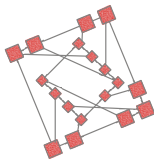


MATKINA BIBLIOTEKA

Podatkovna znanost
uvod u umjetnu inteligenciju za početnike

Tvrtko Tadić

matematika.hr



Intelektualno je vlasništvo, poput svakog drugog vlasništva, neotuđivo, zakonom zaštićeno i mora se poštivati. Nijedan dio ove knjige ne smije se preslikavati niti umnažati na bilo koji način bez pismenog dopuštenja nakladnika.

matematika.hr

CIP zapis je dostupan u nacionalnom skupnom katalogu knjižničnog sustava Bukinet (katalog.bukinet.hr) pod brojem 991005965303409366.

ISBN 978-953-6862-35-1

Tvrtko Tadić

Podatkovna znanost

uvod u umjetnu inteligenciju za početnike

1. izdanje

Zagreb, 2026.

© Hrvatsko matematičko društvo 2026.

GLAVNA UREDNICA
prof. dr. sc. Željka Milin Šipuš

RECENZENTI
doc. dr. sc. Ivan Biočić
izv. prof. dr. sc. Nenad Šuvak

LEKTORICA
Ivana Babić, prof.

ZA NAKLADNIKA
prof. dr. sc. Vesna Županović

SLOG, CRTEŽI I PRIJELOM
dr. sc. Tvrtko Tadić

DIZAJN OVITKA I ILUSTRACIJE
Ninoslav Kunc

NAKLADNIK
Hrvatsko matematičko društvo
Bijenička cesta 30, Zagreb

tel. 01/460 5751

<http://www.matematika.hr/>

e-mail: hmd@math.hr

TISAK
Tiskara Zelina

Sadržaj

0. Podatci danas	7
0.1. Zašto podatkovna znanost?	8
0.2. Koja je veza s umjetnom inteligencijom?	10
0.3. Kome je knjiga namijenjena?	10
0.4. Novi svijet podataka i umjetne inteligencije	11
1. Prikupljanje podataka	13
1.1. Prikrivena slučajnost	14
1.2. Opaženi podatci i (nevjerojatni) zaključci	17
1.3. Pristranost podataka. Kako je promašena anketa uništila tjednik	21
1.4. Slučajni i reprezentativni uzorak	26
1.5. Stratificirani uzorak	30
Zaključak. Izazovi prikupljanja podataka	34
2. Aritmetička sredina i standardna devijacija	37
2.1. Osnovne definicije i primjeri	38
2.2. Teorijska podloga*	44
2.3. Izazovi procjene	46
Zaključak. Prosjek i odstupanja	48
3. Medijan i kvantili	51
3.1. Primjeri, definicija i svojstva medijana	52
3.2. Podatci o letovima	55
3.3. Usporedba vrijednosti medijana i aritmetičke sredine	57
3.4. Teorijska pozadina medijana*	59
3.5. Kvantili i interval pouzdanosti	62
3.6. Procjena na temelju uzorka	70
Zaključak. Razumijevanje podataka kroz medijan i kvantile	70
4. Linearna regresija	73
4.1. Temperatura vrenja vode i tlak zraka	74
4.2. Odabir pravca	76

4.3. Cijene nekretnina	78
4.4. Šum i utjecaj na vrijednost parametara	80
4.5. Intervali pouzdanosti za opažene vrijednosti	83
4.6. Kako ne preplatiti nekretninu?	85
4.7. Teorijska pozadina. Metoda najmanjih kvadrata*	87
4.8. Modeli s više varijabli poticaja	90
Zaključak. Važnost regresijskih modela	93
5. Eksponencijalni rast	95
5.1. Pandemijska i ekspanzijska rast	96
5.2. Grananje kao model za epidemije. Broj reprodukcije R	101
5.3. Štednja, dionički indeksi i nekretnine*	106
Zaključak. Ekspanzijski rast i njegova ograničenja	117
6. Pokusi. A/B testiranje	121
6.1. Online pokusi	122
6.2. Mjerenje uspjeha – metrike	126
6.3. Efikasnost cjepiva i bacanje novčića	128
6.4. Radi li se zbilja o poboljšanju? Zašto testiranje?	132
Zaključak. Kontrolirani pokusi i pouzdano donošenje odluka	135
7. Simulacije	137
7.1. Je li u Lotu 7/35 broj izvučen previše puta?	138
7.2. <i>Let's make a deal!</i> Kako dobiti auto umjesto koze?	142
7.3. Simulacija ankete	146
7.4. Iracionalnost i (pseudo)slučajnost	153
7.5. Udruživanje na izborima za predstavnička tijela	158
Zaključak. Uloga simulacija	168
8. Umjetna inteligencija i uvjetna vjerojatnost*	171
8.1. Uvjetna vjerojatnost kao osnova zaključivanja	172
8.2. Prepoznavanje teksta	177
8.3. Jezični modeli	184
Zaključak. Od vjerojatnosti do inteligentnih sustava	195
9. Društvene i informacijske mreže	197
9.1. PageRank i slučajni posjetitelj internetskih stranica	199
9.2. Modeli simuliranja mreža	204
9.2.1. Erdős–Rényi slučajni graf	205
9.2.2. Model preferencijskog povezivanja	212
Zaključak. Raznolikost društvenih mreža	222
10. Vizualizacija podataka	225
10.1. Sezonalnost podataka i pomični prosjek	226
10.2. Koliko je ljudi više umrlo tijekom pandemije?	230
10.3. Toplinske karte i urbana područja interesa*	237

Zaključak. Vizualizacija podataka u vremenskim i prostornim pojavama	246
11.Privatnost i povjerljivost podataka	249
11.1. Povezivost podataka	250
11.2. Pogled u povijest privatnosti podataka	255
11.3. Pravni propisi i izazovi	260
11.4. Metode kontrole razotkrivanja	263
11.5. Diferencijalna privatnost*	268
Zaključak. Ravnoteža između privatnosti i korisnosti	276
O autoru	277
Od PMF-a u Zagrebu do Microsofta u Seattleu: matematičar koji radi na inteligentnim rješenjima za Office 365	278
Dodatci	282
Upute za upotrebu	282
Oznake	282
Često korišteni pojmovi	283
Dodatni web materijali	289
Tehnički nastanak ove knjige	289
Zahvale	290
Literatura	291
Kazalo	304

matematika.hr

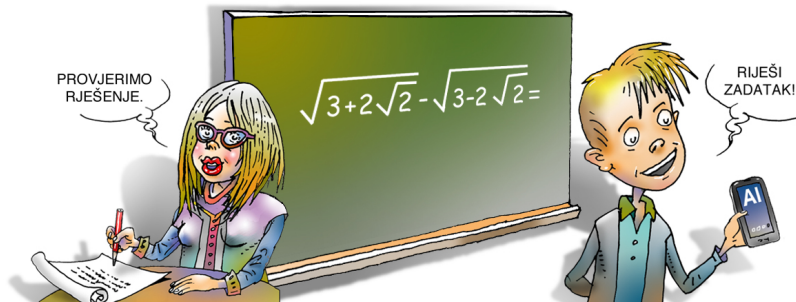
0. Podatci danas

Svaki dan na razne načine stvaramo i koristimo mnoštvo **podataka** od slanja poruka i fotografija do kupovine proizvoda i usluga. Podatci su ključni za razvoj znanosti, tehnologije, umjetnosti i društva. No kako smo došli do toga da na raspolaganju imamo toliko podataka? Kako ih prikupljamo, obrađujemo i analiziramo? I koje su prednosti i izazovi korištenja podataka u različitim područjima?

Ova knjiga namijenjena je čitateljima koji žele doznati više o podacima i njihovoj ulozi u našim životima. Ne trebate imati posebno predznanje o podacima, samo interes i znatiželju. Knjiga će vam objasniti osnovne pojmove i metode vezane uz podatke, kao i primjere iz stvarnog svijeta koji pokazuju kako se podatci koriste za rješavanje problema, donošenje odluka i stvaranje vrijednosti.

Navest ćemo tek neke proizvode koji su nastali zahvaljujući dostupnosti podataka:

- **Chatbotovi** poput ChatGPT-a, Claudea, Copilota, Geminija i drugih razvijeni su zahvaljujući javnoj dostupnosti **golemih količina tekstualnih podataka**. Ta dostupnost omogućila im je da oponašaju ljudsku komunikaciju i služe za brzo dobivanje informacija, automatizaciju zadataka i druge potrebe. Dok ovo čitate, kompanije koje ih razvijaju aktivno traže nove izvore podataka jer su postojeći već u velikoj mjeri iscrpljeni.



- Prije njih, **tražilice** poput Googlea, Binga i DuckDuckGoa omogućile su korisnicima da brzo i jednostavno pronađu informacije na internetu.

- **Društvene mreže** kao što su Facebook, Instagram, TikTok, Twitter/X... omogućile su korisnicima da dijele svoje misli, osjećaje, fotografije i videozapise s prijateljima, obitelji i pratiteljima, stvarajući nove oblike digitalne interakcije i zajedništva.
- **Usluge prijevoza** poput Ubera, Lyfta ili Bolta omogućuju nam da naručimo vožnju s nekim tko je u blizini i platimo je putem aplikacije.
- **Usluge mrežnog emitiranja** poput Netflixa, Spotifyja ili YouTubea omogućuju nam da gledamo ili slušamo sadržaj koji nas zanima kad god i gdje god želimo.
- U području **razvoja lijekova**, računalni modeli i analiza velikih količina podataka pomažu u otkrivanju novih molekula, ubrzavanju kliničkih ispitivanja i personalizaciji terapija.
- U **lancima opskrbe**, napredna analiza podataka omogućuje bolje predviđanje potražnje, optimizaciju zaliha i učinkovitiju distribuciju proizvoda.

Razvoj ovih proizvoda i usluga olakšao nam je i obogatio život na mnoge načine, ali je i otvorio nova pitanja poput privatnosti, sigurnosti, etičnosti i pravde. Kako možemo biti sigurni da naši osobni podatci neće biti **zloupotrijebljeni** ili prodani bez našeg znanja ili pristanka? Kako možemo zaštititi sebe od krađe identiteta, lažnih vijesti, kibernetičkih napada ili diskriminacije? Kako možemo biti odgovorni i kritični korisnici i stvaratelji podataka?

Ova knjiga pomoći će vam da bolje razumijete ove izazove i naučite kako se nositi s njima. Knjiga će vam također pokazati kako možete iskoristiti podatke za vlastitu dobrobit, bilo da je riječ o osobnom ili profesionalnom razvoju, obrazovanju ili zabavi. Dat će vam alate i savjete kako možete prikupljati, obrađivati i analizirati podatke sami ili uz pomoć drugih.

Nadam se da ćete uživati u čitanju ove knjige i da ćete postati dio zajednice ljudi koji se bave podacima. Podatci su budućnost, a vi ste dio nje!

0.1. Zašto podatkovna znanost?

Autor je podlegao trendu? **Podatkovna znanost** (engl. *Data Science*) zvuči modernije od statistike. Mnoga sveučilišta u svijetu preimenovala su svoje studentske programe iz *statistike* u *podatkovnu znanost*.

Što je zapravo podatkovna znanost? Je li to novi, moderniji naziv za statistiku ili se radi o nečemu drugome? Rasprave o tome traju jer postoje različita viđenja stvari.

Krenimo redom. **Statistika** je grana znanosti koja se tradicionalno bavila podacima koje su prikupljali znanstvenici iz raznih disciplina. Obradom tih podataka

0.1. Zašto podatkovna znanost?



Conwayev dijagram iz 2010. godine koji ilustrira što je podatkovna znanost. Prvi ga je predložio **Drew Conway**, američki znanstvenik i poduzetnik. Do ideje za dijagram došao je prilikom rada na svom doktoratu o računalnim metodama u društvenim znanostima.

pokušavalo se utvrditi prate li ti podatci matematičke zakonitosti ili modele. Na temelju toga donose se zaključci i odluke.

