Uvod u dubinsku analizu podataka

Matej Mihelčić

Prirodoslovno-matematički fakultet, Sveučilište u Zagrebu

matmih@math.hr

2. veljače, 2026.

Podaci

Podatak je digitalni zapis svojstava, stanja ili ponašanja određenog objekta.
Skup podataka je digitalni zapis svojstava, stanja ili ponašanja određenog skupa objekata ili sustava.

- Velike količine podataka dobivamo pomoću posebnih uređaja i senzora (npr. fotoaparati, kamere, mobiteli, pametni satovi i drugi nosivi uređaji, medicinski uređaji, meteorološki i drugi tehnički senzori, sateliti, sofisticirani znanstveni uređaji - npr. hadronski sudarač).

Podaci

- Podatke možemo dobiti i digitalnim skladištenjem informacija o događajima ili objektima, npr. trgovački, ekološki, povijesni i umjetnički podaci. Podaci eksperimenata namijenjenih boljem razumijevanju osobina objekata (npr. što razlikuje vrste vina).
- Podatke dobivamo i bilježenjem aktivnosti korisnika, npr. podaci s društvenih mreža, aplikacija za chat, e-mail, povijest pretraživanja i pregledavanja (preglednik, youtube), povijest online kupnji.
- Bilježenje podatkovnog prometa u mreži također daje podatke koje možemo koristiti za određene analize.

Tablični podaci

Opisuju skup objekata (instanci, entiteta) koristeći numeričke, kategorijske ili binarne atribute.

Entitet (osoba)	**Visina**	**Težina**	**Boja kose**	**Vatrogasac**
Marko	182	90	Crna	DA
Luka	178	84	Plava	NE
Ivana	168	66	Smeđa	DA

Višepogledni tablični podaci

Opisuju skup objekata (instanci, entiteta) koristeći dva ili više disjunktnih skupova numeričkih, kategorijskih ili binarnih atributa.

- Višepogledni tablični podaci se sastoje od dvije ili više tablice (pogleda) koji opisuju jedan skup objekata (instanci, entiteta).
- Svaka tablica može opisivati: a) različite aspekte objekata (npr. bioraznolikost u jezeru i čistoća vode u tom jezeru), b) različite vrste mjerenja istih objekata (npr. klinička, bio-medicinska, genetska).

Entitet	Visina	Težina	Boja kose	Spol
Marko	182	90	Crna	M
Luka	178	84	Plava	M
Ivana	168	66	Smeđa	Z

Entitet	Posao	Plaća	Vozilo	Djeca
Marko	Vatrogasac	1228	DA	0
Luka	Violinist	1500	NE	1
Ivana	Vatrogasac	1300	DA	1

Vremenske serije

Opisuju promjene svojstava objekata kroz vrijeme (npr. BDP zemalja, informacije o trgovini zemalja ili populaciji svake godine, cijene dionica, tečaj valuta - dnevni, mjesečni ili godišnji u zadanom vremenskom razdoblju).

Zemlja	2020	2021	2022	2023	2024	2025
Hrvatska	3,953,958	3,924,610	3,907,027	3,896,023	3,875,325	3,848,160
Slovenija	2,102,419	2,113,494	2,115,228	2,118,396	2,118,697	2,117,072
BIH	3,299,349	3,244,907	3,204,082	3,185,073	3,164,253	3,140,095

Tokovi podataka

Opisuju promjene svojstava objekata kroz vrijeme kontinuirano, često u sekundama, minutama ili satima. Npr. mjerenja senzora.

Lokacija	Temperatura	Vlažnost	Tlak	Brzina vjetra
u25kes6b3yt	2.5	83	1012	2.8
u241zykqmzp	2.1	85	1009	1.2
srexw0d9gbw	7.5	48	1005	6.1
u2j71kk6rnq	1.9	99	1012	0.7
srsfeey71bk	9.5	47	1004	4.1

Tokovi podataka

Opisuju promjene svojstava objekata kroz vrijeme kontinuirano, često u sekundama, minutama ili satima. Npr. mjerenja senzora.

Lokacija	Temperatura	Vlažnost	Tlak	Brzina vjetra
u25kes6b3yt	2.4	84	1012	2.7
u241zykqmzp	3.3	76	1009	2.4
srexw0d9gbw	7.4	47	1005	6.5
u2j71kk6rnq	1.6	93	1011	1.7
srsfeey71bk	9.1	47	1004	6.2

Tokovi podataka

Opisuju promjene svojstava objekata kroz vrijeme kontinuirano, često u sekundama, minutama ili satima. Npr. mjerenja senzora.

Lokacija	Temperatura	Vlažnost	Tlak	Brzina vjetra
u25kes6b3yt	2.2	84	1012	1.9
u241zykqmzp	3.2	76	1009	2.2
srexw0d9gbw	7.1	49	1005	5.1
u2j71kk6rnq	0.9	99	1012	1.6
srsfeey71bk	8.9	48	1004	3.6

Tokovi podataka

Opisuju promjene svojstava objekata kroz vrijeme kontinuirano, često u sekundama, minutama ili satima. Npr. mjerenja senzora.

Lokacija	Temperatura	Vlažnost	Tlak	Brzina vjetra
u25kes6b3yt	2.1	85	1012	1.5
u241zykqmzp	1.8	84	1009	1.5
srexw0d9gbw	7.0	49	1005	4.1
u2j71kk6rnq	0.3	100	1012	2.1
srsfeey71bk	8.7	48	1004	3.6

Tokovi podataka

Opisuju promjene svojstava objekata kroz vrijeme kontinuirano, često u sekundama, minutama ili satima. Npr. mjerenja senzora.

Lokacija	Temperatura	Vlažnost	Tlak	Brzina vjetra
u25kes6b3yt	2.0	87	1011	1.6
u241zykqmzp	1.0	88	1008	1.7
srexw0d9gbw	7.0	50	1005	4.8
u2j71kk6rnq	0.0	100	1011	2.2
srsfeey71bk	8.1	45	1004	3.0

Tokovi podataka

Opisuju promjene svojstava objekata kroz vrijeme kontinuirano, često u sekundama, minutama ili satima. Npr. mjerenja senzora.

Lokacija	Temperatura	Vlažnost	Tlak	Brzina vjetra
u25kes6b3yt	1.9	87	1011	1.7
u241zykqmzp	1.7	83	1008	1.8
srexw0d9gbw	6.7	50	1005	3.0
u2j71kk6rnq	-0.4	98	1011	3.6
srsfeey71bk	8.0	45	1004	3.4

Tokovi podataka

Opisuju promjene svojstava objekata kroz vrijeme kontinuirano, često u sekundama, minutama ili satima. Npr. mjerenja senzora.

Lokacija	Temperatura	Vlažnost	Tlak	Brzina vjetra
u25kes6b3yt	1.8	88	1011	1.5
u241zykqmzp	1.3	85	1009	1.1
srexw0d9gbw	6.4	51	1005	5.4
u2j71kk6rnq	-0.6	98	1011	3.6
srsfeey71bk	7.5	48	1004	1.8

Tokovi podataka

Opisuju promjene svojstava objekata kroz vrijeme kontinuirano, često u sekundama, minutama ili satima. Npr. mjerenja senzora.

