Znanost o podacima

Cesarec, Karlo

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Karlo Cesarec

ZNANOST O PODACIMA

ZAVRŠNI RAD

Varaždin, 2022.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Karlo Cesarec

Matični broj: Z-46061/17–1

Studij: Primjena informacijske tehnologije u poslovanju

ZNANOST O PODACIMA

ZAVRŠNI RAD

Mentor:

Prof. dr. sc. Kornelije Rabuzin

Varaždin, rujan 2022.

Karlo Cesarec

Izjava o izvornosti

Izjavljujem da je moj završni rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor potvrdio prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

Ovaj rad proučava znanosti o podacima, od postupka kojim se od podataka dobiva informacija, interpretacija tih istih informaciju te na kraju krajeva korištenje tih informacija.

Između ostalog pobliže pojašnjava put kojim podatak postaje mudrost, kao i različite znanosti od kojih se sama znanost o podacima sastoji i koje je potrebno savladati ne bi li se mogao baviti poslovima koji se mogu raditi po završetku stečenog znanja, a koja su također navedena u radu.

Nadalje se navodi primjena znanosti o podacima, očekivanja od eksperata koji se bave znanošću o podacima i kompetencije kojima oni moraju raspolagati, te prognoze za budućnost tog područja.

Ključne riječi: podatak, informacija, umjetna inteligencija, rudarenje podataka, strojno učenje, baze podataka

SADRŽAJ:

1. Uvod

Prvenstveno, ovaj završni rad će se bazirati na znanosti podataka, te njihovoj primjeni općenito kroz ranije etape razvoja tehnologije na zemlji, također ću spominjati odnose između podataka, koji je prva točka kod raspoznavanja određenog problema, tvrdnje ili činjenice, te kako se od podatka dobije informacija, koja se nakon toga pretače u znanje te u krajnjem slučaju u mudrost sagledavanja tog istog problema ili činjenice. Ovu temu sam birao jer smatram da ću se baviti ovom granom IT industrije nakon završetka svog školovanja na fakultetu.
Umjetnost razotkrivanja uvida i trendova što se tiče podataka je sa nama još od davnih vremena. Drevni Egipćani koristili su primjerice popis podataka što je omogućavalo njihovu bolju učinkovitost kod ubiranja poreza, te su znali točno predvidjeti kada će rijeka Nil poplaviti.

Znanost o podacima obuhvaća skup načela, definicija problema, algoritama i procesa za izvlačenje nejasnih i korisnih uzoraka iz velikih podataka (eng. Big Data) s ciljem poboljšanja donošenja odluka temeljem uvida u velike podatke. Kao područje djelovanja, znanost o podacima obuhvaća skup načela, definicija problema, algoritama i procesa za izvlačenje neočiglednih, ali korisnih uzoraka ponašanja podataka iz velikih podataka. Ona je usko povezana s područjima rudarenja podataka i strojnog učenja, ali je opsegom od njih šira. Danas znanost o podacima upravlja donošenjem odluka u gotovo svim dijelovima modernih društava. Neki od načina na koje znanost o podacima može utjecati na vaš svakodnevni život, između ostalog, uključuju:

- izbor oglasa koji vam se nude putem virtualnog oglašavanja; to su preporuke za filmove koje biste trebali pogledati, knjige koje biste trebali pročitati i prijateljske veze koje biste trebali uspostaviti;
- prijedloge o tome koje poruke e-pošte svrstati u mapu neželjene pošte; pogodnosti koje će vam se ponuditi prilikom obnove pretplate na uslugu mobitela;
- ponudu cijene premije zdravstvenog osiguranja koja vam se nudi; razmještaj i dinamiku funkcioniranja semafora u vašem području;
- sastav lijekova koje trebate; kao i to na kojim mjestima u vašem gradu policija zaustavlja.

2. Metode i tehnike rada

U ovom radu ćemo se pretežito baviti sa znanošću o podacima, znanostima koje su joj bliske te koje se koriste u analizi podataka, procesiranju podataka te stavljanja istih u složene odnose.

Također će kroz rad u dva primjera umjetne inteligencije biti objašnjeno koliko je umjetna inteligencija napredovala do današnjeg dana, na koji način funkcionira, te kako se može koristiti. Isto tako će se pred kraj ovog završnog rada rezimirati sva moguća područja kojima se netko može baviti nakon što završi školovanje u ovom području.

Prilikom pisanja ovog završnog rada upotrijebljeni su razni izvori literature, kao što su knjige s područja poslovne informatike, statistike, filozofije, matematike, znanstveni radovi, te internetski izvori kao stranice na hrvatskom i engleskom jeziku. U radu je opisano što je to znanost o podacima, na čemu ona počiva, njezina primjena, put od podataka do mudrosti, što se očekuje od eksperata koji se njome bave, kakvo je stanje na tržištu glede poslova vezanih za sferu znanosti o podacima, te kakve su prognoze za budućnost.

3. Od podataka do mudrosti

3.1. Podatak

Riječ podatak potječe od množine latinske riječi „datum“ što znači dio informacije. Podatak je činjenica koja se može zabilježiti tekstom, brojevima, bojom ili zvukom.

Podatak je definiran kao tijelo činjenica ili figura, koje su bile sustavno prikupljane za jednu ili više posebnih namjena

Podatak je nematerijalne prirode, on jednostavno postoji u našim mislima i nema značenje unutar ili izvan svog postojanja ili o samom sebi pa se pridružuje značenju kojim opisujemo svojstva objekata.

Struktura podataka je apstraktna i čine ju značenje (naziv i opis značenja određenog svojstva), vrijednost (mjera i iznos) i vrijeme.

Osnovna obilježja podatka koji ga čine korisnim su :

- Bitnost za određenu svrhu
- Potpunost
- Točnost
- Pravovremenost
- Odgovarajući oblik
- Dostupnost

3.1.1. Obrada podataka

Obrada podataka predstavlja proces pretvaranja podataka u korisne informacije. Da bi podatak imao atribute za postati informacija, mora posjedovati neko nove do tad ne znane uvjete koji su do tada nepoznati primaocu. Informaciju čine podaci kojima je dano značenje putem relacijskih veza, odnosno organizirani podaci koji su uređeni za bolje shvaćanje i razumijevanje. Značenje informacije može biti korisno, ali i ne mora.

Proces pretvaranja podataka u informacije se ukratko može usporediti sa tvornicom, gdje imamo sirove materijale u formi podataka, rad u tijeku u formi sabiranja podataka, te gotov proizvod u formi informacije.

3.2. Informacija

Informacija predstavlja najmanje dva podatka u nekom međusobnom odnosu, koji predstavljaju određenu novost i koji nas potiču na neku aktivnost.

Prema Čeriću, V., Vargi, M. i suradnicima, Informacija je stvorena analizom odnosa i veza između podataka. Može odgovoriti na pitanja poput “Tko/Što/Gdje/Kada ”. Informacija je poruka s ponekad unaprijed određenom publikom i svrhom, a njezinu vrijednost određuje sam primatelj.

Informacija odnosno obavijest u pravilu donosi novost, obavještava o nečemu, te otklanja neizvjesnost i općenito služi kao podloga za odlučivanje. Stoga se vrlo često u takvom procesu, informacija smatra temeljnim elementom za donošenje odluke, a ne podatak. Koliko je upotrijebljena informacija relevantnija, točnija i svježija, toliko je vrjednija pri odlučivanju.

Informacija je definirana kao podatak koji je obrađen u obliku koji je smislen njezinom primatelju te je percipirane vrijednosti kod trenutnog ili potencijalnog odlučivanja iako su podaci sastojci informacija, ne čine svi podaci korisne informacije.

Informacija je korisna svojem prijatelju jedino u slučaju kada je :

- pouzdana/lagana za rukovanje i adekvatno zaštićena
- relevantna (s obzirom na namijenjene svrhe i uz odgovarajuću razinu detaljnosti)
- konzistentna (s drugim izvorima informacija)
- pouzdana, precizna i provjeriva (neovisnim načinima)
- razumljiva (tj. razumljiva njezinim primateljima)
- potpuna (u smislu atributne, prostorne i vremenske pokrivenosti)

3.3. Znanje

"Bolje je znati malo ali temeljno, nego mnogo i površno." I. Kant (1787.)

Znanje je možda najteže definirati, a i definicije se mogu odnositi na informaciju koja je obrađena, organizirana ili strukturirana na neki drugi način, te primjene ili stavljanja u neku akciju.

Jedan pogled govori o tome da je znanje proizvod sinteze ljudskog uma, te kao takvo postoji samo u ljudskom umu. To bi značilo da se znanje može podijeliti samo kao informacija, a nakon toga postaje znanje u nekom drugom ljudskom umu. Laudon K.C., Laudon J.P. (2007.)

Općenito se na znanje gleda kao na nešto što je stvoreno korištenjem informacije ili informacija, a rezultira nekom aktivnošću. Znanje bi trebalo odgovoriti na pitanje “Kako?”. Znanje je stavljeno u kontekst, što znači da je osim informacija potrebno imati i kontekst kako bi se proizvele određene akcije.

Znanje je kombinacija podataka i informacija kojemu je dodano profesionalno mišljenje, vještina te iskustvo. Znanje određuje kako koristiti informaciju. Znanje je također vrijedna "imovina", nužna pri donošenju odluka.

3.4. Mudrost

Mudrost je sposobnost razmišljanja i djelovanja korištenjem znanja, iskustva, razumijevanja, zdravog razuma i uvida. Mudrost je povezana sa atributima kao što su nepristrana sudba o nečemu, suosjećanje, nevezanost za određenu stvar i vrline kao što su etika, moralnost i dobronamjernost.

Najbolja definicija mudrosti bi dakle bila, da je to sposobnost razumijevanja informacija i znanja što učiniti sa stečenim znanjem.

