Priprema podataka

Dubinska analiza podataka

2. predavanje

Pripremio: izv. prof. dr. sc. Alan Jović

Ak. god. 2023./2024.

Sadržaj

- Proces pripreme podataka
- Problemi u podacima i njihovo uklanjanje
- Primjer skupa podataka

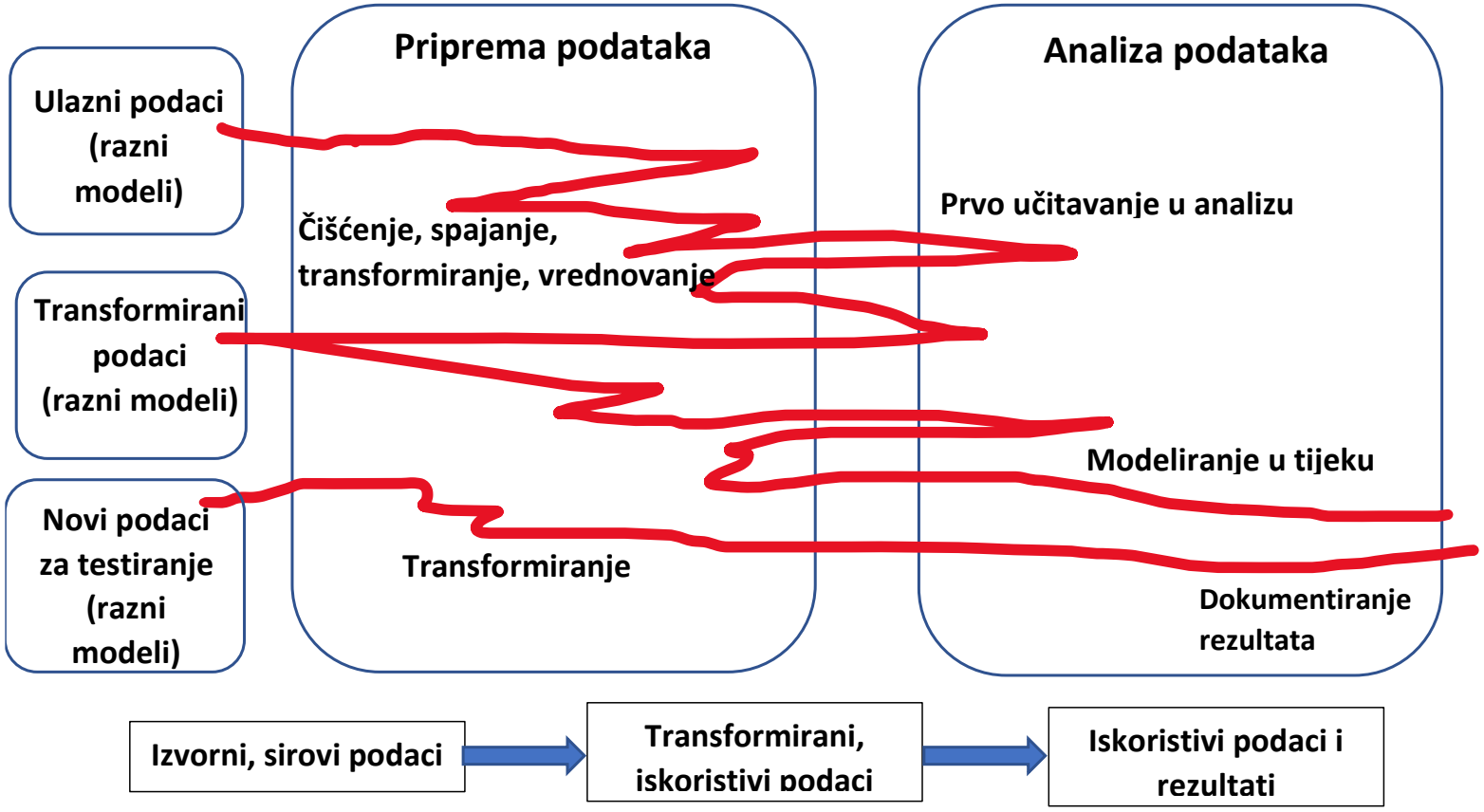
- Iduće predavanje: transformacije podataka, inženjerstvo značajki

Priprema podataka

- Engl. *data preparation, data handling, data wrangling*
- Slijedi **nakon preuzimanja izvornih podataka** s mjesta pohrane sve **do početka analize** statističkim postupcima i postupcima strojnog učenja
- **U širem smislu** uključuje i pristup, pregled, i izbor **izvornih podataka**
- Ukratko: „**Sve što prethodi modeliranju podataka**”
- **50 % – 80 %** vremena na DAP projektima

Proces pripreme podataka

Proces pripreme podataka



Prilagođeno iz: EPFL, ADA, 2020.

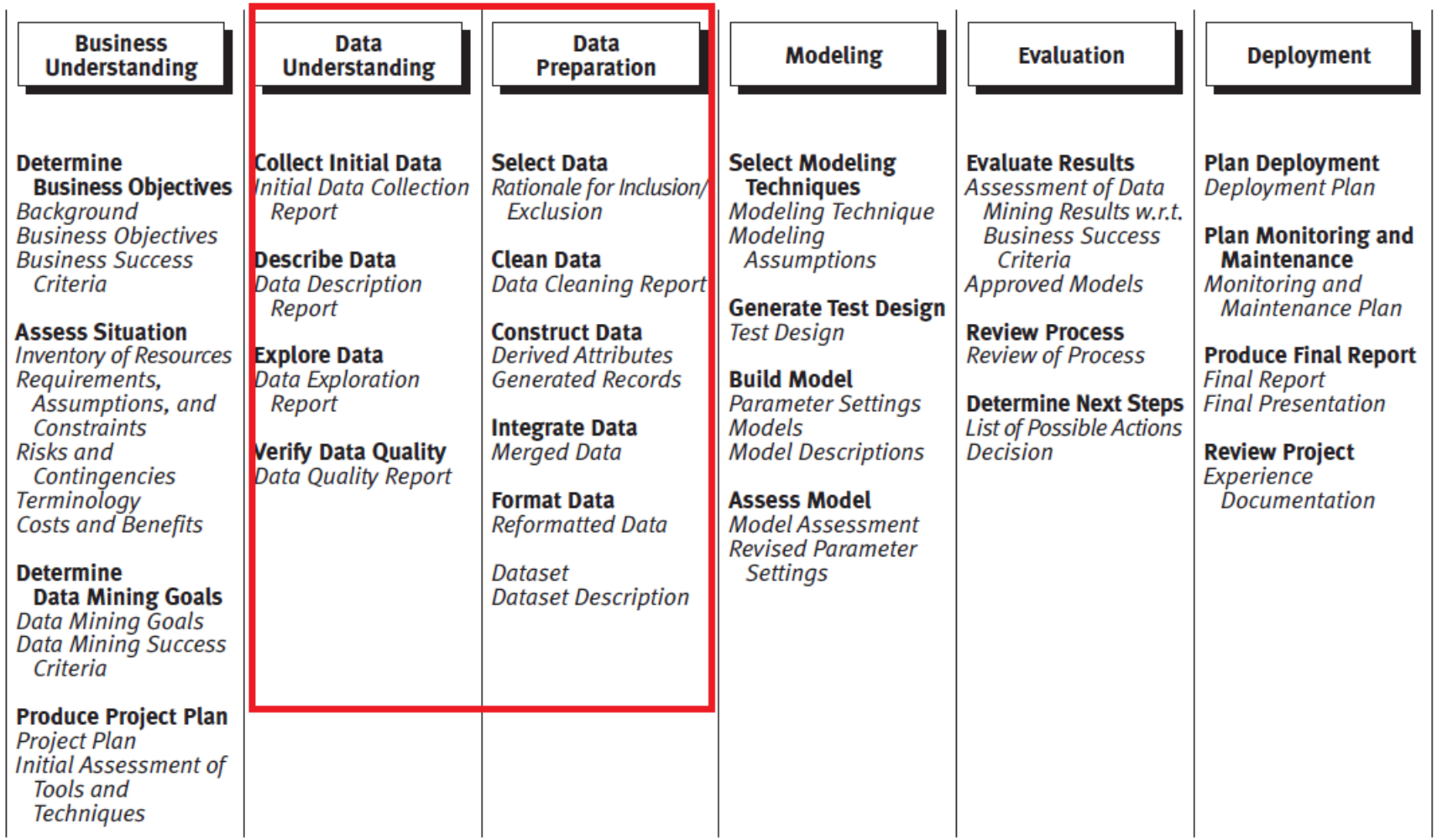
Proces:

- Iterativan
- Ad hoc provedba
- Ovisan o prostoru problema, prostoru rješenja i dostupnim podacima
- Zahtijeva **puno** razmišljanja
- Automatizacija je teška

Proces pripreme podataka

- Postoji više **modela procesa** pripreme podataka različitih autora
- Za sve njih zajednička su tri ključna koraka:
  - **Otkrivanje podataka**
  - **Karakterizacija podataka**
  - **Izgradnja skupa podataka za modeliranje**
- Razmotrit će se ukratko model procesa opisan u metodologiji **CRISP-DM**

Proces pripreme podataka – CRISP-DM



Proces pripreme podataka – CRISP-DM

- 1. faza: **Razumijevanje podataka**
  - Prikupljanje inicijalnih podataka
  - Opis podataka
  - Istraživanje podataka
  - Ispitivanje kvalitete podataka
- 2. Faza: **Priprema podataka**
  - Izbor podataka
  - Čišćenje podataka
  - Izgradnja skupa podataka
  - Integracija skupa podataka
  - Formatiranje skupa podataka

Danas se često obje faze provode putem jedne zajedničke arhitekture u tvrtkama

CRISP-DM – Zadatak: **Prikupljanje inicijalnih podataka**

- Aktivnosti:
  - Osiguravanje pristupa podatcima iz resursa dostupnih projektu
  - Učitavanje podataka u alat kojim se podaci pregledavaju
- Ishodi (izlazi) zadatka:
  - Pobrojeni skupovi s lokacijama i načinom pristupa
  - Navedeni problemi prilikom prikupljanja podataka

CRISP-DM – Zadatak: **Opis podataka**

- Aktivnosti:
  - Okvirni (grubi) pregled skupa podataka
- Ishodi zadatka:
  - Opis formata, kvantitete podataka (npr. broj redaka u tablici, broj i nazivi stupaca)
  - Ustanoviti odgovara li pronađeni skup podataka traženoj specifikaciji

Diskusija: Što ako ustanovimo da nemamo odgovarajuće podatke?

CRISP-DM – Zadatak: **Istraživanje podataka**

- Aktivnosti:
  - Statistički pregled skupa podataka
  - Vizualizacija skupa podataka
  - Ustanovljavanje vrsta varijabli, pronalazak ciljnih varijabli (ako ih ima)
  - Pronalazak temeljnih odnosa između varijabli (npr. korelacije)
- Ishodi zadatka:
  - Opis skupa podataka, uključujući grafikone sa značajnim pronalascima

Diskusija: Što ako je skup podataka prevelik za detaljno istraživanje?

CRISP-DM – Zadatak: **Ispitivanje kvalitete podataka**

- Aktivnosti:
  - Ustanovljavanje kompletnosti skupa podataka
  - Pronalazak problema u podacima
- Ishodi zadatka:
  - Popis pronađenih problema i mogućih rješenja

CRISP-DM – Zadatak: **Izbor podataka**

- Aktivnosti:
  - Odluka oko podataka koji će se koristiti za modeliranje – izbor je i po pitanju značajki i po pitanju primjeraka
- Ishodi zadatka:
  - Popis značajki i primjeraka koji će se koristiti za modeliranje

CRISP-DM – Zadatak: **Čišćenje podataka**

- Aktivnosti:
  - Uklanjanje **problema u podacima** otkrivenih u ranijoj fazi
- Ishodi zadatka:
  - Opis odluka i akcija koje su se napravile kako bi se problemi uklonili
  - Procjena utjecaja napravljenih promjena na ishod analize

CRISP-DM – Zadatak: **Izgradnja skupa podataka**

- Aktivnosti:
  - Izgradnja novih značajki
  - Dodavanje novih primjeraka u skup
  - Transformacije postojećih značajki u skupu
- Ishodi zadatka:
  - Opisati koje su se nove značajke izgradile i kako se pristupilo dodavanju novih primjeraka u skupu
  - Opisati primijenjene transformacije postojećih značajki

CRISP-DM – Zadatak: **Integracija skupa podataka**

- Aktivnosti:
  - Spajanje pripremljenih podataka iz više tablica ili zapisa
- Ishodi zadatka:
  - Opisati zapise koji su integrirani, načine integracije i integrirane zapise

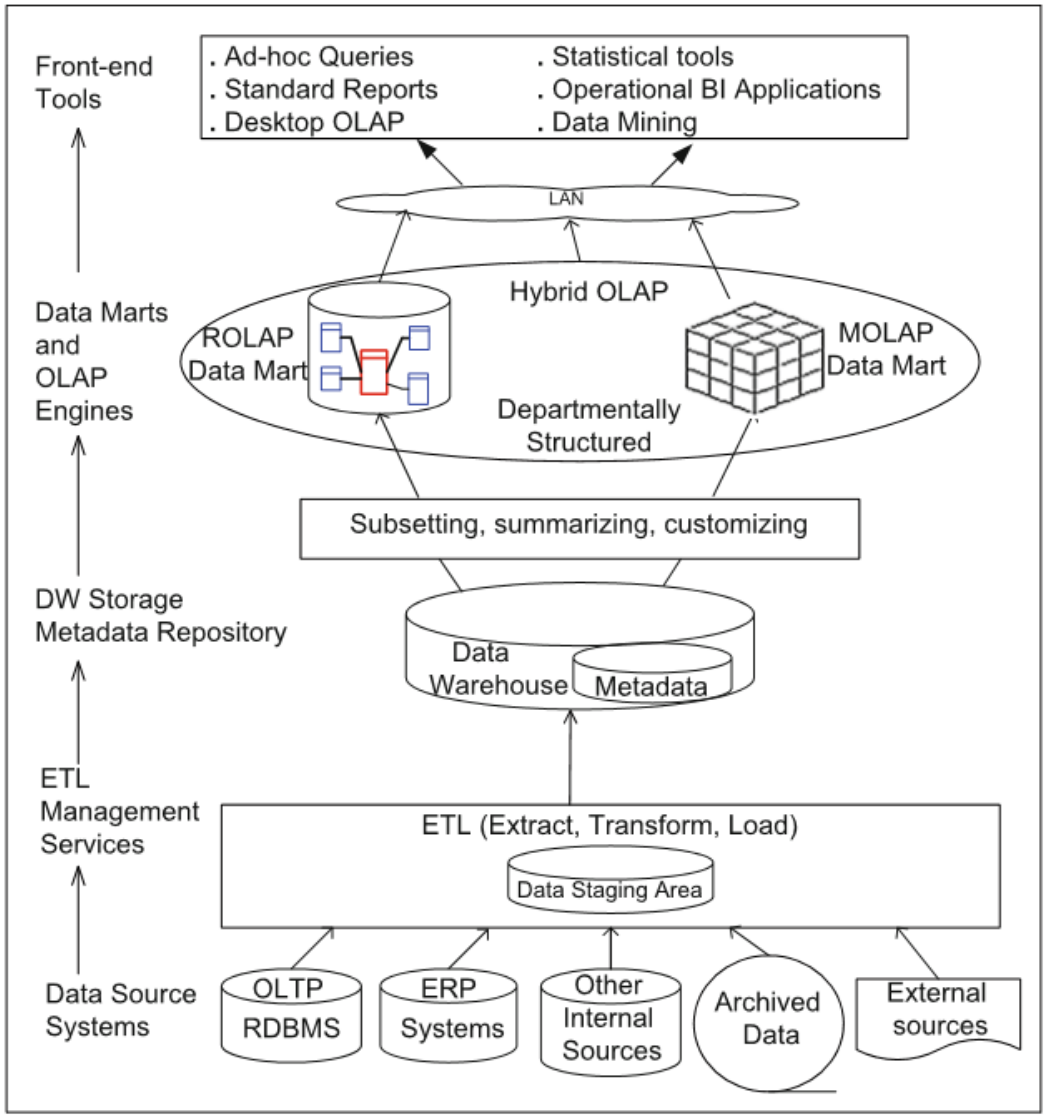
CRISP-DM – Zadatak: **Formatiranje skupa podataka**

