Prikupljanje i predobrada podataka za mašinsko učenje

STUDENT:

MINA NIKOLIĆ 1540

MENTOR:

DOC. DR ALEKSANDAR STANIMIROVIĆ

Sadržaj

- Uvod
- Kvalitet podataka i mere za obradu podataka
- Definisanje pojma izbora atributa
- Traženje podskupa atributa
- Tipovi pristupa za proces izbora atributa
- Određivanje težine atributa
- Statističke metode za selekciju atributa na osnovu tipa podataka
- Interpretabilnost izbora atributa korišćenjem Explainable AI alata
- Zaključak

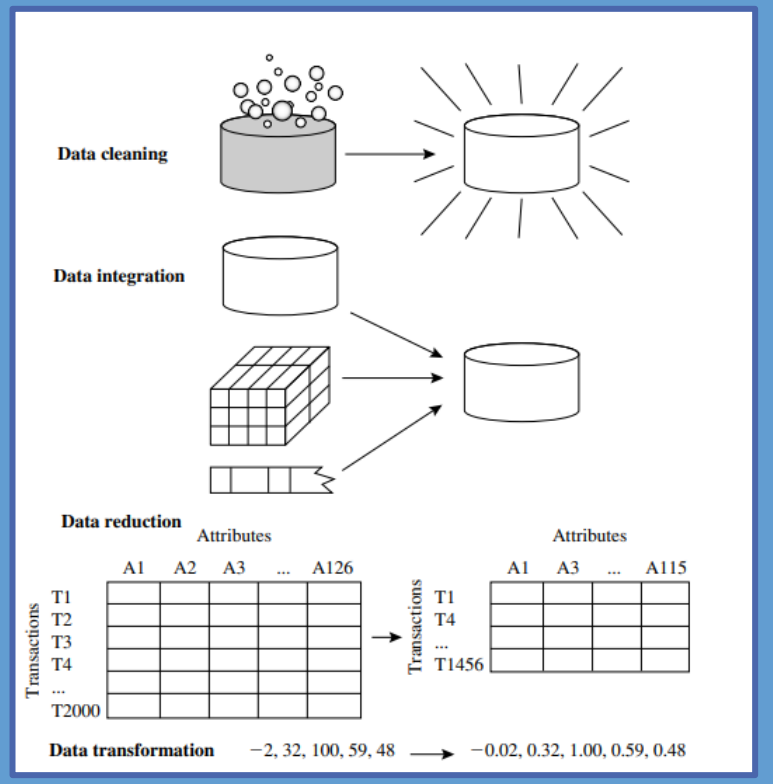
Uvod

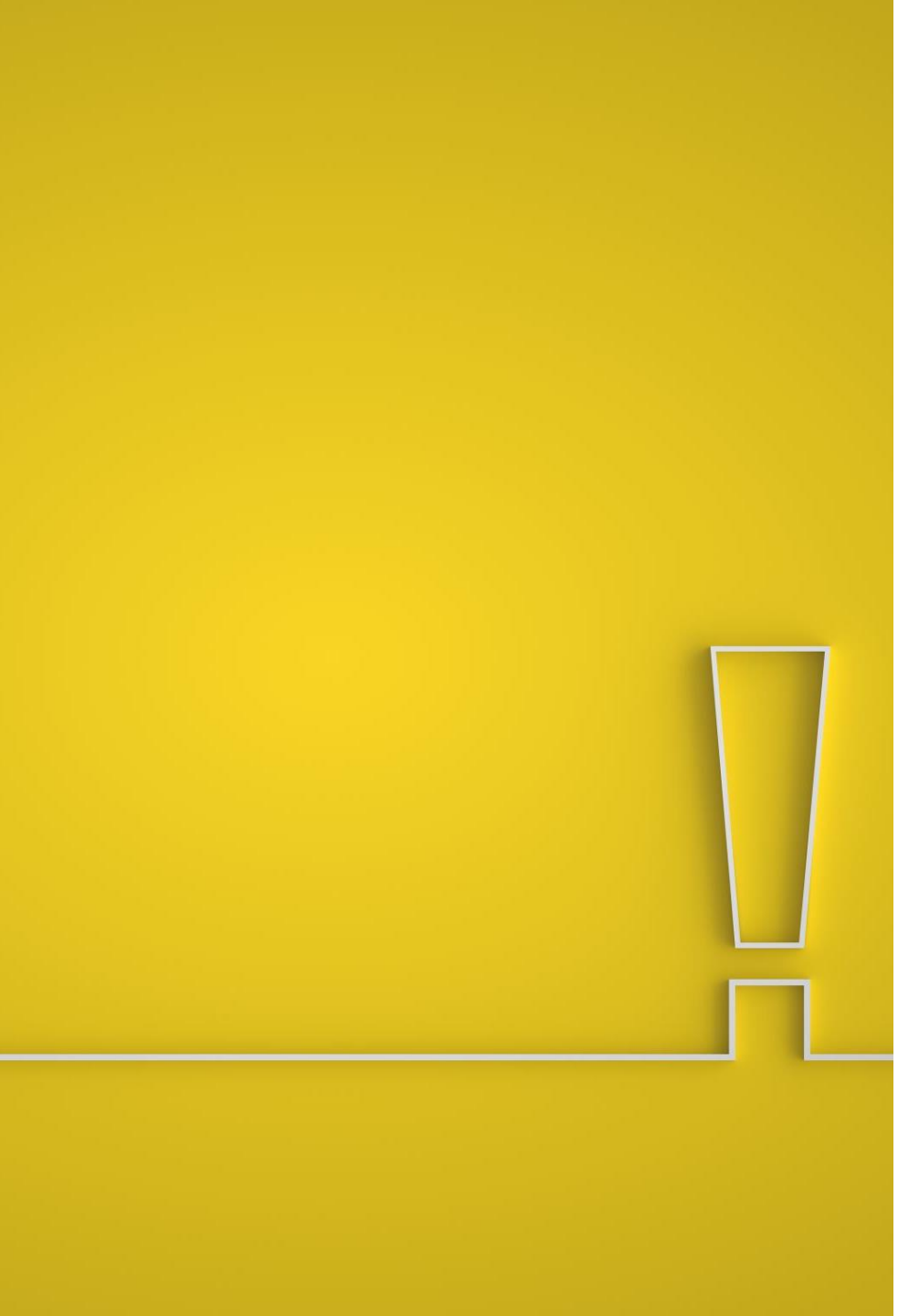
- Tendencija sve masovnije primene algoritama mašinskog učenja i veštačke inteligencije je dovela do razvoja naprednijih i efikasnijih načina na koje je moguće vršiti obradu podataka.
- Fokus istraživačkog rada u oblasti mašinskog učenja je posvećen kreiranju novih i modifikaciji postojećih algoritama, ali i na adekvatnoj pripremi podataka nad kojima će biti vršena obrada.
- Podaci dobijeni iz realnih izvora podataka se retko nalaze u obliku koji je pogodan za direktnu primenu algoritama mašinskog učenja.
- Da bi primena bila moguća, potrebno je izvršiti adekvatnu predobradu datih podataka.

Uvod

Tehnike koje se koriste za predobradu podataka imaju za cilj poboljšanje kvaliteta podataka kroz njihovu:

- **Transformaciju** (normalizacija, skaliranje, ekodiranje kategoričkih atributa),
- Određivanje načina za rad sa **nedostajućim podacima**,
- **Redukciju dimenzionalnosti**,
- Rešavanje problema **šuma u podacima**,
- Izbor **instanci**,
- **Izbor atributa**.





Kvalitet podataka i mere za obradu podataka

- Potreba za predobradom podataka se javlja pre svega zbog prirode realnih podataka nad kojima je potrebno vršiti obradu.
- Podaci mogu biti mašinski generisani, ali i od strane ljudi.
- Faktori kojima se može vršiti procena kvaliteta podataka:
  - Tačnost (accuracy)
  - Kompletnost (completness)
  - Konzistentnost (consistency)
  - Pravovremenost (timeliness)
  - Uverljivost (believability)
  - Interpretabilnost (interpretability)

Definisanje pojma izbora atributa

- Rad sa velikim brojem atributa koji su nerelevantni ili redundantni može imati za posledicu usporavanje celokupnog procesa obrade podataka.
- Veliki broj algoritama mašinskog učenja funkcioniše primetno bolje ukoliko je dimenzionalnost odnosno broj atributa u okviru skupa podataka manji.
- Pogodnost rada sa manjim brojem atributa se ogleda i u mogućnosti lakšeg tumačenja modela, povećanja interpretabilnosti, ali i preglednije vizuelizacije.
- Izbor atributa se može definisati kao ***proces odabira optimalnog podskupa atributa koji zadovoljavaju određeni kriterijum***.

Definisanje pojma izbora atributa

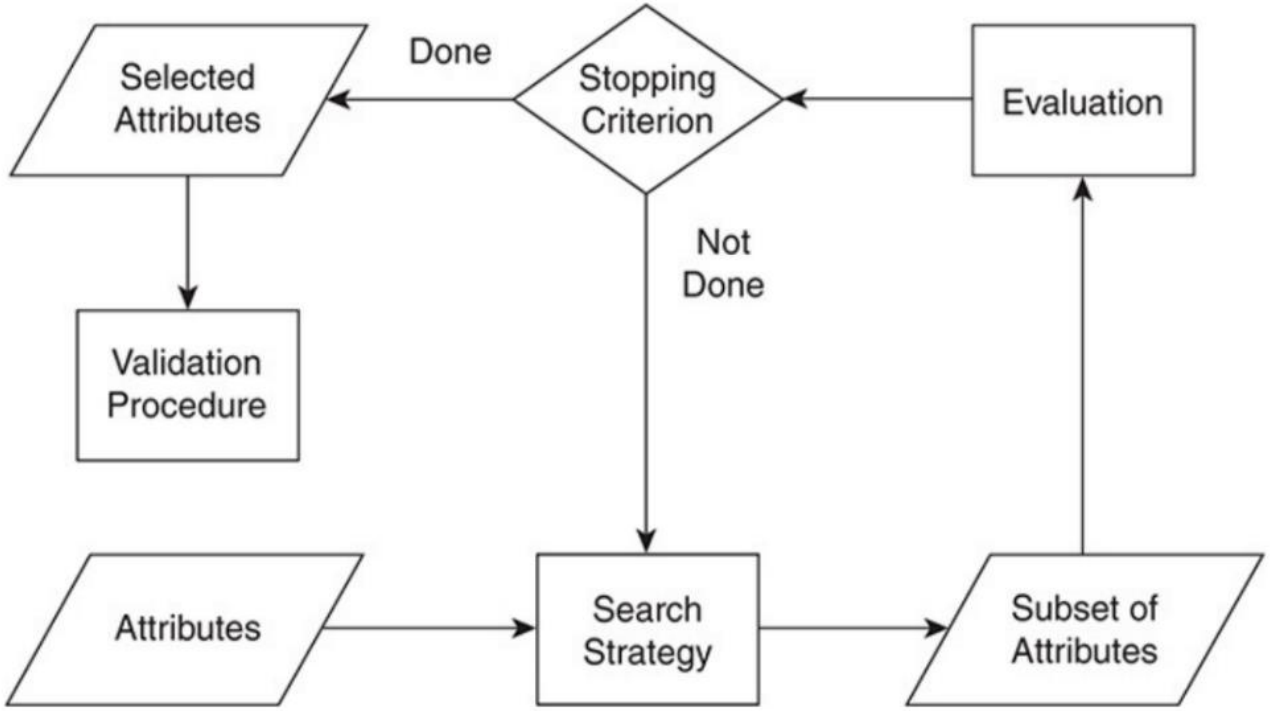
Razlozi zbog kojih se najčešće vrši izbor atributa:

- Odbacivanje nerelevantnih podataka,
- Povećanje tačnosti kod modela mašinskog učenja,
- Redukovanje troškova podataka,
- Poboljšanje efikasnosi učenja kroz smanjenje memorije neophodne za čuvanje podataka i obradu od strane računara,
- Redukovanje kompleksnosti dobijenih rezultata u cilju poboljšavanja razumljivosti modela i podataka odnosno njihove interpretabilnosti.

