

Rukovanje podacima

Uvod u znanost o podacima

2. predavanje

Ak. god. 2021./2022.



Sadržaj

- Rukovanje podacima – koraci procesa
- Problemi skupova podataka
- Inženjerstvo značajki
- Zaključak

Rukovanje podacima



Rukovanje podacima – koraci procesa

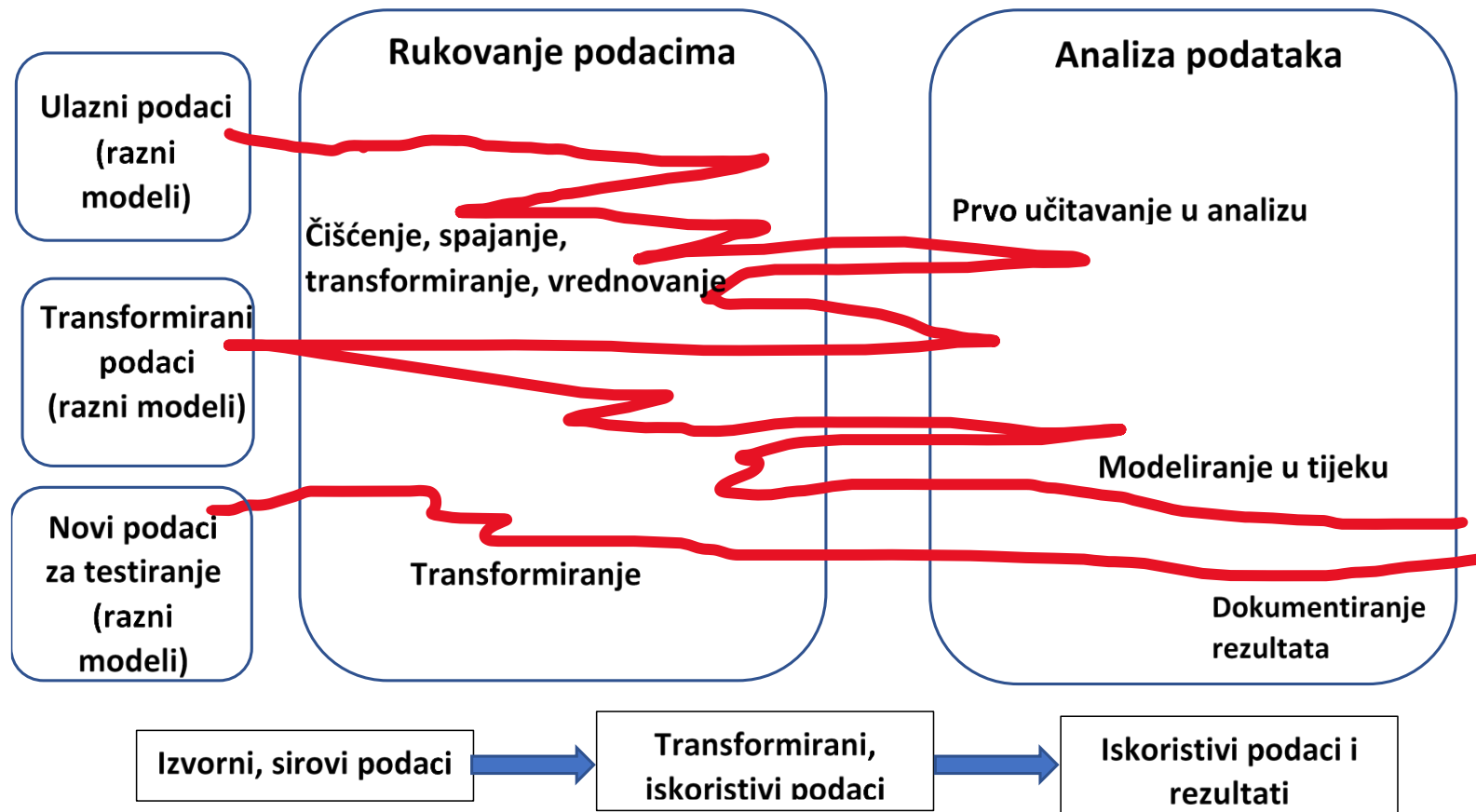
Rukovanje podacima

- Engl. *data handling*
- Općeniti naziv za sve operacije nad podacima koji slijede nakon preuzimanja izvornih podataka s mjesta pohrane sve do početka analize statističkim postupcima i postupcima strojnog učenja
- Alternativni nazivi (suptilne razlike):
 - **Priprema podataka** (engl. *data preparation*) – podaci se pripremaju za neku vrstu analize
 - **Organiziranje podataka** (engl. *data wrangling*) – doslovno: svađanje između podataka
 - **Upotpunjavanje podataka** (engl. *data munging*) – povijesno, „mung” je pojam za progresivnu degradaciju skupa podataka – *backronym* od „mash until no good”
- <https://www.talend.com/resources/what-is-data-preparation/>

Rukovanje podacima

- Korištenje sirovih podataka u daljnjoj statističkoj analizi bez razmišljanja o njima – **recept za katastrofu!**
 - Može onemogućiti ispravno postavljanje cilja analize
 - Može srušiti algoritme strojnog učenja ili davati nevjerodostojne statistike
 - Može dovesti do neispravnih zaključaka
- Na rukovanje podacima odlazi 50% – 80% ukupnog vremena (i novca) tijekom projekta s ciljem otkrivanja znanja u podacima
- Rukovanje podacima je, zajedno s pohranom podataka, temeljno područje kojim se bavi **podatkovni inženjer**
- Cilj rukovanja podacima: **pripremiti podatke da postanu pouzdani i iskoristivi**

Rukovanje podacima



- Crtež koji je teško opisati ali izvršno pogađa u suštinu rukovanja podacima, jer je taj proces:

- Ekstremno *ad-hoc* u svojoj provedbi
- Bez savršenog recepta
- Takav da zahtijeva puno razmišljanja i zdrave logike
- Najčešće necijenjen u tvrtakama, obično se vrednuje samo modeliranje i **rezultati**

Prilagođeno iz: EPFL, ADA, 2020.

