Konzorcijum ADA projekta

Data Science

Osnovi i primene

Konzorcijum ADA projekta

DATA SCIENCE

Osnovi i primene

Beograd, 2020.

Konzorcijum ADA projekta

Data Science

Osnovi i primene

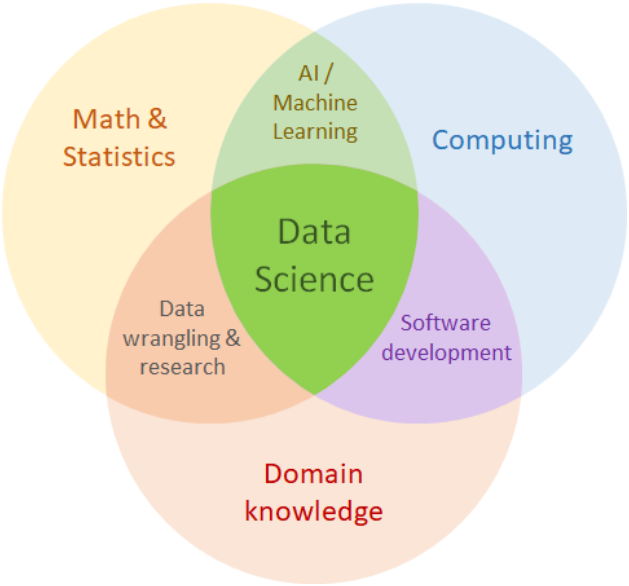
Beograd, 2020.

Ovaj priručnik je izdat kao rezultat projekta **Advanced Data Analytics in Business (ADA)** – EACEA 598829-EPP-1-2018-1-RS-EPPKA2-CBHE-JP, ko-finansiranog od strane Erasmus+ programa Evropske unije*.

Šta je nauka o podacima (Data Science, DS)?

Iako je u današnje vreme DS jedna od najčešće korišćenih reči u akademskim i poslovnim krugovima, kao i oblasti usluga, još uvek ne postoji tačna i široko prihvaćena definicija. Pored toga, kad god ljudi pričaju o DS-u, obično koriste i izraze kao što su analiza podataka, preuređivanje podataka, mašinsko učenje, veštačka inteligencija (AI), velike količine podataka (Big Data),... To su reči koje idu zajedno, ali nisu tačni sinonimi sa DS.

Ipak, mnogi ljudi koji imaju veliko znanje u oblasti DS će se složiti da se radi o stvaranju, pripremi, upravljanju, održavanju i korišćenju različitih skupova podataka (datasets) u cilju analize i pravljenja zaključaka o pojavama i procesima iz kojih se podaci prikupljaju i koje ti podaci predstavljaju. Praksa u oblasti DS zahteva znanje i veštine iz matematike i statistike, kao i znanje u domenima računarstva i aplikacija.



Studenti bi se mogli zapitati: Zašto bih želeo da studiram DS? Pa, zbog sjajnih mogućnosti zapošljavanja. Već postoji velika (i brzo rastuća) potražnja za stručnjacima iz oblasti analize podataka u kompanijama i drugim institucijama, a plate su visoke. Takođe, zbog vrlo obećavajućeg razvoja karijere - od atraktivnih startnih pozicija, preko brze promocije, do onih pozicija na kojima se donose odluke. I poslednje, ali ne i najmanje bitno, raditi u oblasti DS je jako zabavno! Jednostavno je uzbudljivo igrati sa podacima i dobiti neke korisne, do tada nepoznate i često iznenađujuće uvide.

Isto tako, kompanije bi se mogle zapitati: Zašto su nam potrebni stručnjaci iz oblasti nauke o podacima? Najkonkretniji odgovor izneo je časopis Wired koji podatke naziva "pogonskim gorivom digitalne ekonomije". Svuda, u svim organizacijama, bukvalno dolazi do poplave podacima, a za kontrolu i pretvaranje u poslovnu prednost potrebna je sasvim nova radna snaga. Potencijal podataka je ogroman, na svim nivoima, jer kompanije moraju da koriste podatke za vođenje i rast svakodnevnog poslovanja. Kažu da je vreme novac, ali isto se odnosi i na podatke.