Sličnosti i razlike između izbora atributa i srodnih oblasti

Sa jedne strane se može smatrati da izbor atributa predstavlja podskup redukcije dimenzionalnosti, dok se sa druge može posmatrati i kao nezavisna celina.

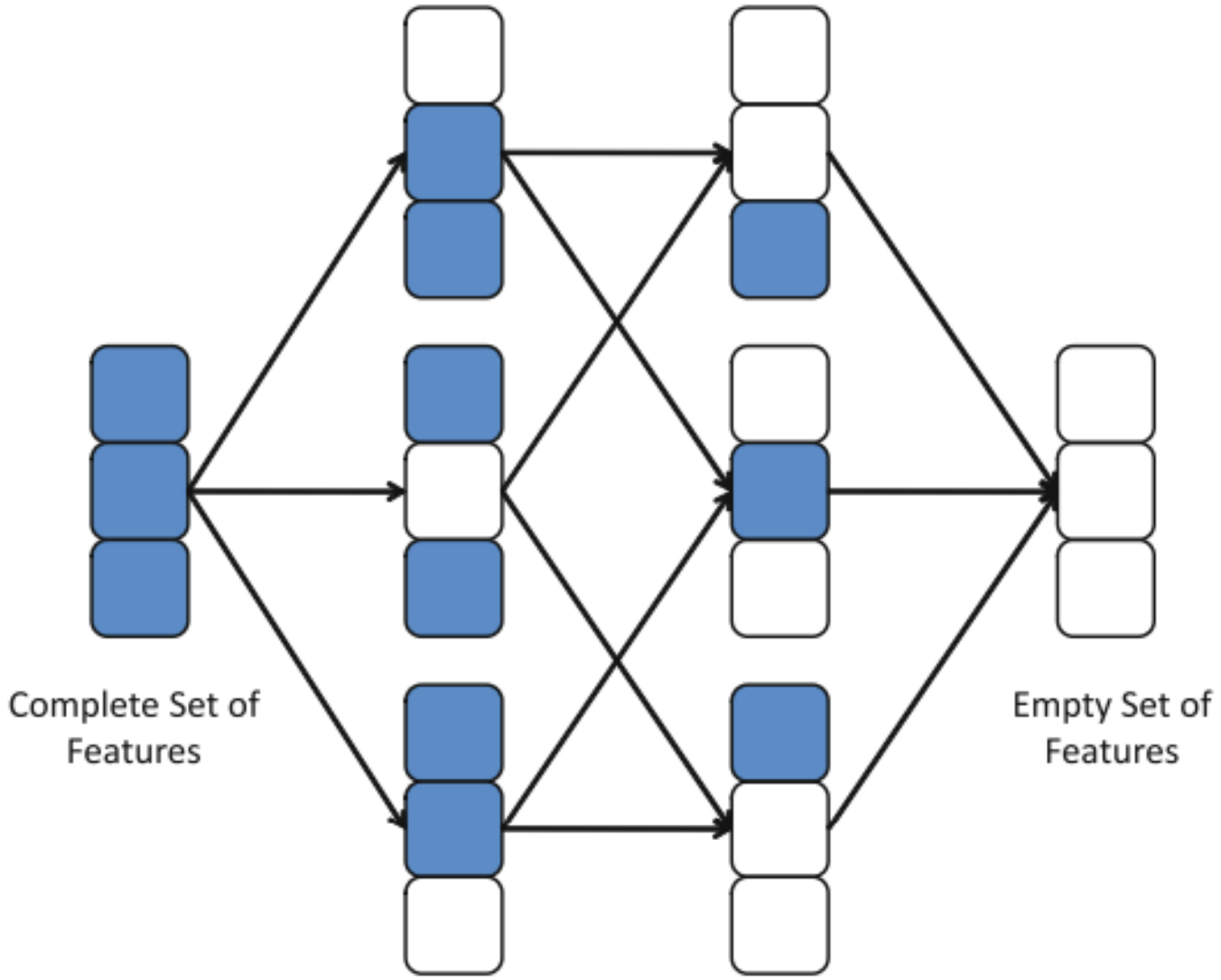
Glavna razlika između redukcije dimenzionalnosti i izbora atributa se ogleda u tome što izbor atributa ne podrazumeva izmenu atributa, već samo izbor najpogodnijih, dok se kod redukcije dimenzionalnosti iz postojećih atributa dobijaju novi.



Arhitektura za proces izbora atributa

Traženje podskupa atributa

- Problem izbora atributa se može posmatrati u okviru problema traženja, pri čemu svako stanje prostora traženja odgovara odabranom podskupu atributa.
- Ukupan broj podskupova će biti veličine 2^M , gde je M broj atributa u okviru seta podataka.
- Realni setovi podataka uglavnom poseduju veliki broj atributa, stoga se proces traženja retko počinje od celokupnog broja atributa pre svega zbog ogromnih troškova procesiranja.



Traženje podskupa atributa

Proces traženja atributa se može posmatrati na različite načine, pri čemu se uzimaju u razmatranje:

- Smer traženja (sekvencijalno generisanje u napred, sekvencijalno generisanje unazad, bidirekciono traženje i nasumično traženje),
- Strategije traženja (strategije grubog traženja, sveobuhvatnog traženja, heurističke pretrage i nedeterminističke pretrage),
- Kriterijumi za selekciju (informacione mere, metrike za distancu, metrike zavisnosti, metrike konzistentnosti, kao i metrike za tačnost) i
- Filter, Wrapper i Embedded metode za izbor atributa.

Traženje podskupa atributa zasnovano na smeru traženja

Prema smeru traženja, podela se može izvršiti na:

- **Sekvencijalno generisanje unapred** (Sequential Forward Generation), pri čemu pretraga počinje sa praznim skupom atributa. Daljim radom algoritma dolazi do dodavanja novih atributa u skup.
- **Sekvencijalno generisanje unazad** (Sequential Backward Generation), gde proces pretrage kreće sa celokupnim skupom atributa, pri čemu se u toku izvršavanja eliminišu jedan po jedan atribut.
- **Bidirekciono generisanje** (Bidirectional Generation), pri čemu se vrši kombinovanje sekvencijalnog generisanja unapred i unazad.
- **Nasumično generisanje** (Random Generation), koje podrazumeva da ne postoji prethodno definisani smer.

Traženje podskupa atributa zasnovano na strategiji pretrage

- Prilikom procesa pretrage generalno važi tvrdnja da što se više resursa potroši, to je veća verovatnoća da se dobiju bolji rezultati, odnosno korisniji atributi.
- Prema tome, postoji nekoliko različitih algoritama koji se koriste tokom procesa pretrage i mogu se podeliti na:
  - **Sveobuhvatne pretrage** (eng. Exhaustive Search), pri čemu se dato rešenje ogleda u nalaženju svih mogućih podskupova rešenja.
  - **Heurističke pretrage** (eng. Heuristic Search), gde se za sam proces pretrage koristi neka heuristika.
  - **Nedeterminističke pretrage** (eng. Nondeterminictic search) koja je kombinacija prethodna dva pristupa.

Traženje podskupa atributa zasnovano na izboru kriterijuma selekcije

- U okviru domena izbora atributa, pored pomenutih metoda potrebno je naći i načine za određivanje njihovog kvaliteta. Međutim, kriterijumi selekcije koje je potrebno odabrati pre svega zavise od namene za koju se vrši predobrada i analiza podataka.
- Kriterijumi selekcije će biti posmatrani kroz sledeće celine:
  - Informacione mere (eng. Information Measures)
  - Mere rastojanja (eng. Distance Measures)
  - Mere zavisnosti (eng. Dependance Measures)
  - Mere konzistentnosti (eng. Consistency Measures)
  - Mere tačnosti (eng. Accuracy measures)

Tipovi pristupa za proces izbora atributa

Kada je reč o različitim pristupima za proces izbora atributa, svi oni se pre svega mogu klasifikovati u odnosu na to da li pripadaju kategoriji nadgledanih ili nenadgledanih algoritama mašinskog učenja.

Različiti tipovi pristupa će biti prikazani kroz:

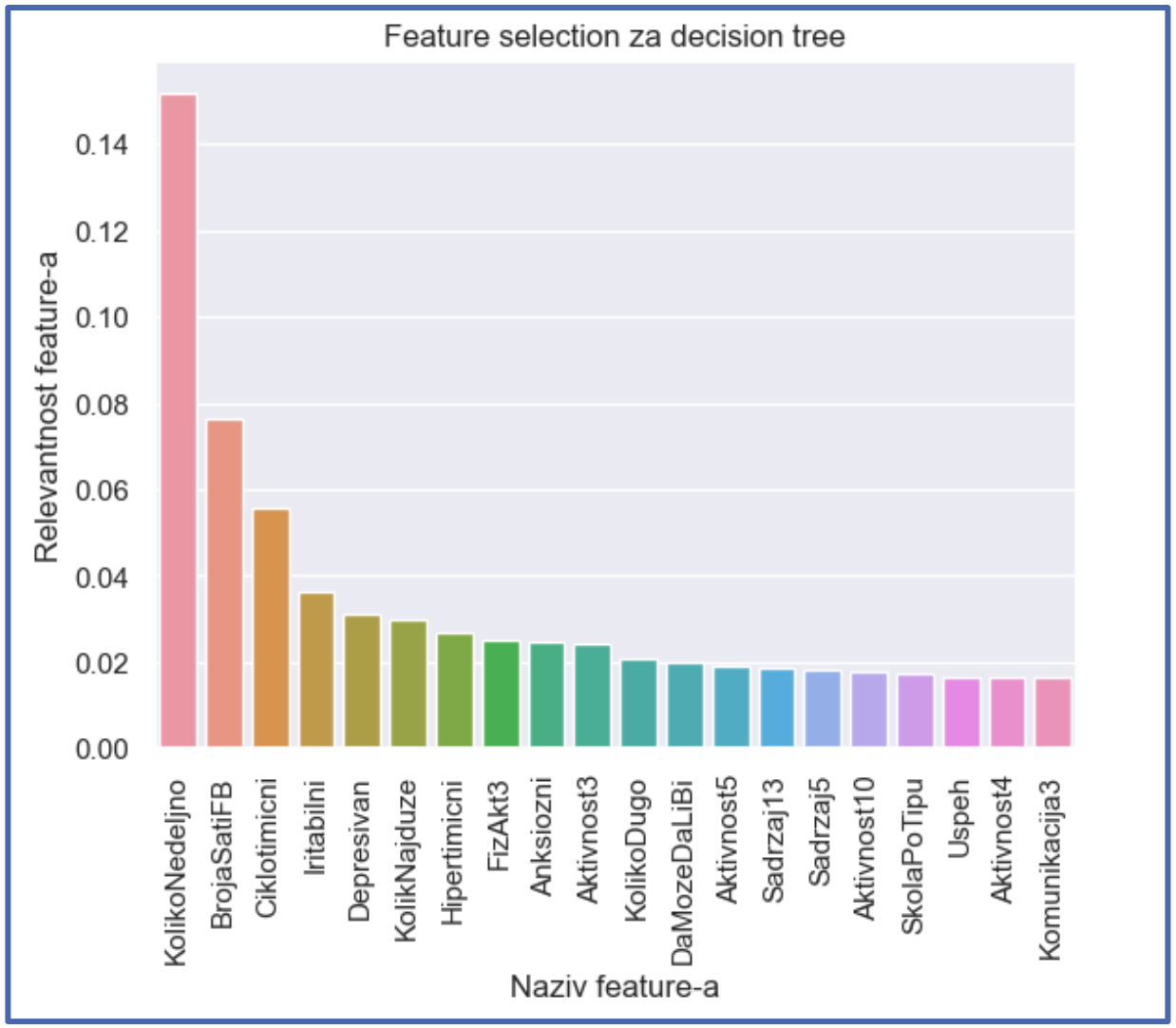
Embedded metode	**Filter** metode	**Wrapper** metode