Proces rukovanja podacima

- Rukovanje podacima sastoji se od sljedećih važnih koraka:
 - 1. Pregled skupa podataka** (engl. *data survey, data exploration, data learning*)
 - Vizualna i statistička dijagnostika skupa podataka (pa i ručno pregledavanje brojaka)
 - Cilj je upoznati se s podacima i ustanoviti njihove nedostatke
 - 2. Transformacija skupa podataka** (engl. *data transformation, data organizing, data assembly*)
 - Transformacija modela, formata i dimenzija podataka u oblik koristan za analizu
 - Najprije transformacija u relacijski, tablični oblik
 - Može uključivati i pronalazak i spajanje s dodatnim izvorima podataka (engl. *data enrichment, data merging*)
 - kada svi podaci nisu na jednom mjestu
 - U slučajevima manjih analiza i lokalno dostupnih podataka, može se preskočiti

Proces rukovanja podacima

- Rukovanje podacima sastoji se od sljedećih važnih koraka:

3. Čišćenje skupa podataka (engl. *data cleaning*)

- Pronalazak i uklanjanje pogrešaka, duplikata, sinonima, stršćih vrijednosti, nedostajućih vrijednosti i drugih problema u podacima

4. Provjera skupa podataka (engl. *data validation, data authentication*)

- Nakon svih prethodnih koraka, provjera jesu li podaci sada ispravni,
- Ponekad implicitno uključeno u sve prethodne korake, ali se često koristi kao zaseban korak
- Detaljnije što se sve provjerav:

<https://corporatefinanceinstitute.com/resources/knowledge/data-analysis/data-validation/>

Proces rukovanja podacima

- Rukovanje podacima sastoji se od sljedećih važnih koraka:
 - 5. Učitavanje skupa podataka** (engl. *data loading*) – opcionalno kao zasebni korak
 - Podaci se učitavaju u strukturu podataka pogodnu za daljnju analizu (ako su ranije mijenjani na nekom drugom mjestu ili u nekom drugom formatu)

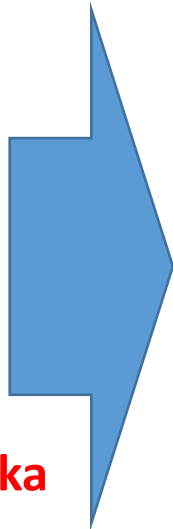
6. Poboljšavanje skupa podataka (engl. *data augmentation*)

- Izmjene u veličini i raznolikosti primjera skupa podataka

7. Inženjerstvo značajki (engl. *feature engineering*)

- Rad na značajkama vezanima uz skupa podataka

Zadnja dva koraka mogu biti **dio rukovanja podataka**, ali već i **analize podataka**



Predobrada podataka (engl. *data preprocessing*)

- „nešto” što se događa s dotad pripremljenim skupom prije „prave” obrade/analize podataka

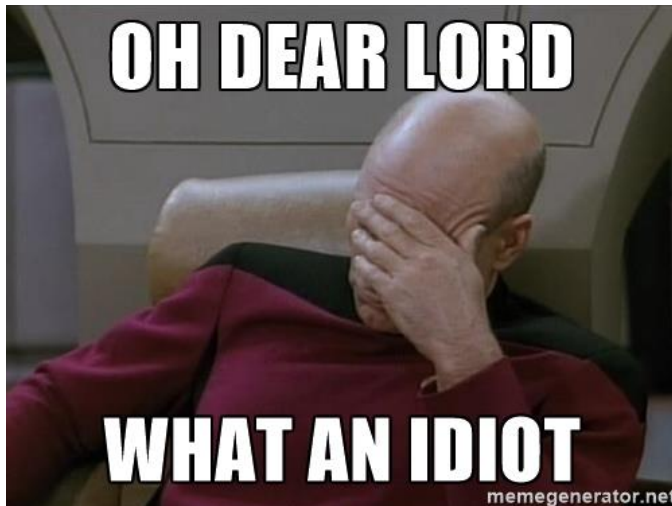
Problemi u podacima

Vrste čestih problema u podacima

- **Nedostajući podaci** (engl. *missing data*)
- **Netočni podaci** (engl. *incorrect data*)
- **Nekonzistentnosti u podacima** (engl. *inconsistent data*)
- **Stršeći podaci** (engl. *outliers*)
- **Rijetki podaci** (engl. *sparse data*)
- **Šumoviti podaci** (engl. *noisy data*)
- **Monotoni atributi** (engl. *monotonic attributes*)
- **Nebalansirani skupovi podataka** (engl. *imbalanced datasets*) → Tema kasnijih predavanja

Oko 75% problema u podacima zahtijeva ljudsku intervenciju da ih se ispravi (npr. stručnjaci u području, *crowdsourcing*)

Horor priče o “prljavim podacima”



- Pismo “Dear Idiot”
- 17,000 muškaraca su trudni
- Direktan put („As the crow flies”)

<https://www.linkedin.com/pulse/dirty-data-horror-stories-when-michael/>

Poanta: značajna količina podataka u tvrtkama je „loša” (10–25%, ovisno o tvrtci, različite procjene)

Nedostajući podaci

Dvije glavne vrste:

- **Nedostajuće (ali poznate) vrijednosti**

- Vrijednosti koje nisu unesene u skup podataka, ali postoje u stvarnom procesu

- **Prazne (nepoznate) vrijednosti**

- Ne može se pretpostaviti vrijednost u stvarnom svijetu i vrijednost nije unesena

- Često nije jasno o kojoj se vrsti radi

- razne konkretne vrijednosti pohranjene na mjestu nedostajućeg podatka
 - " – prazno polje, '-', 'x', 'NULL', 'N/A', 'BLANK', '„' – razne vrste navodnika, '?', '???' ...