Mnoge metode u statistici razvijene su iz **matematike**, a nastale su praktičnom obradom i korištenjem podataka. S razvojem računala pohrana podataka i njihova obrada postali su bitno lakši. Razvojem interneta, mobilne tehnologije i računalnog oblaka (engl. *cloud computing*), prikupljanje, prijenos i obrada podataka dostigli su novu razinu.

Aritmetičku sredinu na papiru možemo izračunati za desetak brojeva, džepno računalo moći ćemo koristiti za do 100 brojeva, nakon čega ćemo se morati koristiti računalom. Danas možemo imati toliko podataka da oni jednostavno ne stanu na jedno računalo, nego se nalaze na više računala, tj. u **računalnom oblaku**¹. U tome slučaju, da bismo izračunali aritmetičku sredinu, trebat će nam poseban algoritam. Izračunavanje drugih jednostavnih statistika, poput medijana, može biti još složenije.

Brojni današnji proizvodi osmišljeni su tako da od korisnika prikupljaju određene podatke na temelju kojih vraćaju *rezultat*, primjerice preporuku sadržaja, rangiranu listu rezultata pretraživanja, predviđanje budućeg ponašanja ili automatski prepoznati objekt na slici. Taj rezultat korisnik mora dobiti **brzo** – u roku od nekoliko sekundi. Pristupi obradi podataka razlikuju se ovisno o tome radi li se o

¹Računalni oblak pojam je koji opisuje skup računala koja su povezana preko interneta ili neke druge mreže. Ta računala mogu dijeliti resurse, podatke i aplikacije međusobno ili s drugim korisnicima.

društvenim mrežama, pretraživanju Interneta ili računalnom vidu. Sve to dovelo je do pojave interdisciplinarnih timova koji se bave **podatkovnom znanosti**.

Statistika se tradicionalno poučava tek na fakultetima, nakon što se studenti upoznaju s osnovama više matematike. Neki te osnove prođu za vrijeme srednje škole i tijekom nje uspiju učiti o statistici. Međutim, takvo tradicionalno poučavanje često ne uključuje nikakvu izloženost **stvarnim podacima**, tako da i oni koji su imali sreću slušati statistiku često ostanu *podatkovno nepismeni*.

Ova knjiga prvenstveno se bavi podacima, a matematika je većinom na razini osnovne i ponegdje srednje škole. Mnogi koncepti za koje bi trebalo znanje više matematike opisani su simulacijama, vizualizacijama i na druge načine na temelju stvarnih podataka.

0.2. Koja je veza s umjetnom inteligencijom?

Umjetna inteligencija grana je računarstva koja se bavi stvaranjem sustava i programa za obavljanje zadataka koji zahtijevaju ljudsku inteligenciju, poput prepoznavanja slika, razumijevanja jezika, donošenja odluka i učenja.

Umjetna inteligencija pokušava imitirati čovjeka na temelju podataka koji su joj dani i stalno joj se daju. Slika, zvuk i jezik pretvaraju se u podatke koje onda računalo obrađuje i vraća neku povratnu informaciju ili obrađuje neki postupak.

U ovoj knjizi koja je, kako smo rekli, ograničena gradivom osnovne i srednje škole, dat ćemo vam jasniju sliku o tome kako funkcioniraju modeli koji stoje iza onoga što danas zovemo umjetnom inteligencijom.



0.3. Kome je knjiga namijenjena?



Knjiga je namijenjena svakome koga zanima kako se podatci koriste. Na pomno odabranim skupovima podataka knjiga ilustrira razne principe koji se koriste u industriji i praksi obrade podataka. U ovoj knjizi naći ćete brojne tablice i grafičke prikaze podataka, dosta simulacija i nešto matematike. Oko 70 % matematičkih sadržaja na razini je **osnovne škole**, dok je preostalih 30 % na razini **srednje**.

Svi ostali napredniji sadržaji ilustrirani su simulacijama i na stvarnim podacima uz pomoć računala. Poznavanje rada na računalu nije nužno za razumijevanje ove knjige, ali je neizbježno ako želite obrađivati podatke. Čitatelji koji to žele, moći će vidjeti kako su nacrtani prikazi podataka ili kako su napravljene simulacije posjetom

0.4. Novi svijet podataka i umjetne inteligencije

GitHub² stranici ove knjige <https://podatkovna-znanost.github.io>, gdje su pohranjeni kôd u Pythonu i Excel bilježnice. Excel je danas standardan alat za jednostavnije skupove podataka, dok se Python koristi za naprednije tehnike i veće skupove podataka. U trenutku pisanja ove knjige barem se dio učenika u svome školskom obrazovanju susreo s jednim ili oba ova alata.

Na kraju ove knjige nalazi se pregled literature koji će vas uputiti na to gdje možete doznati više o ovome području, ovisno o razini obrazovanja na kojoj se nalazite.

Učenici, kojima je knjiga prvenstveno namijenjena, dobit će uvid u nadasve zanimljiv svijet podataka. Također im može dati jasniju sliku o tome žele li nastaviti obrazovanje u tom smjeru. Danas se brojna područja bave podacima ili ih koriste kako bi bolje razumjeli svijet oko nas. Možete ih nastaviti proučavati na studiju matematike, tehnike, prirodoslovlja, biomedicine, društvenih znanosti...

Nastavnici (svih razina) u ovoj će knjizi naći bogat izvor primjera i ideja kako približiti učenicima i studentima ovo područje.

Studentima, koji se često susretnu s teorijskim i primjerima siromašnim kolegijima iz statistike³, ova će knjiga pomoći u boljem razumijevanju sadržaja i pojasniti kako stvari izgledaju u (industrijskoj) praksi.

Podatkovni znanstvenici i drugi koji moraju razumjeti podatke u ovoj će knjizi naći niz zanimljivih tema prikazanih na vrlo pristupačan način. Nadam se da će im to dati ideju o tome kako mogu riješiti probleme, gdje mogu naći korisnu literaturu i kako izbjeći greške koje su drugi napravili prije njih.

0.4. Novi svijet podataka i umjetne inteligencije

Ušli smo u doba u kojem se mnoge odluke, od onih osobnih do onih društvenih, donose uz pomoć podataka. Nekad smo se oslanjali na dojam, iskustvo i autoritet, a danas nas sve češće prate brojevi, grafovi, modeli i **algoritmi** koji *predlažu* što bismo mogli gledati, kupiti, čitati ili komu vjerovati. Tome se posljednjih godina pridružila i **umjetna inteligencija**: sustavi koji iz podataka uče obrasce, a zatim generiraju odgovore, preporuke ili automatiziraju postupke. U takvom svijetu nije dovoljno samo imati podatke; važno je znati **što oni doista znače**, gdje nas mogu zavesti i kako ih odgovorno koristiti.

U poglavljima koja slijede vidjet ćete kako se podatci prikupljaju, kako u njima prepoznamo obrasce i gdje pritom možemo pogriješiti. Upoznat ćete osnovne statističke mjere, brojne modele, pokuse i simulacije. Vidjet ćete kako se na tim

²github.com je mrežno mjesto za pohranu i organizaciju programskog kôda, gdje se projekti mogu sigurno spremati i pregledavati. Koriste ga brojni pojedinci i kompanije kako bi učinili svoj kôd javno dostupnim.

³Unatoč ovom komentaru, dobro razumijevanje sadržaja tih kolegija bitno je i preporučljivo.

temeljima grade sustavi umjetne inteligencije, kako se podatci koriste za analizu **društvenih mreža** te izradu **vizualizacija** potrebnih za donošenje svakodnevnih odluka. Na kraju, vratit ćemo se važnim pitanjima **privatnosti** i društveno odgovornog upravljanja podacima.

Ova knjiga nastoji biti vaš vodič kroz taj novi krajolik. Pomoći će vam da podatke čitate kritički, da iz njih izvučete smisao te da razumijete osnovne ideje iza metoda koje se danas koriste u **znanosti** i **industriji**.

Web-stranica knjige sadrži dodatne sadržaje (snimke povezane s knjigom, poveznice na dodatnu literaturu i izvore podataka,...) pa vas pozivam da posjetite

<https://matematika.hr/knjiga-podatkovna-znanost>.

Ako u tekstu pronađete pogrešku, nejasnoću ili imate prijedlog kako nešto poboljšati, autor će vam biti zahvalan na povratnoj informaciji. Komentare možete poslati e-mailom ili putem formulara na gore navedenoj mrežnoj stranici. Svaka dobra sugestija pomaže da sljedeće izdanje bude jasnije, točnije i korisnije svima.

Želim vam ugodno čitanje!



U Seattleu i Zagrebu 2026. godine.

Tvrtko Tadić
tvrtko@math.hr

1. Prikupljanje podataka



Prikupljanje podataka predstavlja velik izazov u mnogim industrijama i znanstvenim područjima. Glavni problemi uključuju osiguranje **točnosti** i **pouzdanosti** podataka, zaštitu privatnosti i sigurnost informacija te upravljanje velikim količinama podataka. Također, potrebno je osigurati etičko prikupljanje podataka, izbjegavajući pristranost i diskriminaciju.

Korištenje različitih izvora podataka i održavanje njihova integriteta također predstavljaju velik izazov. U konačnici, kvalitetno prikupljanje podataka ključno je za donošenje **informiranih odluka** i **optimizaciju poslovnih procesa**, ali zahtijeva pažljivu strategiju i stalna poboljšanja.

1.1. Prikrivena slučajnost

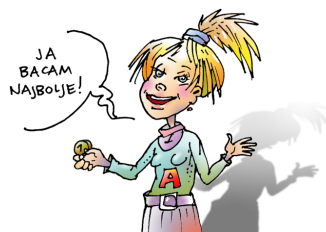
Promotrimo sljedeći primjer. U razredu su 32 učenika. Učitelj je svakome od njih dao kovanicu od jednog eura. Učenici će igrati igru prema sljedećim pravilima:

- U svakome kolu učenici bacaju novčić koji su dobili.
- Ako na bačenome novčiću padne pismo (broj 1), učenik ostaje u igri, a ako padne glava (kuna), učenik ispada iz igre.
- Pobjednik je onaj koji zadnji ispadne iz igre.



Pismo i glava hrvatske kovanice eura iz 2023. godine (Izvor slike: Hrvatska narodna banka)

Nakon pet bacanja pobjeđuje Ana jer je jedina ostala u igri. Bravo, Ana!



Nakon igre prirodno je postaviti ovakva pitanja:

- Zna li Ana bolje bacati novčić od drugih?
- Treba li razred, kako bi povećao izgleda za pobjedu, poslati Anu da igra s predstavnikom susjednog razreda u školi?

Na oba pitanja odgovor je, naravno, ne.

Ana ne zna bacati novčić ništa bolje od svoga razrednog prijatelja Matka koji je ispaio iz igre već u prvom kolu.

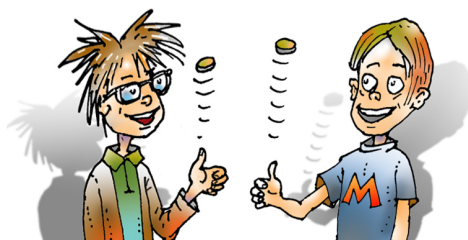
Premda na prvi pogled podatci daju Ani prednost nad Matkom i drugima, u ovom slučaju riječ je o **uobičajenoj realizaciji slučajnosti**.

