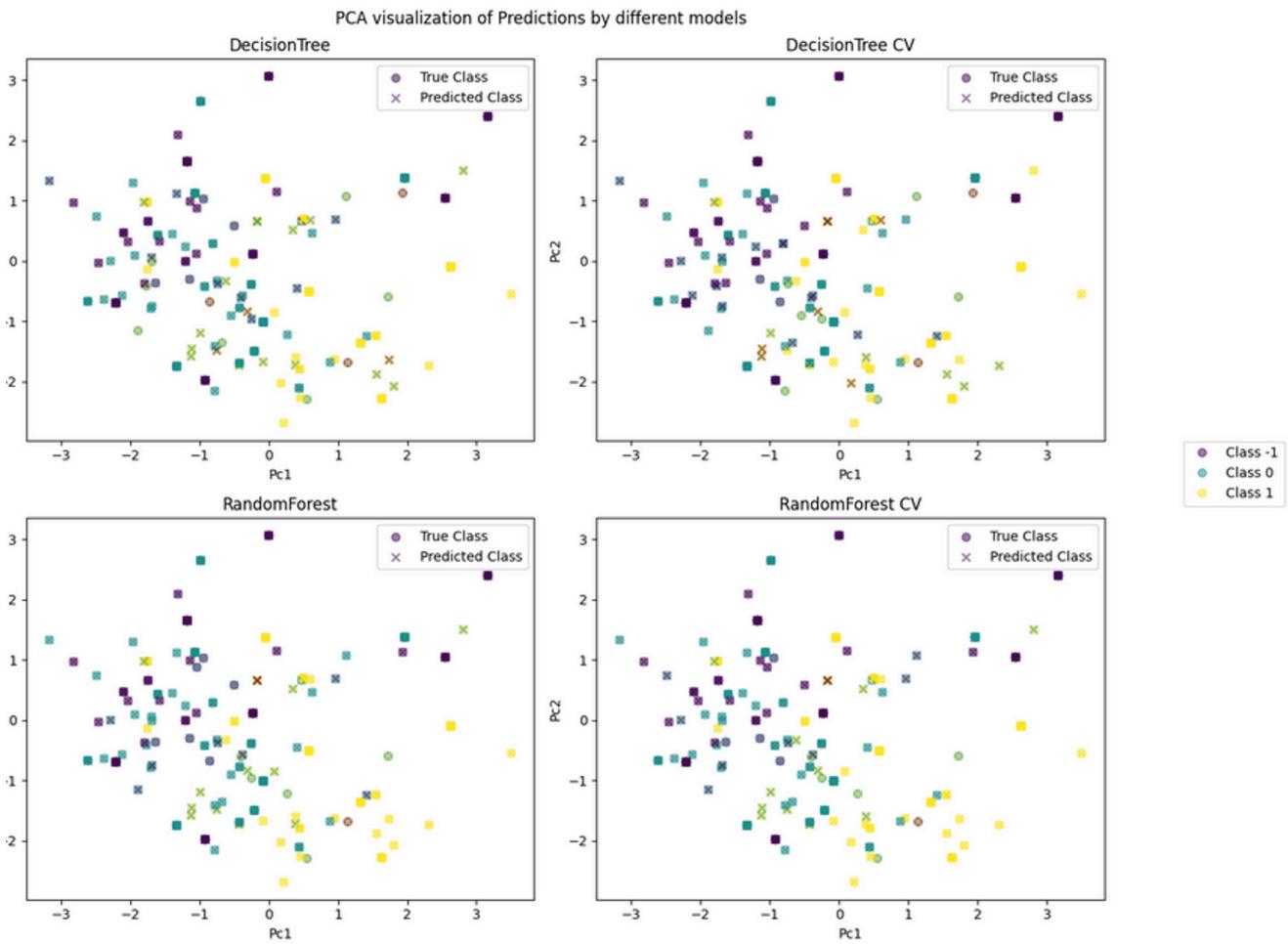
Vizualizacija klasifikacije stablima odlučivanja

PCA (Principal Component Analysis) je statistička metoda koja se koristi za redukciju dimenzionalnosti podataka. Glavni cilj PCA-a je transformisati originalne attribute u novi skup linearne nezavisnih promenljivih nazvanih glavne komponente.

Glavne komponente su linearne kombinacije početnih atributa i sortirane su po smanjenju njihove varijanse, tako da prva komponenta ima najveću varijansu, druga ima drugu najveću varijansu, i tako dalje. Time se postiže smanjenje dimenzionalnosti dok se zadržava što više informacija.

PCA je koristan alat u analizi podataka, posebno kada postoje mnogo atributa i želimo da sačuvamo bitne informacije, eliminuјemo višak redundantnih informacija te olakšati vizualizaciju i analizu podataka.

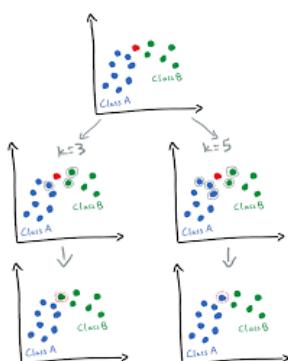




K-NAJBLIŽIH SUSEDА

k-nearest neighbors - KNN

K-najbližih suseda je jednostavan i efikasan algoritam koji se koristi za klasifikaciju podataka. Osnovna ideja je da slični primeri (instance) treba da budu klasifikovani ili predviđeni slično. K-NN spada u grupu "lenjih" klasifikatora, što znači da ne izračunava globalni model tokom treninga, već čuva trening podatke i na osnovu njih identificuje najbliže "susede" trening instanci i koristi ih za donošenje odluka za test instance. Važno je odabrati ispravnu vrednost k (broja susjeda) za najbolje rezultate u primeni.



Koristeći KNeighborsClassifier sa podrazumevanim vrednostima, naš model ima sledeće performanse:

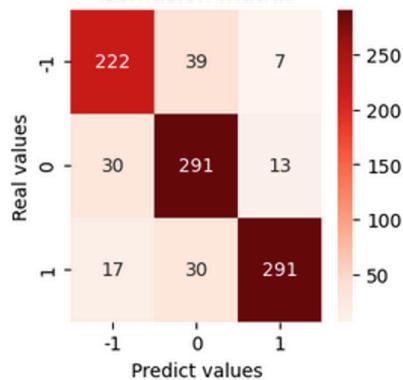
Train data:

	precision	recall	f1-score	support
-1	0.83	0.83	0.83	268
0	0.81	0.87	0.84	334
1	0.94	0.86	0.90	338
accuracy			0.86	940
macro avg	0.86	0.85	0.85	940
weighted avg	0.86	0.86	0.86	940

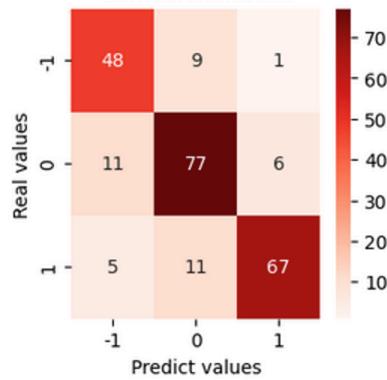
Test data:

	precision	recall	f1-score	support
-1	0.75	0.83	0.79	58
0	0.79	0.82	0.81	94
1	0.91	0.81	0.85	83
accuracy			0.82	235
macro avg	0.82	0.82	0.82	235
weighted avg	0.82	0.82	0.82	235

Confusion matrix



Confusion matrix



Iako model pravi "dosta" grešaka prilikom klasifikacije, uspeva da održi sličan nivo preciznosti i odziva na test skupu kao na trening skupu.

Mere optimizacije:

- Podešavanje hiper-parametara

Slično kao kod stabla odlučivanja, hiper-parametri igraju veliku ulogu u evaluaciji modela i performansi. Korišćenjem GridSearchCV isprobaćemo različite kombinacije hiper-parametara.

```
1 params = {  
2     'n_neighbors': [1, 5, 10, 20, 22],  
3     'weights': ['uniform', 'distance'],  
4     'p': [1, 2]  
5 }
```

Kao najbolji model pokazao se model sa kombinacijom parametara:

```
1 estimator_knn.best_params_  
{'n_neighbors': 22, 'p': 1, 'weights': 'distance'}
```

i on uspeva da ostvari tačnost:

```
1 estimator_knn.best_score_  
0.8404255319148936
```

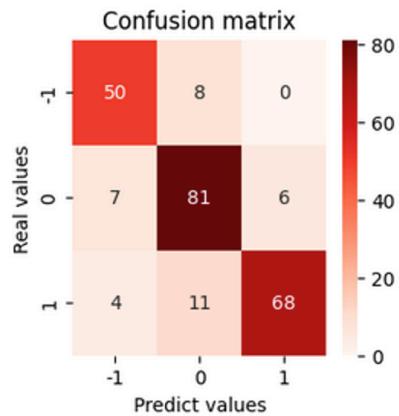
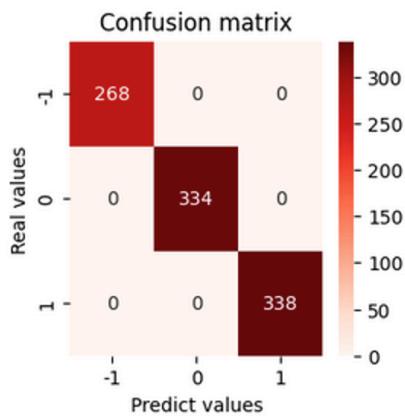
Performanse modela:

Train data:

	precision	recall	f1-score	support
-1	1.00	1.00	1.00	268
0	1.00	1.00	1.00	334
1	1.00	1.00	1.00	338
accuracy			1.00	940
macro avg	1.00	1.00	1.00	940
weighted avg	1.00	1.00	1.00	940

Test data:

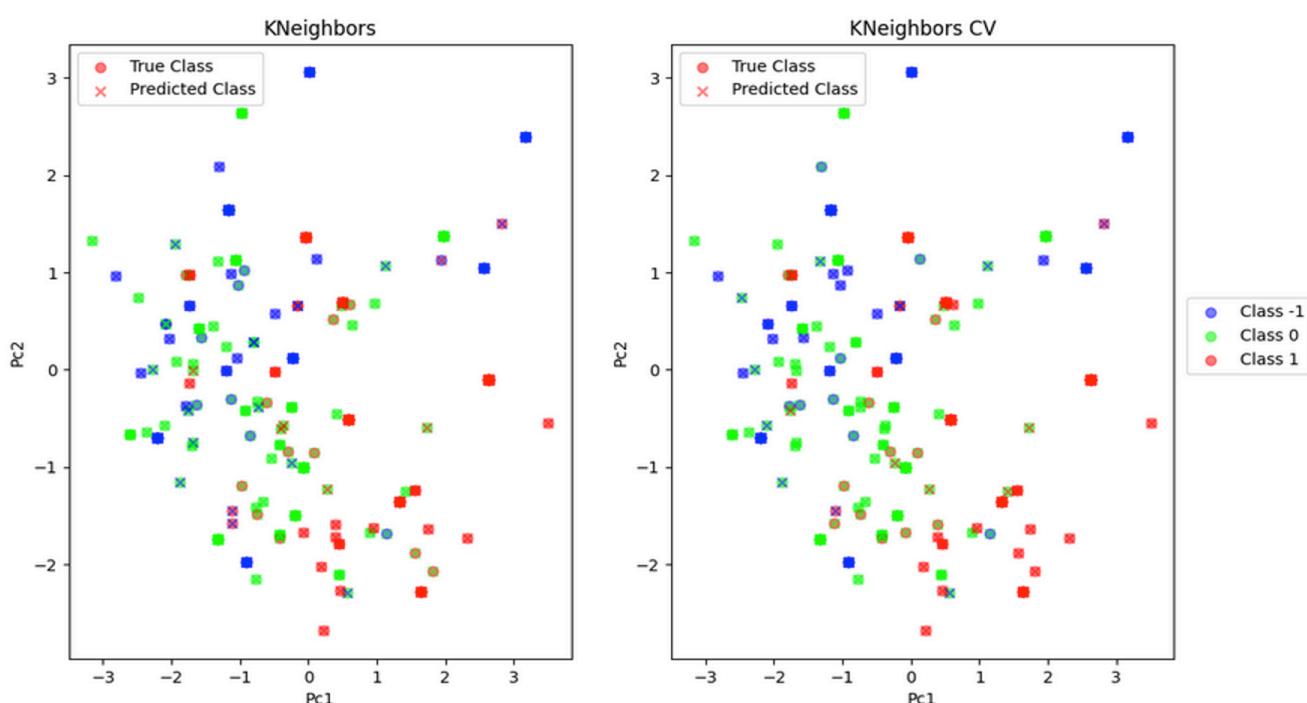
	precision	recall	f1-score	support
-1	0.82	0.86	0.84	58
0	0.81	0.86	0.84	94
1	0.92	0.82	0.87	83
accuracy			0.85	235
macro avg	0.85	0.85	0.85	235
weighted avg	0.85	0.85	0.85	235