Lokacija	Temperatura	Vlažnost	Tlak	Brzina vjetra
u25kes6b3yt	1.7	88	1011	1.5
u241zykqmzp	1.4	83	1009	1.9
srexw0d9gbw	6.0	52	1006	5.3
u2j71kk6rnq	-0.9	98	1011	4.2
srsfeey71bk	8.9	47	1005	1.8

Tokovi podataka

Opisuju promjene svojstava objekata kroz vrijeme kontinuirano, često u sekundama, minutama ili satima. Npr. mjerenja senzora.

Lokacija	Temperatura	Vlažnost	Tlak	Brzina vjetra
u25kes6b3yt	1.9	89	1011	1.4
u241zykqmzp	2.6	73	1009	2.3
srexw0d9gbw	6.3	53	1007	3.7
u2j71kk6rnq	-1.2	98	1012	2.9
srsfeey71bk	8.2	49	1005	2.2

Tokovi podataka

Opisuju promjene svojstava objekata kroz vrijeme kontinuirano, često u sekundama, minutama ili satima. Npr. mjerenja senzora.

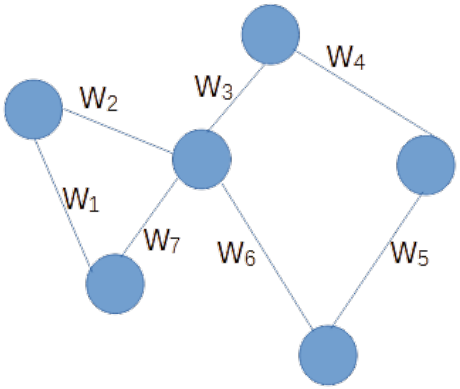
Lokacija	Temperatura	Vlažnost	Tlak	Brzina vjetra
u25kes6b3yt	2.3	84	1011	1.7
u241zykqmzp	4.9	62	1009	1.3
srexw0d9gbw	7.3	49	1007	3.3
u2j71kk6rnq	-0.9	92	1012	3.4
srsfeey71bk	9.0	51	1006	1.3

Tokovi podataka

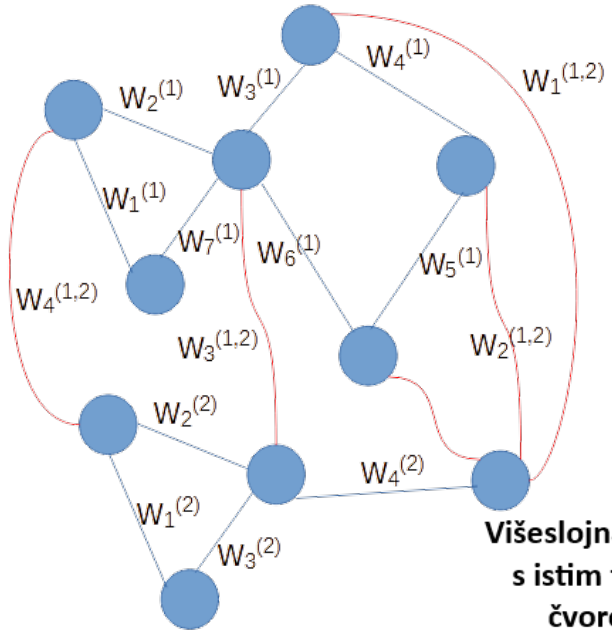
Opisuju promjene svojstava objekata kroz vrijeme kontinuirano, često u sekundama, minutama ili satima. Npr. mjerenja senzora.

Lokacija	Temperatura	Vlažnost	Tlak	Brzina vjetra
u25kes6b3yt	2.5	84	1011	1.0
u241zykqmzp	8.2	56	1009	1.5
srexw0d9gbw	8.5	47	1007	2.7
u2j71kk6rnq	0.7	88	1012	3.7
srsfeey71bk	9.5	53	1006	1.6

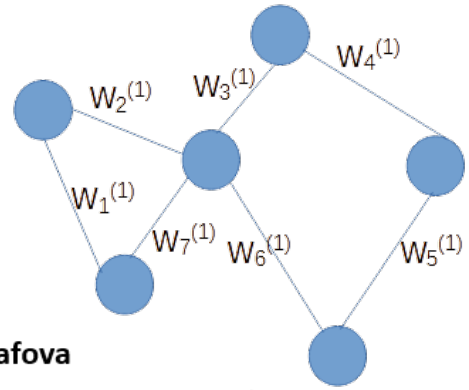
Podaci sa mrežnom strukturom



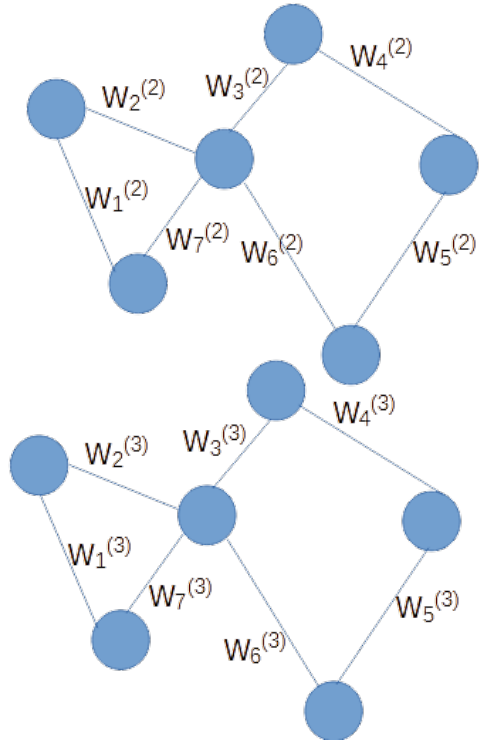
Težinski graf (mreža)