Gotovo da ne postoji domen gde DS nije potreban: od biznisa do energetike, od obrazovanja do demografije, od zdravstvene zaštite do transporta, od finansija i osiguranja do maloprodaje, zabave i još mnogo toga. Na primer, naftni gigant Šel je uspeo da uštedi milione dolara godišnje u upravljanju zalihama zahvaljujući platformi zasnovanoj na prediktivnom modeliranju koja analizira brojne delove mašine za bušenje nafte koji mogu da ispadnu iz funkcije. A Netflix vam preporučuje film koji ćete gledati koristeći Big Data analizu za pretragu i pregledanje podataka poreklom od preko 100 miliona pretplatnika.

Dakle, ko su "naučnici za podatke" (data scientists) i šta oni zapravo rade? Pa, oni znaju kako da identifikuju probleme DS-a sa kojima kompanije moraju da se izbore ako žele da povećaju svoje kapacitete. Takođe koriste različite alate za prikupljanje podataka iz različitih izvora, čišćenje podataka, njihovu validaciju i transformisanje za

analizu. Zatim koriste svoje veštine u modeliranju i ekstrakciji podataka kako bi eventualno otkrili pravilnosti (patterne) u podacima i interpretirali ih radi rešenja nekog problema. Svoja otkrića saopštavaju zainteresovanim stranama koristeći različite alate za vizuelizaciju podataka.

Osnovne komponente DS-a

Matematičke osnove

Matematika je u temelju mnogih nauka pa je tako i u slučaju DS. Metode i tehnike moderne DS su veoma zavisne od matematike, posebno linearne algebre, diskretne matematike, diferencijalnog i integralnog računa (eng. calculus), verovatnoće i statistike. Pored modelovanja realnih problema, matematika takođe pomaže u razvijanju alata za njihovo rešavanje. Iako nije nužno biti matematičar za primenu alata koji rešavaju DS probleme, razumevanje onoga što se dešava u pozadini predstavlja veliku prednost.

Linearna algebra je oblast koja se može nazvati i Matematikom podataka. Ona se bavi operacijama nad tokovima podataka bez obzira na to šta je njihov izvor i interpretacija. Zajedničko za transformaciju slike iz jednog formata u drugi i davanje preporuke filma na prethodno pomenutom Netflix-u je da se obe mogu modelovati preko algebre nad matricama. Ona uključuje raznovrsne teme, poput osnovnih svojstava matrica i vektora – množenje skalarom, linearna transformacija, unutrašnji i spoljašnji proizvodi, množenje matrica i drugi algoritmi, specijalne matrice, vektorski prostori, sopstvene (karakteristične) vrednosti, sopstveni vektori, kao i mnoge druge.

Linearna algebra nije jednostavna i pritom je neophodna za DS, ali – pogodite šta! – ona može da bude i jako zabavna. Primena linearne algebre nad realnim podacima uz upotrebu programa je zanimljiva. A tamo gde linearna algebra ne može da reši problem, *numeričke metode* mogu da pomognu. Numeričke metode su takođe izuzetno polje matematike koje je savršeno usklađeno sa potrebama

računarstva s obzirom da mogu da daju aproksimacije i “dovoljno dobra” rešenja za veliki broj problema.

Moderna Nauka o podacima koristi pomoć računarskih sistema, koji su zasnovani na *diskretnoj matematici* i *diskretnim strukturama podataka*. Neke od ključnih tema u diskretnim strukturama podataka su stekovi, redovi, grafovi, nizovi, heš tabele, stabla, i pojmovi diskretne matematike uopšteno – skupovi, grafovi, rekurentne relacije i jednačine, asimptotsko ponašanje i procena (vremenske, prostorne) složenosti u najgorem slučaju pomoću takozvane O notacije.

Kada su podaci neprekidni (kontinualni, nediskretni), na primer rast deteta, promena brzine auta i slično, *diferencijalni i integralni račun* postaje ključan. U ovoj oblasti su bitni koncepti ograničenosti, neprekidnosti, diferencijabilnosti, izvoda, ubrzanja, brzine, integrala, nagiba, reda... Diferencijalni i integralni račun ima široke primene gde god nastaju podaci. U ekonomiji se ona koristi za računanje stope promene cena. U biologiji se može koristiti za računanje stopa nataliteta i mortaliteta. U fizici se koristi u različitim kontekstima, npr. kod kretanja tela, elektriciteta, itd., dok se u hemiji može koristiti za predviđanje funkcija poput brzine reakcije. Dakle, koncepti diferencijalnog i integralnog računa i njegove primene su prisutne na mnogo raznovrsnih mesta u okviru DS i mašinskog učenja. Oni koji se bave DS-om koriste diferencijalni i integralni račun za implementaciju različitih algoritama, na primer, za logističku regresiju.