Neki ljudi često miješaju pojmove znanje i mudrost. Primarna razlika između te dvije riječi leži u tome što mudrost uključuje zdravu dozu perspektive i sposobnost dobrog prosuđivanja o predmetu dok znati nešto je znanje, i to je to. Svatko može postati upoznat s nekom temom čitajući, istražujući i pamteći činjenice. Mudrost je znati kada što reći od znanja koje smo stekli.

Znanje je nakupljanje činjenica i informacija. Mudrost je sinteza znanja i iskustava u uvidima koji produbljuju nečije razumijevanje odnosa i smisla života. Drugim riječima, znanje je alat, a mudrost je zanat u kojem se alat koristi.

Spomena o mudrosti ima i u Bibliji.
Iz iskustva proizlazi svjedočenje. Jakov (1:5) kaže nam da će, ako zatražite mudrost, Bog je dati velikodušno, ne pronalazeći krivnju. „Ako nekome od vas nedostaje mudrosti, treba zamoliti Boga, koji velikodušno daje svima, a da ne nađe krivnju, i ona će mu se dati.“
Isto tako, mudrost se nalazi među sedam darova Duha Svetoga, uz razumijevanje, savjet, znanje, čvrstinu, pobožnost i strah od Boga. Iako ih neki kršćani prihvaćaju kao konačan popis određenih svojstava, drugi ih shvaćaju samo kao primjere djela Duha Svetoga kroz vjernike.

Mudrost je vrlina koja nije urođena, već se može steći samo iskustvom. Mudrost je dio vrline, koja nije vrsta znanja i stoga se ne može podučavati. Zato se recimo u većini slučajeva za ljude starije životne dobi kaže da su mudraci, Gandalf, stari čarobnjak iz knjige i filma „Gospodar prstenova“ je dobar primjer toga. Svatko koga zanima isprobavanje novih stvari i razmišljanje o procesu ima sposobnost steći mudrost. Učeći što više možete, analizirajući svoja iskustva i stavljajući svoje znanje na test, možete postati mudrija osoba.

4. Kratko o znanosti o podacima i rudarenju podataka

Podatkovna znanost (eng. data science) interdisciplinarno je područje kojim podatkovni znanstvenici proširuju svoja znanja učeći kako da razmišljaju na sistematičan način, integrirajući znanja iz različitih područja (Voulgaris, 2014). Da bi se ti podaci mogli prikupljati i procesuirati, potrebna su znanja iz drugih znanstvenih disciplina kao što su na primjer: matematika, statistika, teorija informacija obrada signala, skladišta podataka, računarstvo visokih performansa, strojno učenje , dubinska analiza podataka, baze podataka te vizualizacija podataka.

Hand i dr. (2001) definiraju rudarenje podataka kao znanost izdvajanja korisnih informacija iz velikih skupova podataka ili baza podataka. To je disciplina koja objedinjuje statistiku, strojno učenje, upravljanje podacima i bazama podataka, prepoznavanje uzoraka, umjetnu inteligenciju i druga područja.

S druge strane, dubinska analiza podataka odnosno rudarenje podacima (eng. Data mining) je način obrade podataka kojem je krajnji cilj izvući korisno znanje i informacije iz podataka. Sam naziv daje naslutiti da se ovdje radi o rudarenju podataka, što znači da se ne uzimaju apsolutno svi podaci koji se nađu, nego se ubiru podaci iz golemih baza, kao što su Amerikanci izvlačili grumene zlata iz blata sredinom 19. stoljeća. No, da se analiza podataka može provoditi, moramo imati neki skup podataka iz stvarnog svijeta koji je konačan. On se dobiva iz neko procesa prikupljanja podataka ili statističke analize određene stvari ili pojave. Bez obzira na to, složenost tih procesa je velika te su zahtjevni za ručnu obradu pa se zbog tog razloga u te svrhe koristi računalo. Prije nego što se kreće u postupak analize podataka mora biti utvrđen točan cilj same analize.

5. Znanosti na kojima Data Science počiva

5.1. Programiranje

U suštini, programiranje je objašnjavanje računalu što da radi kroz operacije. U jednom seminaru koji sam slušao, Ryan Labouve je programiranje opisao kao „izrada nevidljivih kompjutera pomoću riječi“ i mislim da općenito za ljude koji nisu u IT sferi poslovanja, ovo je moguće najbolja i najjednostavnija definicija programiranja. To samo stvaranje programa u sebi sadrži elemente umjetnosti, znanosti, dizajna, matematike te inženjeringa.

5.1.1. Tipovi računalnih jezika

Što se programskih jezika tiče, razlikujemo:

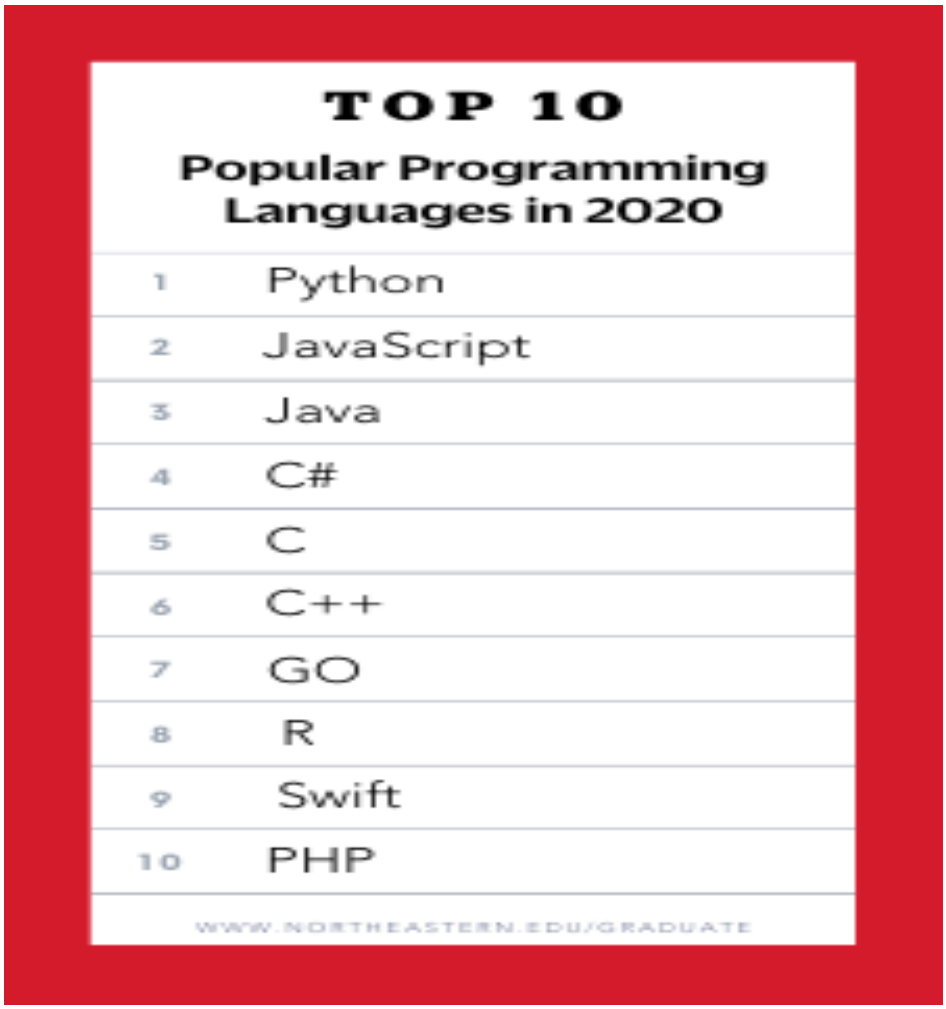
Jezici koji se koriste za izradu desktop programa te mobilnih aplikacija Najkorišteniji programski jezici bili bi C++, C#, Python i Java.

Python se općenito smatra jezikom koji je prilagođen korisnicima upravo zbog svoje jednostavne sintakse i mogućnosti implementacije s jezicima kao što su C i C++.

C++ je moćan programski jezik opće namjene. Može se koristiti za razvoj operativnih sustava, preglednika, igara i slično.

C# je Microsoftov objektno orijentirani programski jezik koji nastoji kombinirati računalnu snagu C++ s lakoćom programiranja Visual Basic-a. C# se temelji na C++ i sadrži značajke slične onima u Javi. Dizajniran je da funkcionira zajedno sa Microsoft-ovom .NET platformom. TechTarget Contributor (2007.)

Java je naširoko korišten objektno orijentirani programski jezik i softverska platforma koja radi na milijardama uređaja, uključujući prijenosna računala, mobilne uređaje, igraće konzole, medicinske uređaje i mnoge druge. Pravila i sintaksa Jave temelje se na jezicima C i C++. IBM Cloud Education (2019.)



Slika 1 prikazuje 10 najpopularnijih programskih jezika 2020. godine, Izvor: https://www.northeastern.edu/graduate/blog/most-popular-programming-languages/

- **Jezici za označavanje**

Jezici koji služe kao prevodioci između ljudskog i kompjutorskog jezika, funkcioniraju na principu da čovjek utipka određeni dio teksta, te nakon toga računalo obrađuje taj jezik u skladu sa zadanim uputama. Najbolji primjer tih jezika su HTML i XML.

- **Upitni jezici**

Koriste se kao komunikacijski alat između programera i baze podataka, najpoznatiji upitni jezik je SQL.

5.1.1.1. Skriptni jezici

Većinom se koriste za izradu sadržaja na internetu (web aplikacije, interaktivne karte i slično). Najkorišteniji skriptni jezici su JavaScript, PHP, Ruby, Bash te Perl.

5.1.2. Kompajleri i interpreteri

Kada komuniciramo sa računalom, većinom to radimo na jeziku koji je razumljiv nama ljudima. To bi se nazivalo izvornim kodom. Međutim, nas računalo ne razumije na našem jeziku, nego čita podatke u obliku binarnih brojeva 1 i 0, te se to naziva strojni kod. Za pretvaranje izvornog koda u strojni kod, služimo se kompajlerima i interpreterima.