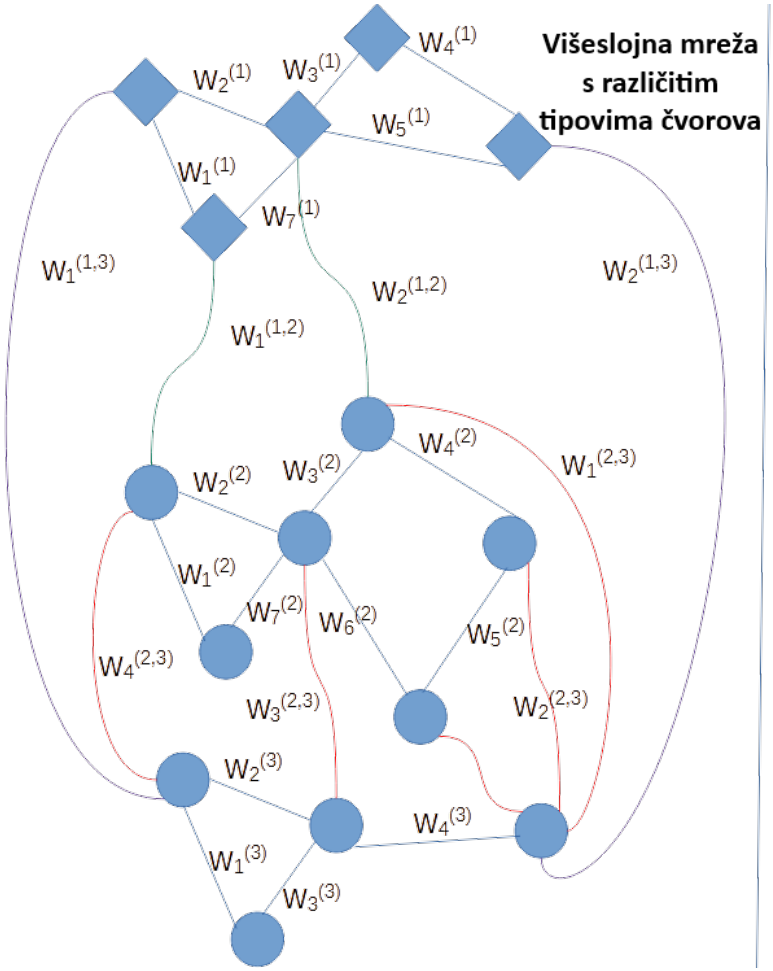
Višeslojna mreža s istim tipom čvorova

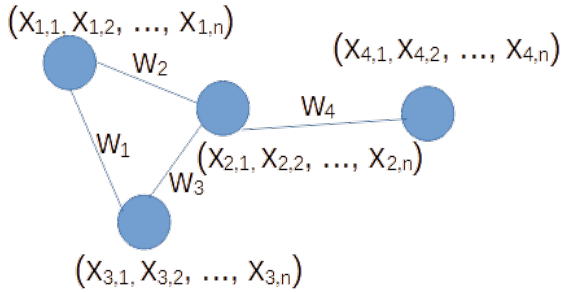


Niz grafova



Podaci sa mrežnom strukturom





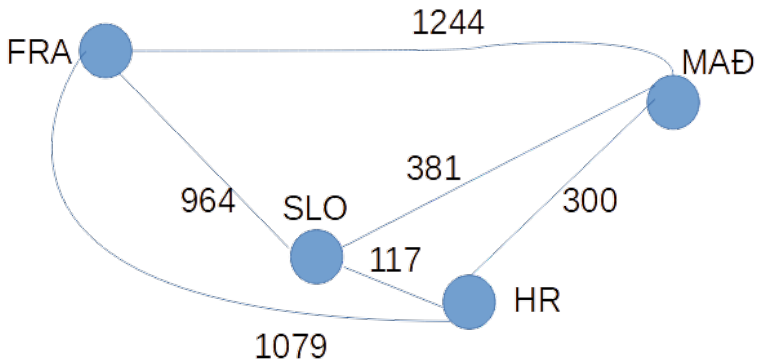
Težinski graf s atributima čvorova

Podaci s pozadinskim mrežnim znanjem

Podaci koji sadrže jedan ili više pogleda tabularnih podataka i dodatnu informaciju o mrežnoj povezanosti objekata (instanci, entiteta).

Entitet	Stanovništvo	%Stan <15	%Stan >64
HR	3,9M	14	23
SLO	2,1M	15	22
FRA	68,6M	17	22
MAĐ	9,6M	14	21

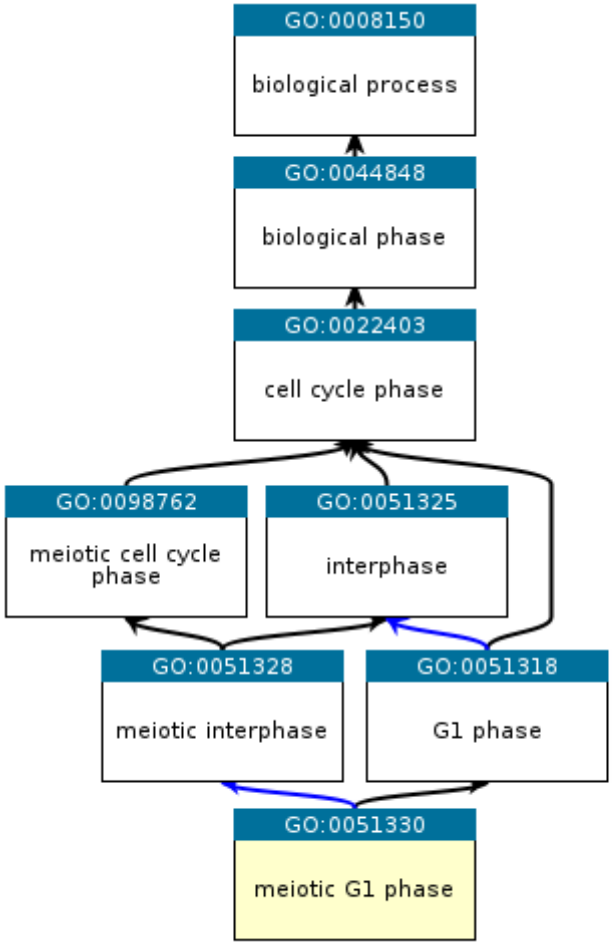
Entitet	More	Površina	Otoci
HR	DA	56,6K	1244
SLO	DA	20,3K	0
FRA	DA	632,7K	1300
MAĐ	NE	93K	0

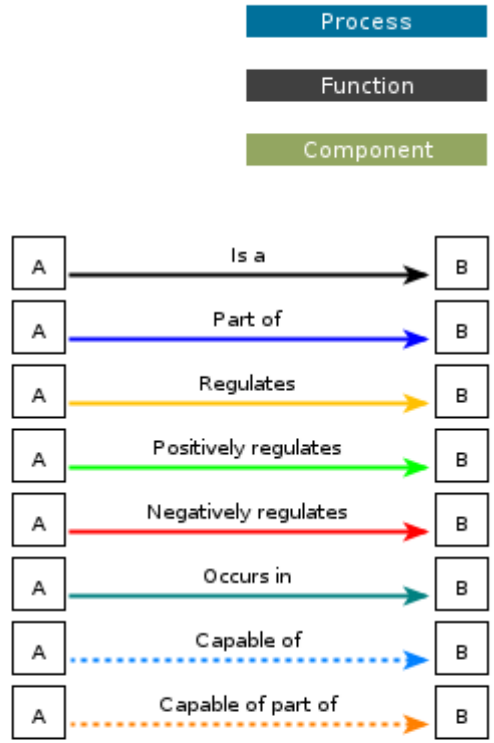


Ontologije

Ontologije definiraju pojmove potrebne za opisivanje znanja u određenoj domeni.

Najčešće imaju oblik hijerarhije, međutim mogu biti i ograničeni grafovi.





Tekstualni podaci

Nestrukturirani tekstualni zapis (npr. odjeljak, tekstualni dokument, knjiga).

```
Teče i teče, teče jedan slap;
Što u njem znači moja mala kap?

Gle, jedna duga u vodi se stvara,
I sja i dršće u hiljadu šara.

Taj san u slapu da bi mogo sjati,
I moja kaplja pomaže ga tkati.",
```

Tekstualni podaci

- Klasični algoritmi strojnog učenja i dubinske analize podataka ne mogu standardno koristiti nestrukturirani ulaz.
- Najnoviji modeli dubokog učenja, mogu koristiti tekstualne podatke u nestrukturiranom obliku.

Tekstualne podatke prije upotrebe u klasičnom strojnom učenju ili dubinskoj analizi podataka prvo strukturiramo koristeći neku od dostupnih metoda. Npr. možemo gornji tekst strukturirati koristeći TF-IDF prikaz.