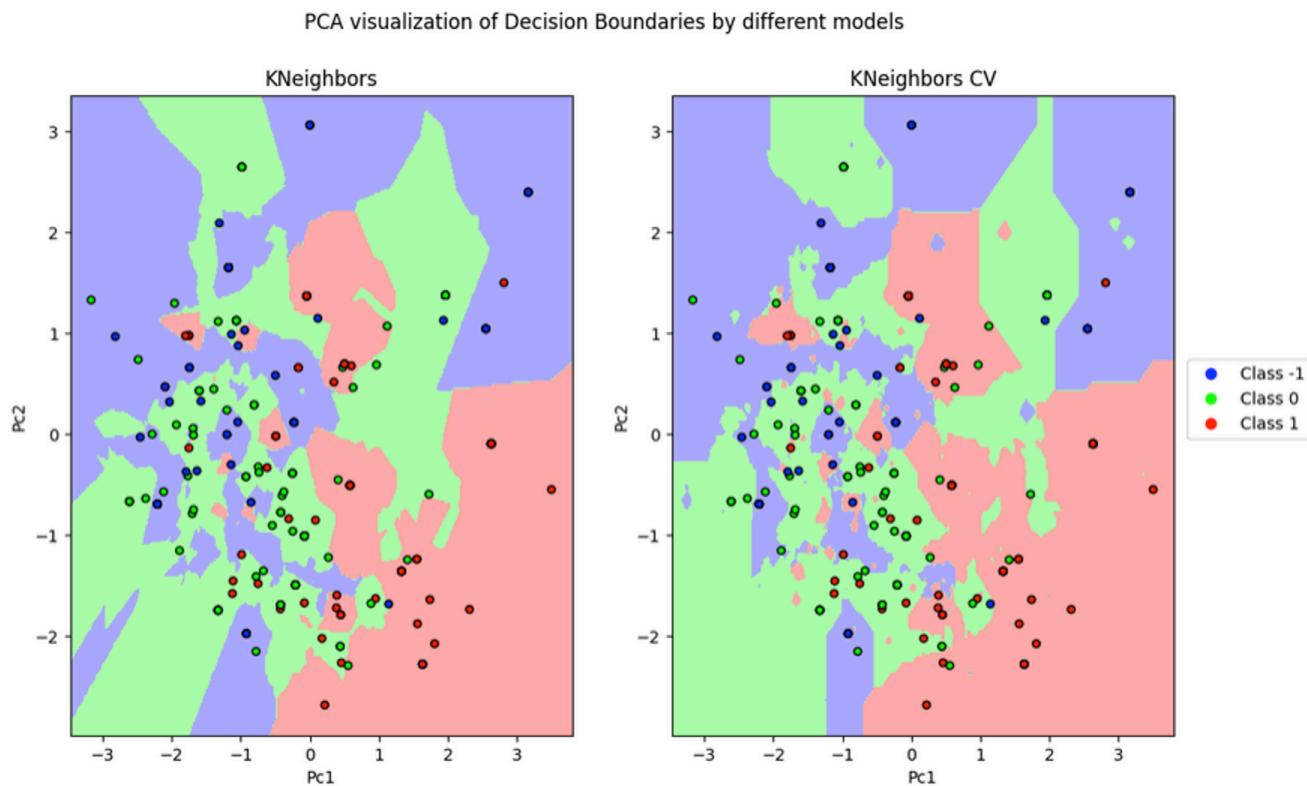
Na trening skupu, model postiže savršene rezultate za sve tri klase (-1, 0, 1), sugerijući da je naučio svaki detalj skupa za obuku i klasifikovao instance s visokom preciznošću i odzivom. Međutim, ovi rezultati nisu reproducibilni na test skupu, gdje dolazi do smanjenja performansi. Naš model se preprilagodion trening podacima.

Vizuelizacija klasifikacije KNN algoritmom

PCA visualization of Predictions by different models



Za lakšu vizualizaciju rasporeda podataka i načina na koji K-NN algoritam razdvaja različite klase ili regije koriste se granice odluka. One omogućavaju jasno prikazivanje kako model donosi odluke o klasifikaciji ili predviđanju na temelju rasporeda podataka, čime olakšava razumijevanje i analizu modela.



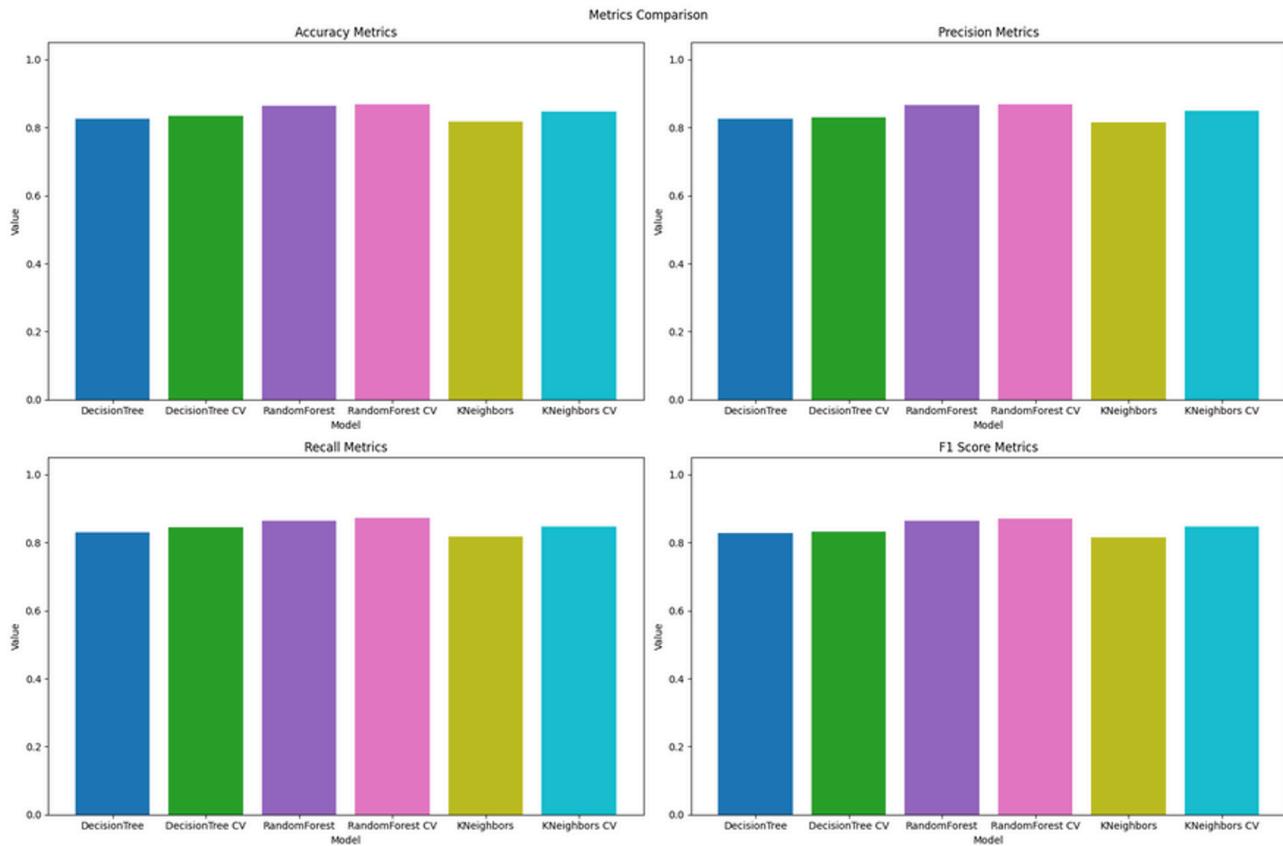
POREĐENJE MODELAA KLASIFIKACIJE

Poređenje modela je bitno jer omogućava identifikaciju najboljeg modela za rešavanje određenog problema. To pomaže u donošenju informisanih odluka o tome koji model ima bolju prediktivnu sposobnost, bolju preciznost i bolje performanse na stvarnim podacima.

Metrike

- Accuracy meri ukupnu tačnost modela u klasifikaciji
- Precision meri koliko je stvarno pozitivnih instanci model pravilno identifikovao
- Recall meri sposobnost modela da pravilno identificiše sve stvarne pozitivne instance
- F1-score je harmonijska srednja vrednost preciznosti i odziva

S obzirom da smo sve metrike detaljno ispisali za svaki model gore, nema potrebe ponovo ih pisati ovde - prikazaćemo samo grafički razlike u metrikama koje smo dobili na test skupu.