5.1.2.1. Kompajleri

Kompajleri skeniraju čitav izvorni kod te ga prevode kao cjelinu u strojni kod. Za razliku od interpretera treba im puno više vremena za analizu koda. Također generiraju objektni kod koji dodatno zahtijeva povezivanje, stoga mu je potrebno više memorije. Najpoznatiji kompajlerski jezici su : Java, C te C++.

5.1.2.2. Interpreteri

Za razliku od kompajlera, interpreteri prevode izvorni kod liniju po liniju. Samo zbog te činjenice potrebno im je manje vremena za analizu izvornog koda, ali kada dolazimo do čitanja cijelog koda, vrijeme izvršavanja je relativno sporije od kompajlera. Također se ne generira nikakav objektni kod, što znači veća ušteda memorije.
Najpoznatiji interpreterski jezici su Python, Ruby i JavaScript.

5.2. Rudarenje podataka (eng. Data mining)

Kao što je već prije bilo spomenuto, rudarenje podataka je proces kojim se iz velikih skupova podataka izvlače korisne, nove informacije. Kroz rudarenje podataka se ne stvaraju nove informacije, nego se izvode vrijednosti već poznatih podataka te se oni stavljaju u složenu cjelinu. Kranji cilj rudarenja podataka jest izvući informacije iz skupa podataka i transformirati ih u razumljivu strukturu za daljnju upotrebu.

Naše svakodnevne aktivnosti djelomično su zabilježene kroz korištenje naših računala, pametnih telefona i kreditnih kartica. Nadalje, svjetske kompanije prikupljaju velike skupove podataka vezane uz prodajne transakcije, trgovanje dionicama, promotivne akcije i povratne informacije korisnika. Domene poput medicine, znanosti, društvenih mreža i web pretraživanja generiraju petabajte podataka na dnevnoj bazi.

„Važno je istaknuti da je rudarenje podataka više umjetnost nego znanost. Ne postoji recept za uspješno rudarenje podataka koje će sigurno rezultirati pronalaženjem vrijednih informacija. Međutim, vjerojatnost uspjeha će se povećati ako se slijede koraci procesa rudarenja podacima“ (Baragoin et.al., 2001.)

Rudarenje podataka je samo po sebi, najkraće rečeno, proces pronalaženja korisnih informacija te stavljanje istih u upotrebu u određenom području. Sam proces rudarenja podataka je vrlo zahtjevan i uključuje više faza, te ga jedna osoba sama jednostavno ne može raditi.

Prema Hand, D. J. (2007). rudarenje podataka je otkrivanje interesantnih, neočekivanih ili vrijednih struktura u velikim skupovima podataka. To bi se moglo podijeliti na dvije strukture. Prva bi bila 'globalna' struktura, a tu je cilj modelirati oblike i značajke oblika distribucije. Druga, 'lokalna' struktura radi na detekciji anomalija te njihovom uklanjanju, ako predstavljaju prijetnju.

Rudarenje podataka je primjena specifičnih algoritama za izvlačenje uzoraka iz podataka. Dodatni koraci u procesu otkrivanja znanja – priprema, odabir i čišćenje podataka, ugradnja odgovarajućeg prethodnog znanja i pravilno tumačenje rezultata rudarenja neophodni su da bi se osiguralo korisno znanje iz podataka (Fayyad i dr. 1996).

Razvitak strojnog učenja započeo je 50-tih godina prošlog stoljeća sa sustavima koji su se bazirali na pravilima da obrade različite vrste informacija, što je zahtijevalo puno vremena, te je njihova upotreba bila ograničena na specifične domene.

S razvojem sve više alata, strojno učenje je postalo jako praktično područje s mnogo upotreba u stvarnome svijetu kroz napredne sustave obrade podataka koji su rješavali

probleme bez pretpostavki. To se općenito naziva prepoznavanje uzoraka koje i dalje igra veliku ulogu u analitici podataka. Kako iz dana u dan sve više napredujemo glede sfere tehnologije, naravno da je i znanost o podacima znatno napredovala u zadnjih 70 godina. Danas na tržištu postoji mnogo besplatnih i široko dostupnih programskih jezika te alata koji omogućuju obradu podataka koja nije bila moguća prije, te se ta ista tehnologija iz dana u dan poboljšava.

5.3. 8 najbitnijih tehnika rudarenja podataka

5.3.2.1. Klasifikacija

Klasifikacija se koristi kada su setovi slučajeva dodijeljeni u nivoe karakterističnih faktora koji se baziraju na njihovim karakteristikama. Trenirani set poznatih slučajeva se koristi za razvoj klasifikacijskog algoritma koji se onda može koristiti za predviđanje nepoznatih slučajeva i kojoj bi oni klasifikaciji mogli pripadati. (Šimec, Lozić 2020.)

Na primjer, ako procjenjujemo podatke o financijskom stanju i povijesti kupnje pojedinačnih kupaca, mogli biste ih klasificirati kao "niski", "srednji" ili "visoki" kreditni rizik. Zatim biste mogli koristiti te klasifikacije da biste saznali još više o tim kupcima.

5.3.2.2. Uzorci praćenja

Jedna od osnovnih metoda rudarenja podataka je učenje prepoznavanja uzoraka u skupovima podataka. Na primjer, možete primijetiti da vaša prodaja određenog proizvoda naglo raste neposredno prije praznika ili primijetiti da toplije vrijeme dovodi više ljudi na vašu web-lokaciju.

5.3.2.3. Asocijativna pravila

Asocijativna pravila su povezana s uzorcima praćenja, ali je specifičnije za ovisno povezane varijable. U ovom slučaju, tražimo određene događaje ili atribute koji su u visokoj korelaciji s drugim događajem ili atributom; na primjer, mogli biste primijetiti da kada vaši kupci kupe određenu stavku, često kupuju i drugu, srodnu stavku. Recimo, kada kupe čajne kekse, često je pretpostavljeno da će kupiti čaj, ili mlijeko. To se obično koristi za popunjavanje odjeljaka "ljudi su također kupili" u internetskim trgovinama.

5.3.2.4. Otkrivanje izvanrednih vrijednosti

U mnogim slučajevima jednostavno prepoznavanje sveobuhvatnog obrasca ne može vam dati jasno razumijevanje vašeg skupa podataka. Izvanredne vrijednosti u ovom slučaju predstavljaju neko odstupanje koje nije uobičajeno. Također morate biti u mogućnosti prepoznati anomalije ili odstupanja u svojim podacima. Na primjer, ako su vaši kupci gotovo isključivo muškarci, ali tijekom jednog tjedna u srpnju, došlo je do velikog porasta broja kupaca ženskog spola, trebali biste istražiti porast i vidjeti što ga je potaknulo, tako da to možete ponoviti ili bolje razumjeti svoju publiku u procesu.

5.3.2.5. Klasteriranje

Klasteriranje je vrlo slično klasifikaciji, ali uključuje grupiranje dijelova podataka zajedno na temelju njihove sličnosti. Na primjer, možete odlučiti grupirati različite demografske podatke svoje publike u različite pakete na temelju toga koliko raspoloživog prihoda imaju ili koliko često kupuju u vašoj trgovini.

5.3.2.6. Regresija

Regresija, koja se prvenstveno koristi kao oblik planiranja i modeliranja, koristi se za utvrđivanje vjerojatnosti određene varijable, s obzirom na prisutnost drugih varijabli. Na primjer, možete ga koristiti za predviđanje određene cijene, na temelju drugih čimbenika kao što su dostupnost, potražnja potrošača i konkurencija. Točnije, glavni fokus regresije je pomoći vam da otkrijete točan odnos između dvije (ili više) varijabli u danom skupu podataka.

5.3.2.7. Stablo odlučivanja

Stablo odlučivanja je grafički prikaz rješenja temeljenog na mnogobrojnim uvjetima koji su zadani. Stablo odlučivanja služi kako bi se na temelju različitih podataka donijela neka odluka. To može biti ili jednostavna analiza kao kvaliteta kave u kafiću, ili kompleksnija analiza poput količine udjela kave po jednoj mjeri, te udio robuste i arabice koje su zapravo dvije vrste kave koje se miješaju da se dobije kava, u serviranoj šalici kave.

5.3.2.8. Predviđanje

Predviđanje je jedna od najvrjednijih tehnika rudarenja podataka jer se koristi za projiciranje vrsta podataka koje ćete vidjeti u budućnosti. U mnogim je slučajevima samo prepoznavanje i razumijevanje povijesnih trendova dovoljno da se donekle točno predvidi što će se dogoditi u budućnosti. Na primjer, možete pregledati kreditnu povijest potrošača i prošle kupnje kako biste predvidjeli hoće li u budućnosti predstavljati kreditni rizik.

5.3.3. Korisnost rudarenja podataka

Može se zaključiti kako je rudarenje podataka sve više prisutno u poduzetništvu zbog velikog gomilanja podataka koje se stvaraju u poduzećima kroz njihovu učestalu komunikaciju sa klijentima preko interneta. U tom velikom skupu podataka sama poduzeća mogu naći neke vrlo korisne informacije za poslovanje, kroz rudarenje podataka pružatelj usluge rudarenja podataka saznaje puno informacija o klijentu, koja su pogodna za daljnji razvoj poslovnog odnosa. Primjer o saznavanju tih korisnikovih informacija bio bi kolačić (eng. cookie), koji predstavlja male datoteke koje internetski preglednici spremaju na stranu klijenta, koja ima svrhu pamćenja podataka klijenta. Kolačić je koristan s obje strane, za klijenta je od koristi jer putem prijave pamti podatke o korisniku te mu tim putem omogućava bržu prijavu na stranicu, a za poslužitelja pruža neke personalizirane informacije koje može iskoristiti u svrhu pružanja bolje podrške u obliku usluge za korisnika.

Specifične prednosti rudarenja podataka uključuju sljedeće:

Učinkovitiji marketing i prodaja - Rudarenje podataka pomaže trgovcima da bolje razumiju ponašanje i preferencije kupaca, što im omogućuje stvaranje ciljanih marketinških i reklamnih kampanja. Slično tome, prodajni timovi mogu koristiti rezultate rudarenja podataka kako bi poboljšali stope konverzije potencijalnih kupaca i prodali dodatne proizvode i usluge postojećim kupcima.