- Detekcija problema **detaljnim pregledom skupa podataka** ili **korištenjem vizualizacije**

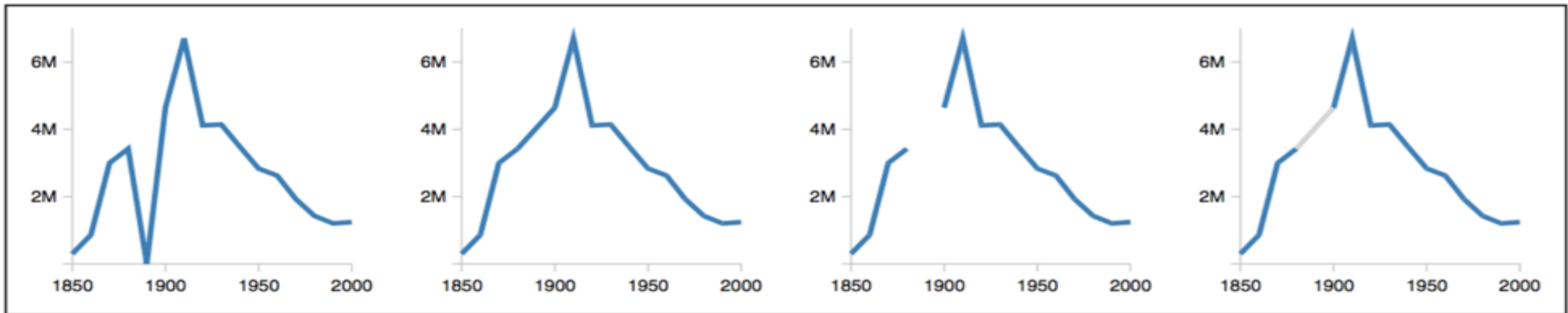
- Rješavanje problema nedostajućih podataka u praksi je najčešće neovisno o vrsti

Nedostajući podaci – rješavanje problema

- **Zanemarivanje svih primjera (objekata) koji ih sadrže**
 - Ponekad nije moguće, npr. kada većina primjera ima vrijednost nekog atributa nedostajuću – tada je možda bolje maknuti takav atribut
- **Zamijeniti nedostajuću vrijednost nekom drugom**
 - **Uz osiguranje da informacijski sadržaj skupa podataka ne degradira**
 - Jednostavni postupci promatraju jedan atribut
 - **Očuvati mjeru sredine** – zamjena sa srednjom vrijednosti, medianom ili najčešćom kategorijom
 - **Očuvati varijabilnost** – po potrebi dodati šum pri zamjeni da se očuva varijabilnost
 - Složeniji postupci promatraju odnos između više atributa i biraju zamjenu koja će najmanje utjecati na čitav skup
 - Npr. regresija, algoritam k -najbližih susjeda

Primjer

- Popis stanovništva SAD-a na kojem su prikazani oni stanovnici koji rade kao „zaposlenici na farmi”, podaci od 1890. nedostaju jer su zapisi izgorjeli



- Postaviti vrijednost na 0?
- Interpolirati na temelju bliskih podataka?
- Posve zanemariti nedostajuće podatke?

Znanje o domeni i načinu prikupljanja podataka treba voditi izbor metode zamjene!

Netočni podaci

- Često kao rezultat zabune pri unosu
- Neki put namjerno unešeni
 - Korisnik ne zna točnu informaciju a ne želi ostaviti prazno
 - Korisnik ne želi da netko drugi sazna točnu informaciju
 - Korisnik ima neku korist unašanjem netočne informacije
- Rijetko rezultat tehničke pogreške sustava (npr. neke izmjene u formatima podataka u bazi podataka)
- U općenitom slučaju, neriješiv problem
- **Zahtijeva detaljan pregled skupa podataka, vizualizaciju i promišljanje o podacima**

Nekonzistentnost u podacima

Dva tipa nekonzistentnosti

- **Različiti atributi** mogu biti predstavljeni **istim imenom** u različitim sustavima
 - Problem se pojavljuje pri povezivanju podataka iz određenog broja različitih sustava u jednu tablicu
 - Pojavljuju se naizgled duplikati vrijednosti ili atributa, koji to **uopće ne moraju biti**
 - Potrebno je ručno uklanjanje / odabir atributa
- Jedan atribut može imati više **sinonima**, u jednom ili više sustava
 - Problem istoznačnica je vrlo često prisutan i ispravlja se samo nakon što se vizualno uoči
 - Teško se otkrivaju ili ispravljaju bilo kojim automatskim postupcima u općenitom slučaju
 - Npr, „zaigrani“ zaposlenici u auto-tvrtki pod atributom *car_type* upisuju vrijednosti: „Merc“, „Mercedes“, „M-Benz“, „Mrcds“, umjesto jednog tipa automobila: „Mercedes“ – ovome je moguće doskočiti ispravno izrađenim korisničkim sučeljem i naputcima za zaposlenike, naknadne promjene su teške

Stršćeći podaci

- Podaci koji odskaču (odudaraju) **daleko izvan uobičajenih vrijednosti** za određene attribute
- Razlozi pojave ovakvih podataka: neispravan unos, greške mjerenja, greške obrade podataka, prirodno stanje
- Problem ako su takvi podaci netočni – ako nisu rezultat prirodnog stanja
- Potrebno ih je pronaći i po potrebi ukloniti (ako se stručnjaci slože da ne prikazuju prirodno stanje)