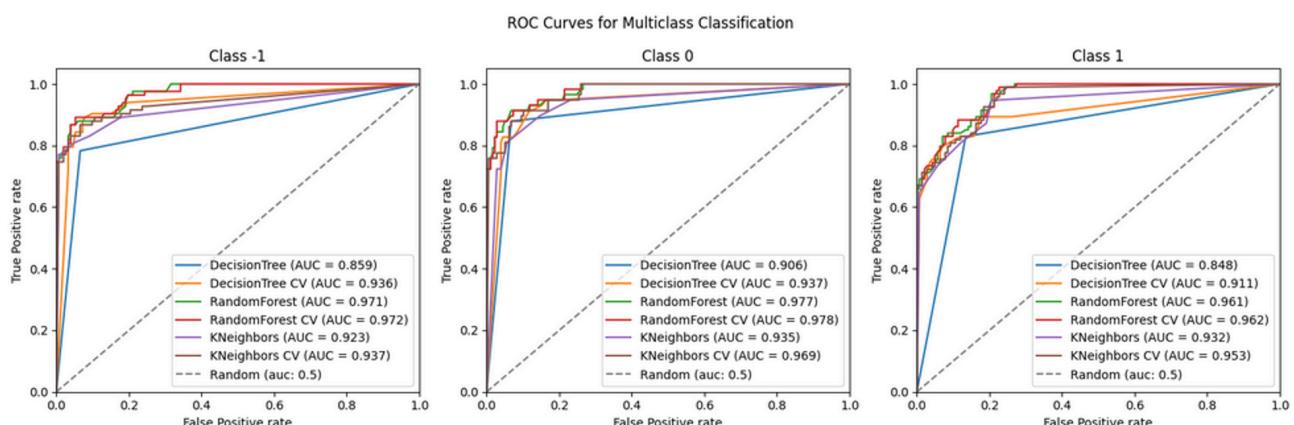


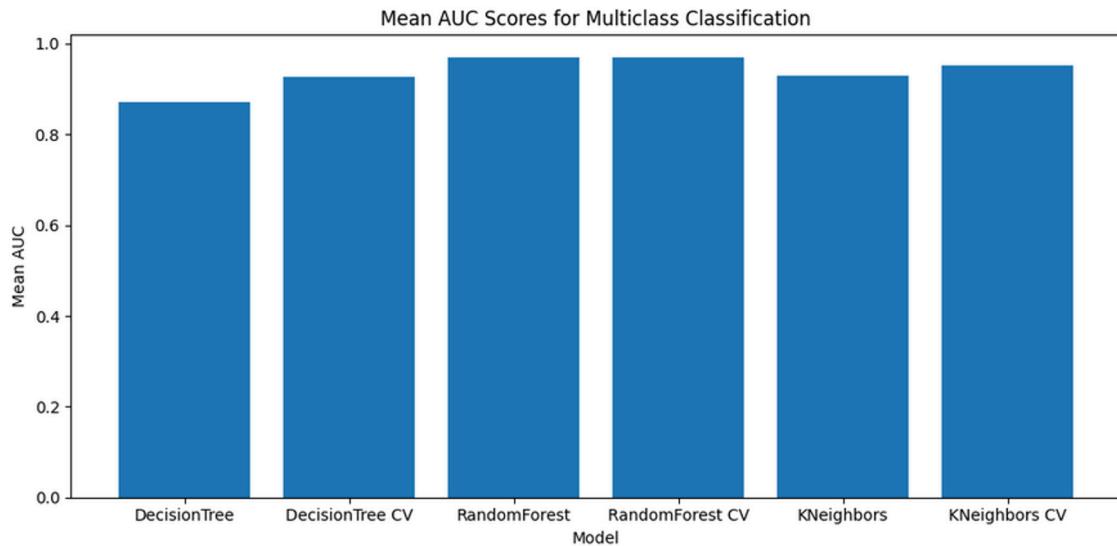
Na osnovu ovih metrika zaključujemo da "Random Forest" bez/са podešenim hiper-parametrima nam daje najbolje rezultate. Važno je napomenuti da su i ostali modeli koji imaju podešene hiper-parametre postigli bolje rezultate u odnosu na modele koji su imali podrazumevane. Međutim, ne zaboravimo da su svi modeli pokazali određenu količinu preprilagođavanju trening podacima.

ROC kriva

ROC kriva prikazuje odnos između stope lažno pozitivnih (FPR) i tačno pozitivnih (TPR). Idealna ROC kriva ide prema gornjem levu uglu, što znači visok TPR i nizak FPR.

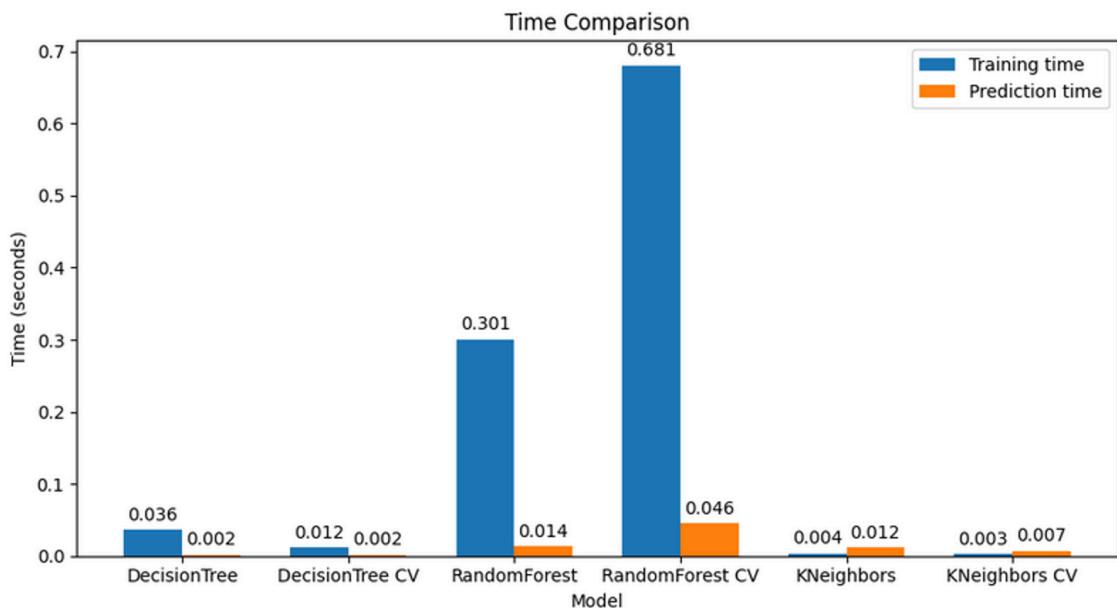
Površina spod ROC krive (AUC - Area Under the Curve) koristi se kao mera performansi modela, gde veća vrednost AUC ukazuje na bolji model. Za $AUC = 0.5$ model nasumično pogaća klase.





Na osnovu ovih vrednosti, opšti zaključak je da, svi modeli (osim "Stabla odlučivanja") imaju slične performanse kada je u pitanju klasifikacija nad test podacima.

Vremenska složenost

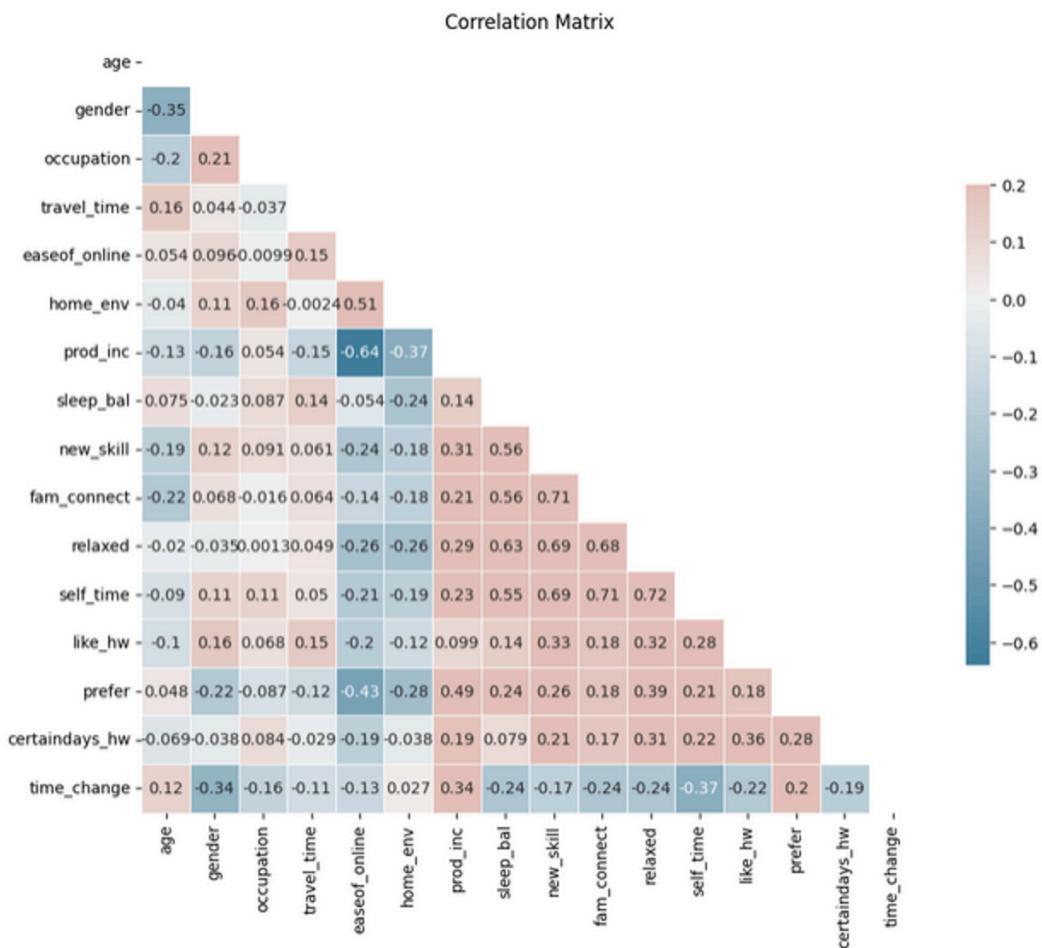


KLASTEROVANJE

Klasterovanje je tehnika analize podataka koja se koristi za grupisanje sličnih podataka zajedno. Ova tehnika je deo nenadgledanog učenja i ima široku primenu. Glavna svrha klasterovanje je otkrivanje grupa, takozvanih klastera, na osnovu sličnosti podataka, bez potrebe za prethodnim znanjem o tome kako bi trebalo da izgledaju te grupe.

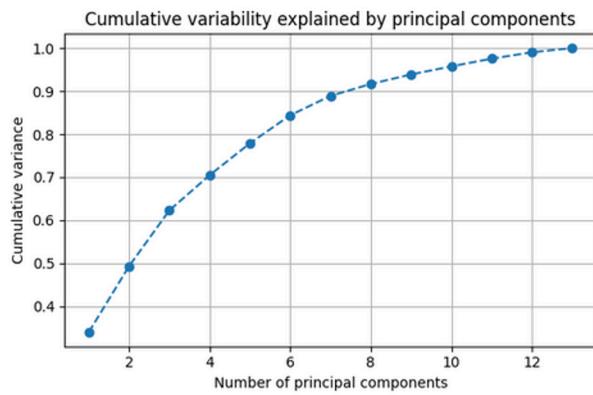
Za potrebe klasterovanja izvršićemo još malo sređivanje baze podataka.

1. Kako bismo jasnije videli kako se promenilo radno vreme pojedinca, napravićemo novi atribut koji će upravo da predstavlja razliku između radnog vremena nakon i pre korone i obrisaćemo pojedinačne kolone.
2. Obrisaćemo kolonu 'dislike_hw' jer jasno je da ako neko da visoku ocenu radu od kuće, odnosno 'like_hw', da će dati nisku ocenu 'dislike_hw' i obrnuto
3. Matrica korelacije:



Primećujemo da atributi 'age', 'gender' i 'travel_time' ne utiču previše na ostale attribute, pa ćemo i ove kolone obrisati radi malo lakšeg klasterovanja.

4. Smanjenjem dimenzionalnosti često se poboljšava efikasnost procesa klasterovanja, što rezultuje bržim izračunavanjima i boljim performansama modela.