Verovatnoća je mera izvesnosti da će se neki događaj desiti. Pošto ponekad ume da bude kontraintuitivna važno je da DS stručnjaci razumeju koncepte verovatnoće. Ilustrujmo ovo kroz nekoliko primera.

- Ako bacite dobro balansiranu kocku 6 puta, da li je verovatnoća da će se desiti sekvenca brojeva 1 5 2 4 6 3 - viša / niža / ista kao verovatnoća da će se pojaviti sekvenca brojeva 6 6 6 6 6 6?
- Ako izvlačimo broj iz skupa prirodnih brojeva N, da li je verovatnoća izvlačenja bilo kog broja – ista? Da li bi suma verovatnoća za sve njih bila 1?
- Pogađamo iza koje od tri zavese se krije Maserati (iza preostale dve su magarci). Ako pre nego što nam se saopšti

da li smo u pravu ili nismo, promenimo našu odluku jer nam se otkrije jedna od zavesa iza koje je magarac, da li time povećavamo ili smanjujemo šansu da pogodimo gde je Maserati?

Standardne teme u teoriji verovatnoće uključuju koncepte verovatnoće, uslovne verovatnoće, slučajne promenljive, Bajesovog pravila, raspodele verovatnoća, normalne raspodele, itd.

DS često koristi *statističko zaključivanje* u cilju predviđanja ili analize tendencija u podacima. Sa druge strane, statističko zaključivanje skoro uvek koristi raspodele verovatnoća nad podacima. Stoga su verovatnoća i njene primene jako značajne u rešavanju DS problema.

Statistika je neizbežna kada je DS u pitanju. Mnogi smatraju da je i mašinsko učenje u opštem smislu zapravo statističko učenje. Iako statistika "svašta otkriva a ništa ne pokazuje", ona ipak pomaže u odlučivanju da li je na primer potrebno primeniti dijetu ili trening u cilju smanjivanja težine, da li će nikotinski flasteri pomoći u prestanku pušenja, da li pušenje tokom trudnoće utiče na IQ deteta i slično.

Za ovladavanje statistikom bitni su koncepti osnovne verovatnoće, uslovne verovatnoće, funkcija raspodela verovatnoća, sumarizacije podataka i deskriptivne statistike, korelacije, testiranja hipoteza, intervala poverenja, p-vrednosti, Analize varijanse (ANOVA), t-testa, linearne regresije.

Današnji DS algoritmi se često oslanjaju na ogromne skupove podataka. U cilju savladavanja tako velikih količina podataka, potrebno je ili prilagoditi algoritme ili sami podaci moraju biti transformisani. Dva standardna pristupa za smanjivanje veličine podataka su *vertikalna i horizontalna redukcija*. Kada su podaci predstavljeni tabelom, kolone odgovaraju atributima podataka dok redovi predstavljaju pojedinačne podatke – vektore atributa. Stoga, vertikalna redukcija (ili dimenziona redukcija) podrazumeva smanjivanje skupa atributa tako da informativnost originalnog skupa podataka nije previše narušena. Polazni skup atributa se može modifikovati na različite načine: atributi se mogu ukloniti ili ponderisati, transformisati pomoću diskretizacije ili agregacije, kombinovati u cilju

formiranja novih atributa i slično. Dobro poznata tehnika dimenzione redukcije je Analiza glavnih komponenti (eng. Principal Component Analysis – PCA). PCA transformiše originalni skup podataka u novi skup podataka manje dimenzije.

Redukcija instanci (redova) je horizontalna, što znači da se originalni skup podataka smanjuje po redovima – eliminacijom pojedinačnih podataka. Jedna od implementacija se zove *uzorkovanje* (*sampling*), tj. odabir podskupa redova koji se ponašaju slično kao ceo skup imajući u vidu određene statističke aspekte. Redukcija instanci se može izvršiti i primenom klaster analize, odnosno formiranja grupa srodnih redova, nakon čega se zadržavaju samo predstavnici grupa (centroidi) koji oslikavaju ceo originalni skup podataka.