- Aktivnosti:
  - Provođenje izmjena u skupu podataka kako bi se mogao modelirati s određenom metodom za analizu; npr. poredati značajke sa zadnjom ciljnom, ukloniti neke znakove u tekstnim podacima i sl.
- Ishodi zadatka:
  - Opisati sve napravljene izmjene

CRISP-DM – razumijevanje i priprema podataka u praksi

Neki ključni pojmovi:

- **ETL (*Extract, Transform, Load*)**
  - Arhitekturni obrazac za rad s podacima koji se ostvaruje dobavljanjem podataka iz izvornog oblika pohrane podataka, transformacijom podataka i ukrcavanjem podataka u novu strukturu za pohranu koja će sadržavati transformirane podatke
- **Data Warehouse** (skladište podataka)
  - Središnji repozitorij integriranih i strukturiranih podataka neke tvrtke
- **OLAP** (***O**nLine **A**nalytical **P**rocessing*)
  - Višedimenzijska organizacija poslovnih tablica u "podatkovne kocke" koja olakšava izvještavanje
    - MOLAP – mnogodimenijske strukture podataka, malo podataka, pogledi u memoriji
    - ROLAP – podaci ostaju u relacijskim tablicama, puno podataka, statički pogledi
- **Data Mart** (tržište podataka)
  - Koncentrirani podskup skladišta podataka koristan za specifičnu primjenu

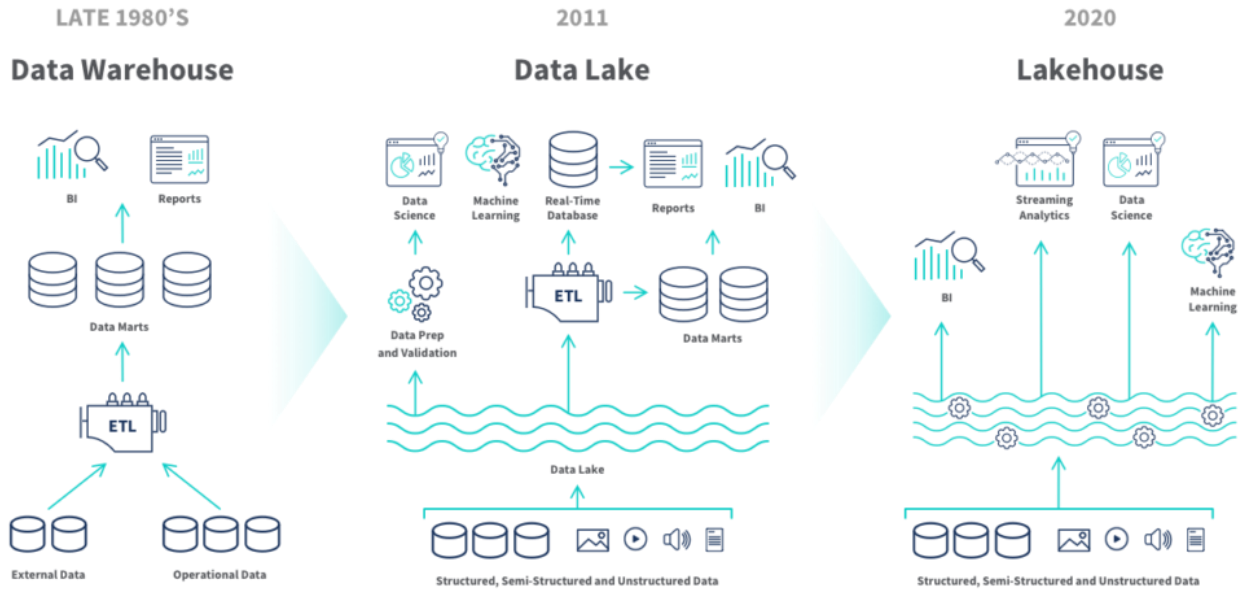


Izvor: I-Y. Song, Data Warehousing Systems: Foundations and Architectures. In: Encyclopedia of Database Systems, Springer, 2017.

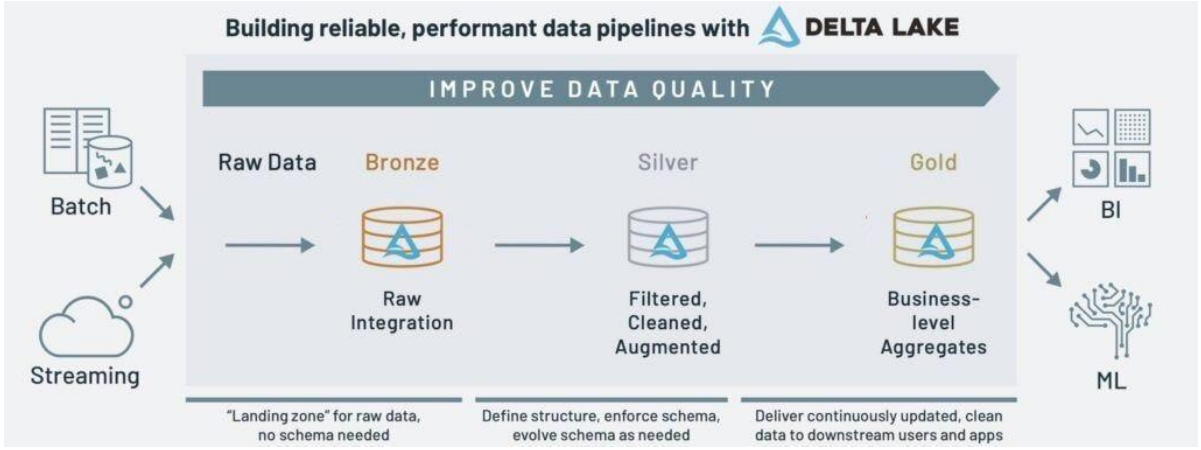
CRISP-DM – razumijevanje i priprema podataka u praksi

Neki ključni pojmovi:

- **Data Lake**
  - Sadrži veliku količinu sirovih, nepripremljenih podataka, u strukturiranom (relacijske baze) ili nestrukturiranom (NoSQL baze) obliku
  - Nad njim se radi ETL
- **Data Lakehouse**
  - Noviji, fleksibilnji hibridni pristup, koristi podatke u svim oblicima
  - Usluga je uvijek u oblaku, s nizom pojedinačnih usluga koje podatke prilagođavaju u strukturirane transakcije za različite primjene
  - Delta Lake – sloj pohrane podataka, služi za pripremu Data Lakehouse podataka