Bolja usluga korisnicima - Zahvaljujući rudarenju podataka, tvrtke mogu brže identificirati potencijalne probleme s korisničkom službom i agentima kontaktnih centara dati ažurne informacije za korištenje u pozivima i online razgovorima s klijentima.

Poboljšano upravljanje opskrbnim lancem - Organizacije mogu uočiti tržišne trendove i točnije predvidjeti potražnju proizvoda, što im omogućuje bolje upravljanje zalihama robe i zaliha. Upravitelji lanca opskrbe također mogu koristiti informacije iz rudarenja podataka za optimizaciju skladištenja, te distribucije.

Povećano vrijeme neprekidnog rada proizvodnje - korištenje operativnih podataka iz senzora na proizvodnim strojevima i drugoj industrijskoj opremi podržava aplikacije za prediktivno održavanje kako bi se identificirali potencijalni problemi prije nego što se pojave, pomažući u izbjegavanju neplaniranih zastoja.

Jače upravljanje rizikom - Menadžeri rizika i poslovni rukovoditelji mogu bolje procijeniti financijske, pravne, i druge rizike za tvrtku te razviti planove za njihovo upravljanje.

Niži troškovi - Rudarenje podataka pomaže u smanjenju troškova kroz operativnu učinkovitost u potrošnji kad se gleda veća slika naravno, jer se mora uzeti u obzir da samo rudarenje podataka također zahtjeva troškove koji se moraju pokriti.

U konačnici, inicijative za rudarenje podataka mogu dovesti do većih prihoda i dobiti, manjih gubitaka vrijednih resursa u vlastitom poduzeću, kao i konkurentskih prednosti koje tvrtke izdvajaju od njihovih poslovnih suparnika.

5.4. Statistika

Riječ statistika dolazi od latinske riječi status, što znači stanje.
Statistika je, ukratko rečeno znanost koja se bavi prikupljanjem te analizom podataka izvođenjem određenih zaključaka.

Statistika je grana primijenjene matematike koja uključuje prikupljanje, opisivanje, analizu i izvođenje zaključaka iz podataka. Matematičke teorije koje stoje iza statistike uvelike se oslanjaju na diferencijalni i integralni račun, linearnu algebru i teoriju vjerojatnosti.(Leksikografski zavod Miroslav Krleža, 2021.)

Statističari, ljudi koji se bave statistikom, također se bave utvrđivanjem kako izvući pouzdane zaključke o velikim grupama i općim događajima iz ponašanja i drugih vidljivih karakteristika malih uzoraka. Ovi mali uzorci predstavljaju dio velike skupine ili ograničeni broj slučajeva općeg fenomena.

„Pojave koje se susreću u društvu i prirodi masovne su pojave. Sastoje se od velikog broja jedinica ili elemenata. Da bi se ove pojave mogle upoznati koristi se specifični znanstveni pristup sadržan u pojmu statistika. Statistikom se opisuju, upoznavaju, istražuju, uspoređuju i analiziraju navedene masovne pojave. Statistika koristi brojčani način izražavanja jer je on precizan, kratak i jasan. Statistika je, zahvaljujući ovoj činjenici, primjenjiva u gotovo svim područjima kako znanstvenog tako i praktičnog istraživanja.“(Kero, Dobša, Bojanić, 2008. str 1.)

„Nezaobilazna je i u teoriji informacija i komunikacija. Govoreći posve općenito informatika se dobrim svojim dijelom temelji na statističkoj teoriji i modelima. Prisutna je u statističkoj kontroli proizvoda, u dijagnostičkim elektroničkim uređajima kod mnogih medicinskih istraživanja i sl.“ (Šošić, Serdar, 1992., 9-10.).

Drugim riječima, statistika predstavlja skup znanstvenih modela i metoda koji se koriste u sveobuhvatnoj analizi masovnih pojava. (Kero, Bojanić-Glavica, 2003., 23.).

5.5. Baze podataka

Kako navode Maleković i Rabuzin, baza podataka u suštini predstavlja skup podataka, ograničenja i operacija koje reprezentiraju neke aspekte iz stvarnog svijeta. Svaka promjena, brisanje ili čitanje obavlja se istim software-om. Riječ je o tehnologiji koja je nastala s namjerom da se uklone slabosti tradicionalne “automatske obrade podataka” iz 60-tih i 70-tih godina 20. stoljeća. Sama po sebi, ta grana IT-a se razvila sa ciljem produktivnosti i pouzdanosti u načinu spremanja i čuvanja podataka na računalima.

Sustav za upravljanje bazom podataka(eng. DBMS) je poslužitelj baze podataka. To se može interpretirati da DBMS zapravo služi kao spremnik i posrednik za podatke koji se kreiraju, ažuriraju, čitaju te brišu iz baze podataka.

Podaci su u bazi podataka organizirani modelom podataka, a model podataka je skup pravila kako bi logički trebala izgledati baza podataka. Kod današnjih DBMS-a su najčešća ova dva modela:

Relacijski model:

Zasnovan na matematičkom pojmu relacije. I podaci i veze medu podacima prikazuju se “pravokutnim” tabelama.

Objektno orijentirani model:

To je sustav u kojem su informacije ili podaci predstavljeni u obliku objekata koji se koristi u objektno orijentiranom programiranju.

Recimo, ako za primjer iz stvarnog svijeta uzmemo tvrtku koja se bavi prodajom i servisiranjem novih i rabljenih automobila, u njihovoj bazi podataka bi se trebali nalaziti svi modeli svih vozila, te svi zamjenski dijelovi koji su potrebni za servisiranje tih istih vozila.

Za izradu tih baza podataka u digitalnom obliku, u većini slučajeva se koristi softver koji je specijaliziran isključivo za tu upotrebu. U toj sferi, najčešće se koriste : MySQL, Oracle, SQL Server, SQLite, PostgreSQL, MS Access, te LibreOffice.

5.6. Umjetna inteligencija (AI - Artificial ingeligence)

5.6.1. OpenAI Five

Umjetna inteligencija se karakterizira kao dio računalne znanosti koja se bavi razvijanjem sposobnosti određenog računala da dio operacija obavlja autonomno, drugim riječima, da računalo samo zna razmišljati do neke mjere, te samo donositi odluke.

Inteligentnim se sustavom smatra svaki sustav koji uči na temelju iskustva. Za primjer tu možemo uzeti program koji razvija Elon Musk sa svojim poduzećem pod nazivom OpenAI Five, koja je osnovana 2015. godine. Naime, u travnju 2019. godine, AI razvijen od poduzeća OpenAI nastupao je na velikoj pozornici protiv najboljih igrača Dote 2 u svijetu.

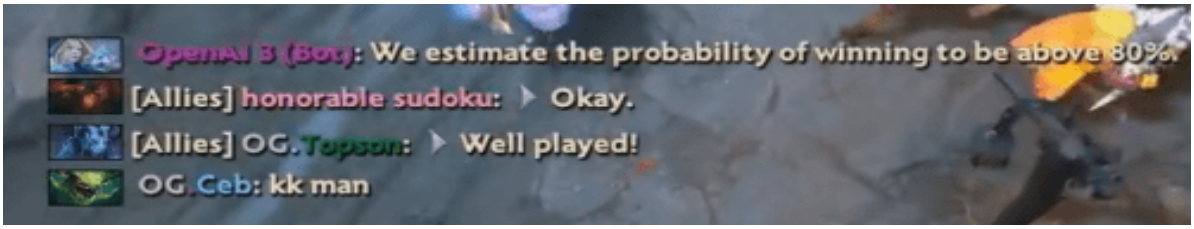
Slika 2 Prikazuje OpenAI na USB disku prije nego što počinje dvoboj u Dota-i 2 (Izvor: OpenAI)

Dota 2 je MOBA(eng. Multiplayer online battle arena, ili na hrvatskom mrežna borbena arena za više igrača) koja je nastala 2013. godine, te izgledom dosta podsjeća na sličnu igru pod nazivom League of Legends, koja je zapravo kopija Dota-e koju mogu savladati i mlađi uzrasti. Dota 2 je vrlo složena i vrlo popularna video igra koju igraju dva tima od pet igrača. Tim koji sruši protivnikovu središnju bazu "Ancient" pobjeđuje u igri. Okruženje igre ima 123 likova i najvažnijih "heroja", 22 obrambena tornja, desetke likova koji nisu igrači, stotine vještina i predmeta te mnogo značajki kao što su rune, drveće, odjeljenja i tako dalje.

Zaposlenici OpenAI-a su koristili Dotu 2 kao platformu za opću namjenu te je AI- igrao sam protiv sebe igrao u vremenskom razdoblju od 10 000 godina za razvoj inicijalnog software-a, te prije igranja protiv najboljih na svijetu, bot je odigrao sam protiv sebe igara u trajanju od 45 000 godina. Tu se naravno ne misli da je igrao pravih 45 000 godina, nego da je više puta paralelno igrao sam protiv sebe, i nakon svakog igranja protiv samoga sebe, AI je naučio nešto više i kroz taj proces je naučio gotovo savršeno igrati. Vraćamo se natrag na 13. 4. 2019. godine, kada je OpenAI igrao protiv dvostrukih prvaka svijeta u doti, klub pod nazivom OG. U relativno brzom vremenskom periodu OpenAI pobjeđuje OG 2-0 u seriji.