Klasifikacija je jedan od najčešće izučavanih problema u DS. Klasifikacione probleme često srećemo i u životu. Na primer, kada odete do banke i zatražite kredit, na vaš zahtev se može primeniti neka klasifikaciona metodologija. Ona može biti realizovana od strane eksperta ili algoritma. Ako je vaša prethodna istorija otplaćivanja kredita čista, ako ste u braku, imate decu, normalne prilive/odlive novca itd., onda postoji veća šansa da ćete vratiti i taj novi kredit koji zahtevate. Ovaj zadatak se može modelovati kao binarni klasifikacioni problem što znači da će izlaz iz klasifikacionog algoritma biti „da, prihvaćen zahtev“ ili „ne, odbijen“. Klasifikacioni algoritmi se često naslanjaju na prethodne, istorijske podatke, a neki od popularnijih su Metoda podržavajućih vektora (eng. Support Vector Mahines – SVM), k Najbližih suseda (k-Nearest-Neighbors – kNN), Veštačke neuronske mreže (eng. Artificial neural networks – ANN), itd.

Regresione metode su slične klasifikacionim, osim što kod njih izlaz koji se predviđa ne mora biti diskretan i/ili konačan. Recimo da umesto predviđanja da li će klijent vratiti kredit ili ne, želite da predvidite koja je šansa da će klijent vratiti kredit, stoga, bilo koja realna vrednost iz intervala [0, 100%] je sada moguća. Regresione metode se mogu koristiti u kontekstu vremenskih serija takođe, na primer za predviđanje atmosferskih uslova poput vazdšnog pritiska i temperature, cena akcija na berzi, itd.

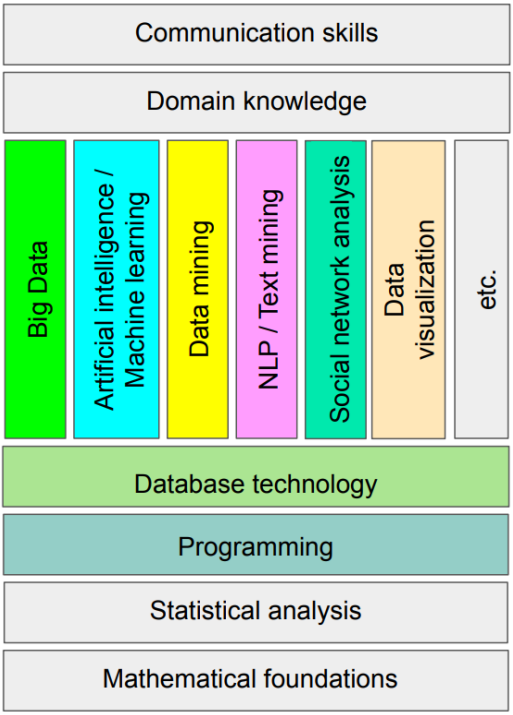
Klaster analiza se tiče pronalaženja grupa podataka koji se mogu smatrati sličnim na neki način. U skladu sa prethodnim primerom, zadatak bi mogao da bude primena klaster analize nad skupom svih klijenata banke kako bi se utvrdile grupe (klasteri) koje unutar sebe imaju slične klijente, dok se klijenti iz različitih grupa jasno razlikuju. Ovaj uvid kasnije može pomoći bankama da bolje razumeju potrebe klijenata u cilju nuđenja adekvatnih tipova kredita ili usluga unutar različitih grupa klijenata.

Optimizacija parametara DS algoritama. Uspešna primena DS algoritama se često oslanja na dobar odabir ulaznih parametara. Na primer, SVM obično zavisi od dva parametra, takozvane regularizacione konstante C i jezgarnog hiperparametra γ. U cilju ostvarivanja visokokvalitetnih klasifikacionih rezultata, SVM mora biti izvršen sa dobrom kombinacijom ovih parametara. S obzirom da oba parametra pripadaju realnom domenu, prostor mogućnosti je teško istražiti. Najjednostavniji pristup, često korišćen kao donja granica prihvatljivog kvaliteta, zove se *Pretraga rešetke* (eng. *Grid search*). U pitanju je optimizacioni algoritam u kojem se dobra kombinacija parametara traži sistematskim proveravanjem svih njihovih mogućih kombinacija sa određenim stepenom preciznosti. Kada je preciznost visoka, broj mogućnosti se takođe drastično povećava. Drugi, efikasniji pristup bi bio da se prati gradijent (nagib) tokom pretrage. Međutim, ovaj pristup često zna da dovede do „zaglavljivanja" u takozvanim lokalnim optimumima. Neki obećavajući pristupi obično izbegavaju iscrpnu pretragu i koriste delimičnu upotrebu generatora slučajnih brojeva. Ekstremni pristup u ovoj grupi je *Slučajna pretraga* (eng. *Random search*). Najbolji rezultati se ipak najčešće dobijaju pristupima koji su između iscrpne i slučajne pretrage, tj. aproksimativnim algoritmima poput metaheuristika (Genetski algoritam, Metoda promenljivih okolina, i drugi) ili verovatnosnim algoritmima poput Bajesove optimizacije.

