OBRADA PODATAKA

Hrvatski studiji

dr.sc. Luka Šikić

Diplomski studij sociologije

13 listopad, 2020

CILJEVI PREDAVANJA

- Strukturirani i ne-strukturirani podatci
- Kvantitativni i kvalitativni podatci
- Diskretni i kontinuirani podatci
- Razine podataka
- Podatci u “praksi”
- Big Data

STRUKTURIRANI I NESTRUKTURIRANI PODATCI

1. **STRUKTURIRANI**

- observacije sa karakteristikama, uglavnom organizirane u tablicu (redovi i kolone)
- znanstveno prikupljeni podatci, telefonski imenik
- manji dio podataka

2. **NESTRUKTURIRANI**

- podatci bez standardne organizacijske hijerarhije
- Facebook objave, Twitter, logovi na server, genetska sekvenca nukleotida, tekstualni podatci
- vjerojatno više od 80% svih podataka
- zahtjevaju prilagodbu prije analize

KVANTITATIVNI i KVALITATIVNI PODATCI

1. **KVANTITATIVNI**

- brojevi, matematičke procedure, prosjek, vremenski trend, threshold efekti

2. **KVALITATIVNI**

- "prirodne" kategorije, jezik
- najčešća observacija, jedinstvene vrijednosti

PRIMJER

```
import pandas as pd

# ucitaj CSV file sa URL
drinks = pd.read_csv('https://raw.githubusercontent.com/sir

# pregledaj prvih 10 redova
#drinks.head(10)
```

```
# prikazi podatke u tablici
py$drinks %>%
  head(10) %>%
  kbl() %>%
  kableExtra::kable_material_dark()
```

DISKRETNI I KONTINUIRANI PODATCI

1. **DISKRETNI**

- prebrojivi
- npr. igraća kocka

1. **KONTINUIRANI**

- postoje na kontinuiranoj skali
- npr. težina ili visina

ČETIRI RAZINE PODATAKA

1. NOMINALNI
2. ORDINALNI
3. INTERVALNI
4. OMJERNI

NOMINALNA RAZINA

- podatci opisani nazivom ili kategorijom (kategorički podatci)
- npr. spol, nacionalnost, biološke vrste
- ne mogu se obavljati matematičke operacije poput zbrajanja ili djeljenja
- računanje prosjeka ili drugih statističkih momenata nema smisla

ODINALNA RAZINA

- kategorički podatci koji imaju hijerarhijsku strukturu
- iako postoji hijerarhija, nije moguće utvrditi relativne razlike među opservacijama
- matematičke operacije kao zbrajanje ili dijeljenje nisu opravdane
- usporedbe i sortiranje podataka su opravdane
- moguće je koristiti medijan (ne i prosjek)

```
import numpy

# anketa o sreci na ljestvici 1-5
results = [5, 4, 3, 4, 5, 3, 2, 5, 3, 2, 1, 4, 5, 3, 4, 4,
4, 5, 4, 3, 2, 4, 4, 5, 4, 3, 2, 1]

# sortiraj rezultate
sorted_results = sorted(results)
```

INTERVALNA RAZINA

- npr. temperatura
- opravdane su matematičke operacije poput zbrajanja i oduzimanja
- opravdano korištenje mjera centralne tendencije i varijabilnosti

```
# temperatura frizidera u fahrenheitima mjerena svakih sat
temps = [31, 32, 32, 31, 28, 29, 31, 38, 32, 31, 30, 29, 30

# pogledaj prosjek i medijan
print("prosjek:",round(numpy.mean(temps),2))
```

```
## prosjek: 30.73
```

```
print("medijan:",round(numpy.median(temps),2))

## pogledaj mjere varijacije
```

OMJERNA RAZINA

- opravdane matematičke operacije množenja i dieljenja
- podatci na ovoj razni ne smiju biti negativni

```
# temperatura frizidera u fahrenheitima mjerena svakih sat
temps = [31, 32, 32, 31, 28, 29, 31, 38, 32, 31, 30, 29, 30

# izracunaj geometrijsku sredinu
num_items = len(temps)
product = 1.

for temperature in temps:
 product *= temperature

geometric_mean = product**(1./num_items)
```

PODATCI U "PRAKSI"

1. **Big Data**

- zbog veličine se ne mogu pohraniti na standardne rrelacijske baze
- nestrukturirani, semi-srtukturirani i strukturirani podatci
- od terabayt-a do zettabayt-a veličine
- osnova za strojno učenje, AI, predviđanje budućnosti

2. **Strukturirani, Ne-strukturirani i Kvazi-strukturirani podatci**

- podatkovni "polymorphism" zbog uspona novih tehnologije (web,mobile, socialNet,IoT, programming)
- MongoDB, RAVENDB,RETHINKDB, ORIENTDB, PostrgreSQL, ArangoDB cassandra

3. **Vremenske serije**