Očekivani ishod igre

Ako su svi novčići simetrični i jednaki u svakom kolu očekujemo da će pola igrača ispasti.

Tako u prvom kolu sudjeluju 32 učenika, a očekujemo da će u drugo kolo ući njih pola, tj. 16 učenika. U trećem kolu očekujemo 8 učenika, u četvrtom 4, u petom 2, a nakon petog kola očekujemo da će u igri ostati samo jedan učenik.

Pukom slučajnošću to je ovaj put bila Ana, a isto tako mogao je to biti Matko, kao i bilo tko drugi iz njihova razreda.



Kolika je vjerojatnost da se ovako nešto dogodi?

Očekivani ishod očito ne znači da će se takav ishod uistinu i dogoditi.

Očito je da igra može prestati i prije petog bacanja jer su svi sudionici dobili glavu u nekome od prethodnih bacanja. Isto tako može se dogoditi da nakon petog bacanja imamo više učenika koji su još u igri.

Ovaj problem istražiti ćemo pomoću simulacija. Simuliranje možemo izvesti stvarnim bacanjem pravog novčića. Budući da želimo provesti postupak veći broj puta, koristiti ćemo se računalom. U programskom jeziku prema vlastitom izboru zapisati ćemo sljedeći postupak – simulaciju igre:

za sve učenike u razredu
 Simuliraj bacanje novčića 5 puta;
ako je palo pet pisama
 Zabilježi;

Zabilježi koliko je učenika imalo pet pisama;

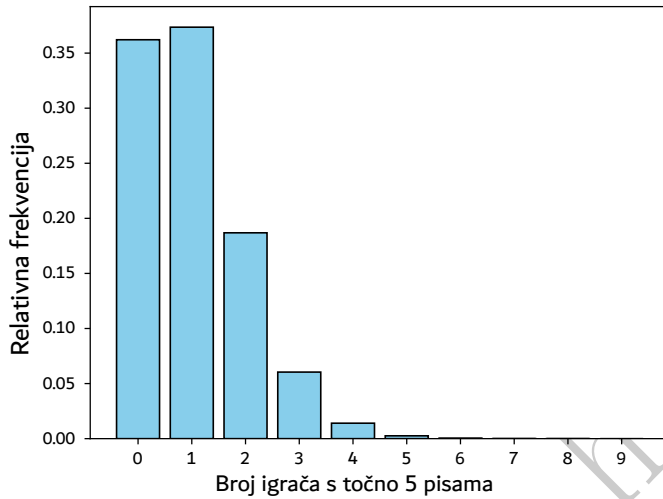
Ovo se može jednostavno isprogramirati i izvesti na računalu. Simulirali smo igru milijun puta, a rezultati simulacija zabilježeni su u Tablici 1.1. i grafički prikazani na Slici 1.1.

Kao što vidimo iz rezultata, ishod gdje točno jedan učenik ostane u igri nakon pet bacanja dogodio se u 374 427 ili 37.4 % slučajeva. Ishod u kojem nakon petog bacanja nitko nije imao niz od pet pisama dogodio se u 361 453 ili 36.1 % slučajeva. Nakon pet bacanja imamo dva ili više igrača u nešto manje od 26.5 % slučajeva.

Rezultati dobiveni simulacijama vrlo su bliski vjerojatnostima za ove događaje. Vjerojatnost da nakon pet bacanja u igri ostane točno jedan učenik može se precizno izračunati te iznosi 37.37 %.

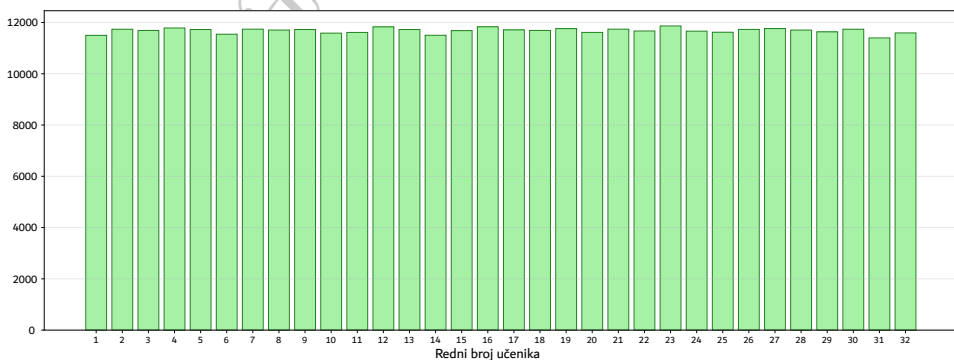
Broj igrača koji je imao pet pisama	Broj simuliranih igara s tim ishodom
0	361 453
1	374 427
2	186 647
3	60 605
4	13 895
5	2 542
6	378
7	43
8	8
9	2
Ukupno	1 000 000

Tablica 1.1. Rezultati simulacija pokazuju kako se u pet bacanja najčešće dogodi da točno jedan učenik ima 5 pisama. No mogući su i drugi ishodi. Naprimjer, u 2 simulacije igre od njih milijun dogodilo se da čak 9 učenika ima 5 pisama nakon 5 bacanja.



Slika 1.1. Dijagram relativnih frekvencija pokazuje koliko je igrača ostalo nakon pet bacanja u milijun simuliranih igara.

Simulacija se može i doraditi. Primjerice, možemo zabilježiti tko je učenik koji je pobjednik nakon pet bacanja. Svakom učeniku dodijelimo broj od 1 do 32 i bilježimo tko je imao pet pisama. Dijagram na Slici 1.2. pokazuje da je u 374 427 simuliranih igara, gdje je točno jedan učenik imao pet pisama, svaki učenik bio *podjednako* mnogo puta pobjednik.



Slika 1.2. Frekvencija koliko je puta koji učenik (označen brojem od 1 do 32) pobijedio u slučaju kad je točno jedan pobjednik nakon pet bacanja. Vidimo da su stupci podjednake visine.

1.2. Opaženi podatci i (nevjerojatni) zaključci

Zaključak

Situacija u kojoj jedan od 32 učenika nakon pet bacanja novčića završi kao jedini koji je dobio pet pisama vrlo je vjerojatna. Međutim, takav ishod **ne znači da je taj pojedinac bolji od ostalih**. Kako je u ovom slučaju netko morao biti izabran, pukom slučajnošću odabran je taj pojedinac.

Promatračima se može učiniti da taj pojedinac zna bacati novčić, dok će taj pojedinac misliti da mu je krenulo. No, uočite da nismo prije početka igre utvrdili tko će ta osoba biti, nego tek nakon nje, što je bitna razlika. Ako ta osoba može ponavljati pobjede u ponovnim izvođenjima ove igre, to bi već značilo da je manje vjerojatno riječ o slučajnosti.

Slični zaključci vrijede ako je ostalo i više od jedne osobe nakon pet bacanja. Sreća nas često može prevariti i odigrati ulogu koje ne moramo biti svjesni.

1.2. Opaženi podatci i (nevjerojatni) zaključci

Svakodnevno se objavljuju različita znanstvena i društvena istraživanja temeljena na prikupljenim podatcima. Čitatelji su mogli uočiti da, ponekad zaključci ovih istraživanja mogu biti u sukobu jedan s drugim. Na Slici 1.3. nalaze se naslovi nekih članaka iz razdoblja pandemije koronavirusa.

Istraživanja se razlikuju:

- prema načinu na koji su odabrani sudionici istraživanja;
- prema postavljenim pitanjima (zabilježenim osobinama sudionika).

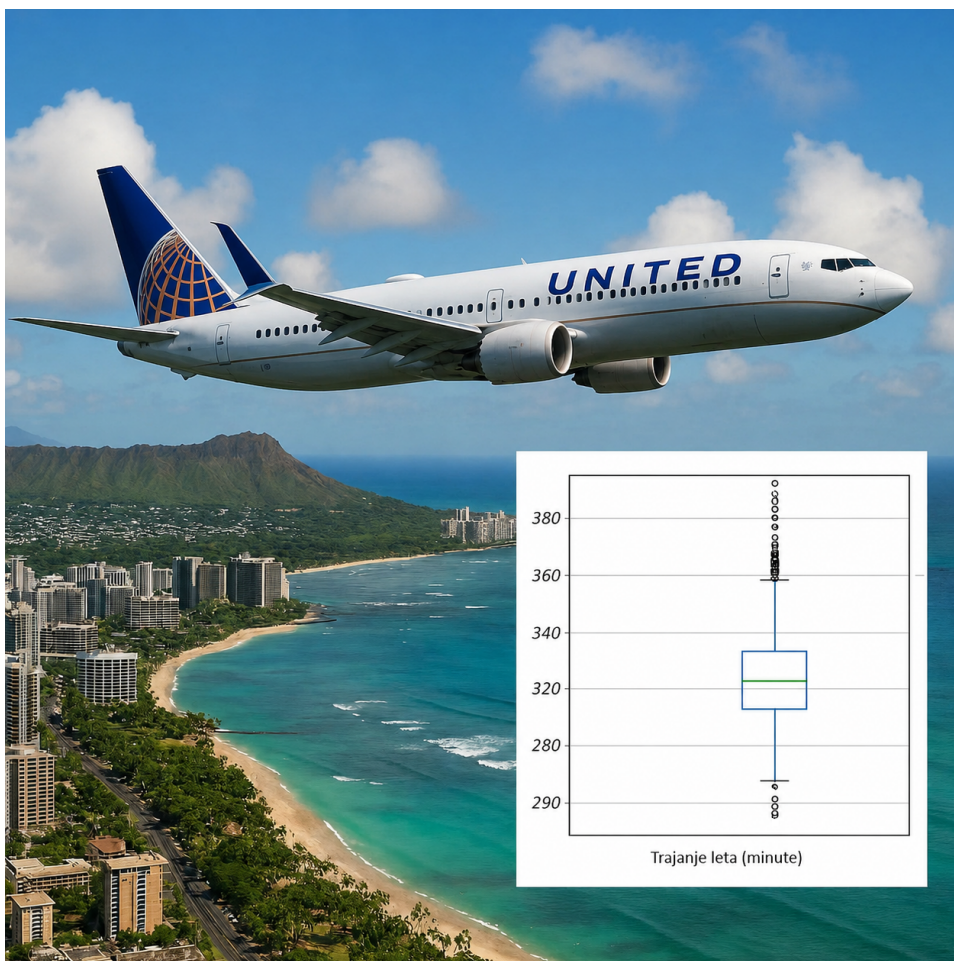
Kontrolirani eksperimenti

Odabir sudionika iznimno je bitan. U §6.3. vidjet ćemo kako je provedeno ispitivanje efikasnosti cjepiva. Ispitanici su bili podijeljeni u dvije podjednako velike skupine. Dodatno se pazilo da ove dvije skupine budu slične općoj populaciji, tj. da sastav skupina po zastupljenosti bitnih obilježja (spol, dob, krvna grupa...) odražava cjelokupno stanovništvo kojemu je cjepivo namijenjeno. Pokazalo se da je skupina koja je dobila cjepivo imala manje oboljelih i manje preminulih od skupine koja je dobila *placebo pripravak*. Ovaj način istraživanja zove se **kontrolirani (A/B) eksperiment**. On predstavlja standard za donošenje zaključaka djeluje li neki postupak ili ne, a primjenjuje se u raznim disciplinama gdje je to moguće i posvećeno mu je cijelo poglavlje §6. No kao što ćemo vidjeti, u nekim slučajevima takve eksperimente nije moguće provesti.