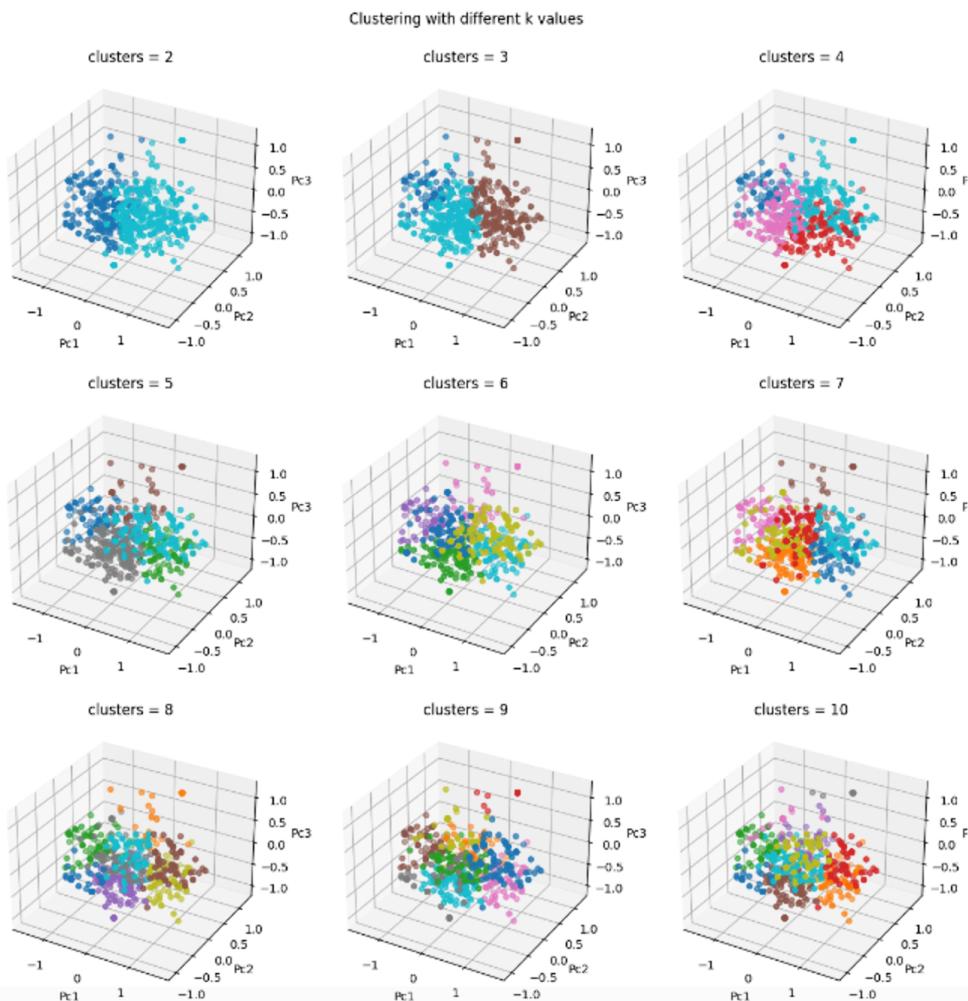


Podatke smo smanjili na 3 dimenzije. Iako je zadržano samo nešto više od 60% varijansi skupa podataka, testiranjem i sa većim brojem atributa dobijeni su slični rezultati kao ovi, pa radi jednostavnosti i lakše vizuelizacije opredelili smo se za ovaj broj.

K-SREDINA

K-means

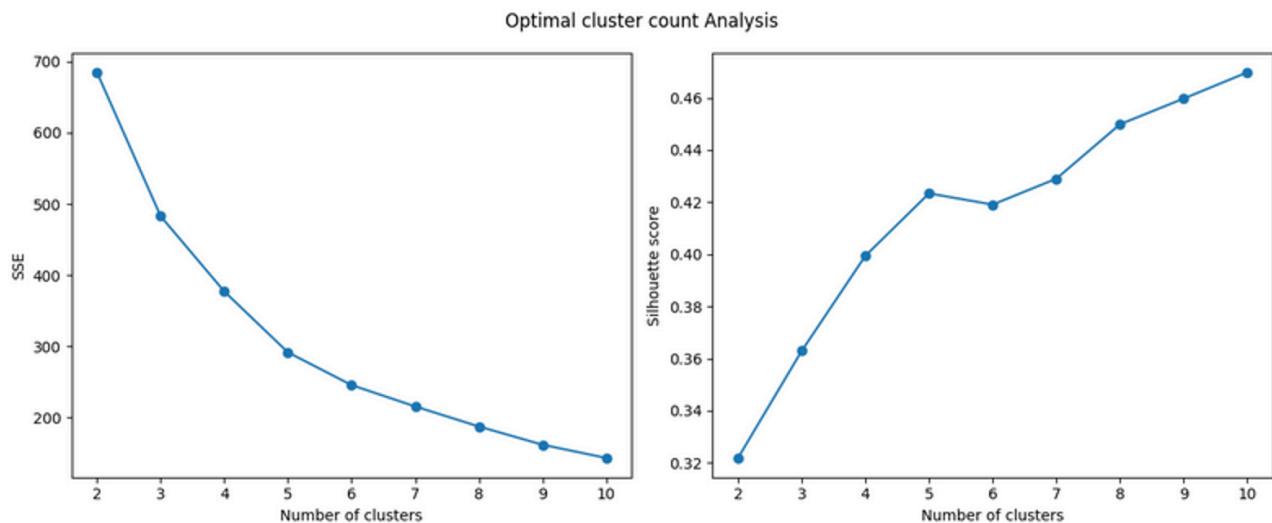
K-sredina je jednostavan algoritam za klasterovanje. Glavna svrha ovog algoritma je grupisanje podataka u k klastera, gde je k unapred određen broj klastera.



Biranje optimalnog broj klastera

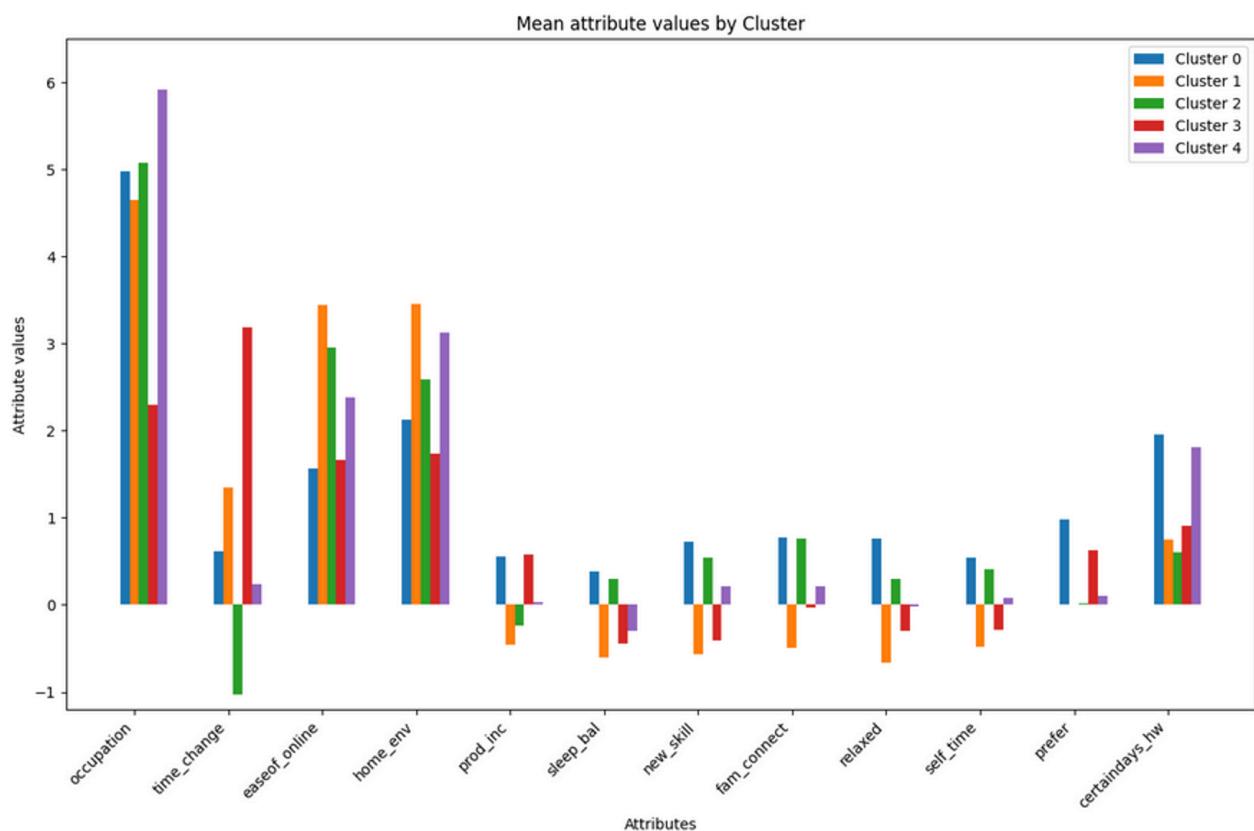
SSE (= Sum of Squared Errors) je metrika koja predstavlja udaljenost između svake instance u klasteru i centra tog klastera.

Silhouette sscore je metrika koja se koristi za merenje kvaliteta klasterovanja. Ova metrika pruža informacije o tome koliko su instance unutar klastera slične međusobno u poređenju sainstancama u drugim klasterima.

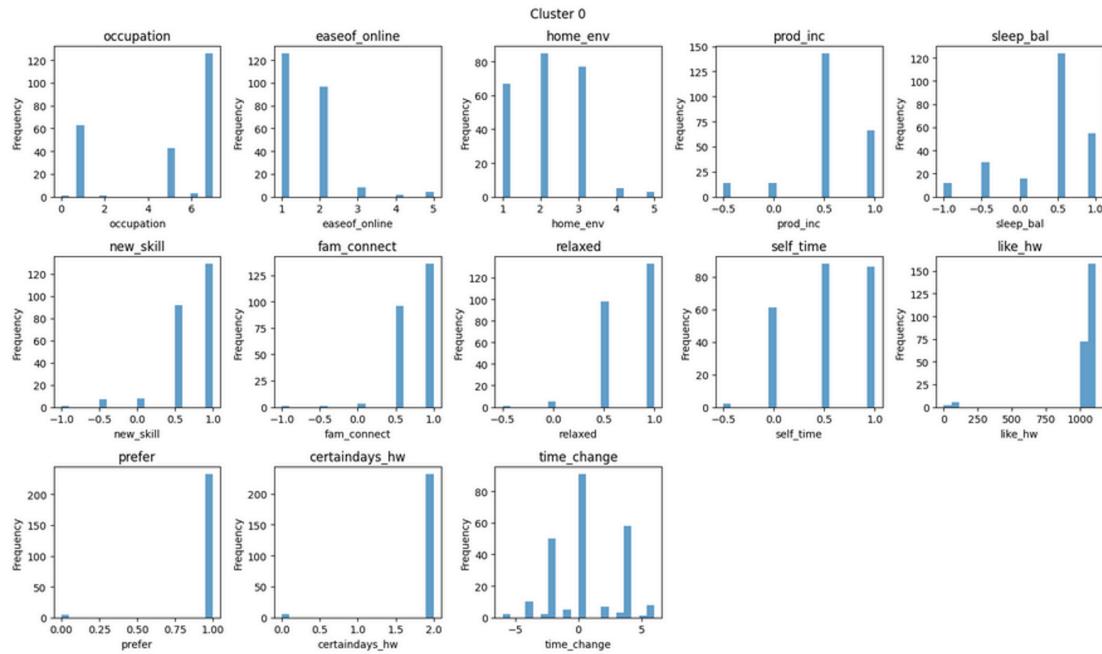


Na osnovu ovih rezultata, model sa 5 klastera pruža najbolje rezultate u smislu smanjenja inercije (Sum of Squared Errors) i povećanja vrednosti siluete.