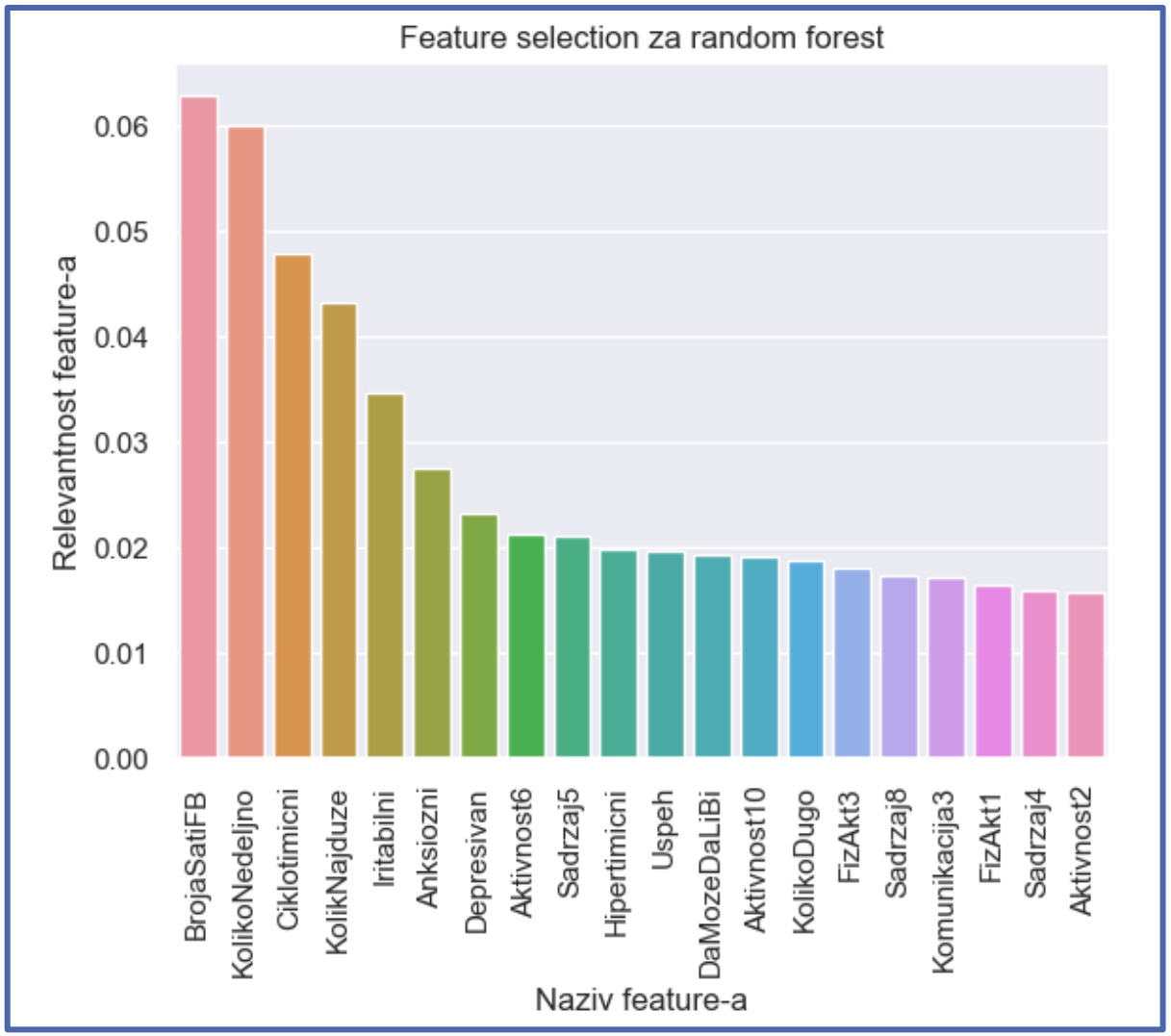
Korišćeni set podataka za prikaz praktične implementacije

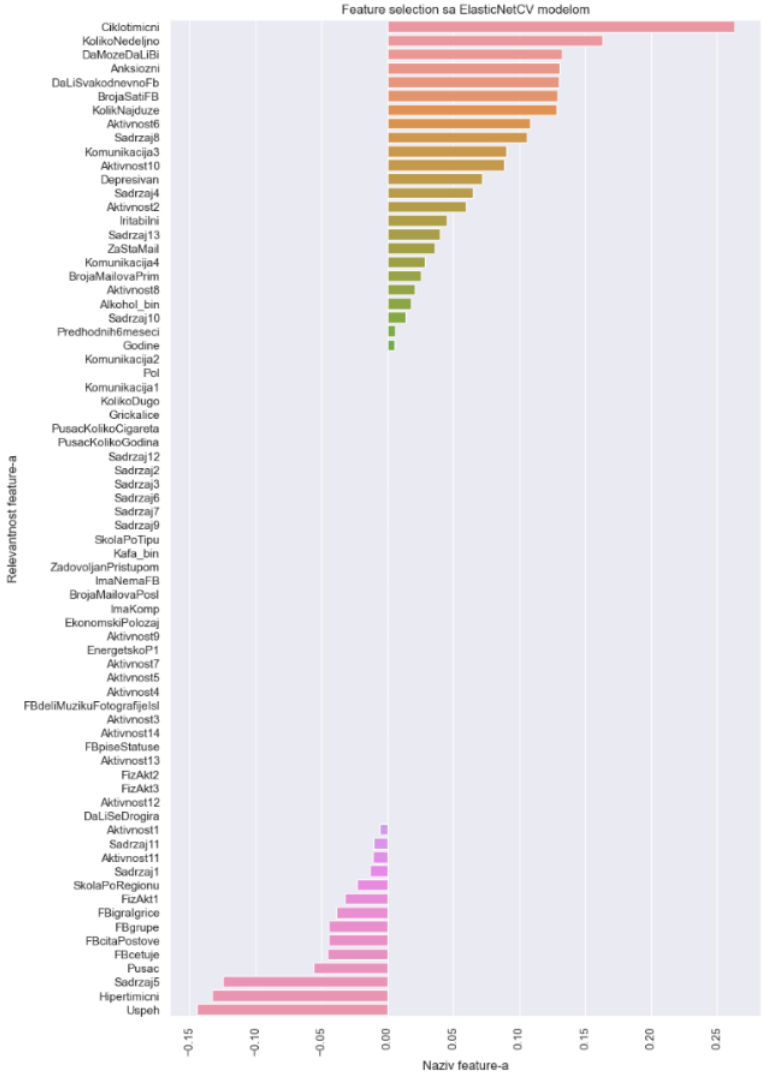
- Praktična implementacija se tiče rada na problemu binarne klasifikacije za određivanje toga da li su korisnici zavisni od upotrebe Interneta ili ne. U setu podataka je inicijalno dostupno 136 kolona koje predstavljaju rezultate dobijene anketiranjem korisnika, dok je broj instanci 2113.
- Što se tiče procesa predobrade podataka, pre svega su odrađene:
  - Provera duplikata,
  - Provera NaN vrednosti uz adekvatno čišćenje podataka,
  - Sređivanje kolona uz eliminisanje onih koje nisu od značaja,
  - Provera balansiranosti seta ali i sam proces balansiranja primenom SMOTE tehnike,
  - Detekcija outliera uz vizuelizaciju i njihova eliminacija Isolation Forest metodom,
  - Normalizacija i standardizacija podataka, kao i
  - Podela na trening i test skup