Izvor: https://insightsoftware.com/blog/remodel-your-oracle-cloud-data-with-a-data-lakehouse/



Izvor: https://www.databricks.com/glossary/medallion-architecture

Problemi u podacima i njihovo uklanjanje

Problemi u podacima

- **Nedostajući podaci** (engl. *missing data*)
- **Netočni i zagađeni podaci** (engl. *incorrect data, poluted data*)
- **Nekonzistentnosti u podacima** (engl. *inconsistent data*)
- **Stršeći podaci** (engl. *outliers*)
- **Rijetki podaci** (engl. *sparse data*)
- **Šumoviti podaci** (engl. *noisy data*)
- **Monotone značajke** (engl. *monotonic features*)
- **Konstantne značajke** (engl. *constant features*)
- **Pristranost podataka** (engl. *data bias*)

U ovom predavanju

- **Nebalansirani skupovi podataka** (engl. *imbalanced datasets*)
- **Prokletstvo dimenzionalnosti** (engl. curse of dimensionality)
- **Pomak koncepta** (engl. *concept drift*)

Sljedeća predavanja

Nedostajući podaci – vrste nedostajućih podataka

- **Nedostajuće ali poznate vrijednosti**
  - Vrijednosti koje nisu unesene u skup podataka, ali postoje u stvarnom procesu
  - Ako je moguće, naknadno unijeti
- **Prazne i nepoznate vrijednosti**
  - Ne može se pretpostaviti vrijednost u stvarnom svijetu i ona nije unesena

- Često nije jasno o kojoj se vrsti radi
- Detekcija problema **detaljnim pregledom skupa podataka** ili **korištenjem vizualizacije**

Nedostajući podaci – rješavanje problema

- Najvažnije je da osoba koja modelira podatke ima **kontrolu nad metodom** rješavanja problema nedostajućih podataka
- Alati ponekad nemaju transparentnu metodu rješavanja ovog problema što može dovesti do distorzije (pristranosti) u skupu podataka
- Obrazac nedostajućih vrijednosti ponekad zadrži važnu informaciju!

Nedostajući podaci – rješavanje problema

- **Zanemarivanje svih primjeraka koji sadrže nedostajuću vrijednost**
  - Često *defaultna* opcija algoritama
  - Ponekad nije dobra metoda
- **Zamjena nedostajuće vrijednosti nekom drugom**
  - Uz osiguranje da informacijski sadržaj skupa podataka ne degradira
  - Postoje jednostavni i složeni postupci

Nedostajući podaci – rješavanje problema

- Jednostavni postupci – razmatraju jednu značajku s jednom ili više nedostajućih vrijednosti
  - **Očuvati mjeru sredine** – zamjena sa **srednjom vrijednosti**, **centralnom vrijednosti** (medijan) ili **dominantnom kategorijom** (mod)
  - **Očuvati varijabilnost** – zamjena treba osigurati da se mjera varijabilnosti (npr. varijanca) značajke ne mijenja – u praksi točnija mjera zamjene od mjere sredine
  - **Proglasiti nedostajuću vrijednost novom kategorijom**
  - **Zamijeniti s konstantnom vrijednošću**

Nedostajući podaci – rješavanje problema

- Jednostavni postupci – primjer

Pozicija	Originalni uzorak	Pozicija 11 fali		Očuvanje srednje vrijednosti	Očuvanje StDev
1	0.0886	0.0886		0.0886	0.0886
2	0.0684	0.0684		0.0684	0.0684
3	0.3515	0.3515		0.3515	0.3515
4	0.9874	0.9874		0.9874	0.9874
5	0.4713	0.4713		0.4713	0.4713
6	0.6115	0.6115		0.6115	0.6115
7	0.2573	0.2573		0.2573	0.2573
8	0.2914	0.2914		0.2914	0.2914
9	0.1662	0.1662		0.1662	0.1662
10	0.4400	0.4400		0.4400	0.4400
11	0.6939	**?**		**0.3731**	**0.6622**
Srednja vrijednost	0.4023	0.3731		0.3731	0.3994
StDev	0.2785	0.2753		0.2612	0.2753
			Abs. pogreška	0.3208	0.0317

Nedostajući podaci – rješavanje problema

- Složeni postupci promatraju odnos između **više značajki** i biraju zamjenu na pojedinoj značajki koja će unijeti **najmanju pristranost** u čitav skup
- Naglasak je na održavanju **zajedničke varijabilnosti** (engl. *joint variability*) dviju ili više značajki
  - **Linearna regresija (obična ili višestruka)** – vidjeti `sklearn.impute.SimpleImputer` i `IterativeImputer`
  - **Algoritam *k*-najbližih susjeda** – vidjeti `sklearn.impute.KNNImputer`
  - **Nelinearna regresija**
  - **…**

Nedostajući podaci – rješavanje problema

- **Linearna regresija (jednostavna) za zamjenu nedostajećih vrijednosti**

$$y = ax + b$$

$$a = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2},$$

$b = \bar{y} - a\bar{x}$, $\bar{x}$ je srednja vrijednost od x

n – broj primjeraka

a – nagib pravca

b – intercept

Stvarne vrijednosti					
n	x	y	x^2	y^2	xy
1	0.55	0.53	0.30	0.28	0.29
2	0.75	0.37	0.56	0.14	0.28
3	0.32	0.83	0.10	0.69	0.27
4	0.21	0.86	0.04	0.74	0.18
5	0.43	0.54	0.18	0.29	0.23
Sum	2.26	3.13	1.20	2.14	1.25
Neke vrijednosti nedostaju					
n	x	y	x^2	y^2	xy
1	0.55	0.53	0.30	0.28	0.29
2	?	0.37	?	0.14	?
3	0.32	0.83	0.10	0.69	0.27
4	0.21	?	0.04	?	?
5	0.43	0.54	0.18	0.29	0.23
Sum	?	?	?	?	?

Pyle D. Data Preparation for Data Mining. Morgan Kaufmann, 1999.

Nedostajući podaci – rješavanje problema

- Zamjena koristi procjene na temelju sume poznatih vrijednosti varijabli
- Omjeri različitih suma ostaju **konstantni** a za procjenu suma koriste se samo poznati parovi vrijednosti i to srednja vrijednost za *x* i *y*
- Ubace se u formule za *a* i *b*

Srednje vrijednosti za procjenu nedostajućih vrijednosti					
n	x	y	x^2	y^2	xy
1	**0.55**	**0.53**	0.30	0.28	0.29
2		0.37			
3	**0.32**	**0.83**	0.10	0.69	0.27
4	0.21				
5	**0.43**	**0.54**	0.18	0.29	0.23
Sum	1.30	1.90	0.58	1.26	0.79
Mean	0.43	0.63			

Omjeri suma prisutnih vrijednosti			
	$sum(x^2)$	$sum(y^2)$	sum(xy)
Omjer prema sum(x):	0.45		0.61
Omjer prema sum(y):		0.66	0.42

Procjene vrijednosti na temelju omjera			
sum(x)	$sum(x^2)$	sum(xy)	
0.43	0.43 x 0.45	0.43 x 0.61	

Dobiva se y = -x + 1.06

Pyle D. Data Preparation for Data Mining. Morgan Kaufmann, 1999.