Tekstualni podaci

Riječ	TF-IDF
bi	0.15617376
da	0.15617376
dršće	0.15617376
duga	0.15617376
ga	0.15617376
gle	0.15617376
hiljadu	0.15617376
jedan	0.15617376
jedna	0.15617376
kap	0.15617376
kaplja	0.15617376
mala	0.15617376
mogo	0.15617376
moja	0.31234752
njem	0.15617376
pomaže	0.15617376
san	0.15617376
se	0.15617376
sja	0.15617376
sjati	0.15617376
slap	0.15617376
slapu	0.15617376
stvara	0.15617376
taj	0.15617376
teče	0.46852129
tkati	0.15617376
vodi	0.15617376
znači	0.15617376
šara	0.15617376
što	0.15617376

Slike

Slike reprezentiramo kao tenzore. Svaki piksel se reprezentira trojkom (r, g, b). Za sliku dimenzija 640×360 piksela, odgovarajući tenzor je dimenzija $(360, 640, 3)$.

Slike

Isječak reprezentacije gornje slike:

```
[[[ 65  64   8]
  [ 64  63   7]
  [ 64  63   7]
  ...
  [ 71  72  14]
  [ 68  69  11]
  [ 64  65   7]]

 [[ 65  64   8]
  [ 64  63   7]
  [ 64  63   7]
  ...
  [149 137 125]
  [159 146 138]
  [125 112 106]]]
```

Slike

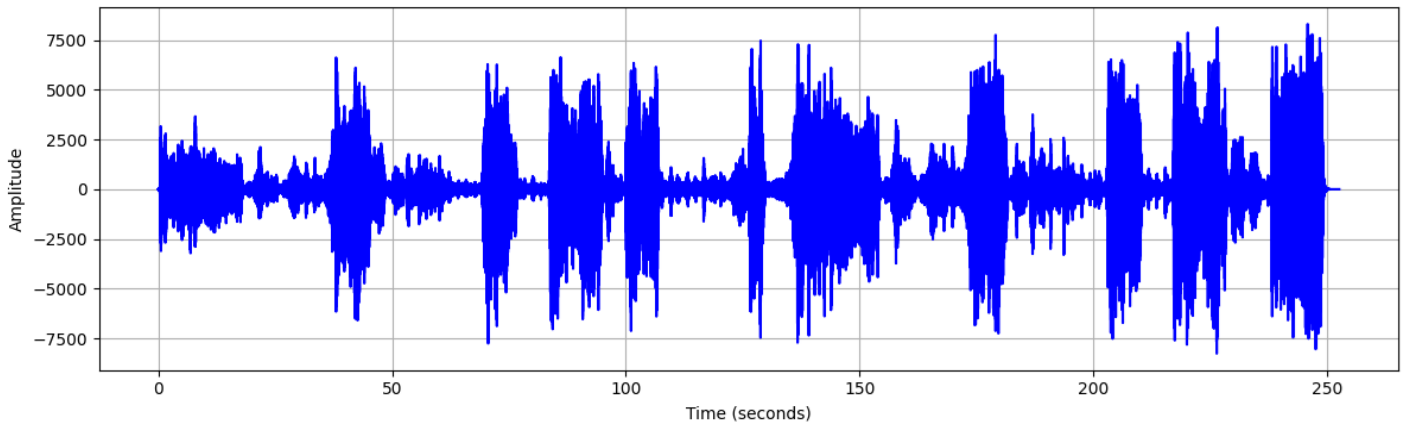
Većina klasičnih algoritama strojnog učenja i dubinske analize podataka ne radi dobro koristeći direktnu reprezentaciju slike (ulazni tenzor).

Relativno poboljšanje možemo dobiti računanjem značajki, npr. korištenjem metode HOG (eng. Histogram of Oriented Gradients).

Audio

Zvuk spremamo u računalo tako da uzorkujemo amplitudu zvučnog vala nekoliko puta svake sekunde. Veći broj uzoraka daje prikaz vjerniji originalnom zvučnom valu.

Pusti glazbu



Uzorak	1	2	3	4	5	6	7	8	9	10	...	11149072
Amplituda	-3	-2	-4	-2	-5	-3	-5	-3	-5	-5	...	0

Video

Pusti video

Video se sastoji od niza sličica (eng. frames) koje prate audio i potencijalno tekst (eng. titles).

Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7	Frame 8	Frame 9	Frame 10

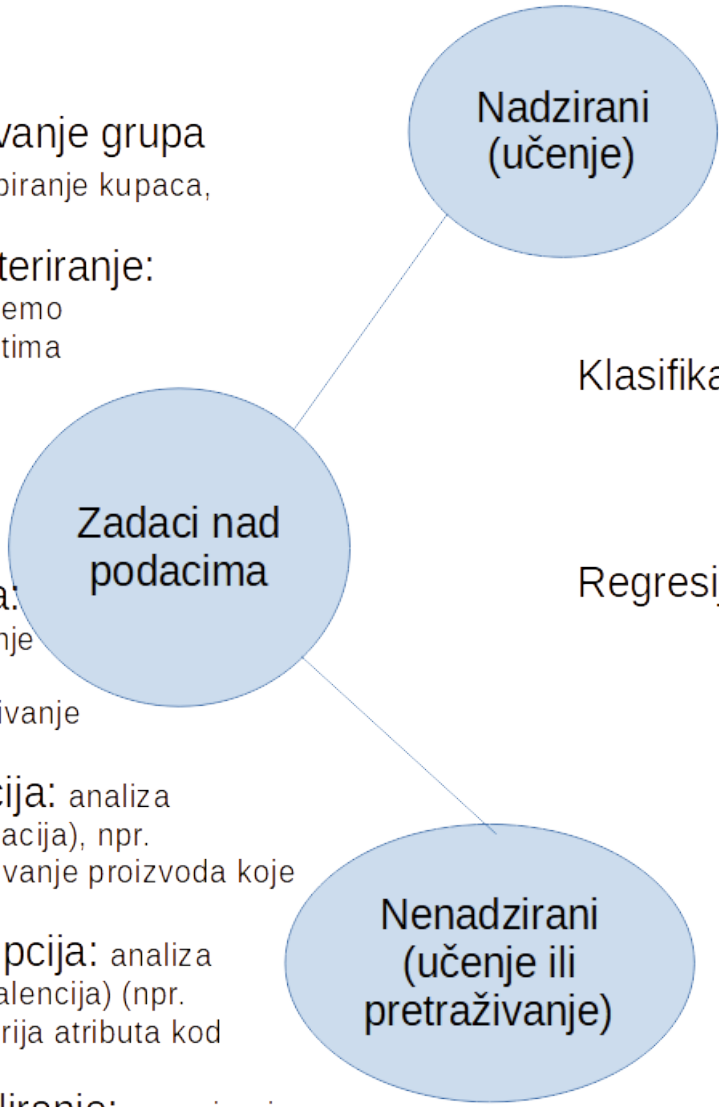
U	1	2	. . .	999999	10^6
A	1	2	. . .	-1	0

Neki tekst koji opisuje video	. . .	(.srt, .vtt formati)

Prva i 10000. sličica u videu:

Vrste zadataka

- Klasteriranje (otkrivanje grupa sličnih objekata): grupiranje kupaca, otkrivanje zajednica
- Konceptualno klasteriranje: otkrivanje grupa koje možemo opisati određenim konceptima
- Redukcija dimenzionalnosti: analiza strukture, konstrukcija značajki, preporučivanje
- Detekcija anomalija: detekcija prijevare, praćenje pacijenata, praćenje performansi sustava, otkrivanje pogrešaka u podacima.
- Otkrivanje asocijacija: analiza asocijacija atributa (implikacija), npr. potrošačke košarice (otkrivanje proizvoda koje kupujemo zajedno).
- Otkrivanje redeskripcija: analiza asocijacija atributa (ekvivalencija) (npr. povezanost raznih kategorija atributa kod ciljanih bolesti)
- Generativno modeliranje: generiranje glazbe, teksta, slika



E	A_1	A_2	. . .	Klasa
E_1	$a_{1,1}$	$a_{1,2}$	. . .	c_1
E_2	$a_{2,1}$	$a_{2,2}$	. . .	c_2
.	.	.	. . .	.
.	.	.		.
.	.	.		.