Stršeći podaci

- **Korišteni postupci otkrivanja**
 - Vizualizacija podataka
 - Statistički postupci – z-skor, vjerojatnosni modeli, linearna regresija
 - Algoritmi nenadziranog strojnog učenja
 - Temeljeni na udaljenosti, gustoći, grupiranju, itd.
 - Normalizacija podataka poboljšava otkrivanje stršećih podataka
 - <https://link.springer.com/article/10.1007/s10618-019-00661-z>

Rijetki podaci

- Slučaj kada za neke attribute samo mali broj primjera ima vrijednost različitu od 0
 - Često kod skupova dobivenih analizom teksta i dokumenata
- Većina algoritama strojnog učenja **loše radi s rijetkim podacima**
 - Prenaučenost modela – loša generalizacija na testnim podacima, davanje prednosti ili zanemarivanje atributa s rijetkim podacima
- pristupi rješavanju problema
 - Uklanjanje atributa s rijetkim podacima
 - Smanjenje dimenzionalnosti – npr. analiza glavnih komponenti
 - Korištenje postupaka strojnog učenja otpornijih na rijetke podatke
 - Npr. *Entropy weighting k-means algorithm* - <https://ieeexplore.ieee.org/abstract/document/4262534>

Šumoviti podaci

- Šum u podacima (engl. *data noise*) je u nekoj mjeri prisutan u svim podacima koji su rezultat mjerenja putem određenih senzora
- **Podatak = pravi signal + šum**
 - Šum je rezultat utjecaja prirodnih procesa
 - Šum je rezultat nesavršenosti mjernih senzora
 - Prisutan i kod 1D, 2D i 3D signala pa i kod drugih, nevremenskih podataka
- Postoje postupci za redukciju šuma u podacima kada je omjer signal/šum nepovoljan
 - **Postupci su iznimno ovisni o konkretnom problemu**
 - Npr. korekcija pomaka nulte linije i gradske strujne mreže kod snimanja elektrokardiograma
- Neki put, šum nije moguće dovoljno ili do kraja ukloniti
 - Skup na kojem se gradi model treba imati ista statistička svojstva kao i skup podataka na kojem će se model kasnije testirati/primijeniti

Monotoni atributi

- **Monotoni atributi su takvi atributi čija vrijednost raste (ili se smanjuje) bez ograničenja**
- Najčešći primjeri
 - Atributi povezani s protjecanjem vremena, npr. datumi u raznim oblicima.
 - Atributi rednih brojeva različitih zapisa i sl.
- Problem je nemogućnost dobivanja korisne informacije iz takve serije
- Rješenja problema:
 - **Zanemariti takav atribut (najčešće)**
 - Transformirati u određeni oblik pogodan za modeliranje
 - Datum se može pretvoriti u godišnje doba ili dan u tjednu, koji se ciklički ponavljaju, ako postoji potreba za takvim podatkom
 - Datumu se može pristupiti kao vremenskoj seriji (nizu), čime se otvara mogućnost korištenja raznih drugih oblika analiza nad podacima

Savjeti prije početka analize podataka

- Upitati se: “Imam li **nedostajućih podataka**?” “Ako neki podaci nedostaju, kako ću to saznati?”
- “Sumnjam li na **iskvarene, loše** podatke?” (zbog grešaka u mjerenju, krivih strategija uzorkovanja, namjernih „pogrešaka” i sl.)
- **Obraditi/transformirati podatke** u odgovarajući format za svoju specifičnu analizu (vidi i 1. predavanje o modelima podataka)
- Ne iznenaditi se ako je potrebno vratiti se na ovaj korak nakon što je analiza već krenula!

Idealni skupovi podataka

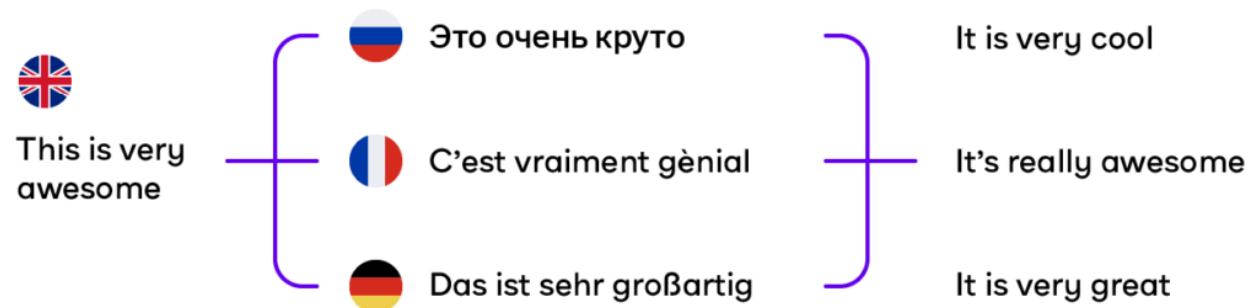
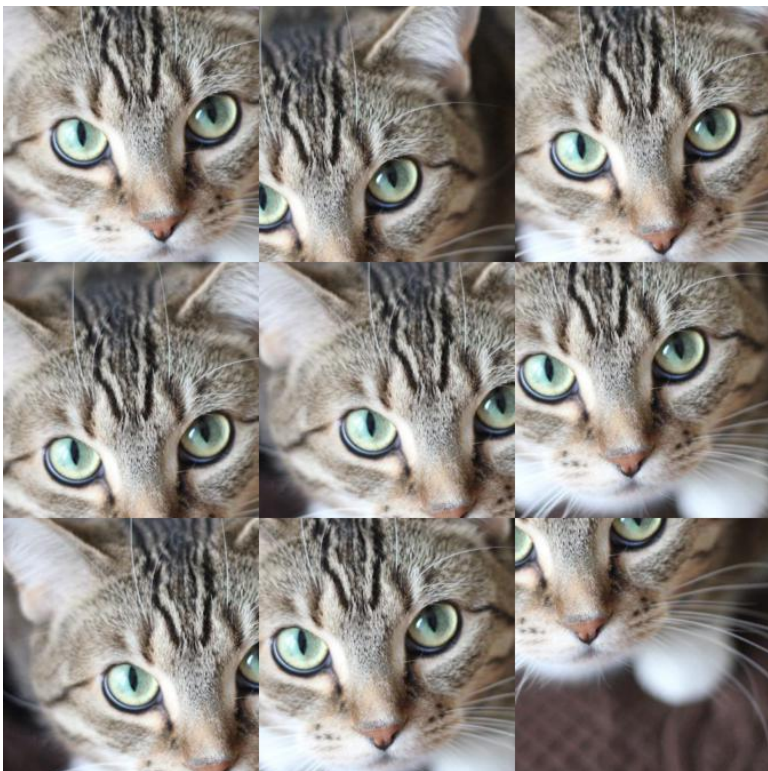
- Skupovi podataka koje prati **dokumentacija / stručni članak i programski kod**
- **Format podataka koji je lako obraditi**
 - Česti slučaj za strukturirane i polustrukturirane podatke, problem je za nestruktuirane (potrebni su regularni izrazi, plaćanje radne snage ili napredne metode ovisne o vrsti podataka)
- **Ranije očišćeni i pripremljeni skupovi podataka**
- U praksi: najčešće nemamo ništa ili vrlo malo od navedenoga