Nedostajući podaci – rješavanje problema

- **Postupak *k*-najbližih susjeda za zamjenu nedostajućih vrijednosti** (engl. *nearest neighbor imputation*)
  - Zamjena se provede tako da se pronađu uzorci u skupu za učenje (njih *k*) koji su **najbliži** određenom uzorku s nedostajućom vrijednosti (i koji sami nemaju nedostajuću vrijednost na istoj značajki)
  - Za kriterij blizine najčešće se koristi **euklidska udaljenost** svih značajki osim one s nedostajućom vrijednosti
  - Vrijednosti odgovarajuće značajke *k* najbližih uzoraka se **uprosječe** (srednjom vrijednosti) i vrijednost se zamijeni

Nedostajući podaci – rješavanje problema

- Nedostajuće vrijednosti u **vremenskim nizovima**
  - Ranije navedeni postupci gotovo nikada nisu dobro rješenje
  - Umjesto toga – interpolacija!
  - Pretpostavka: vremenski bliske vrijednosti **su slične jedna drugoj** (ne vrijedi za sezonske vremenske nizove)
  - Četiri najčešća postupka
    - **Prijenos zadnjeg opažanja unaprijed** (engl. *last observation carried forward (LOCF)*)
    - **Prijenos sljedećeg opažanja unazad** (engl. *next observation carried backward (NOCB)*)
    - **Linearna interpolacija** (engl. *linear interpolation*)
    - **Kubni *spline*** (engl. *cubic spline*)

scipy.interpolate
https://pythonnumericalmethods.berkeley.edu/notebooks/chapter17.03-Cubic-Spline-Interpolation.html

Netočni i zagađeni podaci

- Često kao rezultat **zabune** pri unosu
- Neki put **namjerno** uneseni
  - Korisnik ne zna točnu informaciju a ne želi ostaviti prazno
  - Korisnik ne zna gdje bi unio neku informaciju pa unese bilo gdje (npr. formular za unos je neadekvatan)
  - Korisnik ne želi da netko drugi sazna točnu informaciju ili ima korist unašanjem netočne informacije
- Ponekad tehnička pogreška sustava
- U općenitom slučaju, neriješiv problem, specifični slučajevi jesu riješivi
- **Zahtijeva detaljan pregled skupa podataka, vizualizaciju i promišljanje o podacima**

Nekonzistentnost u podacima

Dva tipa nekonzistentnosti

- **Različite značajke** mogu biti predstavljeni **istim imenom** u različitim sustavima
  - Problem pri povezivanju podataka iz određenog broja različitih sustava u jednu tablicu
  - Semantika nekog naziva je različita (npr. zaposlenik u sustavu plaća ili zaposlenik u sustavu firme su različite tablice!)
- Neka značajka ili vrijednosti neke značajke mogu imati više različitih **sinonima**, u jednom sustavu ili u više njih
  - Npr. „zaigrani" zaposlenici u auto-tvrtki pod značajkom *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

Stršeći podaci

- Podaci koji odskaču (odudaraju) **daleko izvan uobičajenih vrijednosti** za određene značajke
- Razlozi pojave: 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

- Primjer stršećeg podatka*: rođenje djeteta gđe Hadlum dogodilo se 349 dana nakon što je g. Hadlum otišao na služenje vojnog roka – prosječno razdoblje trudnoće kod ljudi je 280 dana (40 tjedana) – statistički gledano, 349 dana je stršeći podatak o trajanju trudnoće

*Barnett, V. 1978. The study of outliers: purpose and model. Applied Statistics, 27(3), 242–250

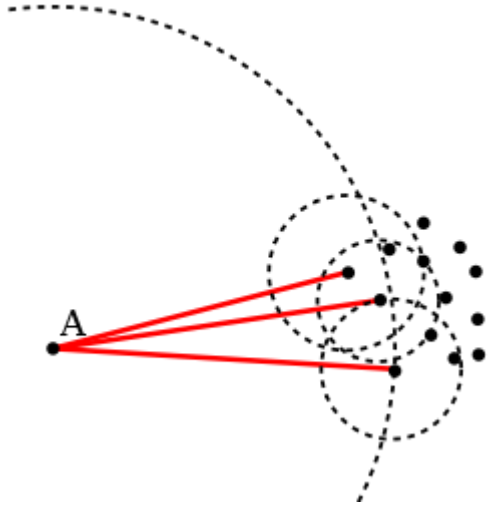
Stršeći podaci

- Korišteni postupci otkrivanja
  - **Vizualizacija podataka i opažanje**
  - **Statistički postupci** – z-skor, vjerojatnosni modeli, linearna regresija
  - **Algoritmi nenadziranog strojnog učenja**
    - Temeljeni na udaljenosti (npr. k-NN),
    - Temeljeni na gustoći (npr. **LOF**)
    - Algoritmi specifični za velike skupove podataka (npr. **IsolationForest**)
    - I dr.

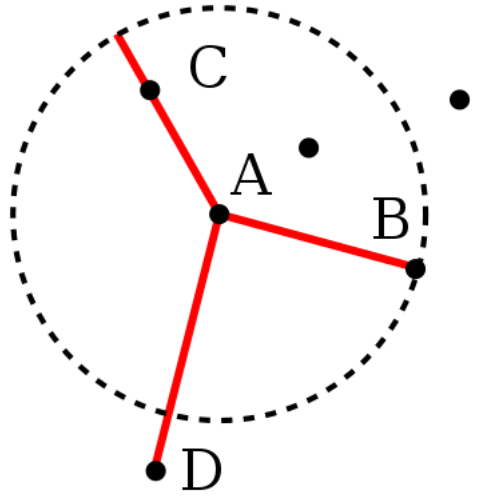
Izvor: Kriegel HP, Kröger P, Zimek A. Outlier Detection Techniques. 13th Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2009.

Stršeći podaci

- Algoritam lokalnih faktora stršećih vrijednosti (engl. *Local Outlier Factor* – LOF)
  - Glavna ideja: uspoređuju se lokalne gustoće točke A s lokalnim gustoćama susjednih točaka
  - Ako se ustanovi da ima manju lokalnu gustoću onda je vjerojatno stršeća



Točka A ima puno manju gustoću nego njezini susjedi, vjerojatno je stršeća



Pojam *k*-udaljenosti: Npr. točke B i C spadaju u 3-udaljenost od točke A, kao i neimenovana točka blizu A (3-udaljenost od točke A je jednaka za sve te tri točke), dok točka D ispada izvan te udaljenosti.

Stršeći podaci

- **Udaljenost dosezljivosti** točke A od točke B
  - *udaljenost-dosezljivosti*$_k$(*A*,*B*)=max{*k*-udaljenost(*B*), *d*(*A*,*B*)}
  - stvarna udaljenost *d* (euklidska) od A do B, *k* je hiperparametar – broj susjeda koji se razmatra

- **Lokalna gustoća dosezljivosti** (engl. *local reachability density*) točke A:
  - Utvrđuje koliko je gusto točka A povezana sa svojim susjedima
  - Inverz prosječne udaljenosti dosezljivosti točke A od njezinih *k* najbližih susjeda
  - $lrd_k(A) = \frac{|N_k(A)|}{\sum_{B \in N_k(A)} udaljenost\text{-}dosezljivosti_k(A,B)}$, $|N_k(A)|$ je broj *k*-najbližih susjeda od A, koji može iznositi *k* ili više, ako više susjeda od točke A ima jednaku udaljenost od nje