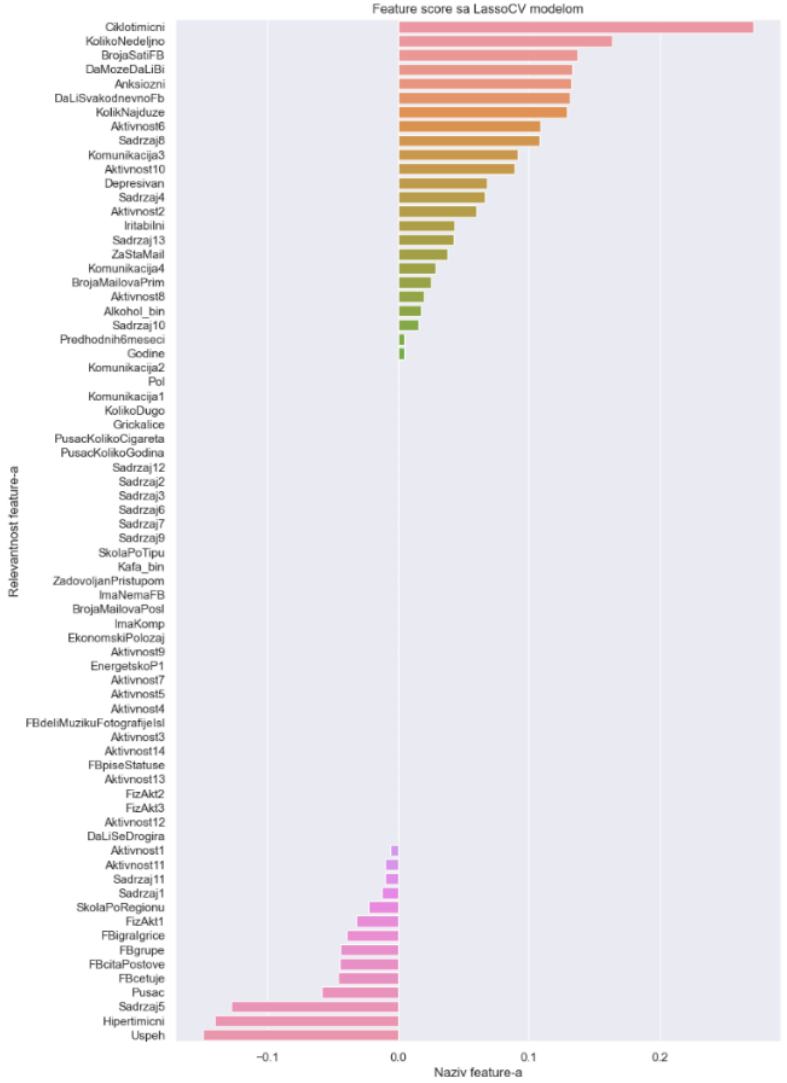
Embedded pristup

- Glavna karakteristika Embedded pristupa za selekciju atributa se ogleda u tome da se proces izbora atributa vrši tokom samog treniranja modela.
- Takav pristup podrazumeva zavisnost izabranih atriubuta od konkretnog modela koji se koristi tokom treniranja, ali i znači da nikakav alat za eksternu selekciju atributa nije neophodan.
- Primenom Embedded pristupa se smanjuje mogućnost overfitting-a, ali se i uglavnom postiže brže određivanje podskupa atributa budući da se postupak dešava prilikom procesa učenja.
- Nedostatak Embedded pristupa je u tome što direktno zavise od modela što može biti naročito problematično ukoliko se radi sa tree-based modelima koji koriste halapljiv (eng. greedy) pristup prilikom selekcije atributa.


Feature selection za decision tree
Relevantnost feature-a
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
KolikoNedeljno
BrojaSatiFB
Ciklotimicni
Iritabilni
Depresivan
KolikNajduze
Hipertimicni
FizAkt3
Anksiozni
Aktivnost3
KolikoDugo
DaMozeDaLiBi
Aktivnost5
Sadrzaj13
Sadrzaj5
Aktivnost10
SkolaPoTipu
Uspeh
Aktivnost4
Komunikacija3
Naziv feature-a


Feature selection za random forest
Relevantnost feature-a
0.00
0.01
0.02
0.03
0.04
0.05
0.06
BrojaSatiFB
KolikoNedeljno
Ciklotimicni
KolikNajduze
Iritabilni
Anksiozni
Depresivan
Aktivnost6
Sadrzaj5
Hipertimicni
Uspeh
DaMozeDaLiBi
Aktivnost10
KolikoDugo
FizAkt3
Sadrzaj8
Komunikacija3
FizAkt1
Sadrzaj4
Aktivnost2
Naziv feature-a

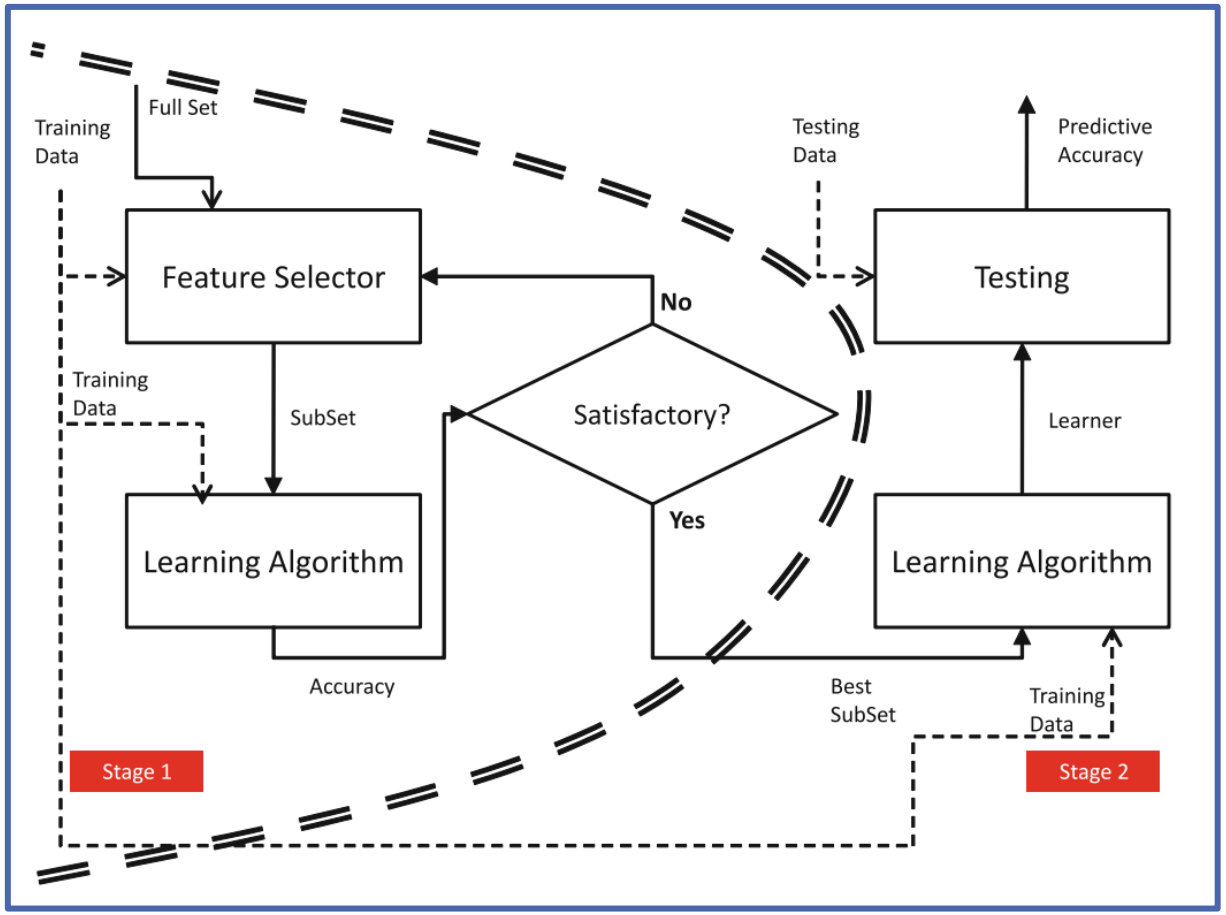

Feature selection sa ElasticNetCV modelom
Ciklotimicni
KolikoNedeljno
DaMozeDaLiBi
Anksiozni
DaLiSvakodnevnoFb
BrojaSatiFB
KolikNajduze
Aktivnost6
Sadrzaj8
Komunikacija3
Aktivnost10
Depresivan
Sadrzaj4
Aktivnost2
Iritabilni
Sadrzaj13
ZaStaMail
Komunikacija4
BrojaMailovaPrim
Aktivnost8
Alkohol_bin
Sadrzaj10
Predhodnih6meseci
Godine
Komunikacija2
Pol
Komunikacija1
KolikoDugo
Grickalice
PusacKolikoCigareta
PusacKolikoGodina
Sadrzaj12
Sadrzaj2
Sadrzaj3
Sadrzaj6
Sadrzaj7
Sadrzaj9
SkolaPoTipu
Kafa_bin
ZadovoljanPristupom
ImaNemaFB
BrojaMailovaPosl
ImaKomp
EkonomskiPolozaj
Aktivnost9
EnergetskoP1
Aktivnost7
Aktivnost5
Aktivnost4
FBdeliMuzikuFotografijeIsl
Aktivnost3
Aktivnost14
FBpiseStatuse
Aktivnost13
FizAkt2
FizAkt3
Aktivnost12
DaLiSeDrogira
Aktivnost1
Sadrzaj11
Aktivnost11
Sadrzaj1
SkolaPoRegionu
FizAkt1
FBigraIgrice
FBgrupe
FBcitaPostove
FBcetuje
Pusac
Sadrzaj5
Hipertimicni
Uspeh
Relevantnost feature-a
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
Naziv feature-a