Slika 3 prikazuje okršaj između OpenAI five-a i tada trenutačno najboljih igrača Dote 2 na svijetu (Izvor: OpenAI)

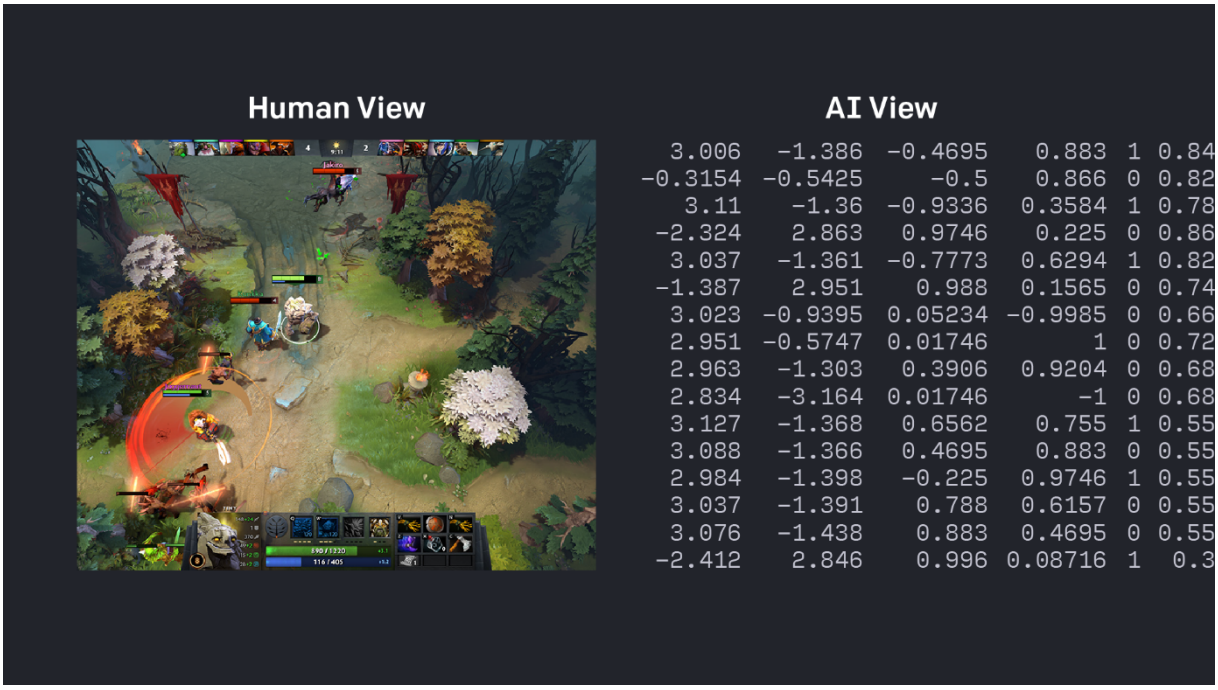
Za razliku od ljudskih igrača, AI sistem je teško omesti u namjeri za pobjedom. Međutim to nije zaustavilo OpenAI da tokom meča s vremena na vrijeme uputi poruku svojim protivnicima. Gotovo sigurno, te u vrlo nasumičnim intervalima, kada se zadovoljio određeni kriterij u igri (broj ubojstava ili količina gold-a, koji predstavlja valutu u igri) OpenAI bi uputio postotak sigurnosti u pobjedu svojim protivnica u pisanom obliku, koji se je također prevodio u text-to-chat. To je započelo na početku igre sa 80% vjerojatnošću za pobjedom, te je broj porasao sve do 99%.



Slika 4 prikazuje odnos komunikacije bot-a OpenAI Five sa najboljim igračima Dota-e (Izvor: OpenAI)

OpenAI je sudjelovao u 7,257 mečeva protiv stvarnih ljudi, a postotci pobjede su pomalo zabrinjavajući. Nakon spomenutog događaja, koji je trajao 3 dana, tko je god htio se je mogao iskušati u pokušaju pobjede OpenAI bota. Brojke nam govore da je OpenAI Five imao postotak pobjede nevjerojatnih 99.4%. To znači da su ljudi pobjedili 42 puta, a OpenAI ostalih 7.215.

Slika 5 prikazuje postotak pobjede OpenAI-Five-a protiv ljudskih protivnika



Slika 6 prikazuje kako OpenAI five procesira informacije dobivane iz videoigre (Izvor: OpenAI)

Zasigurno se pitate zašto Dota, team iz OpenAI-a se je izjasnio, da ako oni uzmu problem koji je nerješiv sa trenutačnim metodama, imaju veće šanse za povećanje sposobnosti svojih alata.

Krajnji cilj ove tehnologije je zapravo, po mome mišljenju, napraviti umjetnu inteligenciju koja će biti u mogućnosti pomagati čovječanstvu, bilo to pomaganje starijim i nemoćnim osobama, ili možda zamjenjivanjem za teške fizičke poslove, te čak kao što vidimo i tu iz primjera na kraju krajeva, u svrhu zabave.

"AI botovi su upravo pobijedili ljude u videoigri Dota 2. To je velika stvar, jer je njihova pobjeda zahtijevala timski rad i suradnju - veliku prekretnicu u napretku umjetne inteligencije." (Bill Gates,2018.)

Nedugo nakon pobjede nad najboljim igračima Dote na svijetu, tim koji je radio na OpenAI-five uzeo je drugi izazov, napraviti Google, ali za slike. Time želim reći da su se fokusirali da naprave software, koji će samo na input preko tipkovnice generirati slike i ilustracije velike rezolucije i odlične kvalitete.

5.7. OpenAI DALL-E, DALL-E 2 i DALL-E mini

DALL-E i DALL-E 2 modeli su transformatora koje je razvio OpenAI za generiranje digitalnih slika iz opisa prirodnog jezika. Ime mu je spoj riječi WALL-E-a i Salvadora Dalíja.

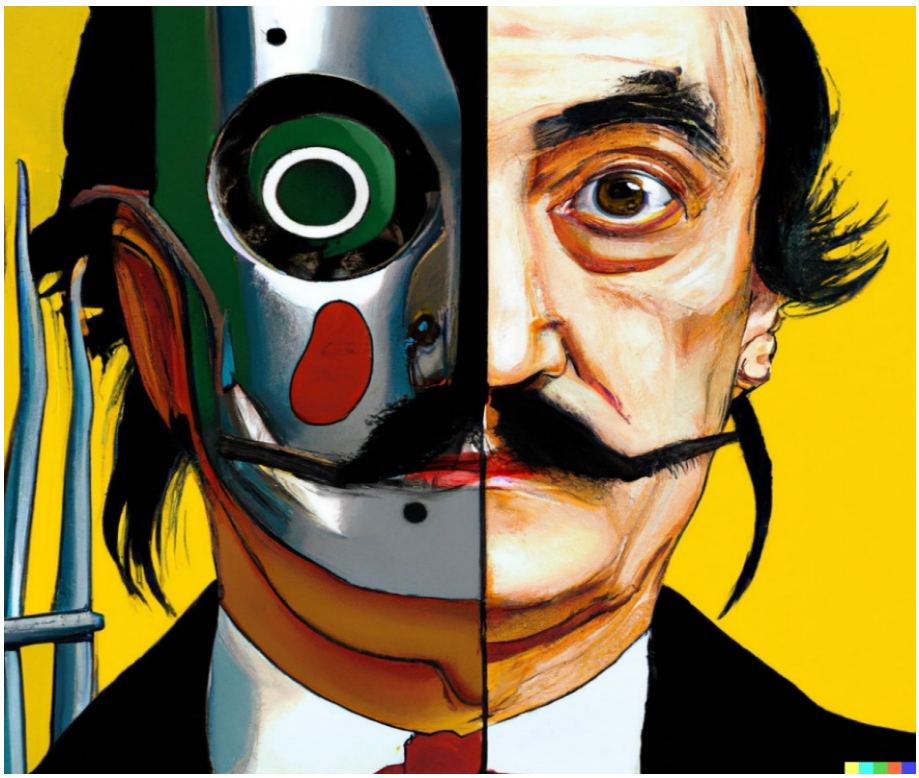
Za početak ovog poglavlja, možda bi najbolje bilo prvo pogledati sliku, a nakon toga se dodatno raspisati, jer ipak se kaže da slika govori više od tisuću riječi.

Slika 7 prikazuje tehnološkog novinara koji piše članak o novom AI sistemu koji može stvarati slike koje su vrijedne pažnje i pomalo čudne (Izvor: OpenAI)

Ovu sliku napravio je program(kada se bolje pogleda u ruke novinara, tada postane jasno). Kao što je i samim opisom slike rečeno, DALL-E 2 je također potekao od tvrtke OpenAI, te zapravo predstavlja njihov najnoviji pothvat u sferi umjetne inteligencije. Najavljen u travnju ove godine, DALL-E 2 je software, koji se bazira na generiranju slike u rezoluciji nevjerojatne kvalitete, koje mogu biti u obliku ilustracija, fotografija, slika, animacija, ili općenito bilo kojeg drugog stila umjetnosti koji se može ručno unijeti u tražilicu programa. Za razliku od svog preteče DALL-E-a, koji je izašao 2021. godine i bio znatno manje razvikan od trenutačnog programa, novi pruža brže procesiranje, bolju rezoluciju, te opciju da korisnik radi promjene na već generiranoj novoj slici. Na primjer, ako su na slici zmajevi koji jedu sladoled na vrhu planine Kilimanjaro, korisnik programa može promijeniti oblik sladoleda sa korneta u sladoled na štapiću. Također se može učitati i vlastita slika s računala te se može i editirati.

Međutim, nije sve tako bajno. OpenAI nije sam program pustio u komercijalnu upotrebu, zbog straha da bi takav program mogao poremetiti širu tehnološku industriju. Jedini trenutačno masi dostupan program je DALL-E mini, koji radi iste stvari kao njegova dva nasljednika, ali u vidljivo slabijoj kvaliteti generiranja slika i sličnih sadržaja.

Sa tim na umu, OpenAI je dozvolio i pozvao određenu skupinu ljudi kao što su znanstvenici, umjetnici, te stvaratelji sadržaja na Youtube-u, da dobiju mogućnost koristiti program na kratki vremenski period ili za određeni broj iteracija.