Poboljšavanje skupa podataka

- Poboljšavanje skupa podataka (engl. *data augmentation*) slijedi nakon faze rukovanja podataka, a prije analize podataka, zajedno s inženjerstvom značajki
- Za razliku od inženjerstva značajki, ovdje je fokus na **primjerima (objektima)**
- **Umjetno povećavanje broja primjera**
- Ne provodi se uvijek, već ovisno o potrebi
 - Češće ako se za analizu podataka koriste duboki modeli koji traže puno podataka
 - Rjeđe ako podataka ima dovoljno
 - Rjeđe ako su podaci dobro izbalansirani između klasa kod nadziranog učenja
 - Češće kod računalnog vida i obrade prirodnog jezika

Poboljšavanje skupa podataka

- Blisko povezano s tematikom „naduzorkovanja” (engl. *oversampling*)
- Generiranje novih sintetskih primjera
 - Izravne kopije starih primjera
 - S dodanim šumom nad starim primjerima
 - Na temelju najbližih susjeda
 - Transformacije starih primjera
 - Kod slika: rotacija, translacija, skaliranje, izvrtanje, izrezivanje, poboljšanje boje, kontrasta, zasićenja...
 - Korištenjem složenih modela – npr. GAN (engl. *Generative Adversarial Networks*)
 - <https://www.nature.com/articles/s41598-019-52737-x>



Source: <https://github.com/xkumiyu/numpy-data-augmentation>

<https://research.aimultiple.com/data-augmentation-techniques/>

<https://iq.opengenus.org/data-augmentation/>

Inženjerstvo značajki

Inženjerstvo značajki

- Značajka: mjerljivo svojstvo objekta (primjera) koje je potrebno uzeti u obzir

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr.	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mr.	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, M	female	26	0	0	STON/O2. 31	7.925		S
5	4	1	1	Futrelle, Mrs	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. W	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. J	male		0	0	330877	8.4583		Q

Source: <https://www.datarobot.com/wiki/feature/>

Inženjerstvo značajki

- Inženjerstvo značajki je **proces** kojim se nastoje **odabrati ili transformirati** najbitnije varijable (značajke) iz pripremljenog skupa podataka s ciljem uspješnog modeliranja
- Razlikuje se:
 - **ručni pristup** inženjerstvu značajki (znanje o domeni je jako bitno)
 - **poluautomatizirani pristup** inženjerstvu značajki (znanje o domeni je manje bitno)
 - **potpuno automatizirani pristup** inženjerstvu značajki (znanje o domeni ne igra ulogu)

Ručni pristup inženjerstvu značajki

- **Izlučivanje (računanje) značajki** (engl. *feature extraction, feature elicitation*)
 - Definiranje, implementacija i računanje značajki iz sirovih podataka
 - Značajke ovisne o domeni primjene, predlaže ih stručnjak (ekspert) u području primjene
 - Potencijalno **beskonačni prostor** značajki
 - U analizi signala razlikujemo
 - Značajke vremenske domene (često statističke značajke)
 - Značajke frekvencijske domene (značajke dobivene iz spektra signala)
 - Nelinearne značajke (značajke faznog prostora, entropije, ...)
 - Različite značajke slike (npr. histogrami boja) i volumnih podataka
 - Značajke se obično računaju nakon prethodne pripreme (npr. uklanjanje šuma, interpolacije nedostajućih vrijednosti i sl.)

Ručni pristup inženjerstvu značajki

- Karakterizira ga **pregled pojedinačnih značajki**, a zatim:
- **Dodavanje novih značajki na temelju postojećih**
- **Uklanjanje nebitnih značajki**

Ručni pristup inženjerstvu značajki

- **Dodavanje novih značajki na temelju postojećih**
 - Obično se provodi nakon izlučivanja značajki iz sirovih podataka
 - Izgradnja temeljem jedne postojeće značajke
 - Diskretizacija numeričkih vrijednosti (engl. ***binning***)
 - Pretvorba jedne kategoričke značajke u više binarnih (engl. ***one-hot encoding***)
 - Normalizacija vrijednosti (engl. ***normalization***)
 - Izgradnja temeljem više postojećih značajki
 - Ručno kombiniranje više značajki u jednu, npr. suma, produkt, kvocijent i sl.