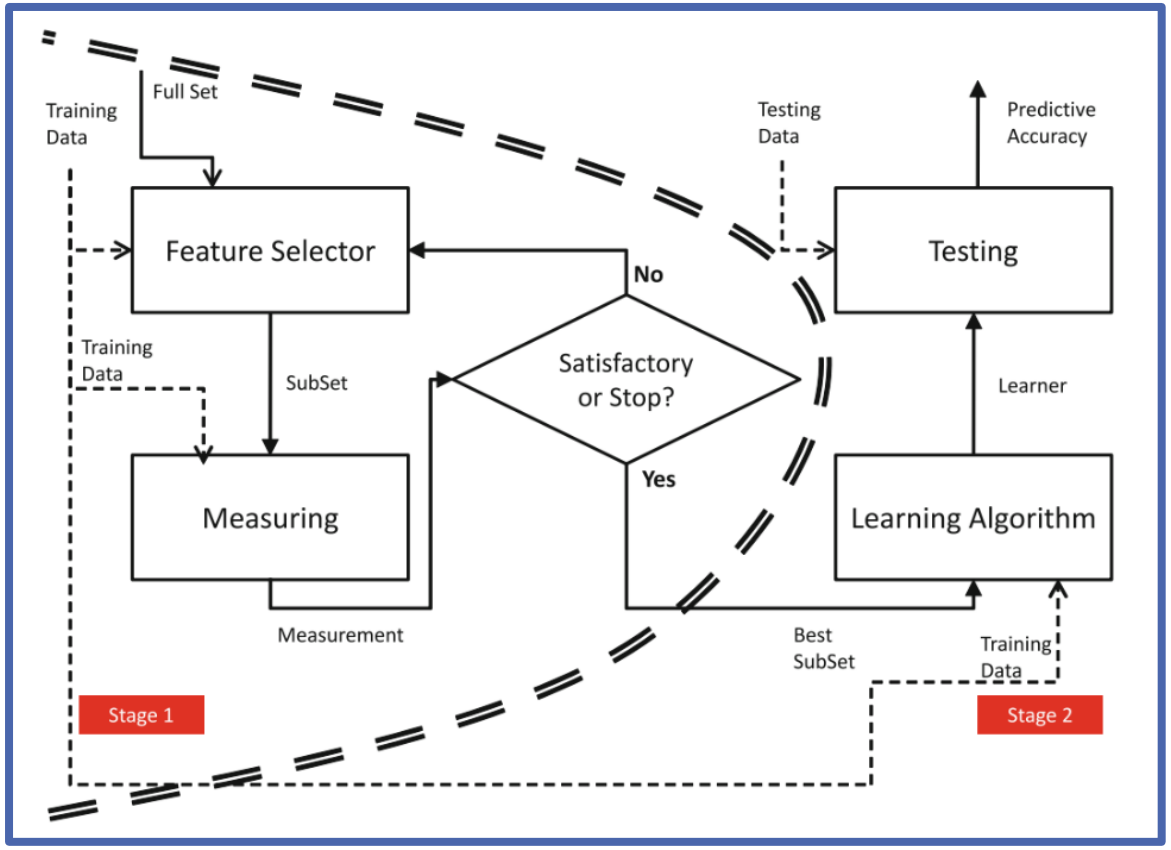

Feature score sa LassoCV modelom
Ciklotimicni
KolikoNedeljno
BrojaSatiFB
DaMozeDaLiBi
Anksiozni
DaLiSvakodnevnoFb
KolikNajduze
Aktivnost6
Sadrzaj8
Komunikacija3
Aktivnost10
Depresivan
Sadrzaj4
Aktivnost2
Iritabilni
Sadrzaj13
ZaStaMail
Komunikacija4
BrojaMailovaPrim
Aktivnost8
Alkohol_bin
Sadrzaj10
Predhodnih6meseci
Godine
Komunikacija2
Pol
Komunikacija1
KolikoDugo
Grickalice
PusacKolikoCigareta
PusacKolikoGodina
Sadrzaj12
Sadrzaj2
Sadrzaj3
Sadrzaj6
Sadrzaj7
Sadrzaj9
SkolaPoTipu
Kafa_bin
ZadovoljanPristupom
ImaNemaFB
BrojaMailovaPosl
ImaKomp
EkonomskiPolozaj
Aktivnost9
EnergetskoP1
Aktivnost7
Aktivnost5
Aktivnost4
FBdeliMuzikuFotografijeIsl
Aktivnost3
Aktivnost14
FBpiseStatuse
Aktivnost13
FizAkt2
FizAkt3
Aktivnost12
DaLiSeDrogira
Aktivnost1
Aktivnost11
Sadrzaj11
Sadrzaj1
SkolaPoRegionu
FizAkt1
FBigraIgrice
FBgrupe
FBcitaPostove
FBcetuje
Pusac
Sadrzaj5
Hipertimicni
Uspeh
Relevantnost feature-a
−0.1
0.0
0.1
0.2
Naziv feature-a

Filter pristup

- Za razliku od Embedded pristupa, Filter metode vrše izbor atributa nezavisno od konkretnog modela koji se koristi prilikom procesa mašinskog učenja.
- Njihov naziv se ogleda u činjenici da se vrši filtriranje atributa pre samog procesa učenja, pri čemu način na koji se vrši filtriranje zavisi od izabrane heuristike.
- Bitno je naglasiti da Filter pristup funkcioniše adekvatno i sa setovima podataka koji imaju veliki broj dimenzija.
- Model za izbor atributa Filter pristupom se sastoji iz dve faze. Prva faza podrazumeva korišćenje metrika kao što su informacione mere, mere rastojanja, zavisnosti i konzistentnosti. U drugoj fazi se vrši proces učenja i testiranja, pri čemu se dobijaju rezultati predikcija.

Filter pristup





Wrapper pristup

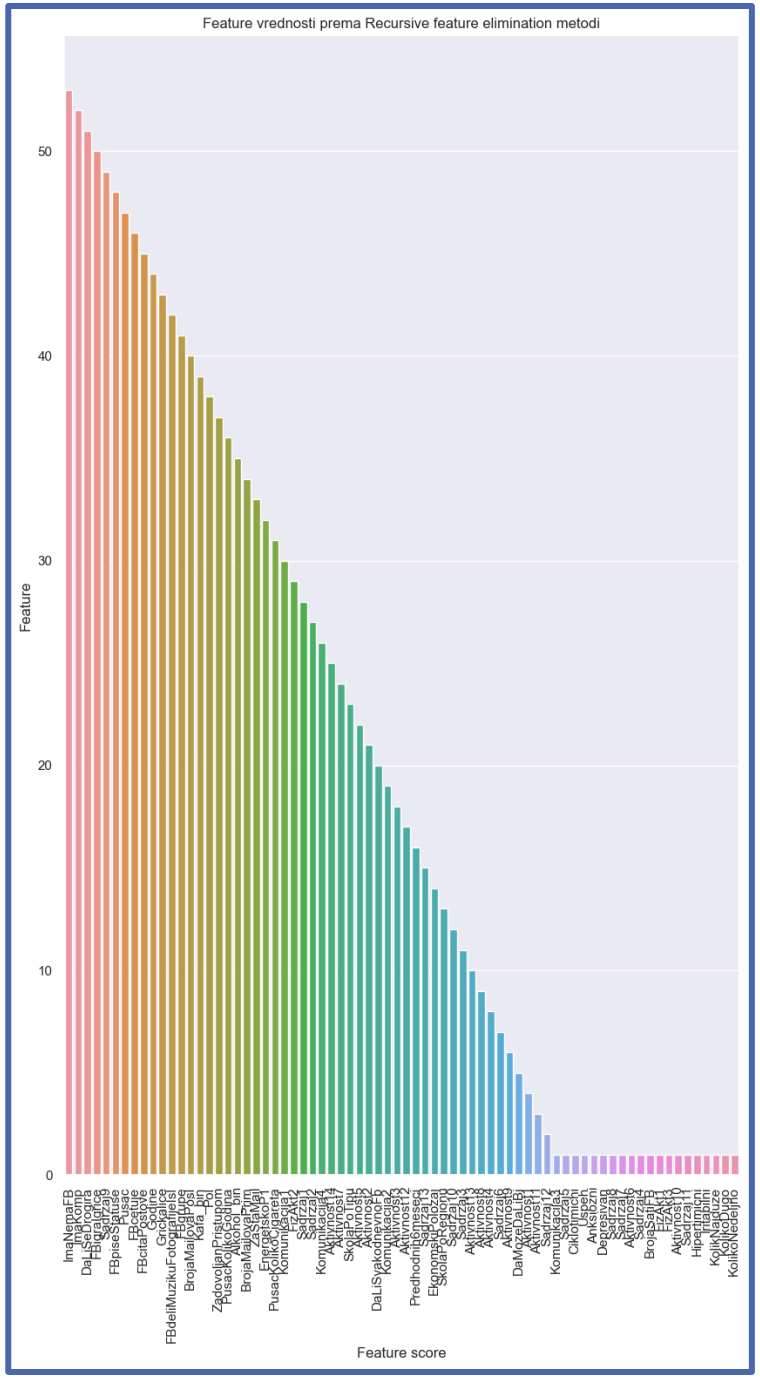
- Izbor atributa primenom Wrapper pristupa se ogleda u ideji da se sam proces izbora vrši korišćenjem klasifikatora i metrika za prediktivne performanse datog klasifikatora.
- Dobrota izabranih atributa će se ogledati kroz performanse samog modela za klasifikaciju.
- Primena Wrapper modela se može posmatrati kroz dve faze. Prva faza obuhvata proces odabira atributa pomoću accuracy mere za dati klasifikator, dok se u drugoj fazi vrši proces učenja i testiranja.

	feature	is_used
0	Aktivnost1	True
1	Aktivnost10	True
2	Aktivnost11	True
3	Aktivnost12	False
4	Aktivnost13	False
5	Aktivnost14	True
6	Aktivnost2	False
7	Aktivnost3	True
8	Aktivnost4	False
9	Aktivnost5	False
10	Aktivnost6	True
11	Aktivnost7	False
12	Aktivnost8	True
13	Aktivnost9	False
14	Alkohol_bin	True
15	Anksiozni	True
16	BrojaMailovaPosl	True
17	BrojaMailovaPrim	True
18	BrojaSatiFB	True
19	Ciklotimicni	True
20	DaLiSeDrogira	True
21	DaLiSvakodnevnoFb	True
22	DaMozeDaLiBi	False
23	Depresivan	False
24	EkonomskiPolozaj	False
25	EnergetskoP1	False
26	FBcetuje	True
27	FBcitaPostove	False
28	FBdeliMuzikuFotografijeIsl	True
29	FBgrupe	False
30	FBigraIgrice	True
31	FBpiseStatuse	False
32	FizAkt1	True
33	FizAkt2	True
34	FizAkt3	True
35	Godine	True
36	Grickalice	False
37	Hipertimicni	False
38	ImaKomp	True
39	ImaNemaFB	True
40	Iritabilni	True

41	Kafa_bin	True
42	KolikNajduze	True
43	KolikoDugo	False
44	KolikoNedeljno	True
45	Komunikacija1	False
46	Komunikacija2	True
47	Komunikacija3	True
48	Komunikacija4	True
49	Pol	True
50	Predhodnih6meseci	False
51	Pusac	True
52	PusacKolikoCigareta	True
53	PusacKolikoGodina	True
54	Sadrzaj1	True
55	Sadrzaj10	False
56	Sadrzaj11	True
57	Sadrzaj12	True
58	Sadrzaj13	True
59	Sadrzaj2	False
60	Sadrzaj3	True
61	Sadrzaj4	False
62	Sadrzaj5	False
63	Sadrzaj6	True
64	Sadrzaj7	True
65	Sadrzaj8	False
66	Sadrzaj9	False
67	SkolaPoRegionu	True
68	SkolaPoTipu	False
69	Uspeh	True
70	ZaStaMail	True
71	ZadovoljanPristupom	True

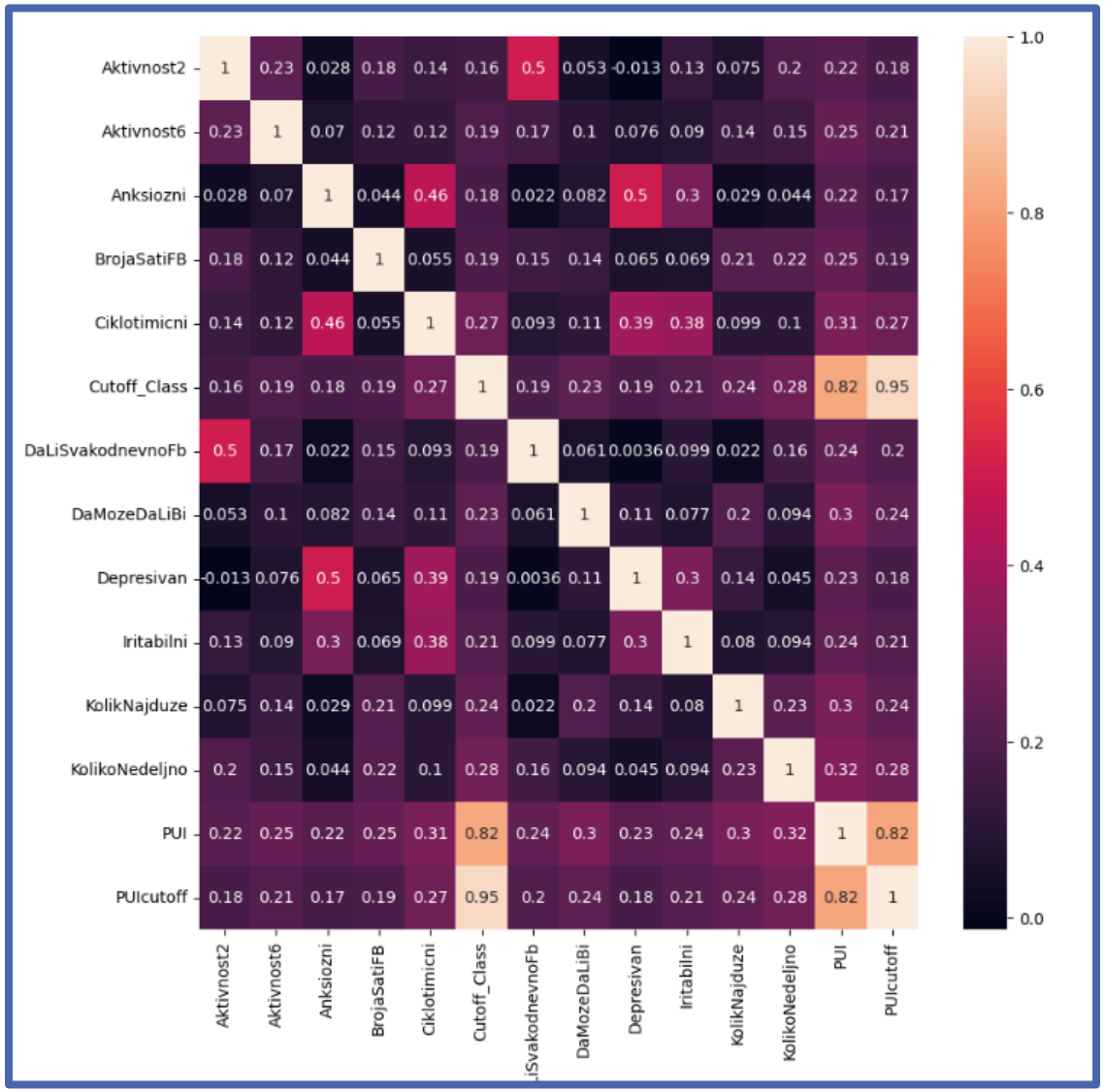
Wrapper pristup Upotreba genetskog algoritma