Karakteristike klastera



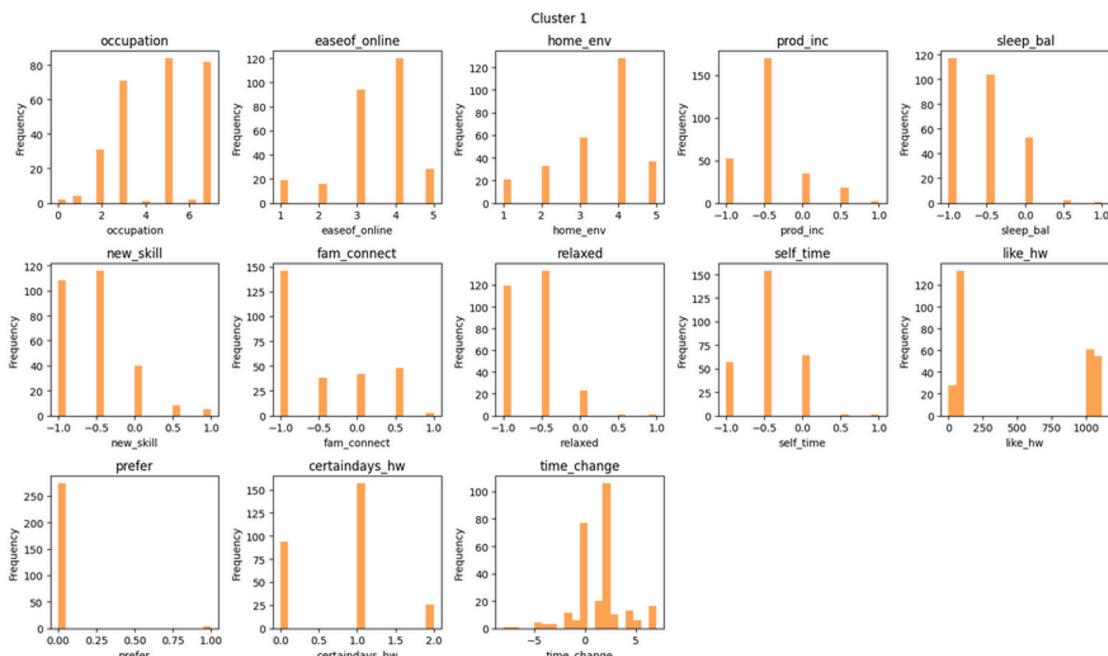
Klaster 0:



U ovom klasteru nalaze se pojedinci koji su prihvatali promene tokom pandemije. Njima je rad od kuće dobro došao, lako su se prilagodili radu na mreži. Većina njih je povećala produktivnost i bolje uskladila režim spavanja. Ovo vreme su takođe iskoristili za povezivanje sa porodicom, sticanjem novih veština.

Sve u svemu, ovo je grupa koja je sigurno doživela pozitivne promene.

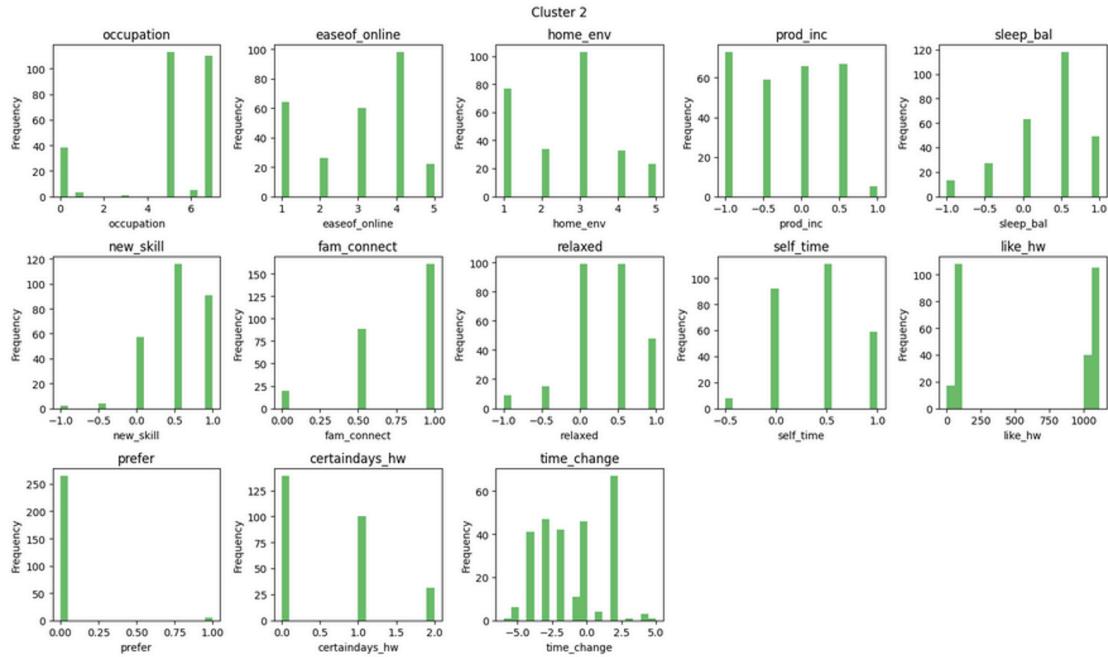
Klaster 1:



Pripadnici ovog klastera su skloniji negativnijem pristupu promenama u pandemiji. Ova grupa ni u čemu nije napredovala, čak je mnogo i nazadovala. Oni više vole rad iz kancelarije, što je direktna posledica smanjenja njihove produktivnosti. Poremetili su režim spavanja, bili su dosta pod sresom, nisu se posvetili sebi. Zanimljiva činjenica je da se ovde nalaze dosta studenata.

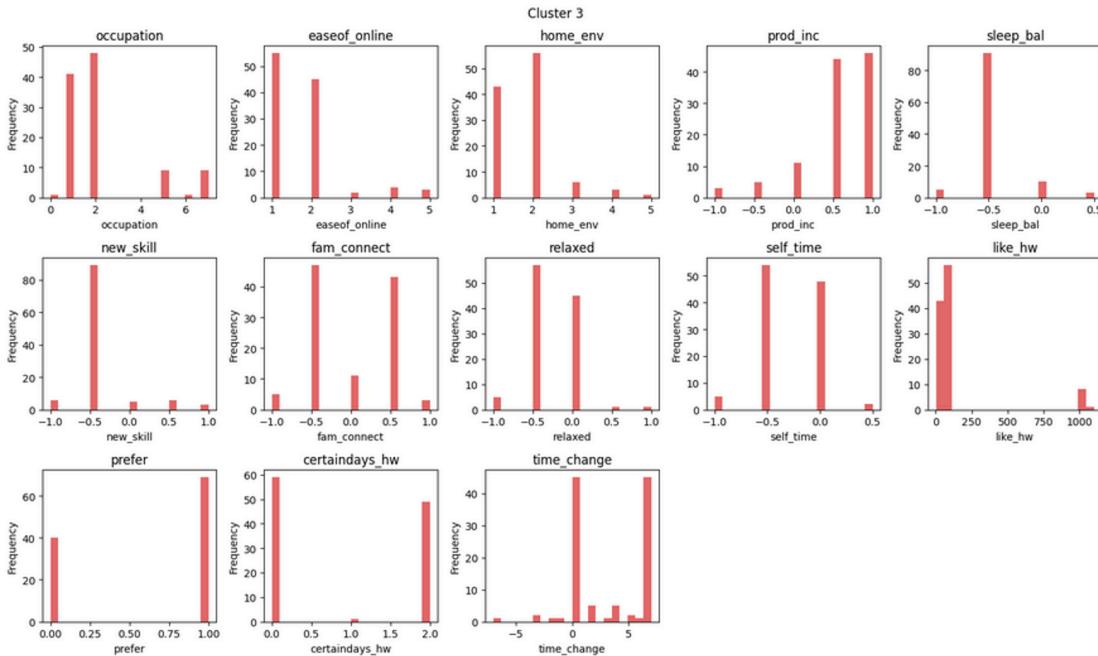
Ovo je grupa ljudi na koju je pandemija dosta loše uticala.

Klaster 2:



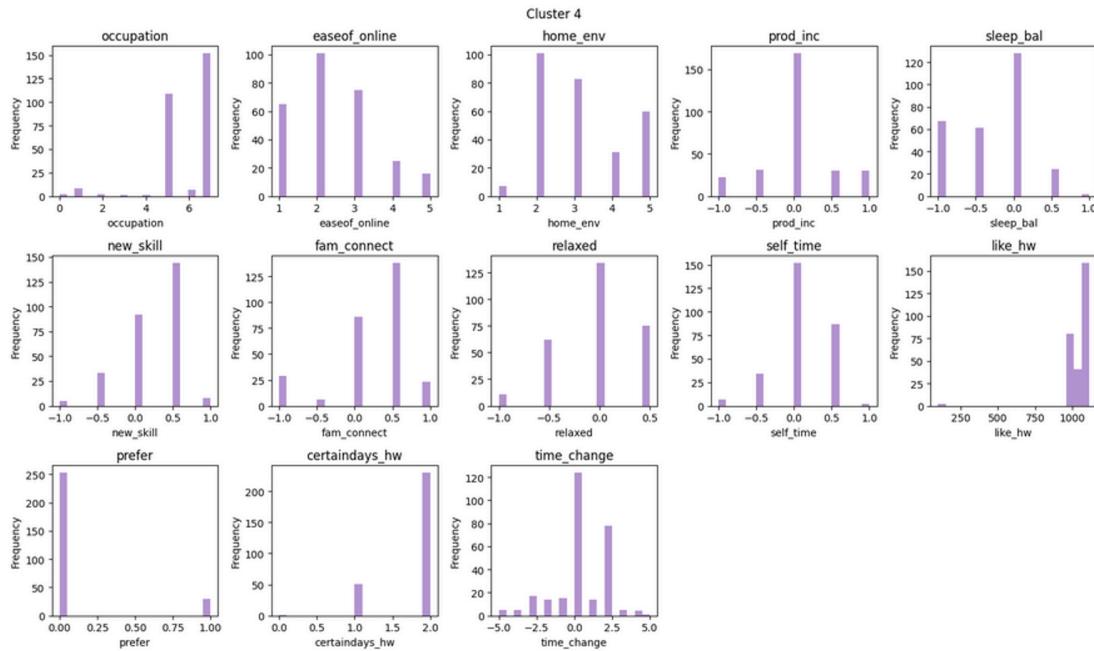
Ovaj klaster obuhvata osobe koje su se lako prilagodile situaciji. lako više vole rad iz kancelarije i nisu se najbolje prilagodili radu online, iskoristili su priliku da rade na sebi i posvete se sebi.

Klaster 3:



Pripadnici ovog klastera su ljudi koji su najviše promenili svoje radno vreme - preduzetnici i domaćice, što je direktno uticalo na njihovu produktivnost. Svi ostali aspekti načina života su manje više ostali isti, s malim akcentom ka negativnim promenama.