Klasifikacija (kategorijska ili binarna klasa): predviđanje funkcije gena, spola, Medicinska dijagnoza, detekcija prijevare, klasifikacija/segmentacija nad slikama, traženje podgrupa.

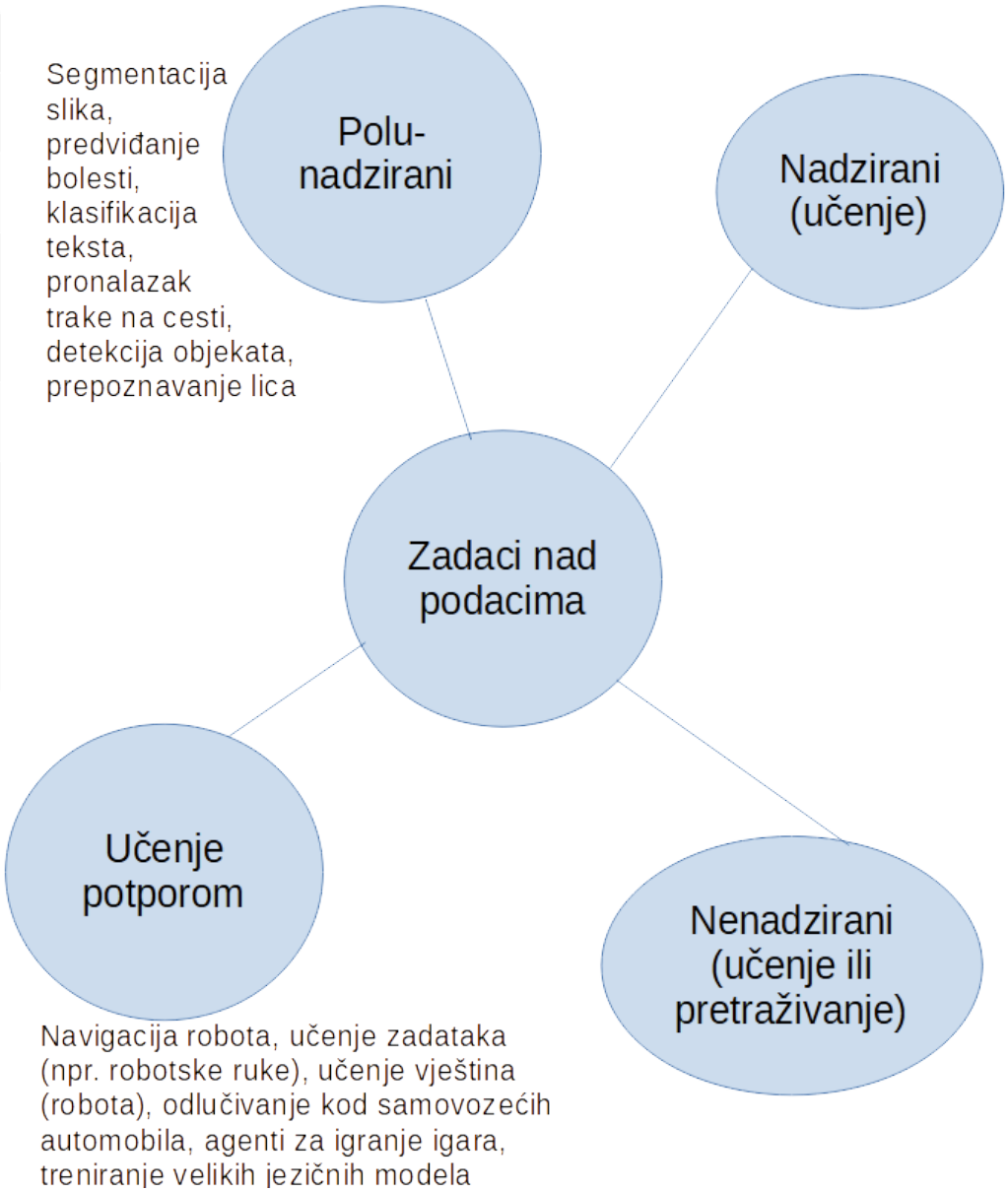
Regresija (klasa iz domene realnih brojeva): Vremenska prognoza, procjena cijene, iznosa trgovine, količine prodanih proizvoda.

E	A_1	A_2	. . .	A_n
E_1	$a_{1,1}$	$a_{1,2}$	. . .	$a_{1.n}$
E_2	$a_{2,1}$	$a_{2,2}$	. . .	$a_{2,n}$
.	.	.	. . .	.
.	.	.		.
.	.	.		.

Vrste zadataka

E	A_1	A_2	. . .	Klasa
E_1	$a_{1,1}$	$a_{1,2}$	. . .	c_1
E_2	$a_{2,1}$	$a_{2,2}$	. . .	c_2
. . .	. . .	. . .	. . .	. . .
E_k	$a_{k,1}$	$a_{k,2}$	. . .	?
. . .	. . .	. . .	. . .	. . .
E_n	$a_{n,1}$	$a_{n,2}$	. . .	?

S	A_t	R_{t+1}	S_{t+1}	K
s_1	a_1	r_2	s_2	NE
s_2	a_2	r_3	s_3	NE
. . .	. . .	. . .	. . .	. . .



Dubinska analiza podataka

Dubinska analiza podataka podrazumijeva **skup** softverskih mehanizama i postupaka čiji cilj je izdvajanje **skrivenih** informacija iz podataka.

- skrivene informacije su teško dohvatljive informacije. U ovoj definiciji sve informacije dohvatljive npr. SQL upitima smatramo jednostavno dohvatljivima.
- pojam informacije interpretiramo u najširem obliku (korisno znanje, može biti tekstualno, slikovno itd.).

Dubinska analiza podataka se kao termin javlja krajm 80-tih godina prošlog stoljeća, a 90-tih se smatra pod-procesom većeg procesa koji se naziva *Otkrivanje znanja iz podataka* (eng. Knowledge discovery from Data - KDD).

Fayyad i ostali 1996. opisuju KDD kao ne trivijalni postupak identificiranja validnih, novih, potencijalno korisnih i konačno razumljivih uzoraka u podatcima.

Dubinska analiza podataka

Otkrivanje znanja iz podataka se sastoji od:

- Pripreme podataka
  - pohrana podataka
  - čišćenje podataka
  - pred-procesiranje podataka
- statističke analiza i obrade podataka (npr. normalizacija)
- dubinska analiza podataka
- analize i vizualizacije rezultata

Detaljnije o tabličnim podacima

Tablični podaci mogu sadržavati **binarne**, **kategoričke** i **numeričke** atribute.