Slika 8 prikazuje Salvador Dalija sa pola robotskog lica, slika također generirana od DALL-E 2 (Izvor: OpenAI)

DALL·E 2 stvara varijacije slike dobivanjem CLIP ugrađenih slika i njihovim provođenjem kroz Diffusion dekoder. Zanimljiv nusprodukt ovog procesa je uvid u to koji su detalji modeli naučeni, a koji detalji propušteni. Vidno je da od svih trenutačno, umjetnici najviše strahuju za svoje buduće zanimanje, jer se kreativnost trenutačno pretvara sa izrade nekog umjetničkog djela, u jednostavno upisivanje riječi u područje za tekst te čekanje AI-a da generira sliku.

5.8. Strojno učenje (ML – Machine learning)

Strojno učenje je zapravo grana AI-a i znanosti o računalima koja se u suštini bavi korištenjem podataka i algoritama za računala, da imitiraju način na koji ljudi uče, sa ciljem postepenog poboljšanja točnosti ishoda realizacije određenog cilja. U zadnjih nekoliko godina strojno učenje doživjelo je nagli razvoj i raširilo upotrebu u specijaliziranim sustavima, ali i u tehnologijama i uređajima kojima se svakodnevno koristimo.

Strojno učenje odnosi se na programiranje računala da mogu „učiti“ iz unosa koji su im dostupni. A sam pojam učenja odnosi se na proces pretvaranja iskustva u stručnost ili znanje. Uneseni podaci u algoritmu učenja su trening podaci koji predstavljaju iskustvo, a izlaz algoritma je neko znanje, koje uglavnom ima formu novog računalnog programa koji može obaviti neki zadatak.

Kroz to se softverskim aplikacijama dozvoljava da postaju preciznije kod predviđanja ishoda bez eksplicitnog programiranja istih da to rade. Strojno učenje koristi povijesne podatke kao ulaz sa ciljem da predviđaju novi izlaz, to jest da daju novu vrijednost van zauzvrat.

Jednostavnije rečeno, strojno učenje dozvoljava korisniku da računalni algoritam hrani sa velikom količinom podataka, te nakon toga računalo te podatke analizira i predlaže odluke temeljene na podacima koje je dobilo.

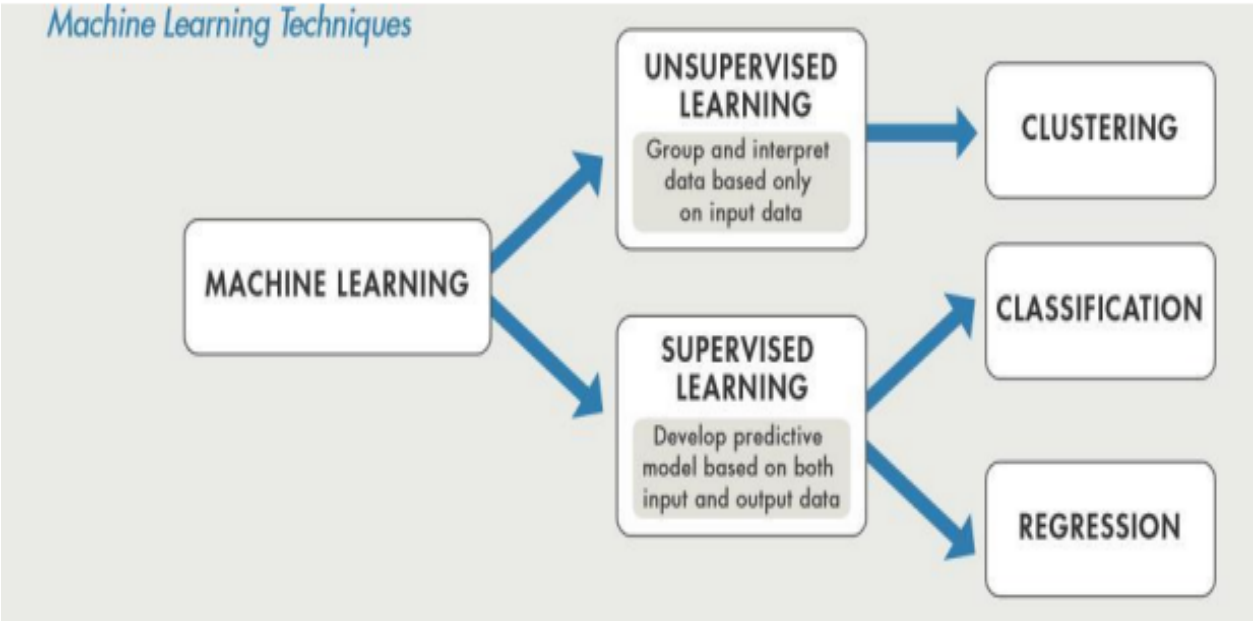
Neki autori, poput Harringtona (2012), smatraju kako se može koristiti u raznim poljima, od politike do geografije. Riječ je naime o alatu koji se može koristiti za mnoge probleme. Svako polje koje zahtjeva interpretaciju i radnju temeljenu na podatcima profitira od tehnika strojnog učenja.

Po navodima Shai Shalev-Shwartz-a i Shai Ben David-a,većina programa su ograničeni u adaptivnosti i ne mijenjaju se previše od trenutka kada su instalirani, no djelovanje programa koji imaju sposobnost učenja ovisi o ulaznim podacima te su takvi programi sami po sebi adaptivni.

5.8.1. Tipovi strojnog učenja

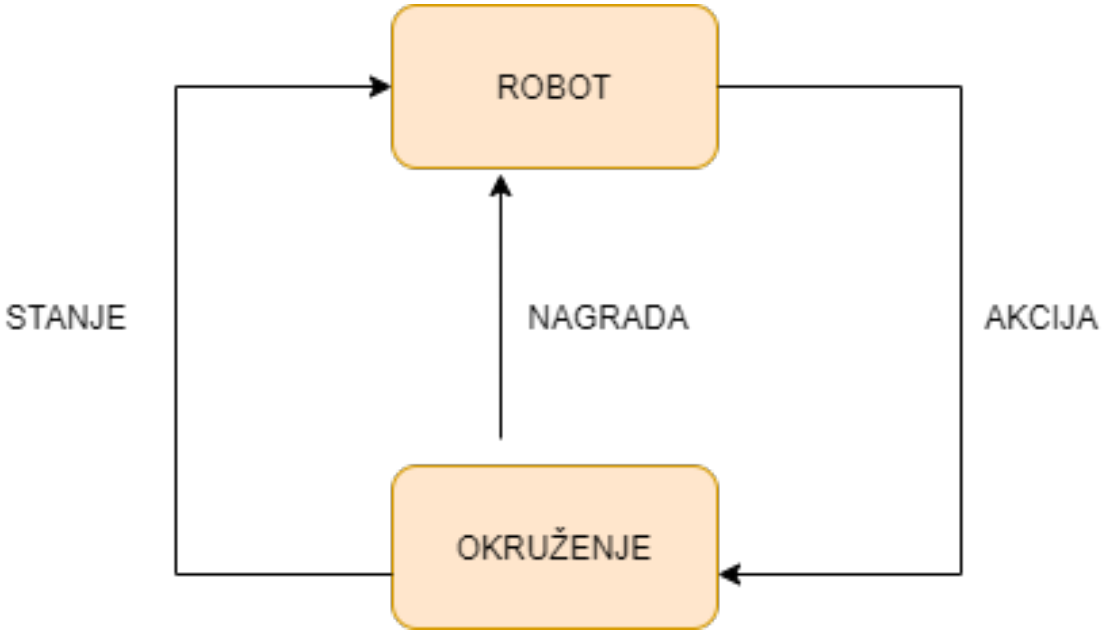
Nadzirano učenje: Učenje pod nadzorom događa se u prisutnosti nadzornika, baš kao i učenje koje malo dijete vrši uz pomoć svog učitelja. Primjer toga bi bio dijete koje se uči prepoznati plodove, boje, brojeve pod nadzorom učitelja.

Nenadzirano učenje: Učenje bez nadzora odvija se bez pomoći nadzornika baš kao što ptica uči sama letjeti. To je neovisan proces učenja.



Slika 9 prikazuje tehnike strojnog učenja (Izvor: https://www.mathworks.com/content/dam/mathworks/tag-team/Objects/i/88174_92991v00_machine_learning_section1_ebook.pdf)

Učenje s potporom ili pojačanjem: U ovoj vrsti učenja algoritam uči pomoću mehanizma povratnih informacija i prošlih iskustava. Uvijek se želi da se svaki korak u algoritmu poduzme za postizanje cilja. Vrsta učenja koja je povezana sa kontroliranjem robota sa svrhom njegovog učenja. To je zapravo AI tehnika koja trenira robota da obavlja zadatke koje želimo, te koji su nagrađivani ako je povratno ponašanje poželjno. Cilj ovog tipa učenja je naučiti pravila, koja su u suštini pridruživanje određene akcije određenim opažanjima. Tijekom učenja robot istražuje okolinu, promatra stanje stvari oko sebe i na temelju tih opažanja poduzima akciju. Opažanje je ono što robot može izmjeriti u svojoj okolini, a akcija u svom najnižem obliku je promjena konfiguracije robota. Ako robot reagira u skladu s našim očekivanjima, nagrađivan je. Proces učenja robota prikazan je na slici ispod.