Istraživanja opažanjem

Naslovi sa Slike 1.3. nastali su za vrijeme pandemije koronavirusa kada je postojao veliki interes da se utvrdi koliko dobro štiti tzv. *prirodni imunitet* koji su pojedinci

3. Medijan i kvantili



U prošlom poglavlju predstavili smo jednu mjeru za sredinu podataka – **aritmetičku sredinu**. Imali smo se priliku upoznati s njezinim svojstvima, te pokazati neke probleme koji mogu nastati njezinom interpretacijom u kontekstu podataka kojima raspolažemo.

U ovom poglavlju upoznajemo se s jednom novom mjerom za sredinu skupa podataka – **medijanom**. Vidjet ćemo da se medijan nalazi u širem skupu podatkovnih funkcija koje zovemo **kvantili**. A čemu služe, imat ćemo priliku vidjeti na primjeru podataka o kašnjenju zrakoplova.

3.1. Primjeri, definicija i svojstva medijana

Ivona provjerava letove za Havaje

Ivona putuje na Havaje preko Los Angelesa. Kako bi znala kad će stići, Ivona je odlučila provjeriti podatke o sljedećemu:

- Koliko letovi iz Los Angelesa za Honolulu kasne u polasku?
- Koliko traje let do Honolulua?

Prije nego što će doći do podataka o kojima ćemo govoriti u §3.2., Ivona se raspitala kod pet prijatelja.

Svi su joj rekli da je let trajao otprilike 5,5 sati, dok je kašnjenje letova u polasku bilo

$$0, 0, 0, 88, 10 \quad (3.1)$$

minuta.



Ivona je razmišljala o tome kako sažeti ove podatke. Prosjek tj. aritmetička sredina od 17.6 minuta ne čini joj se kao podatak koji *informativno* opisuje kašnjenje u polasku. Jer osim jednog leta koji je kasnio puno više, svi ostali poletjeli su na vrijeme.

Ivonu zanima koja bi se mjera mogla iskoristiti za **opisivanje** ovoga skupa podataka.

Osnovno o medijanu

Postoji više mjera koje mogu opisati ovaj niz podataka, no kao što ćemo vidjeti, **medijan** je mjera kojom će Ivona opisati podatke o kašnjenju leta u polasku tako da informacija bude korisna.

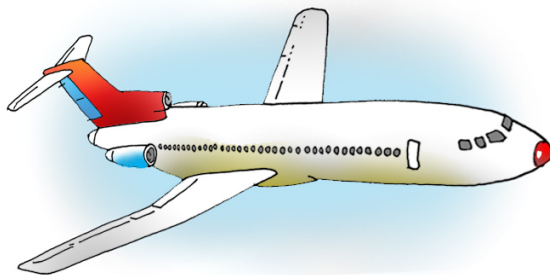
Idejno, medijan je M broj koji brojevnne podatke raspoređuje u **dva podskupa podataka**:

- podatke koji su manji ili jednaki od M ;
- podatke koji su veći ili jednaki od M ;



Honolulu (Izvor fotografije: Wikipedija)

3.5. Kvantili i interval pouzdanosti



Niz podataka / opis vrijednosti	5 %-kvantili	95 %-kvantil
Trajanje leta	300	345
Kašnjenje leta	0	87.75

Tablica 3.5. Izračun kvantila za trajanje i kašnjenje leta

Interval $[q_{0.05}, q_{0.95}]$ ima nekoliko dobrih svojstava:

- Najmanje 90 % članova niza podataka ima vrijednosti u tom intervalu. (Uočimo: ako su sve vrijednosti iste, interval je jedna točka i svi članovi niza imaju vrijednosti u tom intervalu.)
- 5 % najnižih i 5 % najviših vrijednosti nisu u intervalu, osim ako se te vrijednosti ne pojavljuju često. Ovo je dobro jer se ekstremne vrijednosti često postižu u okolnostima koje nisu uobičajene. Primjerice, kašnjenje od 20 sati možda je uzrokovano snježnom olujom koja do tada nikad nije bila zabilježena, ili se let od 280 minuta dogodio jer je taj dan putnike prevezio brži zrakoplov koji nikada ne leti na toj relaciji.

Definicija 3.9. Interval $[q_{0.05}, q_{0.95}]$ zovemo **90 %-interval pouzdanosti**.

Interval pouzdanosti koristi se u praksi tako da Ivona priopći kako:

- let između Los Angelesa i Honolulua u (bar) 90 % slučajeva traje od 300 do 345 minuta,
- isti let u polasku u (bar) 90 % slučajeva kasni od 0 do 87 minuta.

Pogled na buduće podatke

Postavlja se pitanje za što možemo iskoristiti ove zaključke. Ovi zaključci korisni su za planiranje leta. Neka od pitanja na koje Ivona može dobiti odgovor su:

- Kada okvirno stižemo? Koliko bi let mogao kasniti u polasku?
- Koliko hrane da ponesemo u avion? Koliko stranica knjige možemo pročitati ili filmova pogledati u zrakoplovu?

4. Linearna regresija



U prirodnim, tehničkim i društvenim znanostima često se susrećemo s pojavama koje pokazuju određene zakonitosti – pravilnosti koje se mogu opisati matematičkim modelima. Među njima, **linearni modeli** – u kojima se pojave opisuju **linearnim funkcijama** nekih opažanja – zauzimaju posebno mjesto zbog svoje jednostavnosti, razumljivosti i široke primjenjivosti. Oni omogućuju analizu od-

Dokaz. Za sve $i = 1, \dots, n$ vrijedi $y_i = ax_i + b$. Zato je

$$\begin{aligned}\bar{y} &= \frac{1}{n}(y_1 + \dots + y_n) = \frac{1}{n}((ax_1 + b) + \dots + (ax_n + b)) \\ &= \frac{a}{n}(x_1 + \dots + x_n) + b = a\bar{x} + b.\end{aligned}\quad (4.4)$$

Sada je

$$(x_i - \bar{x})(y_i - \bar{y}) = (x_i - \bar{x})((ax_i + b) - (a\bar{x} + b)) = a(x_i - \bar{x})^2 \quad (4.5)$$

za $i = 1, \dots, n$. Dobivamo

$$\hat{a} \stackrel{(4.2), (4.5)}{=} \frac{a(x_1 - \bar{x})^2 + a(x_2 - \bar{x})^2 + \dots + a(x_n - \bar{x})^2}{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2} = a,$$

a iz ovoga, koristeći (4.4), odmah slijedi $\hat{b} = \bar{y} - \hat{a}\bar{x} = (a\bar{x} + b) - a\bar{x} = b$. ■

4.3. Cijene nekretnina



Internetska stranica Redfin⁵ pruža podatke o nekretninama, uključujući oglašene i realizirane prodajne cijene u SAD-u. U ovom ćemo primjeru analizirati kretanje cijena *kuća u nizu*⁶ na području grada Seattlea, unutar poštanskog broja 98122. Imamo skup od 50 nekretnina koje su prodane u razdoblju od veljače do kolovoza 2025. Dio podataka sa svojstvima tih prodanih nekretnina dan je u Tablici 4.2.

Podatci uključuju prodajnu cijenu izraženu u američkim dolarima te površinu u kvadratnim stopama (*sqft*). Za analizu ćemo ponovno koristiti Excel.

⁵Podatci su preuzeti sa stranice [redfin.com](https://www.redfin.com).

⁶engl. *townhomes*

4.3. Cijene nekretnina

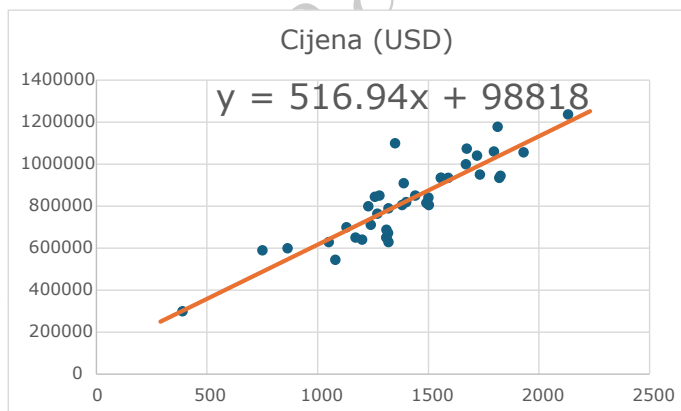
Površina (sqft)	Cijena (USD)	Zemljište (sqft)	Godina izgradnje	Spavaćih soba	Kupaonica
750	590 000	1260	2004	2	1.5
1557	934 950	1739	2016	3	2
1674	1 074 900	1124	2024	4	2.5
1820	935 000	1537	2007	3	3.5
1050	605 000	950	2020	2	2
1400	819 000	1051	2019	3	2
1050	630 000	1300	2007	2	1.5
1796	1 060 000	1276	2009	3	2.5
2500	1 114 500	2521	1891	4	3.5

Tablica 4.2. Pregled nekretnina i nekih njihovih svojstava

Kao što je vidljivo na Slici 4.5., linearni regresijski model procjenjuje da odnos između površine i cijene nekretnine prati pravac

$$y = 516.94x + 98\,818, \quad (4.6)$$

gdje je x površina u kvadratnim stopama, a y prodajna cijena u dolarima.



Slika 4.5. Cijene kuća u nizu (u dolarima) u odnosu na površinu i regresijski pravac.

Zanimljivo je primijetiti da model predviđa cijenu od 98 818 dolara za nekretninu s površinom $x = 0$. Iako se to na prvi pogled može činiti paradoksalnim, takav rezultat ukazuje na činjenicu da **vrijednost nekretnine ne ovisi isključivo o unutarnjoj površini**. Naime, cijenu značajno određuje i zemljište na kojemu se nekretnina nalazi. Nekretnina s malom kvadraturom, ali smještena na atraktivnoj lokaciji ili na velikom zemljištu, može imati visoku tržišnu vrijednost.

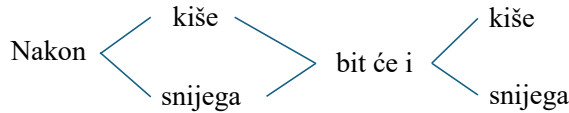
8. Umjetna inteligencija i uvjetna vjerojatnost*



* Ovo poglavlje namijenjeno je čitateljima upoznatima s gradivom srednje škole. Poželjno bi bilo da su upoznati s osnovnim pojmovima iz vjerojatnosti kao što su: vjerojatnosni prostor, svojstva vjerojatnosti... Ovdje ćemo uvesti pojmove koji su

Jednostavni model

Neka je alfabet zadan kao $\Sigma = \{ "Nakon", "kiše", "snijega", "bit će i" \}$.

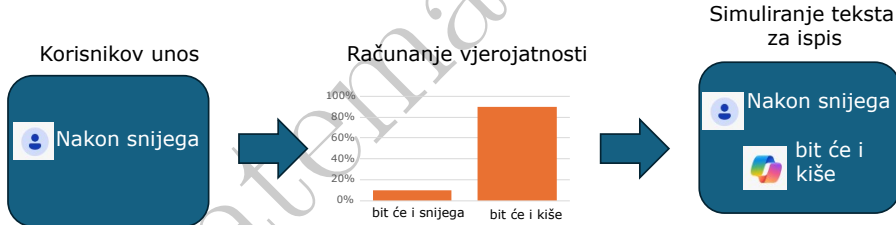


Slika 8.5. Jednostavni jezični model

\mathbb{P} možemo definirati tako da bude ispunjeno:

- tekst ima 4 simbola;
- tekst počinje simbolom "Nakon";
- Drugi i četvrti simbol su iz skupa $\{ "kiše", "snijega" \}$;
- treći je simbol u tekstu "bit će i".

Kao što je prikazano na Slici 8.5, u ovom se modelu zapravo drugi i četvrti simbol pojavljuju po nekom vjerojatnosnom modelu, dok su prvi i treći simbol zadani.