- Binarni atributi reprezentiraju svojstva koja se mogu opisati kratkim "DA" ili "NE" (npr. *Pušač*, *Sportaš*, *Oženjen* itd). U tablici, vrijednosti binarnih atributa najčešće imaju string "DA"/"NE" ili cijelobrojne vrijednosti 1/0.
- Kategorijske atribute koristimo kada atribut sadrži dvije ili više vrijednosti različitih od "DA", "NE" (npr. $boja \in \{$crvena, zelena, žuta, plava, ...$\}$. U tablici, takvi atributi najčešće sadrže string koji označava vrijednost kategorije "crvena"/"zelena"/"žuta"/"plava" ili cjelobrojnu numeričku oznaku koja reprezentira kategoriju 1/2/3/4...
- Numerički atributi sadrže realne vrijednosti i najčešće reprezentiraju rezultate mjerenja (npr. visina, težina, tlak, promjer, koncentracija). U tablici, vrijednosti numeričkih atributa su realni brojevi određene preciznosti (najčešće C-ovskog tipa `double`).

Detaljnije o tabličnim podacima

Entitet	Veza	Boja kose	BMI
Marko	DA	Plava	23.7
Luka	NE	Smeđa	24.6
Ivana	NE	Crna	22.1
Marta	DA	Crvena	21.4

Detaljnije o tabličnim podacima

Moguće je da neki podaci u tablicama **nedostaju**. Pripadne vrijednosti se zovu **nedostajuće vrijednosti** i označavaju se s ?.

Glavni razlozi nedostajanja podataka su:

- Greške pri mjerenjima (npr. kontaminirani uzorci, greške i kvarovi uređaja, iznimna stanja entiteta)
- Gubitak podataka (npr. brisanje baze, hakerski napadi, ne izvršena mjerenja za određene entitete)
- Ne uneseni podaci (npr. postojeća mjerenja koja zabunom ili namjerno nisu unesena u bazu)

Problem **nedostajućih vrijednosti** rješavamo: a) razvojem dediciranih metoda koje znaju baratati s njima, b) postupkom **imputacije** (nedostajuće vrijednosti se nadopunjavaju statističkim ili metodama strojnog učenja na temelju distribucija prisutnih vrijednosti u skupu, kod nekih metoda i iz mnogih drugih skupova).

Detaljnije o tabličnim podacima

Entitet	Veza	Boja kose	BMI
Marko	DA	Plava	23.7
Luka	NE	Smeđa	?
Ivana	?	Crna	22.1
Marta	DA	Crvena	?

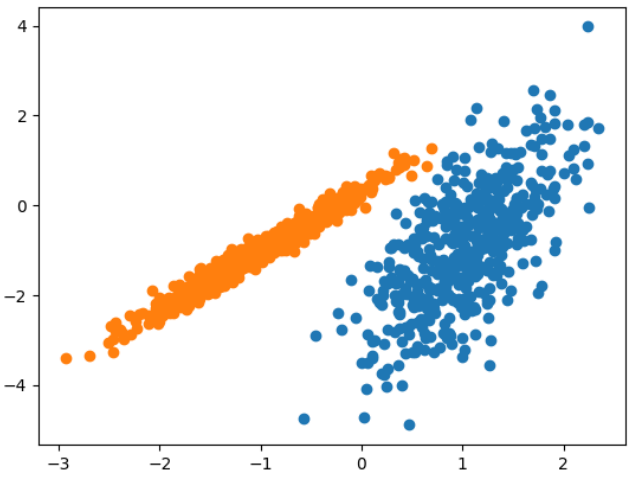
Detaljnije o tabličnim podacima

Ulazni podaci uglavnom sadrže i određenu količinu **šuma** (eng. noise). Šum nastaje zbog nepreciznosti u mjerenjima, interferencija (npr. elektromagnetskih), iznimnih stanja entiteta, pogrešaka pri unosu podataka.

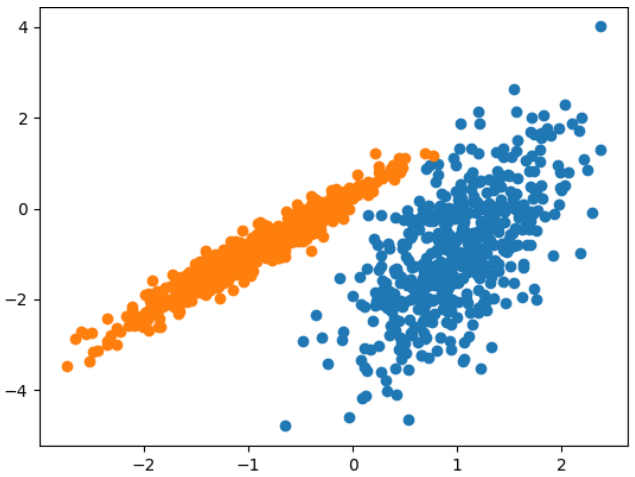
Ukoliko s D označimo čiste ulazne podatke, tada $D' = D + \sigma \cdot \mathcal{X}$ označava ulazne podatke koje uočavamo. $\mathcal{X}$ označava neku distribuciju vrijednosti koje dobivamo zbog šuma, dok σ označava intenzitet šuma, $\sigma \in [0, 1]$.

- Jaki šum u podacima može značajno pogoršati performanse algoritama dubinske analize podataka i strojnog učenja.
- Dodavanje određene količine šuma podacima može poboljšati treniranje i točnost složenijih modela dubokog učenja.