Klaster 4:

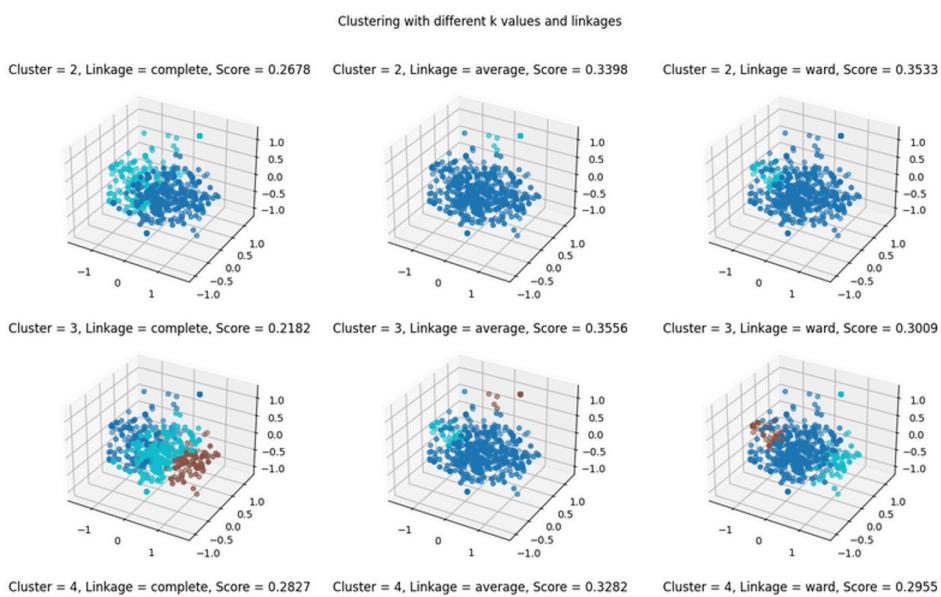


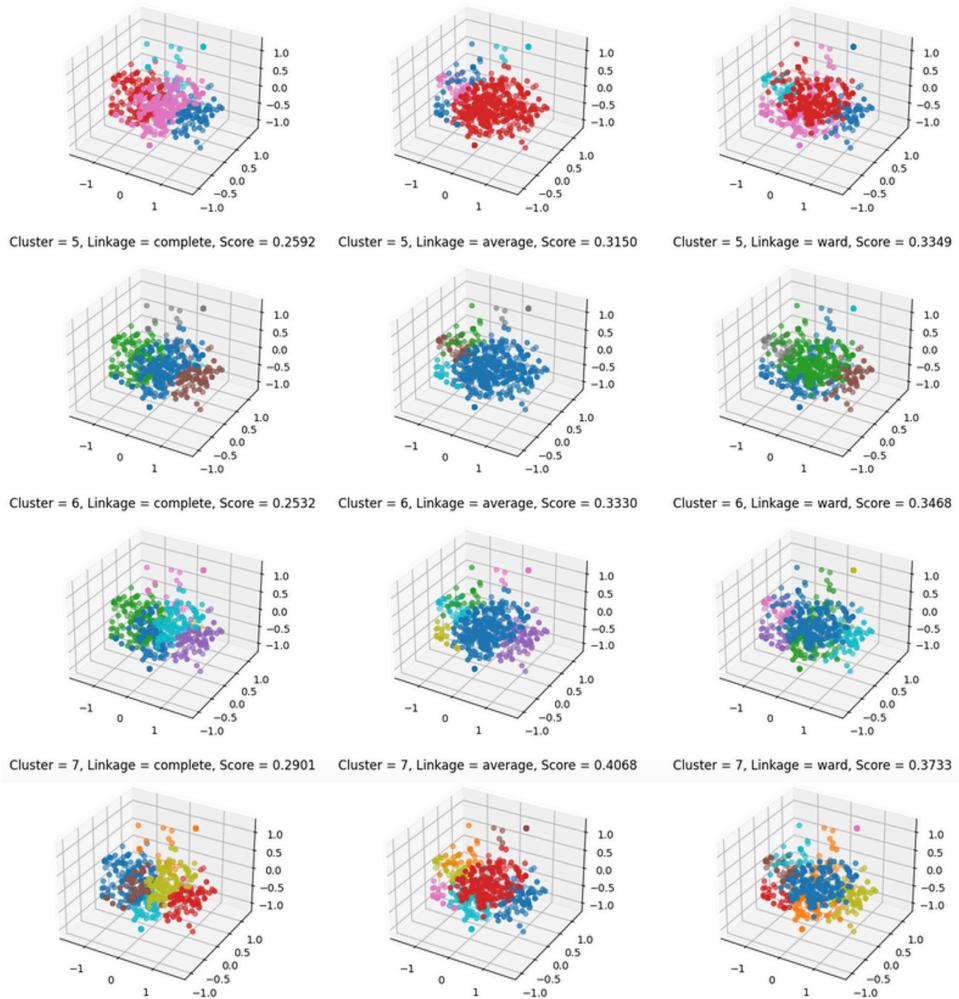
Ovaj klaster predstavlja pojedince koje nisu doživele nikakvu značajniju promenu tokom pandemije.

HIJERARHIJSKO KLASTEROVANJE

Agglomerative Clustering

Hijerarhijsko klasterovanje je tehnika klasterovanja podataka koja organizuje podatke u hijerarhijsku strukturu ili drvo. Ova metoda omogućava vizuelizaciju podataka na način koji olakšava razumevanje odnosa između različitih grupa ili klastera.





Najbolji model:

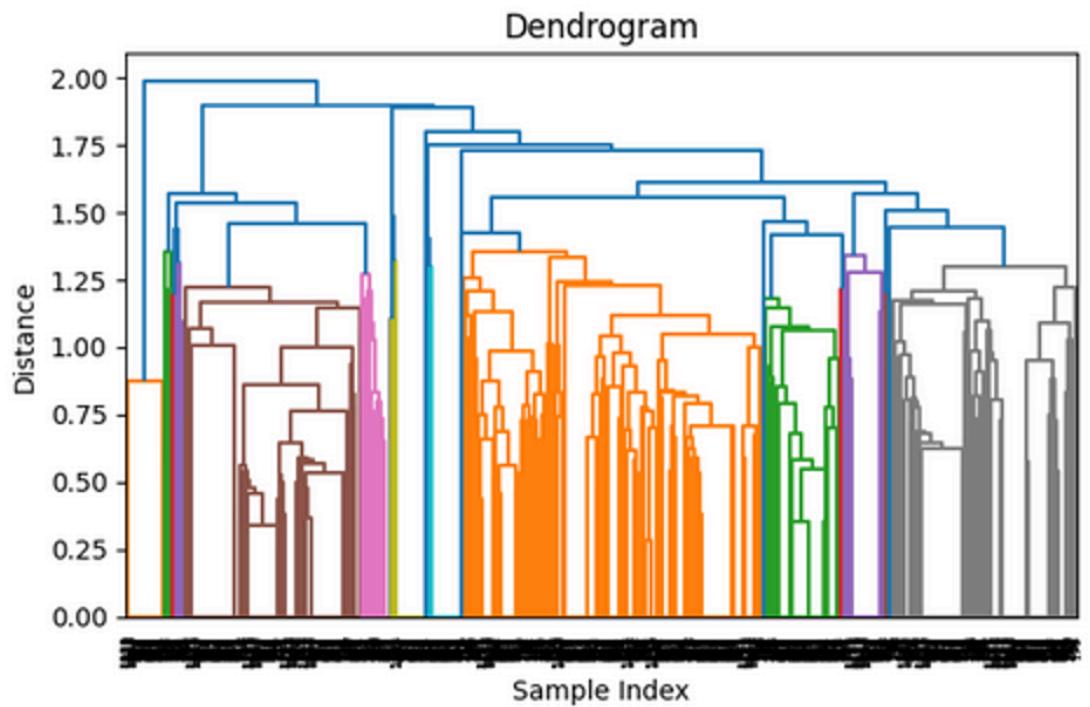
```
1 print(f'score, num of clusters, linkage: {best_score:.3f}, {best_model.n_clusters_}, {best_model.linkage}')
```

score, num of clusters, linkage: 0.407, 7, average

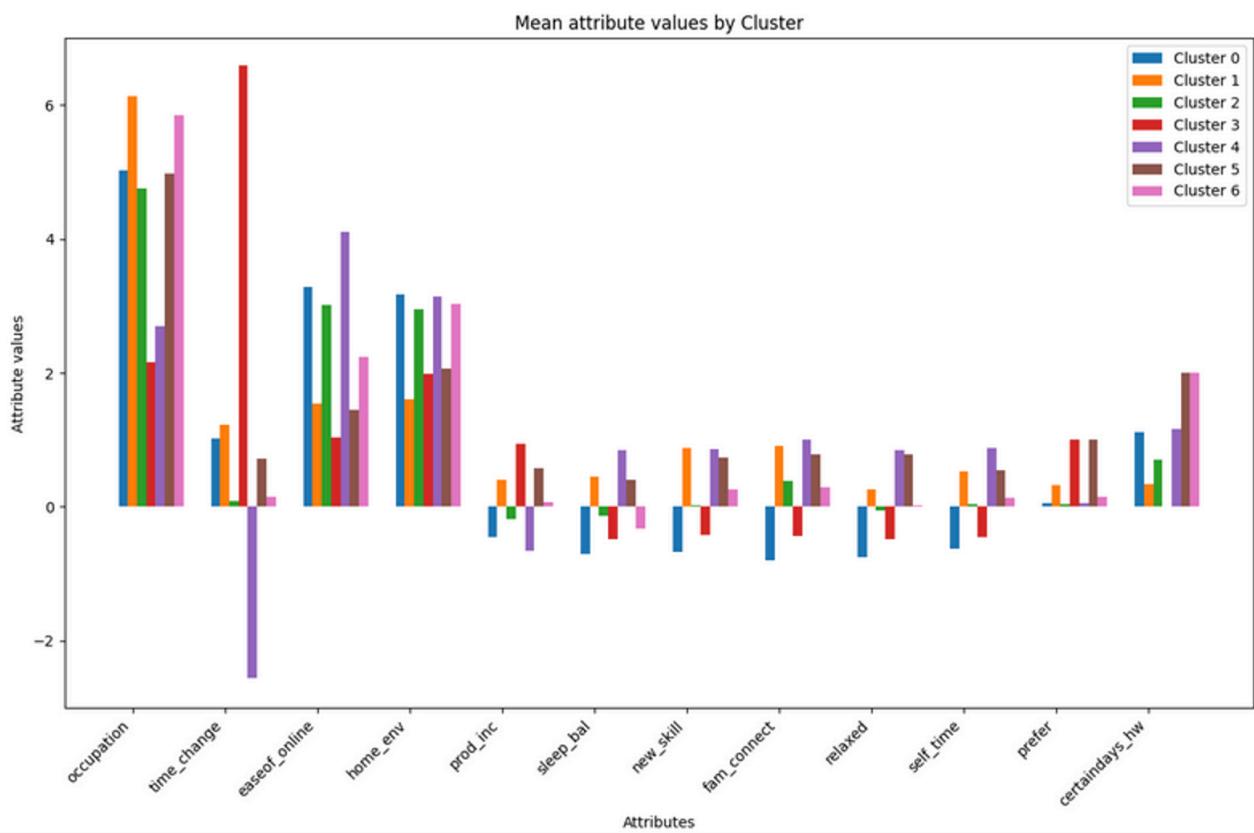
Vizuelna reprezentacija

Dendogram nam služi za vizuelnu reprezentaciju hijerarhijske strukture klastera. To je dijagram u obliku drveta koji prikazuje kako su podaci grupirani u različite grupe ili klastere. U dendrogramu, svaki čvor predstavlja jedan ili više podataka, a grane prikazuju način na koji su podaci grupisani zajedno.