Slika 8.6. Pojednostavljena shema rada jezičnih modela

Jednostavna implementacija

Nakon što korisnik unese svoj tekst, aplikacija će:

- procijeniti **vjerojatnosni model** nastavka teksta **uvjetno** na uneseni tekst. Ovaj model označavamo⁶ s

$$\mathbb{P}(\cdot | \text{korisnikov unos}).$$

- **simulirati uzorak** na temelju ovog modela i ispisati tekst korisniku.

⁶Izraz $\mathbb{P}(\cdot|B)$ označava preslikavanje $A \mapsto \mathbb{P}(A|B)$.

O autoru

Dr. sc. **Tvrtko Tadić** je viši primijenjeni znanstvenik u Microsoftu, aktivist podatkovne pismenosti, autor triju knjiga, sedam znanstvenih članaka i više od pedeset stručnih radova.

Diplomirao je *matematičku statistiku i računarstvo* na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu, gdje je potom radio kao znanstveni novak i asistent na Zavodu za vjerojatnost i statistiku. Tu je imao zadovoljstvo držati vježbe iz više vjerojatnosno-statističkih kolegija, kao i baza podataka.

Doktorski studij završio je na Sveučilištu Washington u Seattleu unutar istraživačke grupe za vjerojatnost. Tijekom doktorskog studija slušao je o raznim temama iz vjerojatnosti, optimizacije, statistike i računarstva.

Prvo industrijsko iskustvo stječe u posljednjem ljetu doktorskog studija, kao *Analytic Science Intern* u Fair Isaac Corporation – FICO, vodećoj kompaniji za kreditno bodovanje potrošača. Nakon obrane doktorske disertacije pod naslovom *Vremenoliki grafovski modeli*, zapošljava se u Microsoftu, gdje prvo radi na razvoju lokalnih pretraga i karata za tražilicu Bing. S razvojem industrije, mijenjaju se i njegovi projekti unutar Microsofta – od digitalnog asistenta Cortane, preko rudačenja znanja, do inteligentne pretrage dokumenata za potrebe umjetne inteligencije koja pokreće M365 Copilota.

Aktivno volontira kao urednik web-stranice Hrvatskog matematičkog društva — matematika.hr. Član je Izvršnog odbora HMD-a, Odbora Inženjerske sekcije te savjeta časopisa *Poučak*. Član je *American Statistical Association*, te jedan od osnivača i prvi tajnik podružnice *Association of Croatian American Professionals* u Seattleu.

Gostujući je znanstvenik na Sveučilištu Washington u Seattleu. Održao je brojna znanstvena izlaganja, seminare i predavanja na sveučilištima i konferencijama u SAD-u, Hrvatskoj i drugim zemljama. Dobitnik je više nagrada i stipendija.



Autor prilikom promocije u doktora znanosti

Literatura



O podacima, vjerojatnosti, statistici i umjetnoj inteligenciji napisano je mnogo. U nastavku, odabranim popisom literature nastojimo dati smjernice o materijalima koji mogu poslužiti za pojedine teme i razine predznanja.

Osnovnoškolci mogu pronaći uvod u vjerojatnost i statistiku u knjizi [2], dok se početnici zainteresirani za umjetnu inteligenciju mogu upoznati s osnovama u knjizi [43].

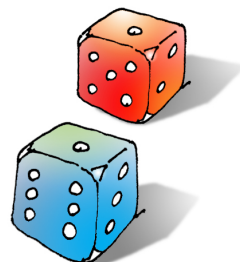
Za srednjoškolce je poželjno da su upoznati sa sadržajem knjiga [3] i [4], jer one pružaju teorijske temelje koji olakšavaju razumijevanje knjige. U vrijeme pisanja ove knjige Python je jedan od najraširenijih programskih jezika u obrazovanju i industriji, pa za njegovo bolje razumijevanje čitateljima preporučujemo korištenje standardnih udžbenika poput [60]. Knjiga [71], namijenjena natjecateljima iz matematike, u dodatcima A i B obrađuje teme iz teorije grafova i vjerojatnosti.

Napredniji čitatelji (studenti, nastavnici i ostali), zainteresirani za dublje razumijevanje tema obrađenih u ovoj knjizi, mogu pronaći prošireni popis relevantnih izvora raspoređenih po temama. Svaka tema uključuje knjige, znanstvene i stručne članke te video materijale.

Većina autorovih članaka uključena je i proširena u ovoj knjizi.

Vjerojatnost i statistika

Popularni udžbenik Sveučilišta Kalifornija u Berkeleyu [6] ostvario je dobru ravnotežu između teorije i stvarnih podataka. S druge strane, knjiga [7] dobro je sažet teorijski udžbenik za matematičku statistiku. Udžbenici Sveučilišta u Osijeku [10] i [11] obrađuju teme iz vjerojatnosti i statistike na raznovrsnim primjerima. Knjige [12] i [13] teoretski obrađuju ovu tematiku za studente tehničkih studija.



[1] Darrell Huff. *How to lie with statistics*. W. W. Norton & Company, 1993.

[2] Petar Vranjković. *Kocka je bačena*. Školska knjiga, 2006.

- [3] Nikola Sarapa. *Vjerojatnost i statistika I. dio: Osnove vjerojatnosti – kombinatorika*. Školska knjiga, 1993.
- [4] Nikola Sarapa. *Vjerojatnost i statistika II. dio: Osnove statistike – slučajne varijable*. Školska knjiga, 1996.
- [5] Jovan Mališić. *Vervovatnoća i matematička statistika: udžbenik za 4. razred Matematičke gimnazije*. Krug, Beograd, 1999.
- [6] David Freedman, Robert Pisani, Roger Purves. *Statistics*. W.W. Norton & Company, 1978.
- [7] George Casella and Roger L. Berger. *Statistical inference*. Wadsworth & Brooks/Cole Advanced Books & Software, 1990.
- [8] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer, 2004.
- [9] Deborah Nolan, Terry Speed. *Stat Labs: Mathematical Statistics Through Applications*. Springer Texts in Statistics. Springer, 2001.
- [10] Mirta Benšić, Danijel Grahovac, Ivan Papić, Nenad Šuvak. *Primijenjena statistika*. Sveučilište J. J. Strossmayera, Fakultet primijenjene matematike i informatike, 2025.
- [11] Mirta Benšić, Nenad Šuvak. *Uvod u vjerojatnost i statistiku*. Sveučilište J. J. Strossmayera, Odjel za matematiku, Osijek, 2014.
- [12] Željko Pauše. *Uvod u matematičku statistiku*. Školska knjiga, 1993.
- [13] Neven Elezović. *Vjerojatnost i statistika*. Element, 2018.
- [14] Zoran Vondraček. *Markovljevi lanci (skripta)*. PMF-MO, 2008.
- [15] Nikola Sarapa. *Teorija vjerojatnosti*. Školska knjiga, 2002.

Osnovnoškolski članci

U sljedećim člancima obrađene su teme koje spadaju u statistiku i vjerojatnost na razini osnovne škole.

- [16] Ivan Budimir. Jedna igra na sreću ili kako osvojiti auto. *Matka* 20(80):230–232, 2012.
- [17] Marina Čičak. Monty Hall Problem. *Matka*, 23(90):80–83, 2014.
- [18] Jelena Gusić. Promjena podataka. *Matka*, 18(70):74–76, 2009.
- [19] Tvrtko Tadić. Efikasnost cjepiva i bacanje novčića. *Matka*, 30(117):2–4, 2021.
- [20] Tvrtko Tadić. Sezonalnost podataka i pomični prosjek. *Matka*, 32(126):134–139, 2023.
- [21] Tvrtko Tadić. Pandemija i eksponencijalni rast. *Matka*, 30(118):74–79, 2021.



- [22] Sanja Varošaneć. Karakteristična petorka i brkata kutija. *Matematika i škola*, 15(73):56–59, 2013.
- [23] Sanja Varošaneć. Peteljka-list dijagram. Medijan i mod. *Matematika i škola*, 15(71):10–13, 2013.
- [24] Suradnici Wikipedije. Dice. *Wikipedia, The Free Encyclopedia*. 2026. Dostupno na <https://en.wikipedia.org/w/index.php?title=Dice&oldid=1337024779>. Pristupljeno 22. ožujka 2026.

Srednjoškolski članci

- [25] James J. Cochran. Pandemics and Exponential Growth. *Significance*, 17(3):18–21, 2020.
- [26] Gavin Freeguard. The story of the R number: How an obscure epidemiological figure took over our lives. Part 1: History. *Significance*, 21(1):6–10, 2024.
- [27] Gavin Freeguard. The story of the R number: How an obscure epidemiological figure took over our lives. Part 2: Modelling. *Significance*, 21(2):6–10, 2024.
- [28] Gavin Freeguard. The story of the R number: How an obscure epidemiological figure took over our lives. Part 3: Media. *Significance*, 21(3):26–29, 2024.
- [29] Gavin Freeguard. The story of the R number: How an obscure epidemiological figure took over our lives. Part 4: Politics of R. *Significance*, 21(4):19–22, 2024.
- [30] Gavin Freeguard. The story of the R number: How an obscure epidemiological figure took over our lives. Part 5: So what did we learn? *Significance*, 21(5):12–15, 2024.
- [31] Tvrtko Tadić. Podatci i uzorak. *Poučak*, 17(67):17–27, 2016.
- [32] Tvrtko Tadić. Aritmetička sredina i standardna devijacija. *Poučak*, 18(69):4–18, 2017.
- [33] Tvrtko Tadić. Medijan i kvantili. *Poučak*, 18(71):21–38, 2017.

Podatkovna znanost

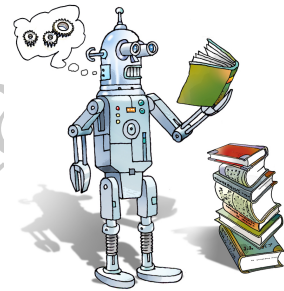
Knjiga [34] daje uvod u podatkovnu znanost za čitatelje koji su prošli neke kolegije sveučilišne matematike, dok je knjiga [35] namijenjena čitateljima koji su završili studij matematike. Zbirka pitanja [36] s intervjuja za radna mjesta podatkovnih znanstvenika daje pregled znanja koje se očekuje na razgovorima za posao.

- [34] Cathy O’Neil, Rachel Schutt. *Doing Data Science: Straight Talk from the Frontline*. O’Reilly Media, Inc., 2013.
- [35] Nathan Carter. *Data Science for Mathematicians*. CRC Press/Chapman and Hall, 2020.

- [36] Kevin Huo, Nick Singh. *Ace the data science interview: 201 real interview questions asked by FAANG, tech startups, & Wall Street*. 2022.
- [37] Petar Babić. Uvod u znanost o podacima kroz iskustvo rada u 3 industrije – bankarstvo, retail i telekomunikacije. *Predavanja Inženjerske sekcije Hrvatskog matematičkog društva*, 2020. Dostupno na <https://youtu.be/kQaShiGtDJI>.
- [38] Lucija Čalić. Vještine za karijeru u tehničkim pozicijama u Big Tech-u. *Predavanja Inženjerske sekcije Hrvatskog matematičkog društva*, 2023. Dostupno na <https://youtu.be/01ejhXgI4K4>.
- [39] Jonathan Auerbach, David Keppinger, Nicholas Rios. What is data science? A closer look at science’s latest priority dispute. *Significance*. 21(2):11–14, 2024.

Strojno i statističko učenje

Statističko učenje fokusira se na modeliranje odnosa među varijablama i procjenu parametara kako bi se objasnili podaci i donijeli zaključci o populaciji. Strojno pak učenje naglasak stavlja na algoritme koji iz podataka uče obrasce i optimiziraju predviđanja, često bez potrebe za eksplicitnim objašnjenjem odnosa među varijablama.