Slika 10 prikazuje ciklus učenja robota uz metodu potpore ili pojačanja

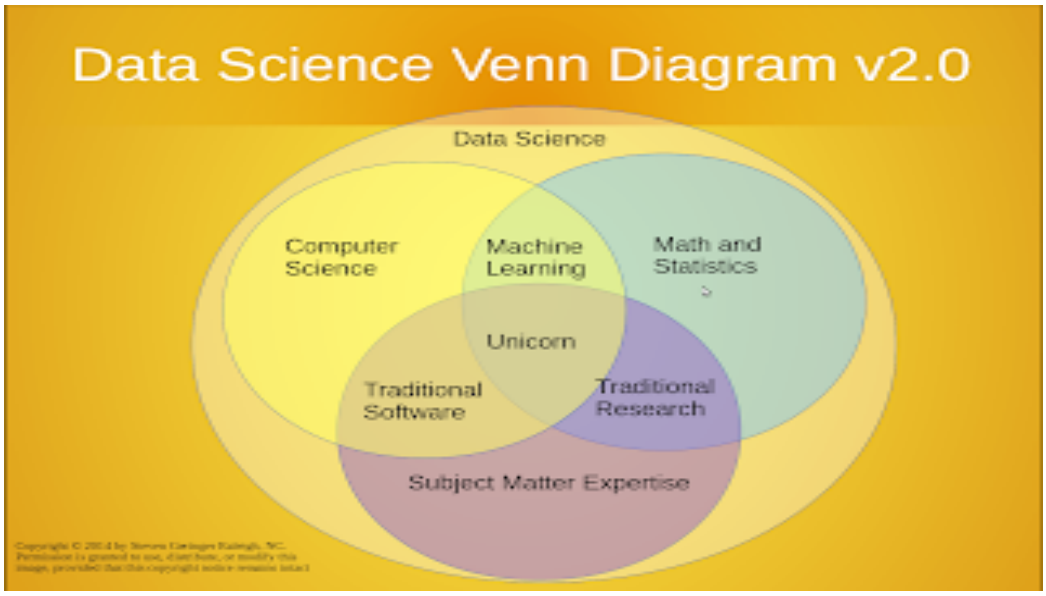
6. Primjena znanosti o podacima

Kad gledamo na znanost o podacima kao cjelinu, uočavamo da to nije samo jedna sfera područja znanja, nego da je to skup znanosti kojima je krajnji cilj stvaranje korisnih podataka i informacija koje se mogu pretočiti u znanje glede nekog poslovnog ili informacijskog sustava.

A te znanosti uključuju statistiku, matematiku, informatiku, strojno učenje, te tradicionalno istraživanje.

Dakle, ukratko rečeno, znanost o podacima je ponešto od svega.
Ovdje u sredini ovog Venn dijagrama vidimo da piše riječ „Unicorn“ što, prevedeno na hrvatski znači jednorog. Tu se implicira da je dobrog znanstvenika podataka teško naći.

Teško je dakle, precizno definirati zanimanje podatkovnog znanstvenika, iz razloga što različita okruženja utječu na definiciju te titule. O'Neil i Schutt (2013) definiraju ga kao znanstvenika koji posjeduje vještine od društvenih znanosti do biologije, koji radi s velikom količinom podataka i suočava se s računalnim problemima strukture, veličine, nereda, kao i kompleksnom prirodom podataka, u isto vrijeme rješavajući aktualne svjetske probleme.



Slika 11 prikazuje Venn-ov dijagram znanosti o podacima(Izvor: https://twitter.com/kdnuggets/status/421678003153039360?lang=hr)

7. STANJE NA TRŽIŠTU GLEDE ZNANOSTI O PODACIMA

Znanost o podacima i dalje je jedna od obećavajućih i traženih karijera za kvalificirane stručnjake. Karijera u znanosti o podacima je korisna i unosna, ali način na koji možete započeti karijeru u znanosti o podacima nije tako jednostavan. Diploma prvostupnika ili magisterija nije potrebna da biste postali stručnjak za podatkovnu znanost. Pravi skup vještina i iskustvo su ono što je potrebno.

Da biste započeli uspješnu karijeru u znanosti o podacima, trebate imati skup vještina koji uključuju analizu, strojno učenje, statistiku, itd. i morate biti sposoban rješavati probleme, kritički misliti i biti dobar pripovjedač da biste bili uspješni u znanosti o podacima.

U nastavku ove cjeline navesti ću i ukratko objasniti, osam zanimanja koja se povezuju sa znanošću o podacima.

7.8.1. Analitičar podataka

Analitičari podataka odgovorni su za niz zadataka uključujući vizualizaciju, brisanje i obradu golemih količina podataka. Također moraju s vremena na vrijeme izvršiti upite u bazama podataka. Jedna od najvažnijih vještina analitičara podataka je optimizacija. To je zato što moraju stvoriti i modificirati algoritme koji se mogu koristiti za izdvajanje informacija iz nekih od najvećih baza podataka bez oštećivanja podataka.

Kako postati analitičar podataka?

SQL, R, SAS i Python neke su od traženih tehnologija za analizu podataka. Dakle, certifikacija u tim alatima može lako dati poticaj vašim prijavama za posao. Također biste trebali imati dobre kvalitete rješavanja problema, što se podrazumijeva kod većine zanimanja povezanih sa IT industrijom.

7.8.2. Inženjer podataka

Prema (Anonymous, 2022). Inženjer podataka radi s skupovima podataka kako bi unaprijedio ciljeve znanosti o podacima. Za razliku od ostalih uloga, poput znanstvenika podataka, inženjer podataka uglavnom nije uključen u sveukupnu stratešku analizu, već je dublje uključen u rad s poslovnim skupovima podataka.

Inženjeri podataka također ažuriraju postojeće sustave novijim ili nadograđenim verzijama trenutnih tehnologija kako bi poboljšali učinkovitost baza podataka.

Kako se postaje inženjer podataka?

Ako ste zainteresirani za karijeru inženjera podataka, tada tehnologije koje zahtijevaju praktično iskustvo uključuju Hive, NoSQL, R, Ruby, Java, C++ i Matlab. Također bi pomoglo ako možete raditi s popularnim podatkovnim API-jima i ETL alatima.

7.8.3. Menadžer podataka i analitike

Menadžer podataka i analitike nadzire operacije znanosti o podacima i dodjeljuje dužnosti svom timu u skladu s vještinama i stručnošću. Njihove snage trebale bi uključivati tehnologije kao što su SAS, R, SQL itd. i naravno upravljanje.

Koje su kvalifikacije potrebne kako bi se postalo menadžer podataka i analitike?

Prvo i najvažnije, da biste krenuli putem karijere menadžera analitike, morate imati izvrsne društvene vještine, kvalitete vodstva i neuobičajen stav razmišljanja. Također biste trebali biti dobri u tehnologijama znanosti o podacima koje smo gore već naveli kao što su Python, SAS, R, Java itd.

7.8.4. Poslovni analitičar

Uloga poslovnih analitičara malo je drugačija od ostalih poslova znanosti o podacima. Iako dobro razumiju kako tehnologije usmjerene na podatke funkcioniraju i kako postupati s velikim količinama podataka, oni također odvajaju podatke visoke vrijednosti od podataka niske vrijednosti. Drugim riječima, oni identificiraju kako se veliki podaci mogu povezati s korisnim poslovnim uvidima za rast poslovanja.

Kako postati poslovnim analitičarom?

Poslovni analitičari djeluju kao poveznica između inženjera podataka i izvršnih direktora. Dakle, trebali bi razumjeti poslovne financije i poslovnu inteligenciju, kao i IT tehnologije poput modeliranja podataka, alata za vizualizaciju podataka i slično.

7.8.5. Statističar

Statističar, kao što ime sugerira, dobro razumije statističke teorije i organizaciju podataka. Ne samo da izvlače i nude vrijedne uvide iz grupa podataka, već također pomažu u stvaranju novih metodologija koje inženjeri mogu primijeniti.

Da biste postali statističar morate imati urođeni smisao za logiku. Također su dobri s raznim sustavima baza podataka kao što su SQL, rudarenje podataka i razne tehnologije strojnog učenja.

7.8.6. Arhitekt podataka

Arhitekt podataka stvara nacrte za upravljanje podacima kako bi se baze podataka mogle lako integrirati, centralizirati i zaštititi najboljim sigurnosnim mjerama. Oni također osiguravaju da inženjeri podataka imaju najbolje alate i sustave za rad.

Karijera u arhitekturi podataka zahtijeva stručnost u skladištenju podataka, modeliranju podataka, transformaciji ekstrakcije i posuđivanju itd. Također morate biti dobro upućeni u Hive, Pig i Spark.

7.8.7. Podatkovni znanstvenik

Znanstvenici koji se bave podacima moraju razumjeti izazove poslovanja i ponuditi najbolja rješenja pomoću analize i obrade podataka. Na primjer, od njih se očekuje da izvrše predviđajuću analizu i prođu kroz "nestrukturirane/neorganizirane" podatke kako bi ponudili korisne uvide. To također mogu učiniti identificiranjem trendova i obrazaca koji mogu pomoći tvrtkama u donošenju boljih odluka.

Da biste postali podatkovni znanstvenik, morate biti stručnjak za R, MatLab, SQL, Python i druge komplementarne tehnologije. Također može pomoći ako imate višu diplomu iz matematike ili računalnog inženjerstva.

7.8.8. Inženjer strojnog učenja

Inženjeri strojnog učenja danas su vrlo traženi. Međutim, profil posla nosi svoje izazove. Osim dubinskog znanja o nekim od najmoćnijih tehnologija kao što su SQL, REST API-ji itd., od inženjera strojnog učenja također se očekuje da provode A/B testiranje, izgrađuju podatkovne kanale i implementiraju uobičajene algoritme strojnog učenja kao što su klasifikacija i grupiranje.

Da bi se postalo inženjerom strojnog učenja, prvo morate dobro poznavati neke od tehnologija kao što su Java, Python, JS itd. Drugo, trebali biste dobro razumjeti statistiku i matematiku. Nakon što svladate oboje, puno je lakše proći razgovor za posao.

7.8.9. Administrator baza podataka

Profil posla administratora baze podataka prilično je jasan sam po sebi - oni su odgovorni za ispravno funkcioniranje svih baza podataka poduzeća i dodjeljuju ili opozivaju svoje usluge zaposlenicima poduzeća ovisno o njihovim zahtjevima. Oni su također odgovorni za sigurnosne kopije i oporavak baze podataka.

Neke od osnovnih vještina i talenata administratora baze podataka uključuju sigurnosno kopiranje i oporavak baze podataka, sigurnost podataka, modeliranje podataka i dizajn itd. Ako ste dobri u upravljanju katastrofama, to je svakako bonus.