Wrapper pristup Upotreba rekurzivne eliminacije atributa



Određivanje težine atributa

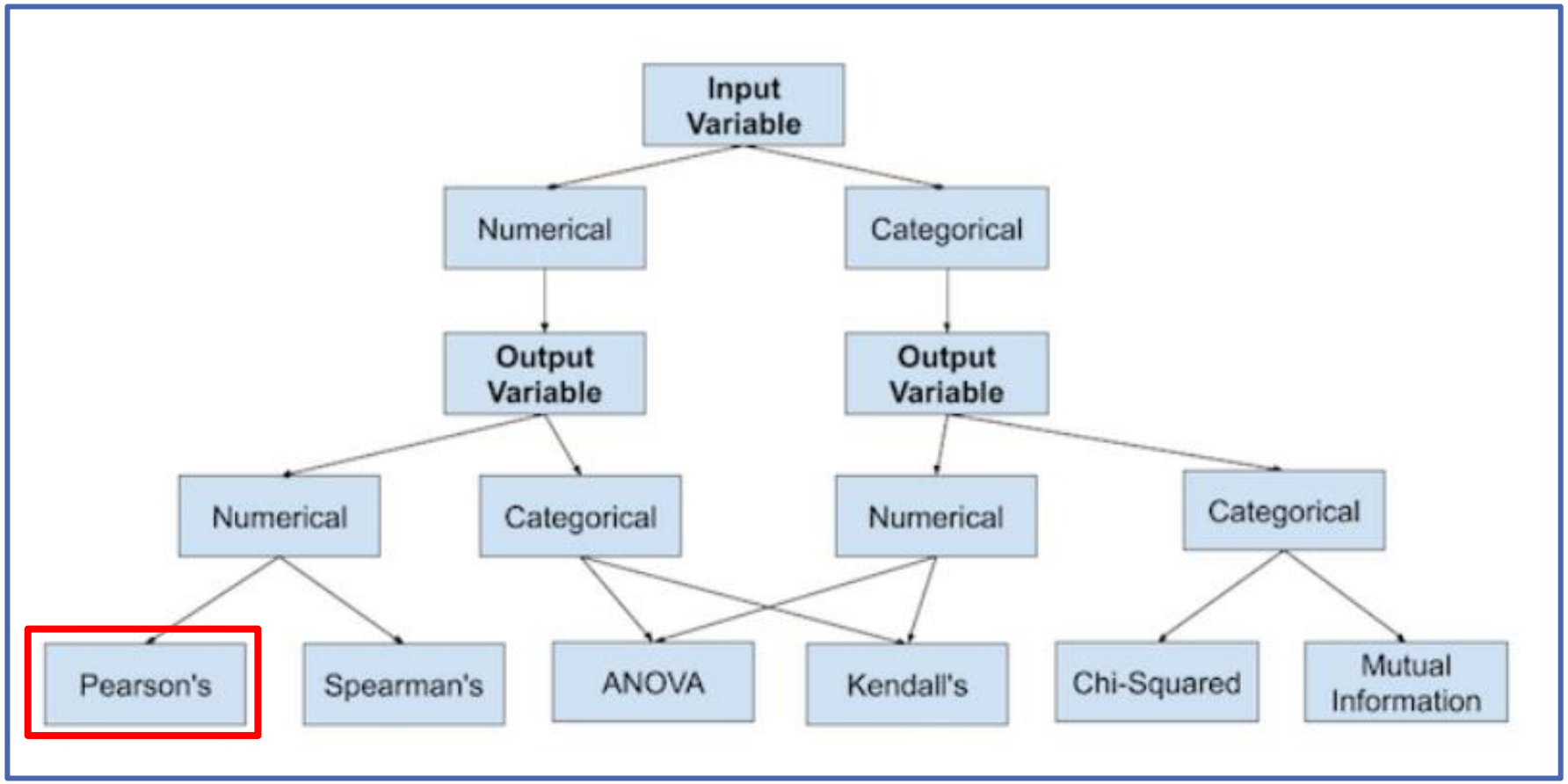
- Kada se radi o postupku odabira najrelevantnijih atributa, može se pribeći i izračunavanju težine atributa nasuprot metodama za njihovu eliminaciju.
- Atributima koji su relevantniji se dodeljuje veća težina, dok se onim manje relevantnim vrši dodela nižih vrednosti.
- Uglavnom je potrebno imati odgovarajuće domensko znanje, kako bi metrici težine bio dodeljen smisao, ali je moguće vršiti i automatsku dodelu težina (na primer Support Vectore Machine).
- Postupak određivanja težine se može izvesti primenom **Relief** algoritma koji ima za cilj da odabere atribute koji su statistički najrelevantniji.

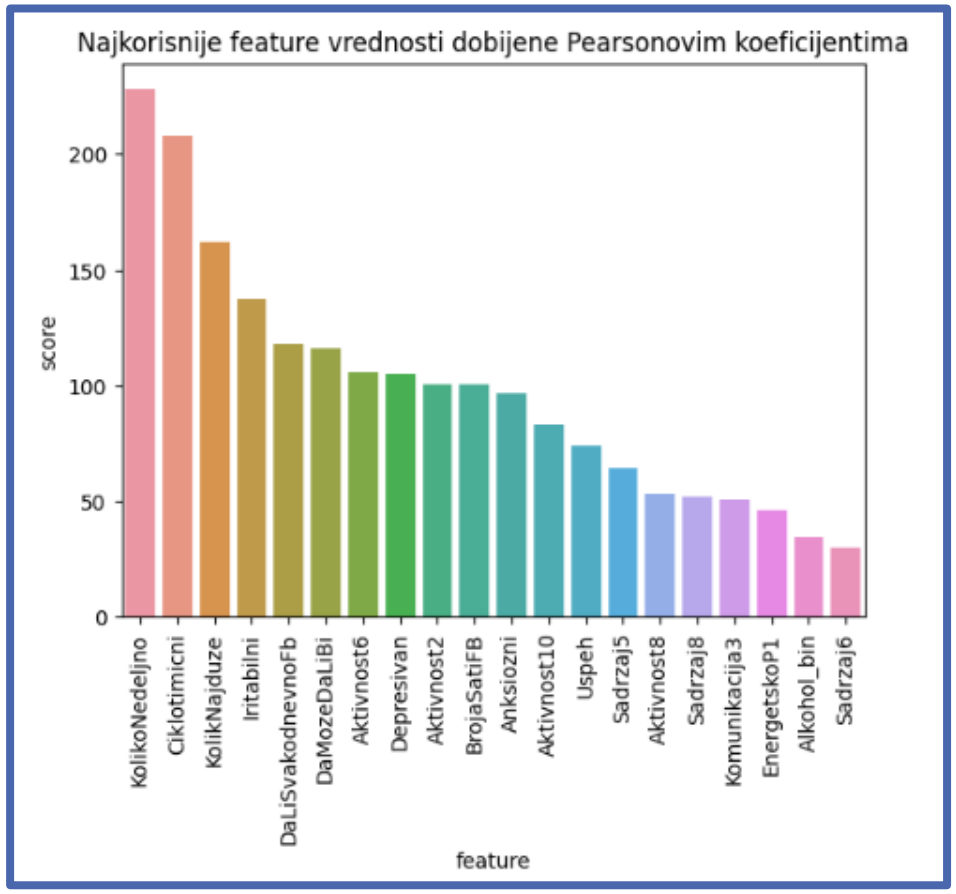


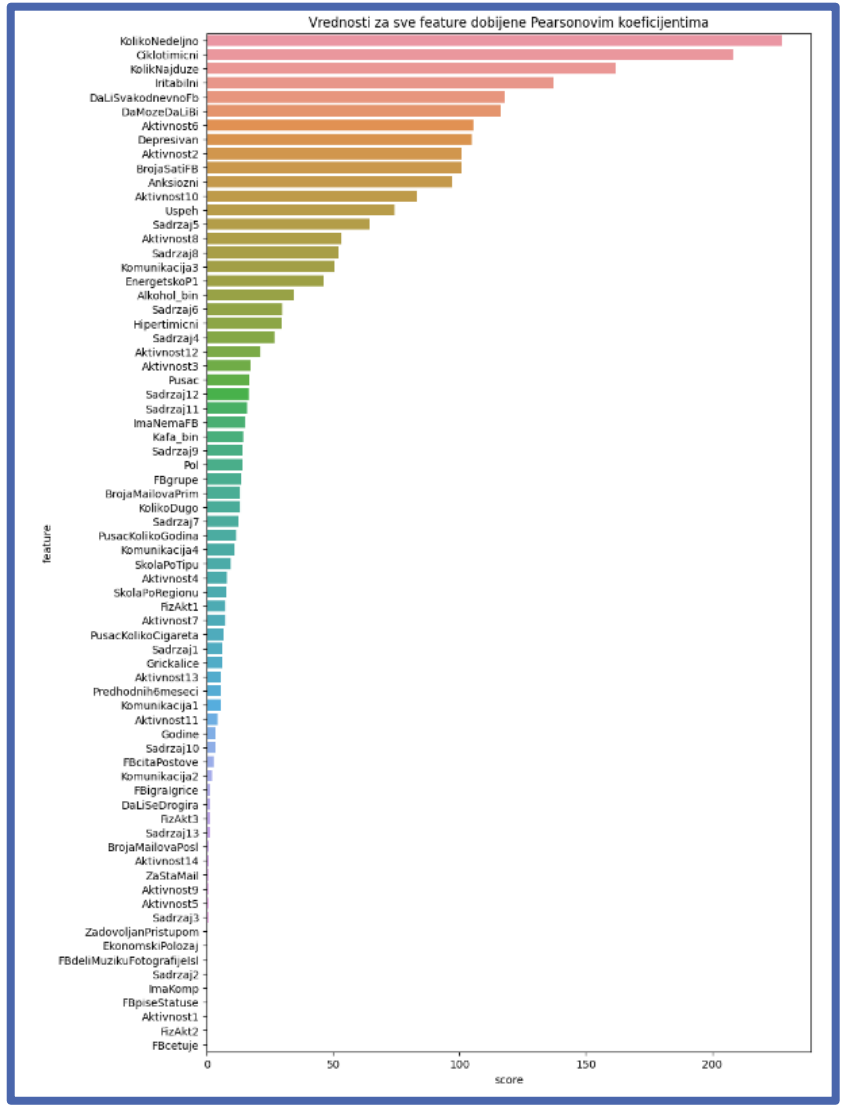
Statističke metode za selekciju atributa na osnovu tipova podataka

- Kada je reč o upotrebi statističkih metoda za proces selekcije atributa, mora se pre svega voditi računa o tipovima podataka nad kojima se vrši obrada.
- Odabir odgovarajuće statističke metode, osim tipa podataka zavisi od toga i da li je reč o ulaznim ili izlaznim podacima.
- Na osnovu matrice korelacije se može uočiti stepen korelacije između ulaznih atributa i target klase.