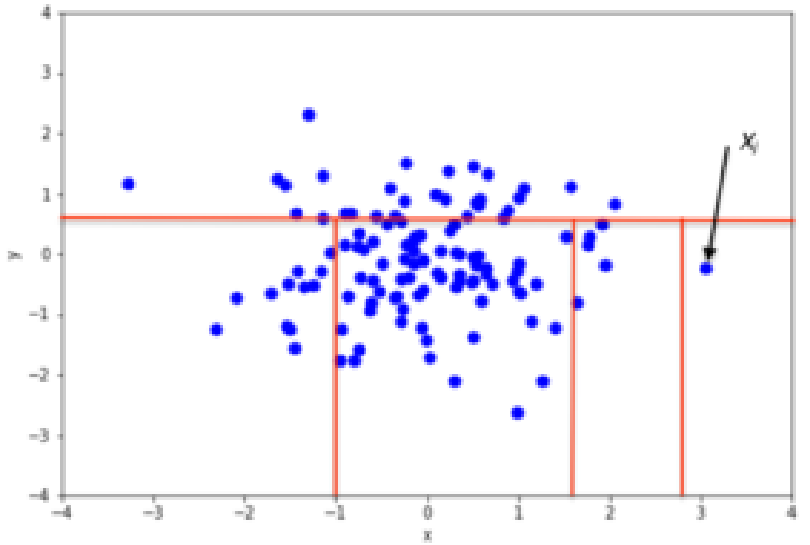
Stršeći podaci

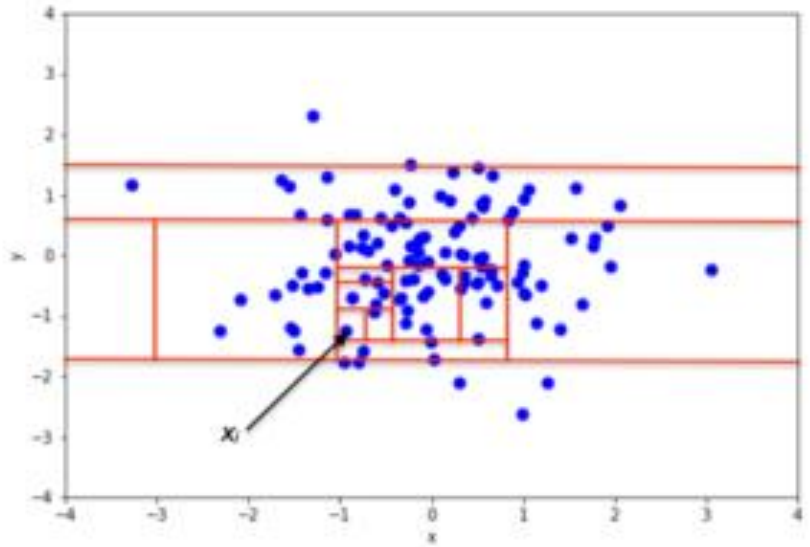
- **Lokalni faktor stršećih vrijednosti (LOF)** točke A:
  - Uspoređuje točku A sa svim njezinim *k*-susjedima da se ustanovi imaju li sličnu lokalnu gustoću dosezljivosti
  - $LOF_k(A) = \frac{\sum_{B \in N_k(A)} \frac{lrd_k(B)}{lrd_k(A)}}{|N_k(A)|}$,
  - LOF = 1 – slična gustoća, nije stršeća točka,
  - LOF > 1 – manja gustoća, vjerojatno stršeća točka,
  - LOF < 1 – točka veće gustoće (*inlier*), nije stršeća točka
- Prednosti: usporediv ili bolji od većine ostalih algoritama
- Nedostaci: LOF vrijednosti ovisne o skupu, nije pogodan za velike skupove podataka

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp. 93-104.
Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2021). A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams. Big Data and Cognitive Computing, 5(1), 1

Stršeći podaci

- **Algoritam IsolationForest** (2008.)
  - Temelji se na pretpostavkama:
    - stršeće vrijednosti je **lakše izdvojiti** (zato: „isolation” pristup) iz skupa nego modelirati normalne podatke
    - **stršećih podataka ima malo i bitno su različite od normalnih**
  - Algoritam rekurzivno generira binarne particije skupa podataka **nasumičnim odabirom značajki** i **raspona vrijednosti te značajke**
  - Rekurzivno particioniranje može se predstaviti strukturom stabla
  - Broj particioniranja potreban za izdvajanje primjerka jednak je **duljini puta** od korijenskog čvora stabla do lista koji predstavlja izoliranu vrijednost
  - Glavna ideja: nasumično particioniranje skupa daje **znatno kraće putove za stršeće** nego za normalne vrijednosti.





Stršeći podaci

- Algoritam IsolationForest
  - 1. faza: izgradnja stabala izolacije (engl. *isolation trees*) kako je ranije opisano
  - 2. faza: testiranje
    - Primjerci se propuštaju kroz izgrađena stabla, nakon čega im se dodjeljuje **mjera stršećih vrijednosti** *s(x,n)*
    - $s(x,n) = 2^{\frac{-E(h(x))}{c(n)}}$, $c(n) = 2H(n-1) - \frac{2(n-1)}{n}$, H(x) = ln(x) + 0.5772156649 (Eulerova konst.)
    - $E(h(x))$ je prosječna duljina puta *h* kroz stablo za sva stabla u šumi za uzorak *x*
    - ***s(x,n)* blizu 1 označava da se vjerojatno radi o stršećem podatku, dok blizu 0.5 je sigurno normalan**
  - Vidjeti: `sklearn.ensemble.IsolationForest`

- Izvor: F. T. Liu, K. M. Ting and Z. Zhou, "Isolation Forest," *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17

Stršeći podaci

- Prednosti Isolation Foresta:
  - Mala vremenska složenost (linearna) i zauzeće memorije
  - Pokazuje se uspješnim i na visokodimenzionalnim skupovima podataka s nebitnim značajkama
- Nedostaci Isolation Foresta:
  - Teško detektira grupirane anomalije i anomalije poravnate s osima
  - Mjera stršećih vrijednosti je dosta heuristička
- Nadogradnje:
  - SCiforest (2010.) – rješava probleme grupiranih anomalija i anomalija poravnatih s osima
  - iForestASD (2013.) – IsolationForest za tokove podataka
  - EIF (2019.) – Poboljšanje relevantnosti mjere stršećih vrijednosti

- F-T. Liu, K-M. Ting, Z-H. Zhou, "On Detecting Clustered Anomalies Using SCiForest". Joint European Conference on Machine Learning and Knowledge Discovery in Databases - ECML PKDD 2010 : Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science. 6322: 274–290. doi:10.1007/978-3-642-15883-4_18
- Z. Ding, M. Fei, "An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window". 3rd IFAC International Conference on Intelligent Control and Automation Science, 2013.
- S. Hariri, C. K. Matias; R. J. Brunner, "Extended Isolation Forest". IEEE Transactions on Knowledge and Data Engineering. 33 (4): 1479–1489, 2019.