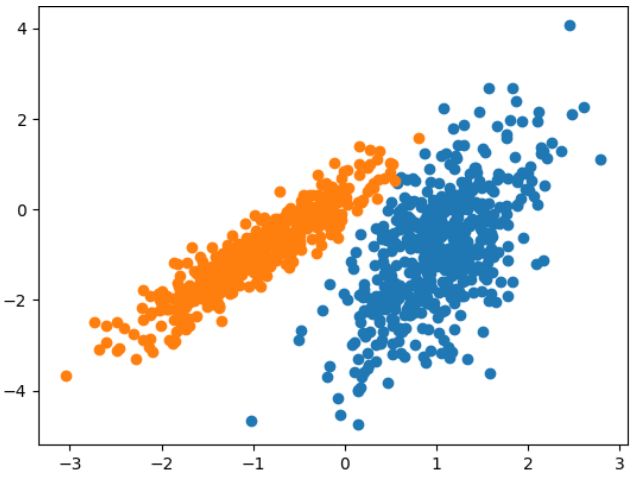
Utjecaj šuma na grupiranje elemenata



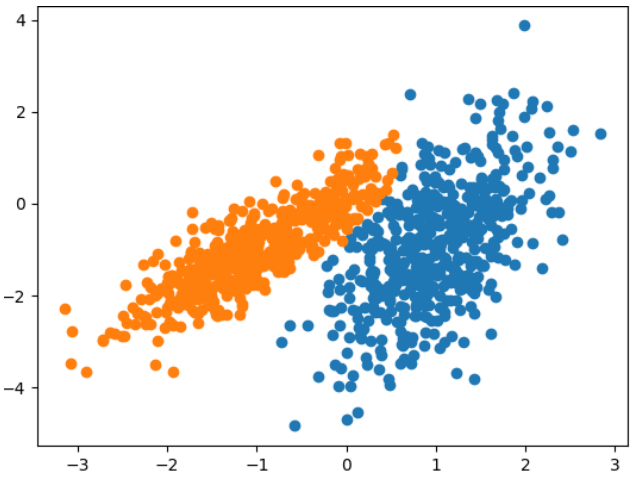
(a) Original



(b) 10% šuma

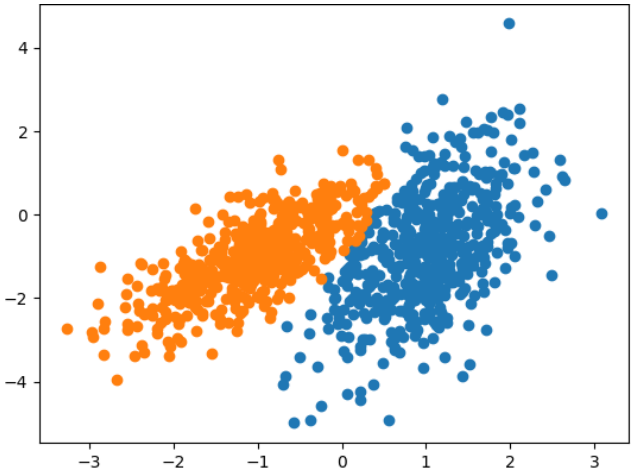


(c) 20% šuma

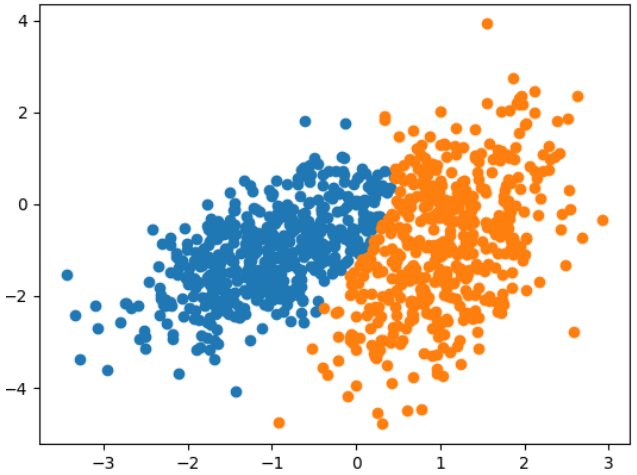


(d) 30% šuma

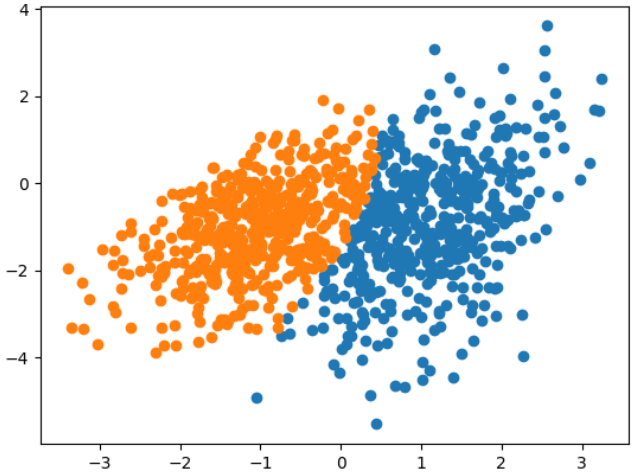
Utjecaj šuma na grupiranje elemenata



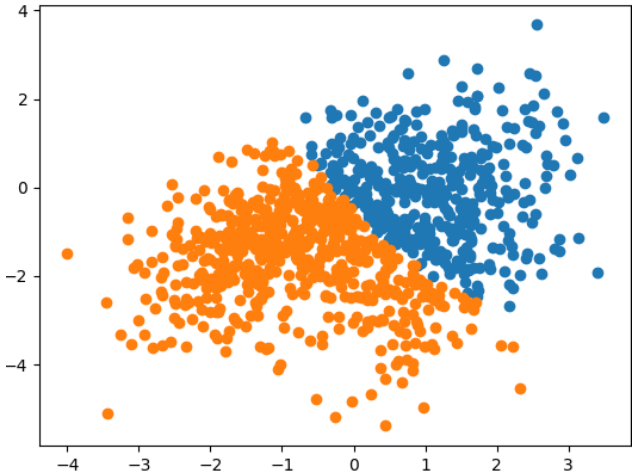
(e) 40% šuma



(f) 50% šuma

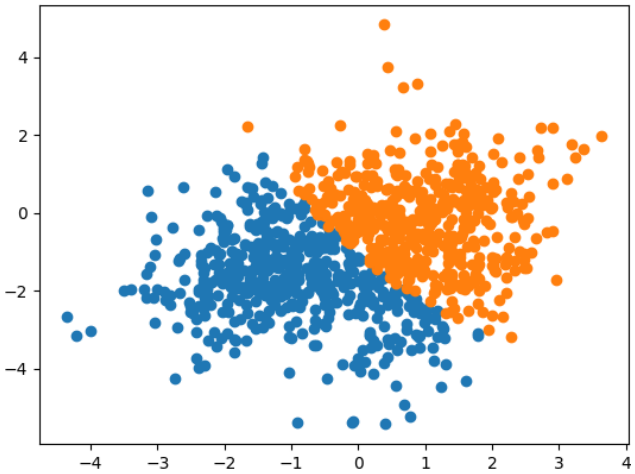


(g) 60% šuma

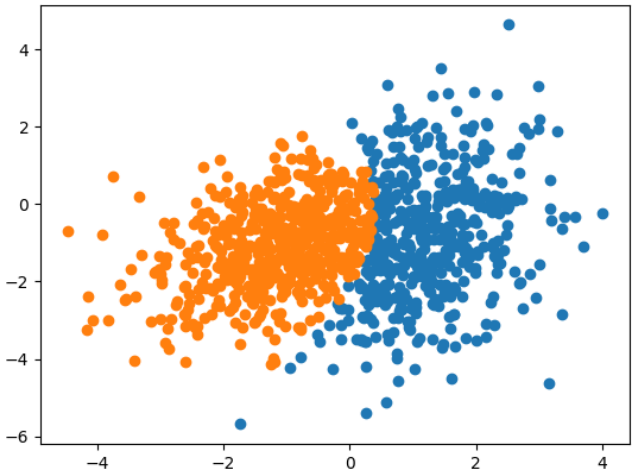


(h) 70% šuma

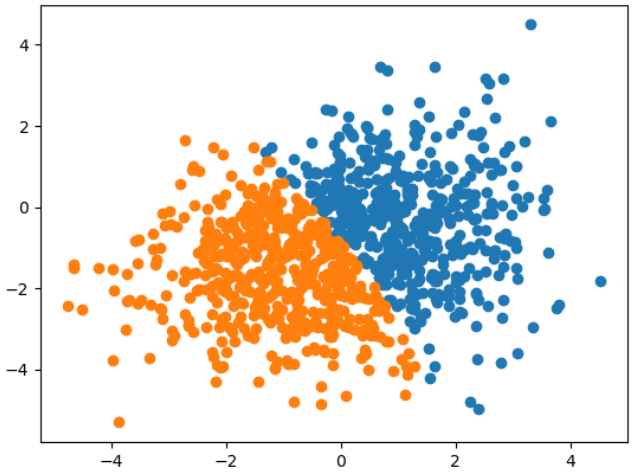
Utjecaj šuma na grupiranje elemenata



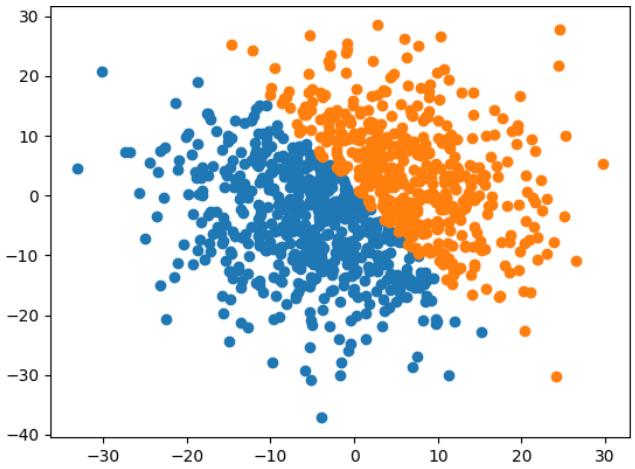
(i) 80% šuma



(j) 90% šuma



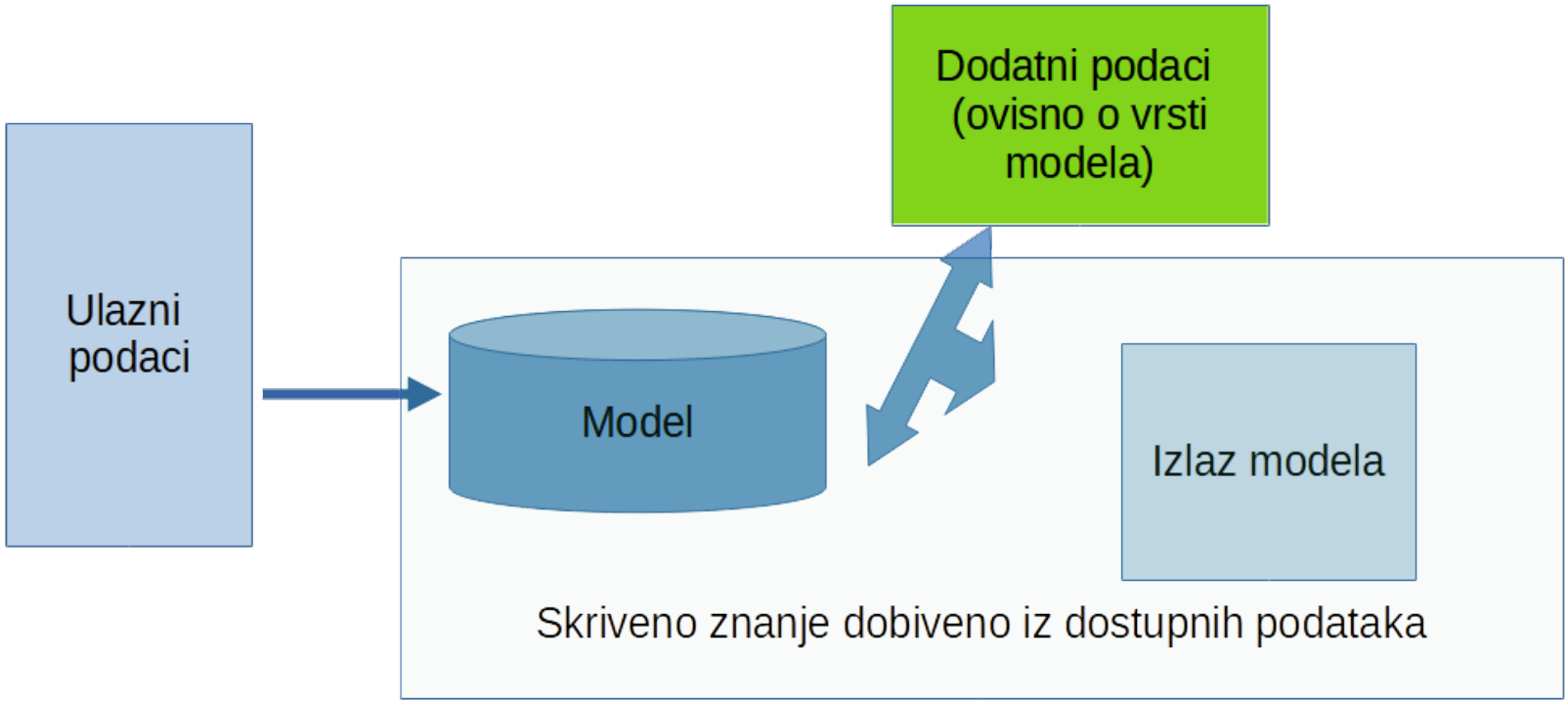
(k) 100% šuma



(l) 1000% šuma

Model

Model je struktura koja sadrži statističke veze, uzorke i/ili pravila, dobivene iz ulaznih podataka. Koristeći navedene informacije, model pruža korisniku određene uvide u skriveno znanje sadržanom u dostupnim podacima.



Vrste modela

- Prediktivni - koristimo informacije iz modela za predviđanje informacija iz do sada neviđenih podataka (npr. informacije modela treniranog na skupu podataka koji opisuje osobe i sadrži informaciju o tome boluje li osoba od dijabetisa, koristimo za predviđanje bolesti kod nove skupine osoba, gdje imamo opise osoba, međutim nemamo informaciju o prisutnosti bolesti). Neke vrste prediktivnih modela:
  - Klasifikacijski modeli - pridjeljuju kategoriju neviđenom entitetu (npr. pas, mačka, krava...)
  - Regresijski modeli - pridjeljuju numerički broj neviđenom entitetu (npr. sutrašnja cijena Intelovih dionica)

Vrste modela

- Prediktivni:
  - Polunadzirani modeli - pridjeljuju kategorijsku ili numeričku oznaku (ovisno o vrsti problema). Za razliku od modela dobivenih iz klasifikacijskih ili regresijskih problema, polunadzirani modeli se treniraju iz skupova podataka kod kojih obično manji dio entiteta sadrži informaciju o ciljnoj labeli (odnosno klasi), dok veći dio entiteta nema informaciju o ciljnoj labeli.
  - Segmentacijski modeli - detektiraju smislene dijelove (cjeline) u ulaznim podacima (npr. ljude, semafore, ceste na slikama ili videu, djelove rečenice u tekstu).

Vrste modela

- Deskriptivni - koristimo informacije iz modela za bolje razumijevanje podataka (npr. otkrivanje sličnih podskupova entiteta, podgrupa, asocijacija između atributa, opisivanje objekata).
  - Klasteriranje - traže disjunktne grupe entiteta, gdje su entiteti iz istih grupa jako međusobno slični s obzirom na vrijednosti atributa, a entiteti iz različitih grupa nisu slični ili je sličnost relativno mala.
  - Konceptualno klasteriranje - ima isti cilj kao i klasteriranje, međutim dodatno teži pronaći i koncept koji opisuje entitete sadržane u grupama.
  - Traženje frekventnih i zatvorenih skupova elemenata - skupovi binarnih atributa koji se često (frekventno) javljaju zajedno, uz to zatvoreni skupovi nemaju nadskupove koji vrijede za identični skup entiteta (transakcija).

Vrste modela

- Deskriptivni:
  - Traženje asocijacija - otkriva implikacijske asocijacije između atributa u skupovima podataka. Koristi se npr. kod analize kupovnih navika.
  - Traženje redeskripcija - otkriva ekvivalencijske asocijacije između atributa u skupovima podataka. Koristi se za bolje razumijevanje domenskih podataka (biologija, medicina, ekologija, ekonomija itd.).
  - Traženje podgrupa - na razmeđi između prediktivnih i deskriptivnih zadataka. Otkriva podgrupe entiteta čija distribucija ciljne labele znatno odstupa od distribucije ciljne labele na cijelom skupu. Uz to opisuje detektirane podgrupe logičkim pravilima (u principu koristeći logičke konjunkcije).