Knjige [40] i [41] popularni su udžbenici iz ovih područja. Knjiga [42] govori o izazovima izgradnje kompleksnih sustava koji izvode algoritme strojnog učenja.

- [40] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An introduction to statistical learning—with applications in R*. Springer, New York, 2021.
- [41] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.
- [42] Chip Huyen. *Designing machine learning systems*. O’Reilly Media, Inc., 2022.

Umjetna inteligencija i jezični modeli

Udžbenik [44] daje opsežan pregled umjetne inteligencije, dok skripta [45] objašnjava matematičke osnove jezičnih modela na naprednoj razini.

- [43] Dalia Kager. *Umjetna inteligencija : razvoj i primjena : priručnik za početno učenje*. Školska knjiga, 2023.
- [44] Stuart Russell, Peter Norvig. *Artificial intelligence : a modern approach*. Prentice-Hall, 2010.

- [45] Ryan Cotterell, Anej Svete, Clara Meister, Tianyu Liu, Li Du. *Formal aspects of language modeling*. *arXiv:2311.04329*, 2023.
- [46] Reni Banov. Veliki jezični modeli – potraga za umjetnom inteligencijom. *Poučak*, 26(102):25–33, 2025.
- [47] Hrvoje Benko. AI asistent u proširenoj stvarnosti. *Predavanja Inženjerske sekcije Hrvatskog matematičkog društva*, 2025. Dostupno na <https://youtu.be/FJI8C3vzq5g>.
- [48] Microsoft. What runs ChatGPT? Inside Microsoft’s AI Supercomputer | Featuring Mark Russinovich. 2023. Dostupno na <https://youtu.be/Rk3nTUfRZmo>.
- [49] Yann Dubois. Stanford CS229 | Machine Learning | Building Large Language Models (LLMs). 2024. Dostupno na <https://youtu.be/9vM4p9NN0Ts>.
- [50] Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1997
- [51] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*. ACM, 2004.
- [52] Kishore Papineni i sur. BLEU: A Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318, 2002.
- [53] Ashish Vaswani i sur. Attention Is All You Need. *NeurIPS*, 2017.
- [54] OpenAI. GPT-4 Technical Report. 2023. Dostupno na <https://arxiv.org/abs/2303.08774>.
- [55] Sebastian Bubeck i sur. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023. Dostupno na <https://arxiv.org/abs/2303.12712>.
- [56] Tvrtko Tadić. Struktura digitalnih asistenata. *Predavanja Inženjerske sekcije Hrvatskog matematičkog društva* 2021. Dostupno na <https://youtu.be/4rsBUgyMBqY>.
- [57] Tvrtko Tadić. Jezični modeli. *Poučak*, 25(99):28–42, 2024.
- [58] Tvrtko Tadić, Marin Smiljanić. Veliki jezični modeli – što se sve promijenilo razvojem ChatGPT-a? *Predavanja Inženjerske sekcije Hrvatskog matematičkog društva* 2023. Dostupno na <https://youtu.be/LZPK34g76No>.

Računarstvo

Premda nismo previše ulazili u teme iz računarstva, one igraju bitnu ulogu u obradi i skladištenju podataka. Tako se, primjerice, u klasiku [61] može naći na koji se način najefikasnije može izračunati medijan niza brojeva i kako se simuliraju slučajni brojevi. Knjiga [63] sveučilišni je udžbenik iz baza podataka.

- [59] Leo Budin, Predrag Brođanac, Zlatka Markučić, Smiljana Perić. *Rješavanje problema programiranjem u Pythonu : za 2. i 3. razred prirodoslovno-matematičkih gimnazija*. Element, 2018.
- [60] Leo Budin, Predrag Brođanac, Zlatka Markučić, Smiljana Perić. *Napredno rješavanje problema programiranjem u Pythonu : za 3. i 4. razred prirodoslovno-matematičkih gimnazija*. Element, 2018.
- [61] Donald Knuth. *Seminumerical Algorithms*, volume 2 of *The Art of Computer Programming*. Addison-Wesley, 1998.
- [62] Andrej Dujella, Marcel Maretić. *Kriptografija*. Element, 2007.
- [63] Robert Manger. *Baze podataka*. Element, 2012.
- [64] Steven Skiena. *Calculated Bets: Computers, Gambling, and Mathematical Modeling to Win*. Cambridge University Press, 2001.
- [65] Tvrtko Tadić. Iracionalnost i (pa)slučajnost. *Matka*, 17(67):198–202, 2009.

Kontrolirani eksperimenti i A/B testovi

Knjiga [66] daje izvrstan pregled online eksperimenata – od tema za početnike do tema za iskusne stručnjake.

- [66] Ron Kohavi, Diane Tang, Ya Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, 2020.
- [67] Lindsey R. Baden i sur. Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *New England Journal of Medicine* 384(5):403–416, 2021.
- [68] Ron Kohavi, Roger Longbotham, Dan Sommerfield, Randal M Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18:140–181, 2009.
- [69] Tvrtko Tadić. Online eksperimenti – iskustva velikih kompanija. *Poučak*, 23(89):70–77, 2022.
- [70] Tvrtko Tadić. Online eksperimenti – iskustva velikih kompanija, *Predavanja Inženjerske sekcije Hrvatskog matematičkog društva* 2022. Dostupno na <https://youtu.be/UTuu0rC4CC0>.

Mreže (grafovi)

Literatura o teoriji grafova i mrežnoj znanosti (*network science*) je opsežna. Knjiga [72] daje temeljiti pregled tematike, dok udžbenik [74], namijenjen studentima matematike, daje uvod u teoriju grafova. Predavanje [76] daje dobar uvid u *PageRank* algoritam, dok knjiga [73] detaljno razrađuje tematiku. [75] daje pregled strojnog učenja na grafovima.

- [71] Tvrtko Tadić. *Pripreme za matematička natjecanja : za 4. razred gimnazije : 4. (dopunjeno) izdanje*. Element, 2014.
- [72] Mark Newman. *Networks*. Oxford University Press, 2018.
- [73] Amy Langville, Carl Meyer. *Google's Page Rank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.
- [74] Darko Veljan. *Kombinatorna i diskretna matematika*. Algoritam, Zagreb, 2001.
- [75] William L Hamilton. *Graph representation learning*. Morgan & Claypool Publishers, 2020.
- [76] Jure Leskovec. Stanford CS224W: Machine Learning with Graphs | 2021 | Lecture 4.1 - PageRank. 2021. Dostupno na <https://youtu.be/TU0ankRcHmo>.
- [77] Benedek Rozemberczki, Ryan Davies, Rik Sarkar, Charles Sutton. Gemsec: Graph embedding with self clustering. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2019*, pages 65–72. ACM, 2019.
- [78] Dušan Mundar, Damir Horvat. Rangiranje ekipa i prognoziranje ishoda u rukometu korištenjem PageRank algoritma. *Poučak*, 17(67):35–42, 2016.
- [79] Monika Štaba. Društvene mreže. *Matka*, 32(128):224–227, 2024.
- [80] Tvrtko Tadić. PageRank i slučajni posjetitelj internetskih stranica. *Poučak*, 19(76):14–19, 2018.
- [81] Damir Horvat, Dušan Mundar. Rangiranje web stranica. *Osječki matematički list*, 17(1):51–62, 2017.
- [82] Andrew Gustar. The Laws of Musical Fame and Obscurity. *Significance* 17(5):14–17, 2020.
- [83] Mark Newman. Power-Law Distribution. *Significance* 14(4):10–11, 2017.
- [84] Noga Alon, Joel H. Spencer. *The probabilistic method*. Wiley-Intersci. Ser. Discrete Math. Optim. Hoboken, NJ: John Wiley & Sons, 4th edition edition, 2016.

Privatnost i povjerljivost

Knjiga [85], namijenjena širem čitateljstvu, daje izvrstan uvod u tematiku privatnosti u današnjem svijetu.

- [85] Claire McKay Bowen. *Protecting Your Privacy in a Data-Driven World*. ASA-CRC Series on Statistical Reasoning in Science and Society. CRC Press, 2021.
- [86] Cynthia Dwork, Frank McSherry, Kobbi Nissim, Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

- [87] Christoph Kurz. Understanding Differential Privacy. *Significance* 18(3):24–27, 2021.
- [88] Sam Smith. Data and Privacy: Now you see me; New Model for Data Sharing; Modern Governance and Statisticians. *Significance* 11(4):10–17, 2014.
- [89] Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, Sergey Yekhanin. Differentially Private Synthetic Data via Foundation Model APIs 2: Text. *ICML*, 2024.
- [90] Úlfar Erlingsson, Vasył Pihur, Aleksandra Korolova. Rappor: Randomized Aggregatable Privacy-preserving Ordinal Response. *ACM CCS*, 2014.
- [91] Marco Geraci, Mario Cortina Borja. Notebook: The Laplace Distribution. *Significance* 15(5):10–11, 2018.
- [92] Apple. *Differential Privacy Overview*. 2017. Dostupno na https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf,
- [93] U.S. Census Bureau. *Disclosure Avoidance System for the 2020 Census*. 2020.

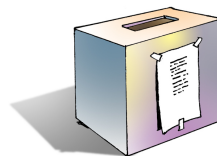
Vizualizacija podataka

Knjiga [94] daje izvrstan uvod u vizualizaciju podataka kroz zanimljive primjere.

- [94] Robert Grant. *Data visualization: charts, maps, and interactive graphics*. Chapman and Hall/CRC, 2018.
- [95] Tvrtko Tadić. Koliko je ljudi više umrlo tijekom pandemije? *Poučak*, 25(97):54–61, 2024.
- [96] Tvrtko Tadić. Toplinska karta u ilustraciji podataka. *Matematičko fizički list*, 76(301):32–39, 2025.
- [97] Beril Sirmacek, Cem Unsalan. Urban area detection using local feature points and spatial voting. *IEEE Geoscience and Remote Sensing Letters*, 7(1):146–150, 2010.
- [98] Suradnici Wikipedije. Multivariate kernel density estimation. *Wikipedia, The Free Encyclopedia*, 2024. Pristupljeno 10. studenog 2024.
- [99] Suradnici Wikipedije. Heat equation. *Wikipedia, The Free Encyclopedia*, 2024. Pristupljeno 10. studenog 2024.

Izbori

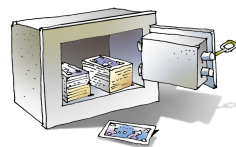
Izbori su zanimljiv primjer prikupljanja podataka u svrhu predviđanja ishoda jer, nakon što završe, znamo ishod. Knjiga [100] daje uvod u tematiku anketa. Kako bismo dali točno predviđanje, trebamo razumjeti i sustave koji se koriste na izborima, o čemu govori knjiga [101].



- [100] Ole J Forsberg. *Understanding elections through statistics: polling, prediction, and testing*. Chapman and Hall/CRC, 2020.
- [101] Tomislav Marošević. *Matematički aspekti izbornih sustava*. Sveučilište Josipa Jurja Strossmayera u Osijeku, Fakultet primijenjene matematike i informatike, 2025.
- [102] Željko Brčić. Matematika na biralištu. *Matka*, 23(91):146–148, 2015.
- [103] Tvrtko Tadić. Matematika iza anketa – primjer izbora. *Poučak*, 11(43):42–52, 2010.
- [104] Tvrtko Tadić. Proporcionalnost D’Hondtove metode i izborni sustav u Hrvatskoj. *Poučak*, 16(63):64–79, 2015.
- [105] Tvrtko Tadić. Simulacija anketa. *Matka*, 73:70–75, 2010.
- [106] Aleksandar Hatzivelkos. Parametri optimalne reprezentabilnosti D’Hondtove metode. *Poučak*, 12(48):4–19, 2011.
- [107] Karsten Schuster, Friedrich Pukelsheim, Mathias Drton, Norman R. Draper. Seat biases of apportionment methods for proportional representation. *Electoral Studies*, 22(4):651–676, 2003.
- [108] Suradnici Wikipedije. Apportionment paradox. *Wikipedia, The Free Encyclopedia*, 2024. Pristupljeno 24. 8. 2024.