Diskretizacija numeričkih vrijednosti

- Pretvorba numeričke vrijednosti primjera neke varijable (broja) u kategoričku vrijednost
- Pretpostavka: broj kategorija \ll broj numeričkih vrijednosti
- Često nužan korak u analizi podataka, budući da:
 - Neki algoritmi funkcioniraju samo koristeći diskretne, kategoričke vrijednosti (induktivna pristranost)
 - Performanse algoritama degradiraju ako varijable nemaju uniformnu razdiobu gustoće vjerojatnosti
- Primjeri:
 - Neki algoritmi stabla odluke (engl. *decision trees*)
 - Neki sustavi temeljeni na induktivnim pravilima (engl. *induction rules, rule-based system*)
 - Sustavi asocijativnih pravila (engl. *association rules*)
- **Diskretizacijom se uvijek gubi određena informacija**, stoga je važno da diskretizacijski postupak bude što bolji

Diskretizacija numeričkih vrijednosti

- Vjerojatno najbolju diskretizaciju nekog numeričkog atributa mogu predložiti stručnjaci iz nekog područja
- U izostanku tog prijedloga, neki češće korišteni postupci diskretizacije su:
 - Podjela u N jednakih intervala (engl. *equal distance binning*),
 - Podjela u intervale s jednakim brojem pojedinaca (engl. *equal frequency binning, percentile binning*),
 - Diskretizacija minimizacijom entropije (engl. *entropy minimization discretization*),
 - Diskretizacija algoritmom k -srednjih vrijednosti (engl. *k-means discretization*)
 - i dr.
- <https://machinelearningmastery.com/discretization-transforms-for-machine-learning/>

Pretvorba jedne kategoričke značajke u više binarnih

- Mnogi algoritmi strojnog učenja ne mogu raditi direktno s kategoričkim vrijednostima, nego zahtijevaju da sve ulazne i ciljne varijable bude numeričke
- Ograničenje koje je uvela **učinkovita implementacija** algoritama strojnog učenja
 - nije nužno ograničenje samog algoritma
- Preslikavanje kategoričke značajke u numeričku: kategorija1 -> 1 ; kategorija2 -> 2 kategorijan -> n **samo u slučaju kada poredak kategorija ima smisla**
- Inače, **svaka kategorija** neke kategoričke značajke **postaje nova binarna značajka**
 - Od n kategorija dobivamo n binarnih značajki, koje imaju vrijednost 1 za one primjere za koje bi dotična kategorija vrijedila, a 0 inače
- <https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>

Normalizacija vrijednosti

- Potrebno kada su različite značajke u skupu podataka **mjerene na različitim skalama**
 - Značajke mjerene na nižim skalama (npr. između 1 i 10) bile bi manje relevantne modelu od onih na višim skalama (npr. između 1000 i 10000), što bi dovelo do lošijih rezultata
- **Najčešća normalizacija je na raspon vrijednosti između 0 i 1**
- Postupci normalizacije
 - **Decimalno skaliranje** (dijeljenje vrijednosti s maksimalnom vrijednosti decimalnog mjesta)
 - Npr. Sa 100, ako su sve vrijednosti do 100, a veće od 10
 - Normalizacija **Min-Max** (linearna transformacija vrijednosti)
 - Normalizacija **z-skorom** (statistička normalizacija putem srednje vrijednosti i varijance), poznato i kao **standardizacija**

Ručni pristup inženjerstvu značajki

- **Uklanjanje nebitnih značajki**
 - Monotone značajke
 - Konstantne značajke
 - Značajke s vrlo rijetkim podacima
 - Duplikati i **statistički redundantne značajke**
 - Najčešće korelacijska analiza, mogu i drugi testovi, npr. hi-kvadrat

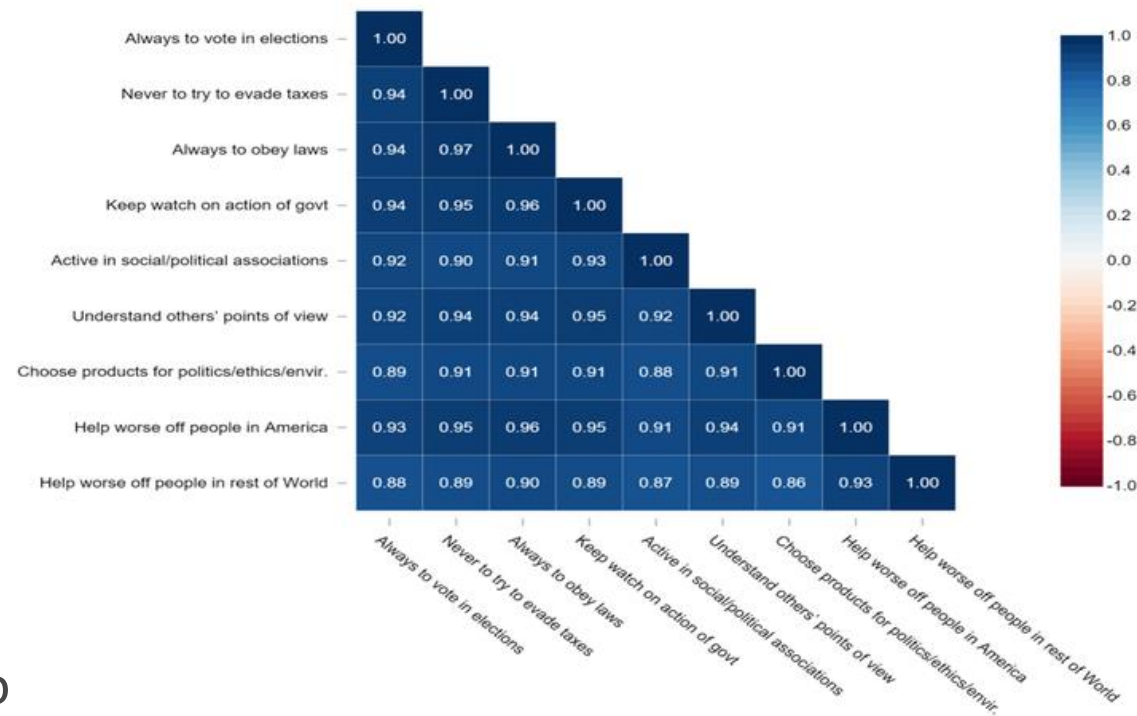
Uklanjanje statistički redundantnih značajki korelacijskom analizom