Stršeći podaci – rješavanje problema

- Što napraviti s otkrivenim stršećim podatkom?
  - Utvrditi je li prirodan ili ne (moguća konzultacija sa stručnjakom)
  - Ako **je** prirodan ali se zna da će smetati prilikom izgradnje modela
    - Koristiti **normalizaciju vrijednosti varijabli** (vidjeti kasnije slajdove)
  - Ako **nije** prirodan
    - Tretirati ga kao nedostajući podatak i potom primijeniti neki od postupaka za nedostajuće podatke
    - Zapamtiti ga (i njegovu poziciju u skupu) radi otkrivanja razloga unosa pogreške

- Vidjeti: S. Kandanaarachchi, et al. "On normalization and algorithm selection for unsupervised outlier detection." Data Mining and Knowledge Discovery 34.2 (2020): 309-354

Rijetki podaci

- Slučaj kada za neke značajke 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
- Pristupi rješavanju problema
  - **Uklanjanje značajke s rijetkim podacima**
  - Transformacije podataka – 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**
- Izvori šuma
  - Šum je rezultat utjecaja prirodnih procesa (npr. elektromagnetske smetnje iz okoline)
  - Šum je rezultat nesavršenosti mjernih senzora (npr. pomak elektroda na površini kože prilikom pokreta ispitanika)

Šumoviti podaci

- Postoje postupci za filtriranje šuma u podacima (engl. *noise filtering*) kada je omjer signal/šum (engl. *signal-to-noise ratio*) nepovoljan
  - **Postupci su iznimno ovisni o konkretnom problemu**
  - Npr. korekcija pomaka nulte linije i gradske strujne mreže kod snimanja elektrokardiograma (EKG), pojasno propusno filtriranje (<0.1 i više od 40 Hz) kod elektroencefalograma (EEG) kako bi se obuhvatile samo značajne frekvencijske komponentne
- Neki put, šum nije moguće dovoljno ili do kraja filtrirati
  - U tom slučaju, skup na kojem se gradi model treba imati ista statistička svojstva kao i skup podataka na kojem će se model testirati, a i onaj na kojem će se u praksi primijeniti

Monotone značajke

- **Monotone značajke su one značajke čija vrijednost raste (ili se smanjuje) bez ograničenja**
- Najčešći primjeri
  - Značajke povezane s protjecanjem vremena, npr. datumi u raznim oblicima.
  - Značajke rednih brojeva različitih zapisa i sl.
- Problem je nemogućnost dobivanja korisne informacije iz takve serije
- Rješenja problema:
  - **Zanemariti takvu značajku (najčešće)**
  - Transformirati u određeni oblik pogodan za modeliranje
    - **Datum se može pretvoriti u godišnje doba ili dan u tjednu**
    - **Datumu se može pristupiti kao vremenskoj seriji (nizu)**

Konstantne značajke

- Vrijednost ovih značajki je ista (kontantna) u cijelom skupu podataka
- Varijanca im je 0 (engl. *zero-variance*)
- Često u skupovima kod znanstvenih pokusa gdje se neke varijable drže konstantnima
- Relativno jednostavno za uočiti, može pomoći vizualizacija (histogram, box-plot)
- Nisu informacijski relevantne, stoga ih se **uvijek treba ukloniti** (i u skupu za učenje i u skupu za testiranje)

Diskusija: Što ako je značajka konstantna samo na skupu za učenje?

Pristranost podataka – pristranost uzorkovanja

- Engl. *sampling bias*
- Uzorak (skup podataka dostupan za analizu) se prikuplja tako da neki primjerci ukupne populacije imaju **manju ili veću vjerojatnost uzorkovanja od drugih**
- **Problematično** iz dva razloga
  - Statistika na dostupnom uzorkom ne mora biti slična statistici na cijeloj populaciji
  - Model izgrađen na skupu za učenje **ne mora dobro generalizirati** na skupu za testiranje
- Neki podtipovi:
  - **Kulturološka pristranost** – skup podataka sadrži samo neke česte obrasce, a ne rijetke (npr. samo muškarci kao vozači kamiona)
  - **Rasna pristranost** – skup podataka sadrži samo podatke o jednoj rasi (npr. samo bijela koža pri detekciji oboljenja kože)

Pristranost podataka – pristranost isključivanja

- Engl. *exclusion bias*
- Pristranost koja nastaje zbog isključivanja vrijednih podataka iz skupa podataka iz razloga što se ne smatraju relevantnima
- Često pri formiranju skupa podataka ako se smatra da bi uključivanje svih podataka dovelo do prevelikog skupa
- Često kod pripreme podataka ako se neispravno uklanjaju značajke ili primjerci (npr. uklanjanje redundantnih značajki koji ustvari nisu redundantne, uklanjanje primjeraka iz neke manjinske skupine koja se previdi ili se ne smatra važnom)
- Kako se ne bi uvela pristranost isključivanja, bitno je razumjeti postupke poboljšavanja skupa podataka i inženjerstva značajki (tema idućih predavanja)

Pristranost podataka – pristranost promatrača

- Engl. *observer bias, confirmation bias*
- Pristranost koja se uvodi zbog subjektivnosti promatrača skupa, promatrač vidi ono što očekuje vidjeti u podacima
- Može biti problematično ako se podaci ne sagledaju objektivno, može dovesti do uvođenja ostalih vrsti pristranosti
- Čest slučaj pristranosti promatrača je pri **označavanju** (labeliranju) skupova podataka, gdje se označavanje treba provesti krajnje pažljivo te što je više moguće objektivno

Koraci pripreme podataka – pojednostavljeno

1. Sudjelovati u procesu prikupljanja i isporuke skupa podataka (po mogućnosti)
2. Pomno pregledati skup podataka kako bi se ustanovili svi problemi u skupu
    1. Za velike skupove uzorkovati podskup podataka koji se može vizualno proučiti
    2. Razmotriti svaku značajku pojedinačno, vizualizirati njezinu razdiobu (po mogućnosti na cijelom skupu)
    3. Proučiti dokumentaciju kako je skup dobiven (mjerenja možda imaju poznati šum)
3. Raspisati sve uočene probleme i predložiti rješenja za njihovo uklanjanje
4. Prodiskutirati uočeno i predložena rješenja s relevantnim dionicima (kolege inženjeri, nadređeni, klijenti)
5. Provesti usvojena rješenja

Za sve navedeno koristiti odgovarajuće tehnologije i arhitekturna rješenja – *warehouse, OLAP, lake, lakehouse*…

Primjer skupa podataka

Skup podataka Missing people

- https://www.kaggle.com/datasets/arjoonn/missing-people
- Sadrži podatke o 2127 nestalih i pronađenih osoba u Indiji.
- Skup s puno problema u podacima
- Značajke koje su dostupne su:

Name – osobno ime osobe
Gender – spol nestale osobe (MALE, FEMALE)
Relative – osobno ime rođaka koji je prijavio nestanak
Address – adresa rođaka koji je prijavio nestanak
AgeStart – dob u godinama kada je osoba nestala
AgeEnd – dob u godinama kada je osoba pronađena
HeightStart – visina u cm kada je osoba nestala
HeightEnd – visina u cm kada je osoba pronađena
Built – tjelesna građa osobe (thin, normalmedium, strong)
Date – datum nestanka osobe
Dist – četvrt u kojoj je osoba nestala (više mogućih vrijednosti)
State – regija Indije gdje je osoba nestala (DELHI, WEST, EAST, NORTH, SOUTH, CENTRAL)