7.9. Svijetla budućnost

Da biste razumjeli budućnost znanosti o podacima, prvo morate razumjeti njezinu povijest – a ako mislite da je znanost o podacima relativno nov izum, razmislite ponovno. Znanost o podacima ima svoje korijene u polju analize podataka, koju je akademik John Tukey prvi predstavio 1985. godine.

Kako navodi Anonymous(16.5.2022), do 1998. neki su akademici tražili da znanost o podacima postane vlastita disciplina. Pojavili su se i novi pojmovi koji opisuju kako se postupa s podacima. Sada već dobro poznata fraza 'data-mining' počela se koristiti u to vrijeme.

Odbor za podatke za znanost i tehnologiju službeno je pokrenuo Data Science Journal. Oko 2008. godine znanost o podacima prešla je iz uglavnom akademske djelatnosti u uobičajenu profesiju, s poslovima za podatkovne znanstvenike koji su se počeli pojavljivati i od tada eksponencijalno rasti.

Budućnost podatkovne znanosti je svijetla, a s povećanom primjenom u različitim domenama, njezini izgledi su golemi. Uz umjetnu inteligenciju i strojno učenje, podatkovna znanost pridonijet će višoj razini i inteligentnijem donošenju poslovnih odluka.

Prema navodima Reinsel, Gantz i Rydning (2017).globalna pohrana podataka će porasti sa 45 zetabajta na 175 zetabajta 2025. godine, tvrtkama će trebati više podatkovnih znanstvenika sa stručnošću za rukovanje ogromnom količinom i složenošću podataka s kojima rade.

Tvrtke sve više preferiraju rješenja umjetne inteligencije (AI) koja klijentima nude višestruke usluge, u biti sve na jednom mjestu. Iz tog razloga, end-to-end rješenja AI vjerojatno će nastaviti rasti u popularnosti. AI startupovi pomažu korisnicima očistiti velike skupove podataka i izgraditi modele učenja podataka, kao i automatizirati druge zadatke upravljanja podacima.

I da na kraju odgovorimo na pitanje, da li se isplati postati znanstvenik o podacima?

Ukratko, odgovor je da, znanost o podacima je vrlo dobra karijera s ogromnim mogućnostima napredovanja u budućnosti. Već sada je potražnja velika, plaće su konkurentne, a povlastice su brojne - zbog čega je znanstvenika o podacima nazvao "najperspektivnijom karijerom" od strane LinkedIna i "najboljim poslom u Americi" od strane Glassdoor-a, američke tvrtke koja anonimno recenzira tvrtke.

8. Zaključak

Sa razvojem tehnologije se povećava broj podataka te se nameće pitanje kako i gdje spremati te sve novonastale podatke. Znanost o podacima je nastala kao odgovor na ta pitanja. Teško je zapravo, još i danas, imati točnu predodžbu o definiciji znanosti o podacima, s obzirom na širok spektar znanosti koje znanost o podacima obuhvaća. Jedno što je svim definicijama slično je činjenica da je znanost o podacima skup vještina kojima znanstvenik o podacima mora znati baratati kako bi došao do nekog konačnog produkta filtriranih i korisnih informacija koje može dalje koristiti ili koje može dati nekome drugome na korištenje. Rad u ovom polju nije lagan, dapače zahtjeva znanja iz područja statistike, matematike, ekonomije, te informatike.

Ovu temu sam odabrao, kao što sam već u uvodu naveo, iz razloga što me iznimno zanima, i nadam se da ću po završetku svog obrazovanja imati priliku raditi u tom polju, naime prošle godine sam se prijavio na razgovor za posao isto glede znanosti o podacima u jednom poduzeću, ali nisam zadovoljio uvjete jer nisam bio dovoljno kvalificiran. Naravno se i to smatra uvjetom za prijavu za posao znanstvenika o podacima. Godine rada barem u Python-u ili u programskom jeziku R, te znanje i baratanje upitima sa bazama podataka, isto kao osnove iz područja matematike i statistike. Teško bi bilo za povjerovati da bi tvrtka primila nekog tko ima minimalno teorijsko iskustvo bez ikakve prakse u istom polju. Izrazito mi je drago što sam imao mogućnost pisanja o temi koja me zanima, jer je sama po sebi zanimljiva, te također glede moje daljnje karijere u IT industriji. Po mom samom istraživanju za vrijeme pisanja ovog završnog rada naučio sam mnogo stvari, ali sam također primijetio da je ovo područje iz kojeg se još mnogo toga može naučiti.

Popis literature

Knjige i znanstveni članci

1. Hand, D., Mannila, H., Smyth, P. (2001): Principles of data mining, The MIT Press, Cambridge, Massachusetts, London
2. Kornelije Rabuzin: Uvod u SQL (2011.) FOI Varaždin
3. Kornelije Rabuzin: SQL: napredne teme(2014.) FOI Varaždin
4. Data Scientist: The Definitive Guide to Becoming a Data Scientist, (2014) Zacharias Voulgaris, PhDTechnics Publication
5. Baragoin, C., Andersen, C.M., Bayerl, S., Bent, G., Lee, J., Schommer, C. (2001). Mining Your Own Business in Banking Using DB2 Intelligent Miner for Data. Dostupno na: http://www.redbooks.ibm.com/
6. 6. Krsto Kero, Dobša Jasminka, Bojanić-Glavica, Benedikt (2008.):Statistika (deskriptivna i inferencijalna) i vjerojatnost, Varaždin
7. Ivan Šošić, Vladimir Serdar: Uvod u statistiku. VII. izmijenjeno izdanje (1992.) Zagreb
8. 8. Krsto Kero, Dobša, Jasminka, Bojanić-Glavica, Benedikt (2003.): Statistika u primjerima, Varaždin
9. Mirko Maleković, Kornelije Rabuzin: Uvod u baze podataka (2016,), Fakultet ´ organizacije i informatike, Varaždin
10. Harrington, P. (2012). Machine learning in action. Shelter Island, NY: Manning Publications.
11. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996): From Data Mining to Knowledge Discovery in Databases, AI Magazine, Vol. 17, 3, 37–54.
12. Immanuel Kant, Kritika čistog uma, 1787., Njemačka
13. Biblija, Novi zavjet
14. Giuseppe Primiero, Information and Knowledge 2007., Belgija
15. O'Neil, C. i Schutt, R. Doing data science. 1st ed. Sebastopol: O'Reilly Media, Inc. (2013).
16. Hand, D. J. Principles of data mining. *Drug safety*, *30*(7), 621-622. (2007). SAD
17. The Collected Works of John W. Tukey: Graphics 1965-1985, Volume V, SAD

18. The Future of Data Science 16.5.2022. preuzeto 21.6.2022. s https://online.jcu.edu.au/blog/future-data-science
19. The Future of Data Science 16.5.2022. preuzeto 22.6.2022. s https://online.jcu.edu.au/blog/future-data-science
20. David Reinsel, John Gantz, John Rydning, Data Age 2025,The Evolution of Data to Life-Critical, Don't Focus on Big Data; Focus on the Data That's Big, Travanj 2017. , SAD
21. Čerić, V., Varga, M. i suradnici: Informacijska tehnologija u poslovanju, Element, Zagreb, 2004.
22. Alen Šimec, Davor Lozić, Nove tehnologije u primjeni, Zagreb, 2020.
23. Laudon K.C., Laudon J.P.: Management Information Systems, Prentice-Hall, USA, 2007.

Web stranice

1. Ryan LaBouve, (24. travnja 2018.) preuzeto 25.5.2022. sa *Twittera* - https://twitter.com/RyanLaBouve/status/988588702321475584
2. Bill Gates(26.6.2018.) preuzeto 21.7.2022. sa *Twittera* - https://twitter.com/billgates/status/1011752221376036864
3. Data Science 101 27.2.2022., 4.3.2022. preuzeto s https://builtin.com/data-science
4. Big Data, for better or worse: 90% of world's data generated over last two years 22.5.2013., preuzeto 4.3.2022 s https://www.sciencedaily.com/releases/2013/05/130522085217.htm
5. Interpreter Vs Compiler : Differences Between Interpreter and Compiler preuzeto 25.6. s https://www.programiz.com/article/difference-compiler-interpreter
6. Definicija statistike i deskriptivne statistike, 2021., preuzeto 22.5.2022. s http://www.enciklopedija.hr/Natuknica.aspx?ID=57896
7. The Future of Data Science 16.5.2022., preuzeto 21.7.2022. s https://online.jcu.edu.au/blog/future-data-science
8. Understanding Machine Learning: From Theory to Algorithms, preuzeto 24.6.2022. s http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning ,
9. Is Data Science a Good Career? Preuzeto 29.7.2022 s https://brainstation.io/career-guides/is-data-science-a-good-career
10. Što je podatkovni inženjer? - definicija iz tehopedije 2022., preuzeto 29.7.2022. s https://hr.theastrologypage.com/data-engineer#
11. OpenAI Five Defeats Dota 2 World Champions 15.4.2019., preuzeto 29.5.2022. s https://openai.com/blog/openai-five-defeats-dota-2-world-champions/#fn1 https://openai.com/five/
12. OpenAI Five Finals 26.3.2019.,preuzeto 15.5.2022. s https://openai.com/blog/openai-five-finals/
13. Dall-e-2, preuzeto 28.8.2022. s https://openai.com/dall-e-2/
14. Definicija vezana za programski jezik C#, preuzeto 6.9.2022. s https://www.techtarget.com/whatis/definition/C-Sharp
15. Definicija jezika Java, preuzeta 6.9.2022. s https://www.ibm.com/cloud/learn/java-explained
16. Slika popisa najkorištenijih programskih jezika 2020. godine, preuzeto 9.6.2022. s https://www.northeastern.edu/graduate/blog/most-popular-programming-languages/

17. Definicija pojma statistike, preuzeto 6.9.2022. s https://www.enciklopedija.hr/natuknica.aspx?ID=57896

Popis slika