			HIGHLY CORRELATED ATTRIBUTES
person_name	is_male	is_female	
Aman	1	0	One attribute can be removed without any information loss. As one attribute can easily determine the other.
Abhinav	1	0	
Ashutosh	1	0	
Dishi	0	1	
Abhishek	1	0	
Avantika	1	0	
Ayushi	0	1	

Source: <https://www.geeksforgeeks.org/redundancy-and-correlation-in-data-mining/>

Uklanjanje statistički redundantnih značajki korelacijskom analizom

- Računa se korelacija između svakih dviju varijabli u skupu i gradi se korelacijska matrica
- Za one dvije varijable za koje je vrijednost korelacije vrlo visoka (idealno 1) odabire se jedna od njih koja se uklanja iz skupa – ona je redundantna
- Prag vrijednosti korelacijskog koeficijenta za odbacivanje neke značajke ovisi o domeni i cilju analize, ali obično je viši od 0.9
- Ponekad je bolje ne ukloniti značajku ako nismo sigurni da bi to bilo ispravno



Source: <https://www.displayr.com/what-is-a-correlation-matrix/>

Poluautomatizirani pristup inženjerstvu značajki

- Odabir značajki (engl. *feature selection*)
- Izgradnja značajki (engl. *feature construction*)
- Redukcija dimenzionalnosti (engl. *dimensionality reduction*) – razmatramo u kasnijim predavanjima

Odabir značajki

- Fokus na **smanjenje dimenzije** skupa podataka
- **Zadržava se interpretacija značajki**, jer se one koje se zadržavaju **ne mijenjaju**
- Želi se **zadržati rezultat** modeliranja početnog skupa značajki ili ga **poboljšati**
- Postupci:
 - **Filterski postupci (filteri)** (engl. *filters*)
 - **Postupci omotača** (engl. *wrappers*)
 - **Ugrađeni postupci** (engl. *embedded methods*)
 - **Hibridni postupci** (engl. *hybrid methods*)

Odabir značajki

- Optimalan podskup značajki = **najmanji mogući broj značajki koji daje najbolje rezultate** (za klasifikaciju, predikciju...)
- Potraga za optimalnim podskupom značajki je **NP težak problem**
 - Pretraga 2^M podskupova značajki, gdje je M broj značajki
- Postojeći empirijski postupci rješavanja obično rade u polinomnom vremenu i **ne garantiraju pronalazak optimalnog podskupa**
 - Obično pronalaze lokalni optimum
- Radi za:
 - Nadzirano učenje – kriterij je određen s obzirom na odnos vrijednosti značajke prema vrijednosti klase ciljne varijable (ili numeričkoj vrijednosti ciljne klase u slučaju regresijskih problema)
 - Nenadzirano učenje – kriterij je određen s obzirom na kompaktnost grupa (klastera)

Filterski postupci

- Filterski postupci definiraju **kriterij** koliko je određena značajka bitna za opis ciljne varijable
- Obično se **značajke rangiraju** s obzirom na taj kriterij
 - Korisnik može onda odabrati prvih n značajki
- Različiti filteri (svaki ima svoju matematičku formulaciju):
 - Zajednička informacija (engl. *mutual information*)
 - hi-kvadrat, χ^2 (engl. *chi-square*, χ^2)
 - Simetrična nesigurnost (engl. *symmetrical uncertainty*)
 - Relief (Relief, ReliefR, ReliefC...)
 - Korelacijski koeficijent (uglavnom za regresijske probleme) (engl. *correlation coefficient*)
- Neki filteri mogu istovremeno određivati bitnost i redundantnost značajki
 - Npr. mRMR (engl. *minimum redundancy maximum relevance*)

Postupci omotača

- Koriste **algoritam strojnog učenja za evaluaciju** određenog podskupa značajki kako bi donijeli odluku o tome je li taj podskup bolji / isti / lošiji od nekog nadskupa
- Algoritam strojnog učenja često nije onaj koji se kasnije koristi za izgradnju modela
 - Preferiraju se brzi algoritmi kako bi se što više skupova značajki evaluiralo – npr. Naivni Bayes
- Pretraživanje prostora podskupova značajki može početi od punog skupa ili od praznog skupa i koristiti različite strategije (naivni pristup je slučajno pretraživanje)
 - Pohlepne strategije (npr. najbolji prvi)
 - Unaprijednu selekciju i eliminaciju unazad, zrakasto pretraživanje
 - Evolucijske algoritme
- **U pravilu: sporiji, ali točniji postupci od filtera**

Ugrađeni postupci

- Izbor značajki koji se **temelji na nekom algoritmu strojnog učenja**
- Unutarnja struktura izgrađenog modela oslikava važnost značajki, bilo zbog broja pojavljivanja određene značajke u modelu ili njezine težine (značaja) u modelu
- Mogu se koristiti za dobivanje rangirane važnosti pojedinačnih značajki prema određenom kriteriju ili samo za dobivanje podskupa bitnih značajki
- Primjeri
 - Slučajna šuma
 - Logistička regresija s penalizacijom (ridge, LASSO, elastic net)
 - Stroj s potpornim vektorima
- <http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/>