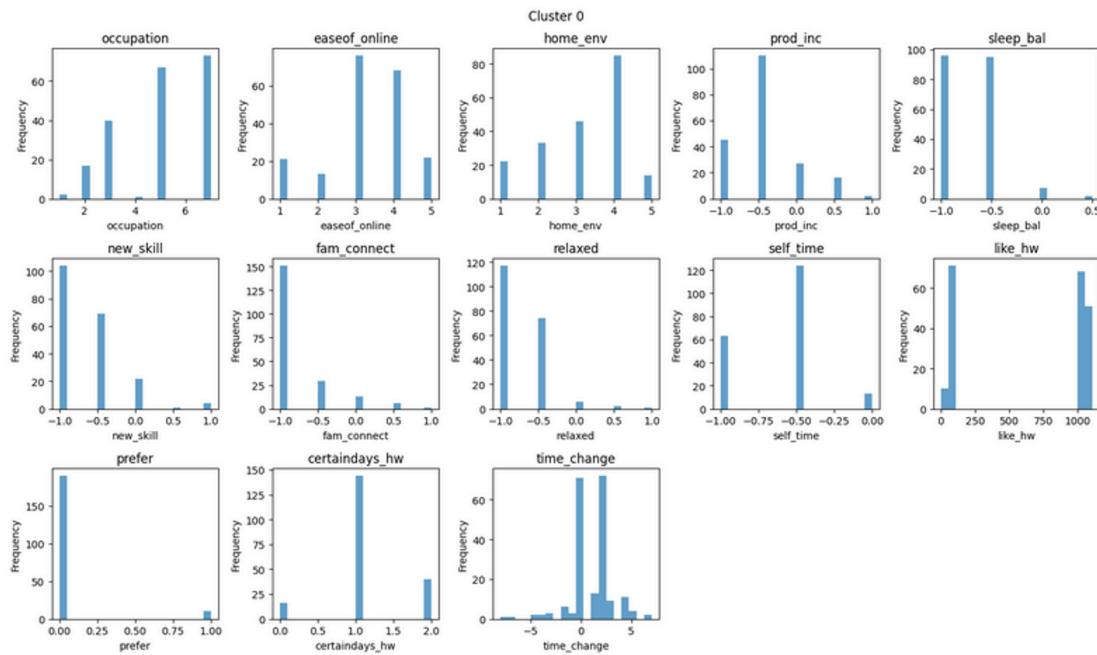
Ukratko, dendrogram se koristi za vizualno razumevanje sličnosti između podataka i kako se podaci grupišu na različitim nivoima hijerarhije.



Karakteristike klastera

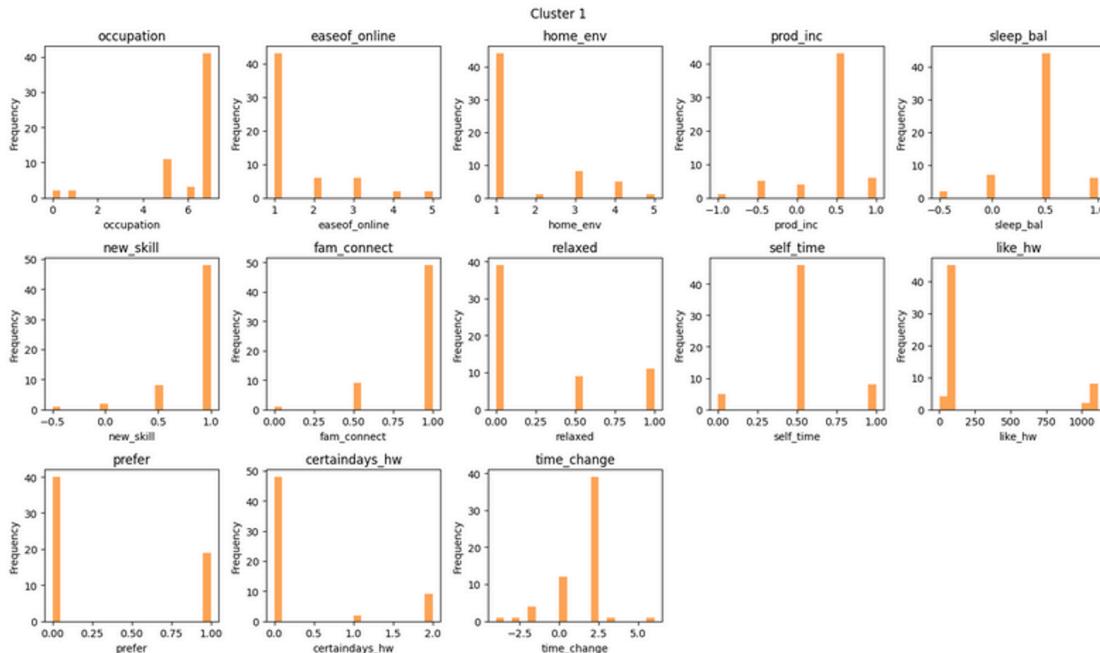


Klaster 0:



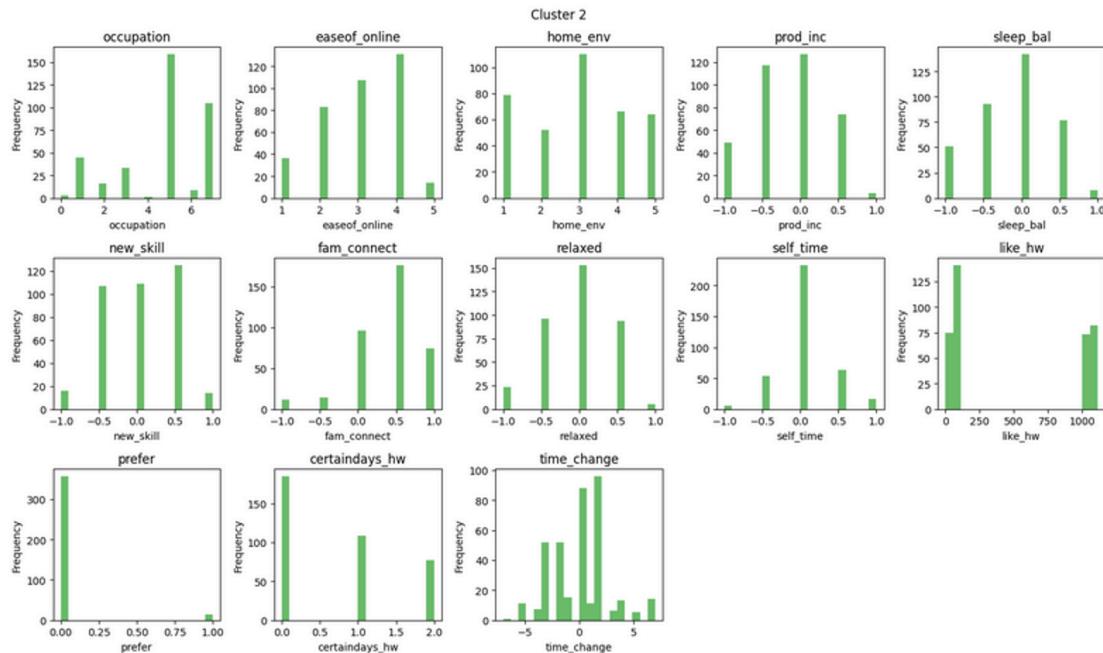
Pripadnici ovog klastera su dali niže ocene za opuštenost, što ukazuje na potencijalno veći nivo stresa i napetosti. Takođe, njihove ocene za nove veštine, produktivnost, spavanje, povezanost sa porodicom su niske. Ovo je grupa ljudi koja je definitivno doživela negativne promene tokom pandemije.

Klaster 1:



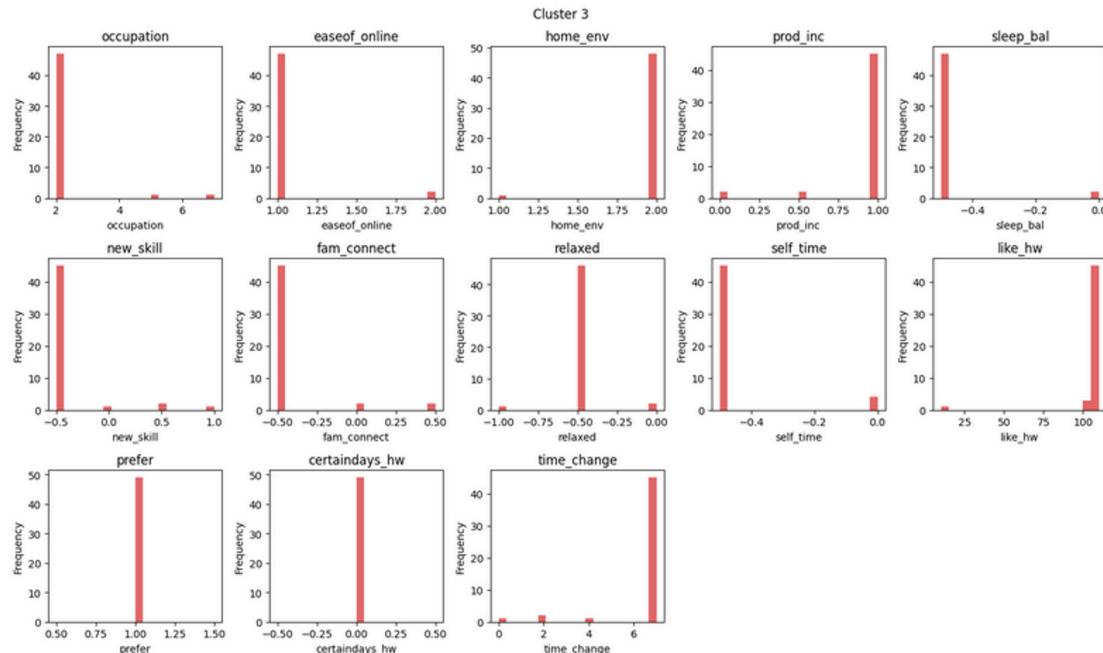
Članovi ovog klastera su se lako prilagodili radu online, sticali nove veštine, povezivali se sa porodicom,... što sugerira da su njihovo bolje psihološko blagostanje.

Klaster 2:



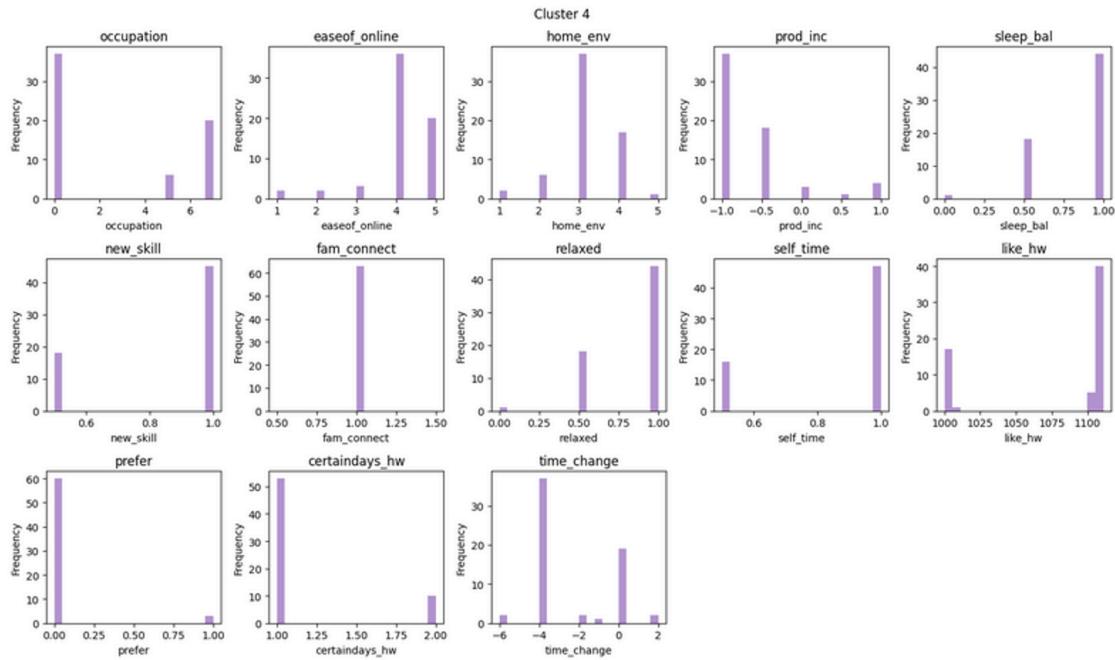
Pripadnici ovog klastera nisu doživeli neku značajniju promenu tokom pandemije.