Financije

Financije su široka tema koju smo u ovoj knjizi tek spomenuli. Dobar uvod u ovu tematiku daje knjiga [109]. Knjiga [110] 50 godina nakon izlaska i dalje se potvrđuje kao jedan od najboljih vodiča u svijet investicija.



- [109] Toni Milun, Sandra Ferenčak, Toni Vitali. *Budi financijski fit : praktičan vodič za zdrav financijski život*. Meritum, 2025.
- [110] Burton G. Malkiel. *A random walk down Wall Street : the best investment guide money can buy*. W.W. Norton & Company, 2023.
- [111] Paul Wilmott, David Orrell. *The Money Formula*. John Wiley & Sons, 2017.

- [112] Katarina Justić, Petra Posedel Šimović. Ljerka Domac, Ajka Kaleb-Kovačević. *Uvod u osobne financije i financijsku pismenost : udžbenik za srednje strukovne škole*. Školska knjiga, 2026.
- [113] David H. Bailey, Marcos López de Prado. How “Backtest Overfitting” in Finance Leads to False Discoveries. *Significance* 18(6):22–25, 2021.

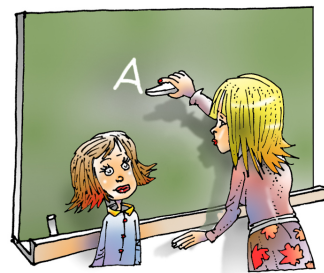
Društvo, znanost i tehnologija

Razvoj tehnologije i upotrebe podataka otvorio je brojna društvena pitanja. Knjiga [114] govori o problematičnoj upotrebi algoritama čiju osnovu čine podatci. Naslov [115] daje pregled izazova razvoja umjetne inteligencije na primjeru kompanije OpenAI, dok knjiga [116] govori o ovisnosti društva o nekoliko tehnoloških kompanija te o tome koje izazove to sve donosi.

- [114] Cathy O’Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2017.
- [115] Karen Hao. *Empire of AI: Dreams and nightmares in Sam Altman’s OpenAI*. Penguin Group, 2025.
- [116] Marietje Schaake. *The Tech Coup: How to Save Democracy from Silicon Valley*. Princeton University Press, 2025.
- [117] Andrew Gelman. Ethics in Statistical Practice and Communication: Five Recommendations. *Significance* 15(5):40–43, 2018.

Poučavanje podatkovne znanosti i statistike

Poučavanje o podacima predstavlja izazov zbog stalnih promjena u tehnološkim mogućnostima njihove obrade i prikupljanja. Izdanje [118] donosi preporuke o tome kako pristupiti poučavanju u osnovnim i srednjim školama, a [119] sličan je dokument za visoko školstvo. Ovi dokumenti periodično se dopunjavaju i proširuju.



- [118] Anna Bargagliotti, Christine Franklin, Pip Arnold, Rob Gould, Sheri Johnson, Leticia Perez, Denise A. Spangler. *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II)*. American Statistical Association, National Council of Teachers of Mathematics, 2020.
- [119] Robert Carver, Michelle Everson, John Gabrosek, Nicholas Horton, Robin Lock, Megan Mocko, Allan Rossman, Ginger Holmes Roswell, Paul Velleman, Jeffrey Witmer i sur. *Guidelines for Assessment and Instruction in Statistics*

Education (GAISE) College Report 2016. American Statistical Association, 2016.

Ostale knjige i članci

- [120] Žarko Dadić. *Ruđer Bošković*. 2. izdanje. Školska knjiga, 1990.
- [121] Martin Gardner. *Aha! A Two Volume Collection: Aha! Gotcha Aha! Insight*. Mathematical Association of America, 2006.
- [122] Tomislav Pejković. *Iracionalni brojevi*. Matkina biblioteka. Hrvatsko matematičko društvo, 2001.
- [123] Steven M. Stigler. Studies in the History of Probability and Statistics XL: Boscovich, Simpson and a 1760 Manuscript Note on Fitting a Linear Relation. *Biometrika*, 71:615-620, 1984.
- [124] Tanja Soucie. Francis Galton. *Matka*, 19(76):240-241, 2011.
- [125] Roger Koenker, Gilbert Bassett. On Boscovich's Estimator. *Annals of Statistics*, 13(4):1625-1628, 1985.



Novinski članci

U ovoj knjizi citirali smo brojne novinske članke koji su pratili razvoj tehnologije.

- [126] BBC News. Coronavirus: What is the R number and how is it calculated?. 26. travnja 2021. Dostupno na <https://www.bbc.com/news/health-52473523>.

- [127] Michael M. Grynbaum, Ryan Mac. The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. *The New York Times*, prosinac 2023.
- [128] Stuart A. Thompson, Charlie Warzel. Twelve Million Phones, One Dataset, Zero Privacy. *The New York Times*, prosinac 2019.
- [129] Luke Jones. Google Gains Exclusive Access to Reddit Data for AI, Restricting Rival Search Engines. *WinBuzzer*, srpanj 2024.
- [130] Wall Street Journal. What is ChatGPT, the AI Chatbot That Everyone Is Talking About? 2022. Dostupno na <https://www.wsj.com/podcasts/google-news-update/what-is-chatgpt-the-ai-chatbot-that-everyone-is-talking-about>. Pristupljeno: 30. 12. 2024.
- [131] Kaitlyn Tiffany. This Instagram Story ad with a fake hair in it is sort of disturbing / Yes? *The Verge*, prosinac 2017.
- [132] John Tierney. Behind Monty Hall's Doors: Puzzle, Debate and Answer? *The New York Times*, 21. srpnja 1991.
- [133] Maria Yagoda. Airline held liable for its chatbot giving passenger bad advice - what this means for travellers. *BBC Travel* 23. siječnja 2024. Dostupno na <https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know>

Izvori podataka

Ovdje su navedeni detalji nekih izvora podataka korištenih u ovoj knjizi.

- [134] Centers for Disease Control and Prevention. Excess Deaths Associated with Covid-19. Dostupno na https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm. ožujak 2023. Pristupljeno 3. 3. 2024.
- [135] Centers for Disease Control and Prevention. CDC COVID Data Tracker. Dostupno na <https://covid.cdc.gov/covid-datatracker>. ožujak 2020. Pristupljeno 21. 5. 2022.
- [136] United States Department of Transportation. Airline On-Time Statistics. Bureau of Transportation Statistics. Dostupno na <https://www.transtats.bts.gov/ONTIME/Airborne.aspx>. Pristupljeno 10. 2. 2017.
- [137] Državno izborno povjerenstvo. Službene stranice. Dostupno na <https://www.izbori.hr>.
- [138] OpenStreetMap. Dostupno na <https://www.openstreetmap.org/>.
- [139] Jure Leskovec, Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. Dostupno na <http://snap.stanford.edu/data>. June 2014.

LITERATURA

- [140] Suradnici Wikipedije. Covid-19 pandemic in Croatia. *Wikipedia, The Free Encyclopedia*. Dostupno na https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Croatia. Pristupljeno 20. 2. 2024.
- [141] Our World in Data. Coronavirus (Covid-19) Deaths. Dostupno na <https://ourworldindata.org/covid-deaths>. Pristupljeno 20. 2. 2024.
- [142] Seattle Police Department Crime Data: 2008-present. Dostupno na https://data.seattle.gov/Public-Safety/SPD-Crime-Data-2008-Present/tazs-3rd5/about_data. Pristupljeno: 30. 11. 2024.
- [143] Državni zavod za statistiku. Prirodno kretanje stanovništva. Dostupno na <https://podaci.dzs.hr/media/meehktks/msi-stanovnistvo.xlsx>. Pristupljeno 31. 1. 2024.
- [144] Hrvatski zavod za javno zdravstvo. Mortalitet statistika. Dostupno na <https://www.hzjz.hr/tag/mortalitet-statistika/>. Pristupljeno 10. 4. 2026.
- [145] The Atlantic. The Covid Tracking Project. Dostupno na <https://covidtracking.com/>. ožujak 2021. Pristupljeno 21. 5. 2022.
- [146] World Health Organization. Global Excess Deaths Associated with Covid-19, January 2020 - December 2021. Dostupno na www.who.int/data/stories/global-excess-deaths-associated-with-covid-19-january-2020-december-2021. svibanj 2022. Pristupljeno 3. 3. 2024.

Kazalo

- 3-sigma pravilo, 42, 46
- 68-95-99 pravilo, 42, 46

- A/B eksperimenti, 17
- A/B test, 122, 279
- Air Canada, 193
- Albert, Réka, 212
- algoritam, 161
- Alibaba, 191
- Amazon, 241, 252
- Američki ured za popise (United States Census Bureau), 23, 31, 34, 258
- ankete, 29, 30, 146–153
 - izlazne, 147
- Apple, 271
- apsolutna vrijednost, 59–60
- aritmetička sredina, 37–49, 57, 58, 62, 148, 236
 - definicija, 39
- asimptotska analiza, 219
- astronomska mjerenja, 60
- autorsko pravo, 188

- bacanje novčića, 14–17, 130–131, 133
- Barabási, Albert-László, 212
- Barabási-Albert model, 212–214, 216–221, 223
- baze podataka, 251, 271
- BERT, 191
- bezrazmjerne mreže, 212
- Bing, 7, 124, 187
- Bing AI, 254
- Bing Chat, 193
- biomedicinska istraživanja, 123, 128
- BLEU, 192

- Bolt, 8
- Bošković, Ruđer, 60
- Brin, Sergey, 199, 203
- brkata kutija, 67–68
- brzina vožnje, 20–21
- burza, 113, 230

- Census Office (SAD), 257
- Centar za kontrolu i prevenciju bolesti (CDC), 100, 236
- ChatGPT, 7, 184, 188
 - cijepivo
 - efikasnost, 129
 - testiranje, 128
- Claude, 7
- clustering (grupiranje), 246
- Conway, Drew
 - dijagram, 9
- Covid-19, 18, 96–100, 128–132, 229
 - broj umrlih, 230–237
- CROBEX (dionički indeks), 113
- csv (comma-separated values) format, 238

- Čebiševljeva nejednakost, 43, 45

- D’Hondt, Victor, 158
- D’Hondtova metoda, 158–168
 - monotonost, 165
 - očuvanje većine, 167
 - proporcionalnost, 165
- DAX (dionički indeks), 113
- DeepSeek, 191
- diferencijalna privatnost, 268–276
- dionički indeks, 112
- distribuirani sustavi, 55