Hibridni postupci

- Kombiniraju najbolja svojstva filtera i postupaka omotača
- Primjena dvaju ili više različitih postupaka filtera, omotača i ugrađenih postupaka
 - Najčešće najprije primijenjen filter kako bi značajno smanjili prostor značajki
 - Potom primijenjen postupak omotača kojim se nastoji pronaći optimalni podskup značajki
 - Moguće i drugačije kombinacije
- Nema garancije niti da su filterom zadržane sve bitne značajke niti da se postupkom omotača dobiva najbolji skup
- U praksi se pokazuju točnijima od filterskih postupaka i bržima od postupaka omotača
- <https://heartbeat.comet.ml/hands-on-with-feature-selection-techniques-hybrid-methods-b93b1b06d3a5>

Izgradnja značajki

- Fokus na **poboljšanju performansi**
- **Iterativna primjena različitih operatora za izgradnju novih značajki**
- Dobivene značajke po potrebi se dijelom uklanjaju korištenjem algoritama za odabir značajki
- Nastoji se izbjeći veliko povećanje broja značajki
- Nije toliko često u praksi kao odabir značajki i redukcija dimenzionalnosti
- Uobičajeni pristupi izgradnji novih značajki su temeljeni na:
 - Stablima odluke
 - Genetskom programiranju
- Za detalje vidjeti: <http://sifaka.cs.uiuc.edu/~sondhi1/survey3.pdf>

Potpuno automatizirani pristup

- **Učenje značajki** (engl. *feature learning, representation learning*)
 - Pristup kojim zaobilazimo ekspertno izlučivanje (računanje) značajki
 - Pristup je nezavisan o poznavanju domene
 - Sve češće korišteno u različitim područjima primjene (biomedicina, računalni vid)
 - Pretpostavka je da se radi **nad sirovim ulaznim podacima** (očišćenim, pripremljenim) i to najčešće:
 - Signalima (1D vremenskim nizovima)
 - Slikama – 2D signalima
 - Volumnim podacima – 3D signalima
 - Sirovi podaci se transformiraju unutar algoritma u unutarnji model koji je opisan značajkama niske razine (engl. *low-level features*)
 - Značajke koje imaju jasnu matematičku formulaciju ali nejasnu semantiku

Potpuno automatizirani pristup

- Koristi se određeni **algoritam strojnog učenja** koji interno uči nove značajke
 - Ideja je da nove značajke budu **visoko diskriminatorne i korisne** za problem koji se rješava
 - Nove značajke su dobivene **transformacijama ulaznih podataka** ili početnog skupa značajki
 - Nove značajke često se nazivaju reprezentacijama (engl. *representations*)
 - Koriste se i algoritmi nadziranog i nenadziranog strojnog učenja
- Neki poznati algoritmi za učenje značajki
 - Tradicionalni: **analiza glavnih komponentata (PCA)**, **analiza neovisnih komponentata (ICA)**
 - Duboko učenje: **višeslojni perceptron**, **konvolucijske neuronske mreže**, **autoenkoderi** i **ograničeni Boltzmanovi strojevi**
- <https://towardsdatascience.com/unsupervised-feature-learning-46a2fe399929>

Zaštita privatnih podataka

- Rukovanje osjetljivim, privatnim podacima krajnjih korisnika je često u praksi
- Tvrtke obično s klijentima potpisuju različite sporazume o zaštiti privatnih podataka
 - Ugovor između poslovnih subjekata
 - Sporazum o neotkrivaju informacija (engl. *Non-disclosure agreement*, NDA)
 - Različiti sporazumi na internetu (npr. sporazum o licenci za krajnjeg korisnika, engl. *End-user license agreement*, EULA)
- Opća uredba o zaštiti podataka (engl. *General Data Protection Regulation*), 2016.
 - Osigurava da osjetljivim podacima upravlja samo ovlaštena osoba – pozitivno!
 - U praksi, veća sigurnost otežava životnost – protok informacija
 - Problem identifikacije osobe putem privatnih podataka – **što je sve privatni podatak?**

Zaštita privatnih podataka

- Najosjetljiviji podaci:
 - medicinski podaci
 - financijski podaci
 - osobni podaci (npr. OIB, broj osobne iskaznice, broj putovnice...)
- Malo manje osjetljivi podaci: socio-demografski podaci (dob, spol, obrazovanje, narodnost, vjenčani status, prihodi kućanstva...)
- Neki put je jednostavna anonimizacija podataka najučinkovitije rješenje
- Tvrtke rade na rješenjima za sveobuhvatnu zaštitu privatnih podataka
- <https://inteligencija.com/poslovna-inteligencija-i-fer-uz-pomoc-eu-fondova-razvijaju-platformu-za-klasifikaciju-osobnih-podataka-i-njihovo-kontrolirano-uklanjanje-9712/>

Literatura

- Alice Zheng, Amanda Casari (2018.), *Feature Engineering for Machine Learning*, O'Reilly Media
- Dorian Pyle (1999.), *Data Preparation for Data Mining*, Morgan Kaufmann
- Alan Jović, Karla Brkić, Nikola Bogunović (2015.), A review of feature selection methods with application, *MIPRO 2015*, <https://ieeexplore.ieee.org/abstract/document/7160458>

Zaključci

- Rukovanje podacima je **složen proces** kojim podatke pripremamo za analizu
 - Sastoji se od niza koraka i transformacija podataka
- Veliku ulogu u tom procesu igra priroda podataka – veličina i značajke skupa podataka
- Podaci mogu imati različite probleme, od kojih su neki lako, a neki teško rješivi
 - Savršenog rješenja za sve probleme nema
 - Zahtijeva puno inženjerskog rada
- Inženjerstvo značajki naglašava važnost koje imaju značajke za korisnost daljnje analize podataka
 - Ručni pristup, poluautomatizirani pristup, potpuno automatizirani pristup
 - Cilj je pronaći optimalni skup značajki za dani problem