Skup podataka Missing people

	A	B	C	D	E	F	G	H	I	J	K	L
1	Name	Gender	Relative	Address	AgeStart	AgeEnd	HeightStart	HeightEnd	Built	Date	Dist	State
2	JYOTI	Female	GEETA	, , E 129 A SHOK NAGAR F	16	17	122	183	normalmedium		G.T.B. ENCLAVE/NORTH EAST	DELHI
3	ABHISHEK	Male	MUKE SHKUMAR	, , JHUGGI NO. N- 78/102	19	20	153	183	thin		PUNJABI BAGH/WEST	DEL
4		Male	RAJE SHKUMAR	, NO, 79, GALI NO. 3 KON	19	20	153	183	thin		NEW ASHOK NAGAR/EAST	DE
5	JAIPRAKASH	Male	KAMA LKISHORE	, , 15/286, KALYANPURI D	24	25	153	183	thin		KALYANPURI	EAS
6	SADDAM	Male	MUNN A	, , 19-20, KABOOTAR MAF	21	22	153	183	thin		WELCOME/NORT EAST	DELHI
7	RAMENDERSING Male	Male	G-185/A	RAMENDERSING , RADH A	47	48	5	6	thin		JAIT PUR/SOUTH- EAST	DELHI
8		Male	DHAN NO	, A2/198, A MAR COLONY	28	29	153	183	thin		JYOTI NAGAR/NORTH EAST	DELHI
9	UMASHANKAR	Male	ANILK UMAR	, , B-116, GALI N O. 12, JO	21	22	153	183	thin		YAMUNA VHR/GOKULPURI/ EAST	DELHI
10	SHRIRAM	Male		„C-18,GALINO.16,MATAV	16	17	153	183	normalmedium		YAMUNA VHR/GOKULPURI/ EAST	DELHI
11		Male	RAJA NI	, 73 RAMA MARKET PRITA	13	14	122	153	thin		RANI BAGH/NORTH WEST	DELHI
12	FAIZAN	Male	SHAB ANA	, , JHUGGI E 48/A25 JHUG	8	9	92	153	thin		SEEMAPURI/NOR EAST	DELHI
13	VARUN VOHRA	Male	ANIL VOHRA	, , H NO CA 64 TAGORE GA	30	31	153	122	normalmedium		UTTAM NAGAR/WEST	DE
14	SINGH	Male	MALT I	, , 1A/5 S AINIK ENCLAVE	30	31	153	305	strong		RANHOLA/WEST	
15	ARJUN MANDAL	Male	ASI RA J KR	, , H NO 81 A KALLU MOH	45	46	153	244	thin		AMAR COLONY/SOUTH- EAST	DELHI
16	MADHURI	Female	SANJEE VKUMAR	, , 2/44, GALI NO. 1, HARI	25	26	153	183	thin		SARAI ROHILLA/NORTH	
17	Female	Female	VAIDANTA ENTER	, MANTU CHAUDH 276-2	16	17	122	153	normalmedium		SHAHBAD DAIRY/OUTER DISTRICT	DELHI
18	AYESHA	Female	RAJESH THAPA	, , D-181, JJ COLONY SHAH	24	25	153	183	normalmedium		SUBHASH PLACE/NORTH WEST	DELHI
19	ARUN RAMANUJAN	Male	DR SU NDARA V M	, , FLAT NO C 1 RIDGE CAS	23	24	183	31	thin		MEHRAULI	SOUT
20		Female		,ARYAKANYASADAN,1488	26	27	4664	5578	normalmedium		DARYA GANJ/CENTRAL	
21	PANKAJ	Male	RAKE SH	, , H NO 191 9 BASTIJULAI	9	10	92	31	strong		SADAR BAZAR/NORTH	D
22	ARTI		PREMLA TA	, Fema le, JHUGGI NO. F-	15	16	122	153	thin		NARELA/OUTER DISTRICT	DELHI
23	KUMAR	Male	SARV ESHWAR KU	, , SARV ESHWAR , A 267 F	17	18	153	153	normalmedium		FATEHPUR BERI/SOUTH	DEL
24	SIMRAN	Female	SONIA	, , D-49, H ASTSAL VIHAR I	19	20	153	183	thin		UTTAM NAGAR/WEST	DE
25	AVNISH KUMAR SINGI	Male	NEET U SINGH	, , FLAT NO 6814 KHASRA	15	16	153	153	thin		MAURYA ENCLAVE/NORTH WEST	DELHI
26		Female	INDRA	, KH NO. 14/11/1, PARKAS	16	17	153	183	fat		SHAHBAD DAIRY/OUTER DISTRICT	DELHI
27	SHIVANI	Female	GOPAL	, , D-5/35, SHAHBAD DAIR	18	19	122	153	normalmedium		SHAHBAD DAIRY/OUTER DISTRICT	DELHI

Skup podataka Missing people

- Neinformativne značajke
  - *Date* – prazna značajka – sve vrijednosti su nedostajuće – ukloniti odmah
  - *Name*, *Relative*, *Address* – preveliki broj različitih vrijednosti – razmotriti za uklanjanje
- Značajka *Gender*
  - Ponegdje fali informacija, ali se nalazi skrivena u drugim značajkama, npr. u *Address* – prije uklanjanja neinformativnih značajki izvući vrijednost o spolu

Skup podataka Missing people

- Značajke *AgeStart*, *AgeEnd*, *HeightStart* i *HeightEnd*
  - Značajke s nelogičnostima i stršećim vrijednostima
  - Provjeriti nelogičnosti za svaki primjerak:
    - *AgeEnd* je manja od *AgeStart* ili *HeightEnd* je manja od *HeightStart*
    - *HeightStart* ili *HeightEnd* viši od 230 (cm) ili manji od 50 (cm)
    - Razlika između *HeightEnd* i *HeightStart* veća od 10 cm po svakoj godini razlike između *AgeEnd* i *AgeStart* (ubrzani rast?)
    - Još nešto?
  - Ponegdje i nedostaju vrijednosti – kako to riješiti?
  - Stršeće vrijednosti mogu se vizualizirati histogramom ili *box*-plotom

Skup podataka Missing people

- Značajka *Built*
  - Velika većina primjeraka ima vrijednosti *thin*, *normalmedium*, *strong*
  - Manji broj primjeraka ima neke druge vrijednosti: *fat*, *veryfat*, *verylanky*, *muscular*...
  - Neke vrijednosti nedostaju
  - Rješenje: izraditi histogram mogućih vrijednosti, razmotriti rješavanje nekonzistentnosti u podacima – jesu li *fat* i *strong* sinonimi, ili *strong* i *muscular*?

Skup podataka Missing people

- Značajka *State*
  - Veliki broj nekonzistentnosti – sinonimi su uneseni kao kratice punog naziva, npr. umjesto DELHI: DELH, DEL, DE, D ; umjesto WEST: WES, umjesto EAST: EAS
  - Rješenje: pobrojiti nizove znakova koji se pojavljuju, napraviti pravila zamjene i provesti zamjenu
  - Kako riješiti nedostajuće podatke za ovu kategoričku značajku?

Skup podataka Missing people

- Značajka *Dist*
  - Što napraviti s ovom značajkom?
  - Postoji veliki broj vrijednosti koje se pojavljuju, ali možda sadrži neku relevantnu informaciju
  - Možda ako je *State* = DELHI, izvući barem informaciju o dijelu grada (EAST, SOUTH-EAST, OUTER DISTRICT, itd.), a za ostale vrijednosti značajke *State* ostaviti postojeću vrijednost u *Dist*?
  - Pripaziti: dijelovi grada ponovno imaju problem sa sinonimima – EAS, WES...

Skup podataka Missing people

- Što na kraju?
- Napraviti neku statistiku nestajućih ljudi – ima li više npr. mladih žena koje nestaju u odnosu na ostale kategorije žena? – DA
- Modelirati nekim od postupaka za grupiranje podataka (npr. *k*-means) za izabrane parove varijabli radi bolje vizualizacije (npr. odnos između *AgeStart* i *State* – koristiti LabelEncoder)
- Vidi li se neki uzorak u ovim podacima? ☺

Zaključak

- Priprema podataka je važan korak koji prethodi modeliranju podataka
- Model procesa CRISP-DM obuhvaća faze razumijevanja podataka i pripreme podataka, svaku sa svojim generičkim zadacima koje se u praksi provode u određenoj zajedničkoj arhitekturi
- Razmatrali smo veći broj problema koji se pojavljuju u podacima – nedostajući podaci, stršeći podaci, nekonzistentni podaci...
- Metode za rješavanje problema imaju svaka svoje prednosti i nedostatke
- Za neke skupove podataka potrebno je kombinirati više metoda kako bi se podaci uspješno očistili