- dizajn eksperimenta, 34, 135
dobne skupine, 34
Drugi svjetski rat, 258
društvene mreže, 8, 12, 197–223, 251, 252
Državni zavod za statistiku, 110, 226, 231
DuckDuckGo, 7
duljina trudnoće, 40
Dwork, Cynthia, 270
- efikasnost cjepiva, 17
eksponencijalni rast, 96–101, 103
ekstremi, 39
epidemija, 101, 104
Erdős, Paul, 206
Erdős-Rényi slučajni graf, 205–212, 223
Eugenij Onjegin, 188
Euler, Leonhard, 198
Europska unija (EU), 255, 260
Eurostat, 260
evaluacija modela, 85
Excel, 36, 41, 55, 78, 94, 97, 100, 142, 196
- Facebook, 8, 214–216
graf, 222
frekvencija, 26
definicija, 283
frekvencijski dijagram, 26, 140
- Gallup institut za istraživanje javnog mnijenja, 24
Gallup, George, 24, 26
Galton, Francis, 101
Galton-Watsonov proces, 101
Gardner, Martin, 19, 20
Gates, Bill, 43, 58, 272
Gauss, Johann Carl Friedrich, 46
Gauss-Markovljevi teorem, 83
Gaussova krivulja, 45, 57, 151, 287
Gaussovska jezgra, 238, 242
GDPR, 255, 260
GEMSEC (skup podataka), 214
geodetska mjerenja, 60
- geografska dužina i širina, 239
geografska lokacija, 34
geometrijska vjerojatnost, 80, 84, 103, 154, 206, 288
Geopandas, 240
GitHub, 11, 36, 84
Google, 7, 187, 190, 199, 203
Chrome, 271
Gemini, 7
GPS koordinate, 251, 260
GPT-2, 191
GPT-3, 191
graf (mreža), 197–223, 279, 296
definicija $G = (V, E)$, 206
heterogen(a), 222
homogen(a), 222
usmjeren, 203
grafika, 178
grupiranje (clustering), 246
- hackathon, 124
Hall, Monty
dilema, 143
halucinacije, 192, 254
histogram, 285
Hooker, Joseph, 74
Hrvatski zavod za javno zdravstvo (HZJZ), 232
- identifikacija urbanih područja, 244
inflacija, 112
influenceri, 212
Instagram, 8, 128, 198
internet, 122, 188, 199
internetska anketa, 30
interval pouzdanosti, 65, 83–84, 86
iracionalni brojevi, 154
izbori, 22–25, 147–153, 158–168, 298–299
izborni prag, 168
izlazne ankete, 29
- Jefferson, Thomas, 158
jezični modeli, 184–195, 275
 n -gram modeli, 189

- alfabet, 185
matematičke reprezentacije prirod-
nog jezika, 185
rekurzivni, 189
simboli, 185
tekst, 185
Jupyter bilježnica, 36, 84, 94, 142,
196, 223
- Königsberški mostovi, 198, 222
karakteristična petorka, 67
karte, 237–246
kategorijski podatci, 284
kategorizacija podataka, 266
KDE (kernel density estimate), 241
klasifikacija, 178
kockarnica, 175
Kongres SAD-a, 257
kontrolna grupa, 123, 129
koordinate, 238
koordinatni sustav, 75
koreliranost, 21
koronavirus, 17, 128–132
kultura eksperimentiranja, 125
kvadratna funkcija, 44–45, 87, 90
 minimum, 44
kvadratna jednadžba, 219
kvantil, 62–71
kvartil(i), 67
- lanac opskrbe, 8
Laplace, Pierre-Simon, 61
Laplaceov model, 175
Laplaceova raspodjela, 272
Let's Make a Deal (televizijska emi-
sija), 142
linearna funkcija, 88
linearna regresija, 74–94
 s više varijabli poticaja, 90–93
linearne jednadžbe, 202–204
 sustav, 202, 204
LinkedIn, 254
Literary Digest, 22–25
 fijasko, 23
logistička regresija, 180
- lokacijski podatci, 251
Loto 7/35, 138–142
Lyft, 8
- maksimum, 57, 62
Markov, Andrej, 188
Markovljevi lanci, 202
 granične vrijednosti, 202
Markovljevi modeli, 188
Markovljevo svojstvo, 189, 202
masa djeteta, 40
matematika, 9
matematička indukcija, 108, 165
medijan, 52–60
Meta (kompanija), 191, 222
metoda najmanjih kvadrata, 76, 77,
88
metoda potpornih vektora, 184
metrike, 126–127
 cilja, 126
 izigravanje, 127
 pokretačke, 126
 zaštitne, 126
Microsoft, 97, 124, 187, 254, 270, 278
 Copilot, 7, 254
 graf, 222, 223
mikrolokacija, 244
minimum, 44, 57, 62
 funkcije, 89
 kvadratne funkcije, 90
Ministarstvo prometa (SAD), 55
Ministarstvo rata (SAD), 258
mirovinska štednja, 117
Mistral, 191
mnogokut, 245
model, 87
 parametri, 88
model grananja, 101
Moderna (farmaceutska kompanija),
128
mostovi Königsberga, 198
mreža karate kluba (skup podataka),
208, 219–221
mrežna znanost (network science), 296
multigraf, 222

- multiskup, 53
- nekretnine, 78–80, 85–86, 90–93, 110
 - indeks cijena stambenih objekata, 110
- nesreće u prometu, 20–21
- Netflix, 8, 125
- networkx (Python paket), 206
- neuronske metode, 185
- neuronske mreže, 184
- New York Times, 188, 251, 252
- nezaposlenost, 34
- Njujorška burza, 113
- normalna raspodjela, 286–288
 - očekivanje, 287
 - standardna devijacija, 287
- normalni model, 45
- normalno raspodijeljeni podatci, 42, 56, 151, 232
- normirani histogram, 285
- numeričke metode, 180
- Nvidia, 191
- Office 365, 19–20, 223
- oglasi, 124, 127–135
- OIB (osobni identifikacijski broj), 252
- online oglasi, 124
- online pokusi, 279
- opaženi podatci, 17–21
- OpenAI, 184, 187
- OpenStreetMap, 246
- optimizacija, 180
- osobno identificirajuće informacije, 251
- overfitting, 181
- Page, Larry, 199, 203
- PageRank, 198–204
 - jednadžbe, 203
- Pandas, 238
- parametri modela, 46
- patent, 203
- perplexity, 192
- placebo sredstvo, 129
- podatkovna pismenost, 11
- podatkovna znanost, 8
- podatkovni centar (data center), 191
- podzastupljenost, 30
- POI (points of interest), 238
- pokusna grupa, 123, 129
- polinomi, 88
- pomak u objavi podataka, 268
- pomični prosjek, 228–229
- popis stanovništva, 26
 - austrijski dio Austro-Ugarske 1910., 259
 - Belgija 1846., 259
 - Engleska 1801., 259
 - Francuska 1801., 259
 - Hrvatska 1857., 259
 - Rusija 1897., 259
 - SAD 1790., 255
 - SAD 1850., 257
 - Švedska 1749., 259
- populacija, 21, 30
- potzdani interval, 152
- povezivost podataka, 250–255
- povjerljivost podataka, 250
- pravac (u koordinatnom sustavu), 76
- pravo na zaborav, 261
- preferencijsko povezivanje, 212
- prekomjeran broj umrlih, 232
- prepoznavanje teksta, 177–184
- pretrage, 244–245
- preučenje modela, 181
- prezastupljenost, 30, 31
- prikupljanje podataka, 13–35, 122
- pristranost
 - ponuđenog izbora, 25
 - preživjelih, 25
 - zbog dobrovoljnosti, 24
 - zbog odbijanja sudjelovanja, 24
 - zbog prigodnosti, 24
- privatnost, 12, 238, 249–276
- procjena parametara modela, 85, 179
- programski jezik, 15
- pseudoslučajni brojevi, 153
- Python, 30, 34, 36, 55, 94, 139–141, 196, 223, 238
- R (programski jezik), 246

- racionalni brojevi, 154
 raspodjela podataka, 286
 raspon, 39, 67
 raspon podataka, 62
 računalni oblak, 9, 122, 187
 rekurzivno zadani niz, 107, 109
 relativna frekvencija, 134
 definicija, 283
 reprezentativnost, 30
 reprodukcijski broj R , 103
 Roosevelt, Franklin D., 23, 258
 ROUGE, 192
 Rényi, Alfréd, 206
- S&P 500 (dionički indeks), 113
 semantička reprezentacija teksta, 275
 semantička sličnost teksta, 276
 sezonalnost podataka, 66, 68, 226–228
 period, 230
 sigurnost podataka, 250
 simetrična kocka, 157, 175
 simetrično razdijeljeni skup podataka,
 61
 simulacije, 15–16, 28, 31, 36, 80, 130,
 137–169, 200–201
 bacanja kocke, 157
 grafova (mreža), 204–221
 ispis velikog jezičnog modela, 186
 sintetički podatci, 267
 skup (u matematici), 53, 143
 skup elementarnih događaja, 173
 skup podataka
 multiskup, 53
 Slack, 125
 sloj (stratum), 32
 slučajni brojevi, 153
 slučajni procesi, 219
 Spotify, 8
 standardna devijacija, 37–49, 231–233,
 243
 definicija, 39
 Stanford (sveučilište), 203
 statistika, 8
 statistički test, 126, 129, 155
 značajnost, 131
- statističko učenje, 294
 strojno učenje, 179, 180, 294
 stršće vrijednosti (outliers), 68–69,
 264
 stupanj vrha (u grafu), 207
 super računalno, 191
 Svjetska zdravstvena organizacija (WHO),
 236
- štednja, 106–117
 šum (noise), 80–84, 91, 244, 271
- telefonsko istraživanje, 30
 telemetrija, 271
 temperatura vrelnja vode, 74, 75
 teorija grafova, 198, 221
 terensko istraživanje, 30, 34
 testiranje modela, 85
 TikTok, 8, 198
 Title 13, 258
 tlak zraka, 74, 75
 toplinske karte, 237–246
 točke (u koordinatnom sustavu), 76
 točke interesa, 238
 train/test split, 85, 181
 transformatori (u jezičnim modelima),
 190
 treniranje modela, 85, 181, 187, 191
 trokut (u grafu), 212, 217
 tržišta kapitala, 113
 turistički dolasci, 226–229
- Uber, 8
 umjetna inteligencija, 11, 171–196, 253–
 254, 275
 United Airlines, 55
 Ustav SAD-a, 256
 uvjetna vjerojatnost, 146, 174, 186
 uzorak, 21, 22, 26–34, 47, 70, 146
 reprezentativni, 28, 70
 slučajni, 28
 stratificirani, 30–34
 uzročno-posljedične veze, 21
 učenje modela, 179
- varijabla odaziva, 87

KAZALO

varijabla poticaja, 87
veliki jezični modeli, 278
veličina uzorka, 34
vizualizacija podataka, 12, 225–247
vjerojatnosna metoda, 221
vjerojatnosni model, 179
vjerojatnost, 101, 145–146, 172, 199
 događaja, 173
vremenski niz, 226

Wald, Abraham, 25
Wall Street, 113
Watson, Henry William, 101
web-trgovina, 124
Wikipedija, 188, 204

X (društvena mreža), 8

YouTube, 8

Zachary, Wayne, 208
zajednički uzrok, 19
zakon potencije (u grafu), 212, 217–
 219
zdravstvene usluge, 251
zemljišne knjige, 252
zvonolika krivulja, 287

matematika.hr

Hrvatsko matematičko društvo (HMD) osnovano je 1949. godine s ciljem promicanja matematičke znanosti, obrazovanja i kulture. Kao središnja strukovna udruga matematičara u Hrvatskoj, HMD okuplja stručnjake iz akademske zajednice, obrazovnog sustava i industrije, potičući razmjenu znanja i razvoj matematičkog mišljenja, te predstavlja Hrvatsku u Međunarodnoj matematičkoj uniji i Europskom matematičkom društvu.

Društvo je aktivno u izdavaštvu stručnih i znanstvenih publikacija, organizaciji konferencija, predavanja, natjecanja i edukativnih programa. Ova knjiga objavljena je u sklopu izdavačke djelatnosti HMD-a, s namjerom da doprinese razumijevanju i širenju matematičke pismenosti među različitim generacijama čitatelja.

Informacije o izdanjima i aktivnostima HMD-a možete naći na web-stranici

matematika.hr

i pratiti putem društvenih mreža.