Podela statističkih metoda prema tipu korišćenih podataka







Pearsonov koeficijent

Interpretabilnost izbora atributa korišćenjem Explainable AI alata

Ubrzani razvoj veštačke inteligencije i mašinskog učenja u prethodnim godinama je rezultirao njihovom sve većom upotrebom u domenu nauke, istraživačkog rada, ali i primene u okviru raznih grana industrije.

Potencijalna opasnost i neizvesnost prilikom primene se ogleda u tome što način funkcionisanja većine modela nije u potpunosti jasan, već se smatra da rade po principu crne kutije (eng. Black box).

Kako bi ti problemi bili prevaziđeni, predloženi način je razvoj alata koji se bave interptretabilnošću i objašnjivošću (eng. Explainable AI – XAI) rezultata dobijenih prilikom primene algoritama mašinskog učenja.

U okviru datog domena je predložen veliki broj implementacionih rešenja, međutim su se kao rešenja sa najvećom primenom izdvojile biblioteke **LIME, SHAP i ELI5**.

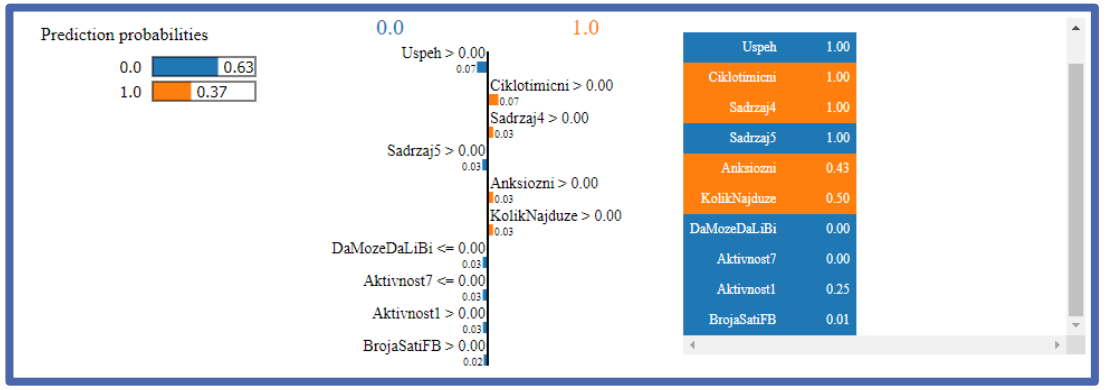
LIME
(Local Interpretable Model- Agnostic Explanations)

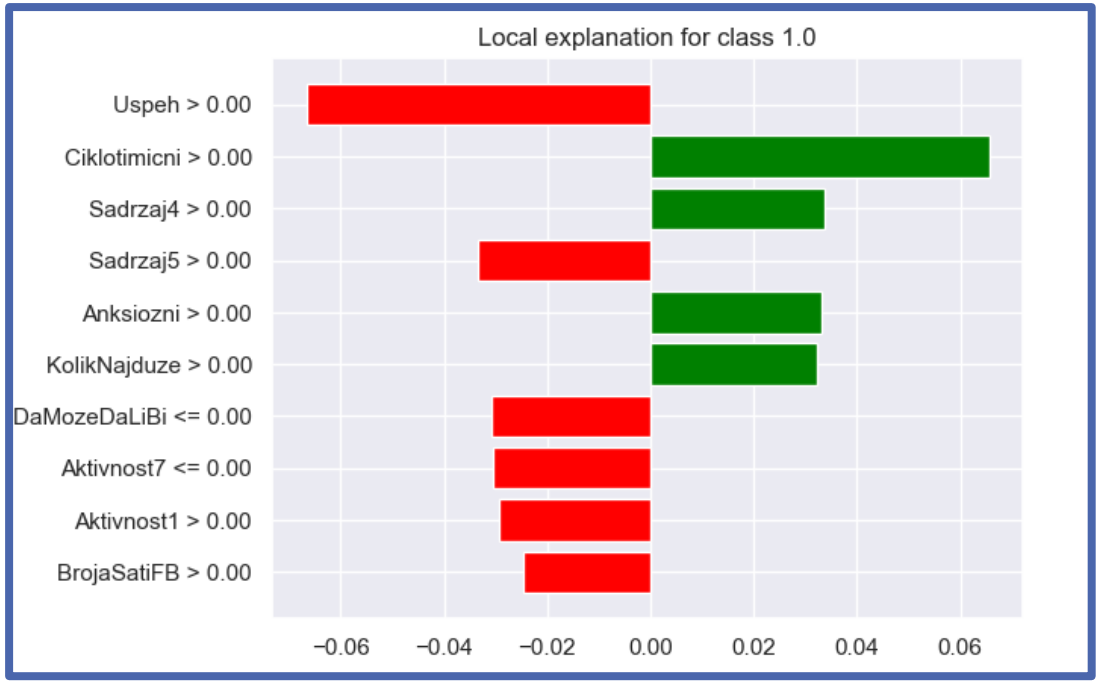
LIME (Local-Interpretable Model-Agnostic Explanations) se može definisati kao Explainable AI alat koji ima mogućnost da odredi predikcije bilo kog klasifikatora ili regresora, korišćenjem globalne i lokalne perspektive.

Bitna odlika LIME-a je pre svega interpretabilnost, odnosno mogućnost određivanja kvalitivne mere za razumevanje odnosa između ulaznih atributa i izlaza.

Pored interpretabilnosti je potrebno definisati i lokalnu vernost (local fidelity), koja se ogleda u tome da objašnjenje (eng. prediction) mora odgovarati onome kako se model ponaša u blizini instance koja se predviđa, odnosno koju je potrebno objasniti.

Još jedna bitna karakteristika LIME-a je mogućnost pružanja objašnjenja za bilo koji modela, odnosno drugačije rečeno model-agnostički pristup.

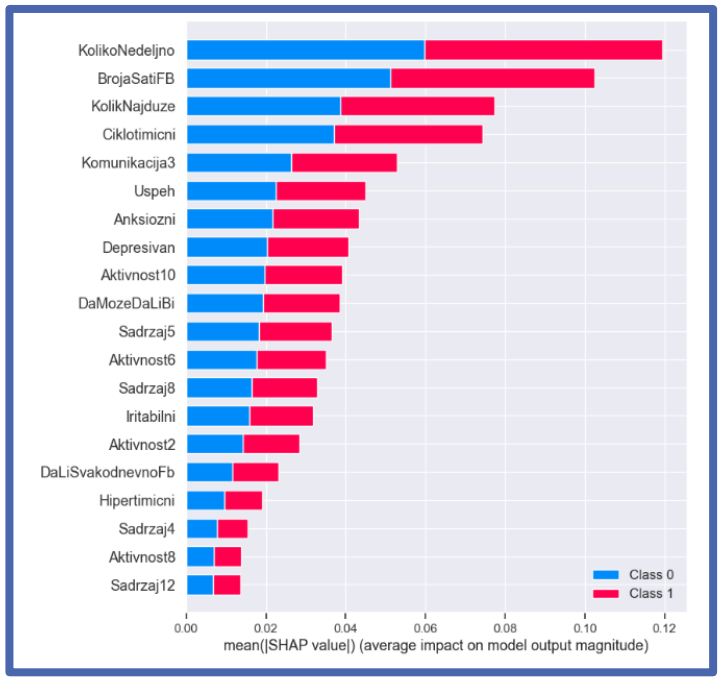


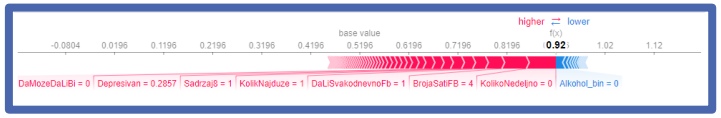


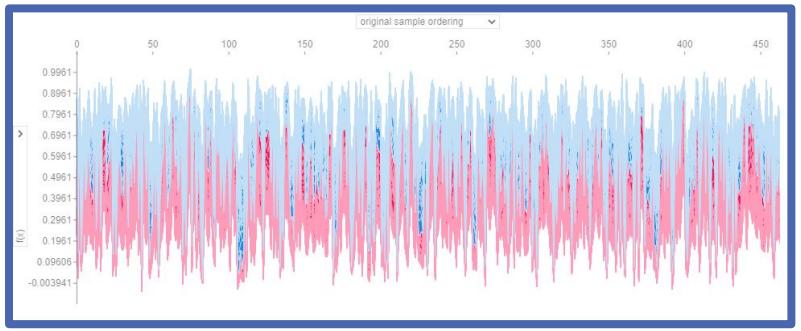
LIME
(Local Interpretable Model- Agnostic Explanations)

SHAP (Shapley Additive Explanations)

- SHAP (Shapley Additive Explanations) je biblioteka koja ima za cilj da prikaže predikciju za pojedinačnu instancu na način da vrši izračunavanje doprinosa svakog atributa prilikom donošenja odluke za predikciju vrednosti.
- Shapley vrednosti se koriste da odrede efekte koje pojedinačni atributi imaju na određivanje output-a za dati model mašinskog učenja, pri čemu se vrši dodela vrednosti, odnosno definisanje težine za svaki atribut u datom setu podataka.