Klaster 3:



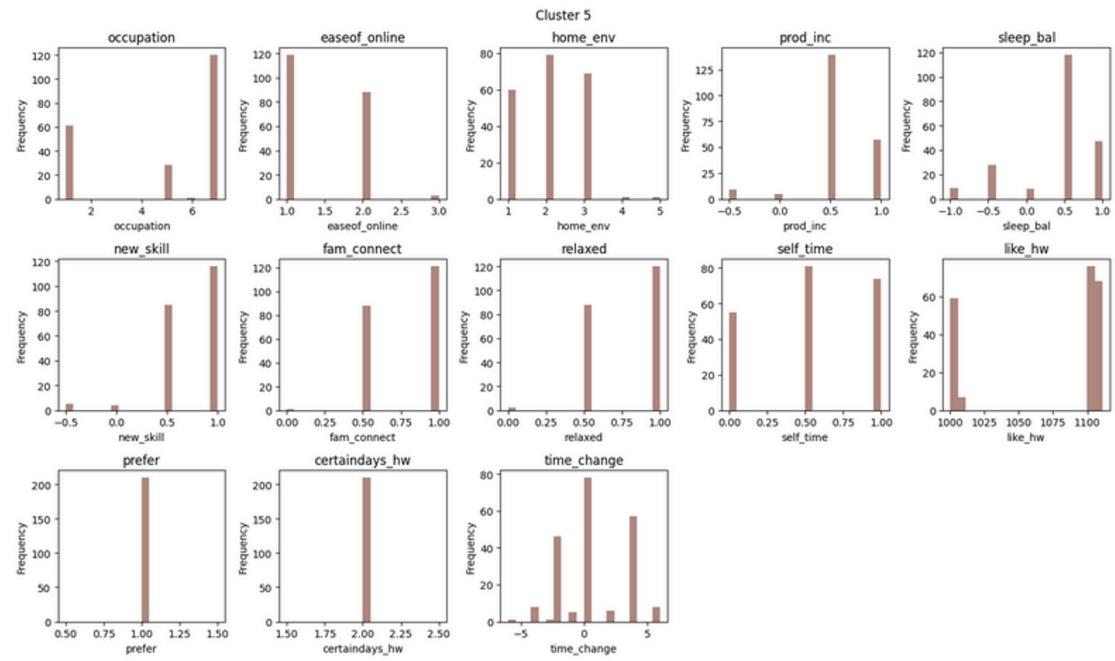
Pripadnice ovog klastera su domaćice koje su najviše promenile svoje radno vreme, što je direktno uticalo na produktivnost. Međutim, u ostalim apsektima života su doživele blage negativne promene.

Klaster 4:



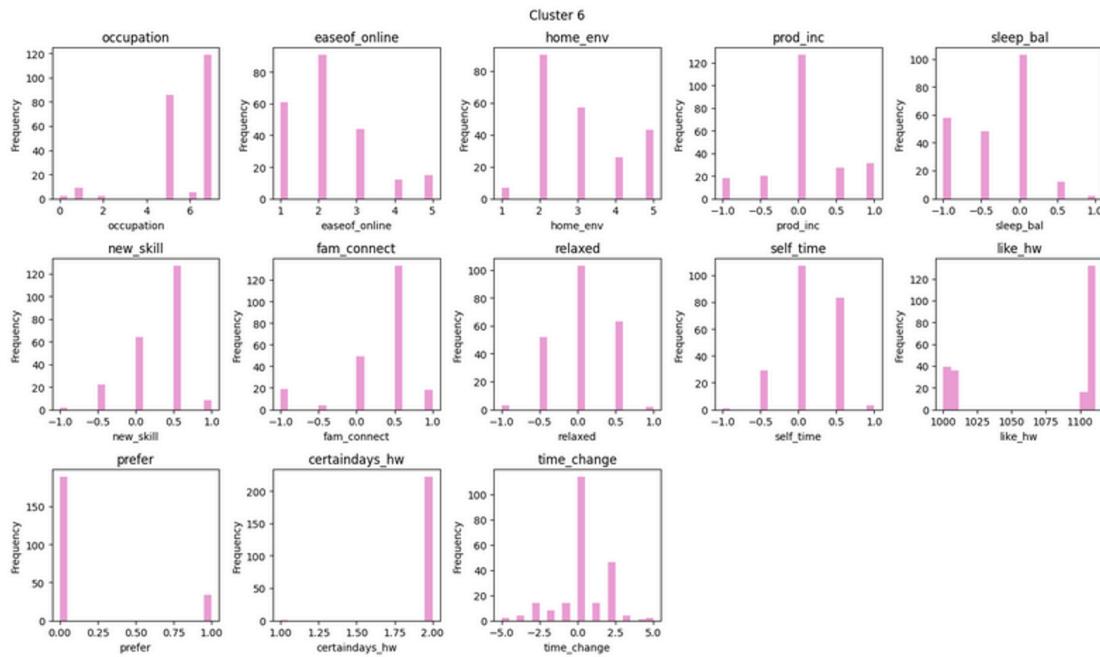
Ovaj klaster obuhvata pojedince koji su tokom pandemije ostali bez posla, što je u direktnoj vezi sa smanjenjem njihove produktivnosti. Međutim, to su dobro iskoristili za rad na sebi, zблиžavanje s porodicom,...

Klaster 5:



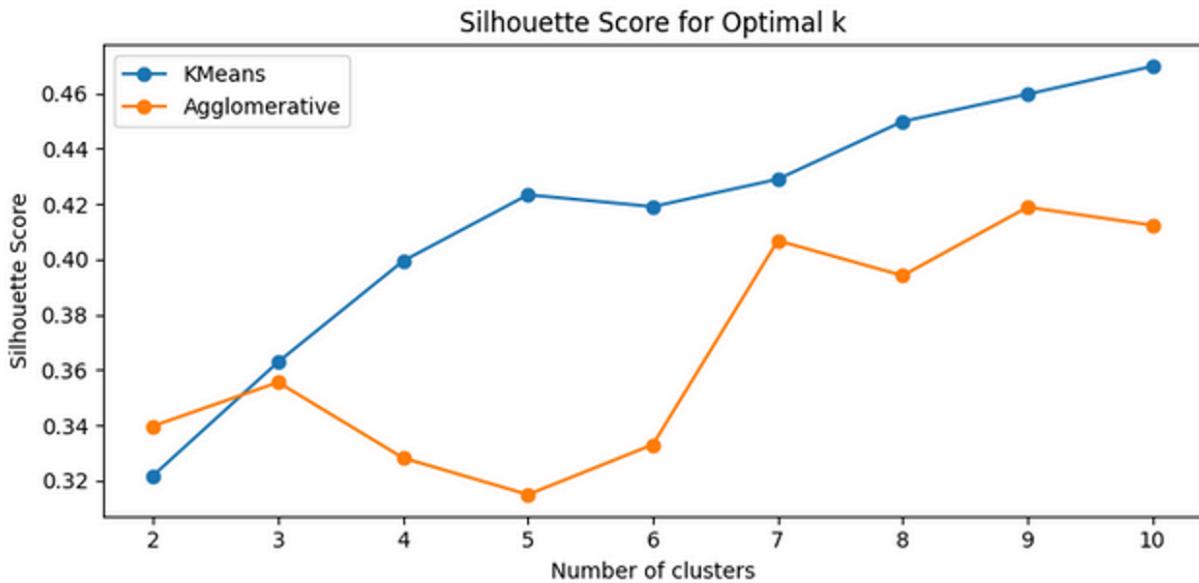
Članovi ovog klastera često daju više ocene za online aktivnosti i sticanje novih veština, što ukazuje na otvorenost za promenе и учење. Takođe su opušteni i skloni druženju.

Klaster 6:



Ova grupa pojedinaca je doživela blage pozitivne promene tokom pandemije.

POREĐENJE MODELA KLASTEROVANJA



Na osnovu analize silhouette skorova, možemo zaključiti da je model K-Means verovatno bolji od modela Agglomerative Clustering za klasterizaciju ovih podataka.

PRAVILA PRIDRUŽIVANJA

Pravila pridruživanja, u kontekstu analize i istraživanja podataka, predstavljaju formalne izraze koji identifikuju odnose, uzorke ili asocijaciju između atributa i skupa podataka. Ova pravila obično imaju formu "Ako se javi određeni uslov u atributima A, B, C,... onda se može očekivati ili predvideti određena vrednost ili događaj u atributu X".

Pokušala sam identifikovati pravila koja bi otkrila kako promena načina života utiče na psihološko stanje pojedinaca, međutim, nisam uspela pronaći nijedno takvo pravilo. Ovo može ukazivati na kompleksnost veza između promena načina života i psihološkog stanja, koje možda ne može biti jasno izraženo jednostavnim pravilima prodrživanja. Zbog toga, odlučila sam se da vidim kako su promene na psihološkom nivou uticale na različite grupe.

Pravila koje sam tako dobila su:

Consequent	Antecedent	Support %	Confidence %	Lift
age = 50-60	relaxed = -0.5 sleep_bal = -0.5 lifestyle_change = -1	10.128	82.353	5.692
age = 50-60	new_skill = -0.5 self_time = -0.5 relaxed = -0.5 travel_time = 0.5	10.043	82.203	5.682
age = 50-60	new_skill = -0.5 self_time = -0.5 lifestyle_change = -1 travel_time = 0.5	10.128	81.513	5.634
age = 50-60	new_skill = -0.5 self_time = -0.5 relaxed = -0.5 lifestyle_change = -1	10.298	80.992	5.598
age = 50-60	new_skill = -0.5 self_time = -0.5 travel_time = 0.5	10.298	80.165	5.541
occupation = Working Pro...	relaxed = 0.5 self_time = 0.5 new_skill = 0.5 time_change = 0	10.638	94.4	2.316
occupation = Working Pro...	relaxed = 0.5 self_time = 0.5 new_skill = 0.5 time_change = 0 lifestyle_change = 1	10.553	94.355	2.315
occupation = Working Pro...	self_time = 0.5	***		

Najjače pravilo je da je pandemija negativno uticala na osobe između 50 i 60 godina i pozitivno uticala na osobe čije je zanimanje označeno kao 'Working Professional'.

ZAKLJUČAK

U sklopu ovog istraživanja analizirani su različiti aspekti psiholoških promena kod pojedinaca tokom pandemije, uzimajući u obzir različite faktore, kao što su pol, godine, zanimanje, produktivnost, slobodno vreme,...

Iz analize podataka možemo zaključiti da su psihološke promene jako složene i nije ih lako predvideti. Različiti faktori igraju ključnu ulogu u oblikovanju psihološkog stanja pojedinca, a njihove međusobne veze su veoma kompleksne.

Najvažniji faktori koji su se izdvojili su: koliko lako su se ljudi prilagodili online radu, produktivnost, kvalitet spavanja i odnosi sa porodicom. Osobe koje su dale visoke ocene ovim atributima, češće su bili povezani sa boljim psihološkim stanjem.

Osim toga, analiza je pokazala da različite ciljne grupe (starosne grupe, zanimanje, pol) mogu doživeti različite psihološke promene u istim uslovima i da sve to zavisi od specifičnih karakteristika pojedinaca.

Ovo istraživanje o psihološkim promenama tokom pandemije naglašava kompleksnost ovog fenomena. Promene u načinu života tokom pandemije ne mogu se jednostavno svesti na jedno pravilo ili obrazac. Različiti faktori i specifične karakteristike osobe igraju ključnu ulogu u oblikovanju psihološkog stanja pojedinca.

LITERATURA

1. <http://poincare.matf.bg.ac.rs/~nenad.mitic/ip1.html>
2. https://github.com/MATF-istrazivanje-podataka-1/materijali_2022-2023
3. <https://www.kaggle.com/datasets/hemanthhari/psychological-effects-of-covid>
4. <https://www.kaggle.com/code/mansi0123/eda-on-psychological-effect-of-covid/notebook>