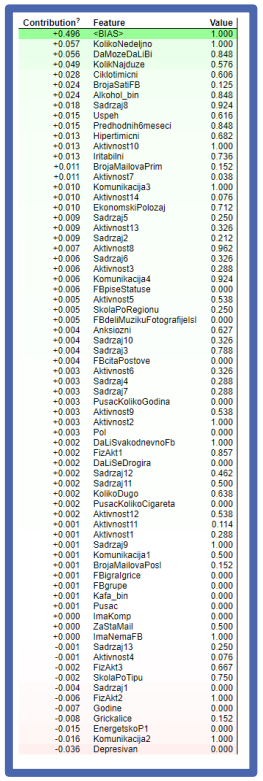
SHAP (Shapley Additive Explanations)

ELI5 (Explain like i am five)

Eli5(Explain like im 5) je open-source Python biblioteka koja omogućava korisnicima pregled objašnjenja predikcija dobijenih odabranim modelima mašinskog učenja.

Koristi se u svrhe debagiranja modela mašinskog, ali i dubokog učenja, pri čemu je upotreba ograničena na linearne i tree-based modele.

Najkorišćenije funkcije u okviru eli5 biblioteke su **eli5.show_weights()**, čija je namena inspekcija parametara datog modela, dok **eli5.show_prediction()** ima za cilj da otkrije zbog čega je model odredio takvu predikciju za odabranu instancu.

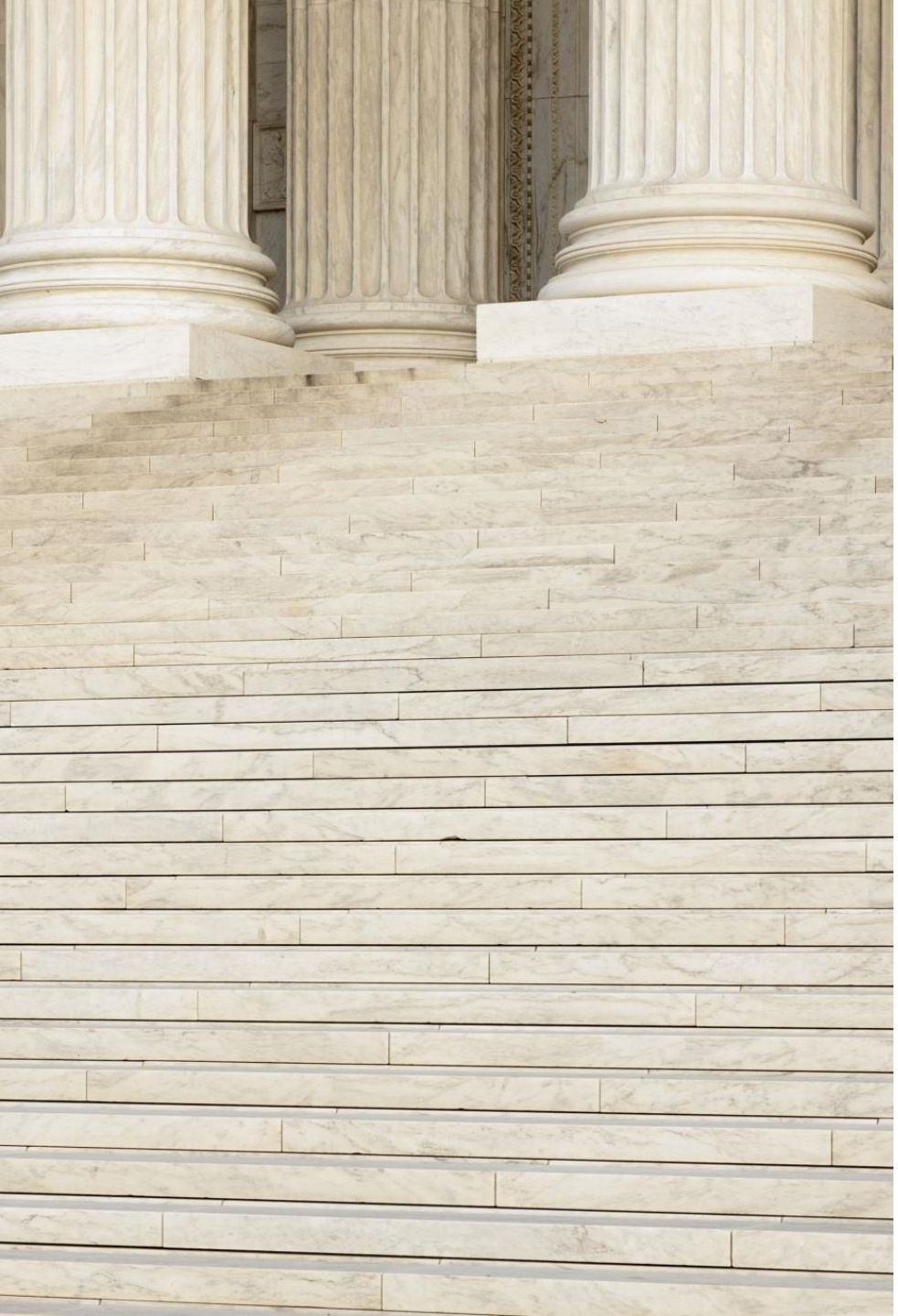


Contribution[?]	Feature	Value
+0.496	<BIAS>	1.000
+0.057	KolikoNedeljno	1.000
+0.056	DaMozeDaLiBi	0.848
+0.049	KolikNajduze	0.576
+0.028	Ciklotimicni	0.606
+0.024	BrojaSatiFB	0.125
+0.024	Alkohol_bin	0.848
+0.018	Sadrzaj8	0.924
+0.015	Uspeh	0.616
+0.015	Prednodnih6meseci	0.848
+0.013	Hipertimicni	0.682
+0.013	Aktivnost10	1.000
+0.013	Iritabilni	0.736
+0.011	BrojaMailovaPrim	0.152
+0.011	Aktivnost7	0.038
+0.010	Komunikacija3	1.000
+0.010	Aktivnost14	0.076
+0.010	EkonomskiPolozaj	0.712
+0.009	Sadrzaj5	0.250
+0.009	Aktivnost13	0.326
+0.009	Sadrzaj2	0.212
+0.007	Aktivnost8	0.962
+0.006	Sadrzaj6	0.326
+0.006	Aktivnost3	0.288
+0.006	Komunikacija4	0.924
+0.006	FBpiseStatuse	0.000
+0.005	Aktivnost5	0.538
+0.005	SkolaPoRegionu	0.250
+0.005	FBdeliMuzikuFotografijeIsl	0.000
+0.004	Anksiozni	0.627
+0.004	Sadrzaj10	0.326
+0.004	Sadrzaj3	0.788
+0.004	FBcitaPostove	0.000
+0.003	Aktivnost6	0.326
+0.003	Sadrzaj4	0.288
+0.003	Sadrzaj7	0.288
+0.003	PusacKolikoGodina	0.000
+0.003	Aktivnost9	0.538
+0.003	Aktivnost2	1.000
+0.003	Pol	0.000
+0.002	DaLiSvakodnevnoFb	1.000
+0.002	FizAkt1	0.857
+0.002	DaLiSeDrogira	0.000
+0.002	Sadrzaj12	0.462
+0.002	Sadrzaj11	0.500
+0.002	KolikoDugo	0.638
+0.002	PusacKolikoCigareta	0.000
+0.002	Aktivnost12	0.538
+0.001	Aktivnost11	0.114
+0.001	Aktivnost1	0.288
+0.001	Sadrzaj9	1.000
+0.001	Komunikacija1	0.500
+0.001	BrojaMailovaPosl	0.152
+0.001	FBigralgrice	0.000
+0.001	FBgrupe	0.000
+0.001	Kafa_bin	0.000
+0.001	Pusac	0.000
+0.000	ImaKomp	0.000
+0.000	ZaStaMail	0.500
+0.000	ImaNemaFB	1.000
-0.001	Sadrzaj13	0.250
-0.001	Aktivnost4	0.076
-0.002	FizAkt3	0.667
-0.002	SkolaPoTipu	0.750
-0.004	Sadrzaj1	0.000
-0.006	FizAkt2	1.000
-0.007	Godine	0.000
-0.008	Grickalice	0.152
-0.015	EnergetskoP1	0.000
-0.016	Komunikacija2	1.000
-0.036	Depresivan	0.000

Weight	Feature
0.0129 ± 0.0028	BrojaSatiFB
0.0093 ± 0.0022	KolikoNedeljno
0.0053 ± 0.0020	Ciklotimicni
0.0017 ± 0.0008	DaMozeDaLiBi
0.0009 ± 0.0011	Iritabilni
0.0008 ± 0.0005	Sadrzaj8
0.0006 ± 0.0008	Depresivan
0.0004 ± 0.0004	Sadrzaj5
0.0003 ± 0.0005	Aktivnost6
0.0003 ± 0.0005	Aktivnost10
0.0003 ± 0.0013	Anksiozni
0.0002 ± 0.0005	KolikNajduze
0.0002 ± 0.0005	Alkohol_bin
0 ± 0.0000	EnergetskoP1
0 ± 0.0000	FBpiseStatuse
0 ± 0.0000	FBigralgrice
0 ± 0.0000	FBcetuje
0 ± 0.0000	FBdeliMuzikuFotografijeIsl
0 ± 0.0000	FBcitaPostove
0 ± 0.0000	EkonomskiPolozaj
… 52 more …	

Weight	Feature
0.0576 ± 0.0690	BrojaSatiFB
0.0553 ± 0.0912	KolikoNedeljno
0.0424 ± 0.0476	Ciklotimicni
0.0396 ± 0.0707	KolikNajduze
0.0294 ± 0.0337	Iritabilni
0.0273 ± 0.0276	Anksiozni
0.0259 ± 0.0389	Aktivnost6
0.0246 ± 0.0296	Depresivan
0.0234 ± 0.0386	Sadrzaj5
0.0217 ± 0.0340	Sadrzaj8
0.0204 ± 0.0426	DaMozeDaLiBi
0.0202 ± 0.0186	Hipertimicni
0.0198 ± 0.0172	KolikoDugo
0.0197 ± 0.0204	Aktivnost10
0.0188 ± 0.0147	FizAkt1
0.0180 ± 0.0169	FizAkt3
0.0175 ± 0.0201	Sadrzaj4
0.0175 ± 0.0258	Komunikacija3
0.0172 ± 0.0230	Uspeh
0.0152 ± 0.0176	Sadrzaj12
… 52 more …	

ELI5 (Explain like i am five)



Zaključak

- U okviru domena prikupljanja i predobrade podataka postoji veliki broj tehnika i metoda čiji je cilj da omoguće rad sa što kvalitetnijim podacima, radi dobijanja boljih i preciznijih rezulata.
- Korak predobrade podataka je neophodan i potrebno je odabrati odgovarajuće pristupe koji pre svega zavise od prirode podataka nad kojima se vrši dalja analiza.
- Bitno je napomenuti da upotreba kvalitetnih algoritama sama po sebi neće dovesti do dobrih predikcija i rezultata budući da najčešće ne može da nadomesti rad sa podacima koji prethodno nisu adekvatno preprocesirani i obrađeni.
- Stoga, potrebno je staviti poseban akcenat na korišćenje i implementaciju neophodnih metoda u okviru domena prikupljanja i predobrade podataka.

HVALA NA PAŽNJI!