Tehnološke osnove

Oni koji se bave DS-om često koriste izraz stek DS tehnologija. Intuitivno, to uključuje širok spektar alata, tehnologija i veština koje

pomažu ljudima da vode DS projekat. Iako su neki od ovih alata i tehnologija alternative jedni drugima, potrebno je imati širu sliku o steku DS tehnologija. Ona pruža širi tehnološki kontekst kada su u pitanju zadaci i procesi poput prikupljanja podataka, skladištenja, čišćenja, transformacije, istraživanja, analize i prezentacije, kao i integrisanja tih procesa i njihovih rezultata u koherentne DS aplikacije.



Sav DS polazi od *podataka*. Sami podaci potiču iz raznih izvora. To mogu biti Web stranice, aplikacije, mobilni uređaji i aplikacije, različiti servisi, senzorski sistemi, društveni mediji, velike količine teksta, velike kolekcije audio i video snimaka itd. Neki podaci su delimično obrađeni i delimično strukturirani, a drugi mogu biti u velikoj meri nestrukturirani. Neke podatke treba ipreuzeti sa Interneta, a neki mogu da dođu iz sistema Interneta inteligentnih uređaja (Internet of Things, IoT).

Odakle god da podaci dolaze, oni se na kraju skladište u baze podataka i/ili skupove podataka (datasets). I baze podataka i skupovi podataka sadrže podatke, ali postoji razlika među njima. *Baza podataka* je organizovana kolekcija podataka, koju obično kontroliše i kojoj se pristupa pomoću sistema za upravljanje bazama podataka. To je softver koji upravlja skladištenjem, pronalaženjem i ažuriranjem podataka, kao i višekorisničkim pristupom bazi podataka. Podaci u bazi podataka mogu da se čuvaju kao skup tabela koje obično upućuju jedna na drugu, ili mogu da se čuvaju kao dokumenti (u različitim formatima), grafikoni, parovi ključ-vrednost, vektori i tako dalje. Postoji mnoštvo tehnika koje omogućavaju brzi pristup podacima u velikim bazama podataka.

Dataset je samo skup podataka, obično organizovanih kao tabela, tj. raspoređenih u redove i kolone za obradu statističkim softverom. Podaci u skupu podataka su možda došli iz baze podataka, a možda i nisu. U DS-u, podaci koji se analiziraju često dolaze iz skupova podataka. Redovi u skupovima podataka često se nazivaju opažanjima/opservacijama ili slučajevima. Kolone se obično nazivaju varijablama ili obeležjima (features). Kažemo da podaci u skupu podataka predstavljaju različita opažanja/slučajeve neke pojave, a svaki slučaj je opisan istim skupom varijabli/obeležja.

Ali šta ako ti skupovi podataka postanu preveliki ili previše složeni da bi se obrađivali i predstavljali tradicionalnim aplikativnim softverom za obradu podataka poput aplikacija za rad sa tabelama, paketa za vizuelizaciju ili široko korišćenih sistema za upravljanje bazama podataka? Tada govorimo o *velikim količinama podataka* (*Big Data*). U svetu Big Data, količina podataka je ogromna (ne govorimo o terabajtima, već petabajtima ili egzabajtima), vrste podataka su vrlo raznolike, brzina promene podataka (brzina kojom se podaci menjaju i kojom kojom treba da budu obrađeni) mnogo je veća nego kod tradicionalnih izvora podataka, a verodostojnost podataka (količina "šuma" u podacima) je kritična.

Međutim, treba imati na umu da velike količine podataka moraju na kraju krajeva da imaju određenu vrednost za poslovanje i krajnje korisnike – ako ne možemo da pretvorimo ogromne količine podataka

u vrednost, onda su oni beskorisni. U tome je presudno razumevanje kako troškova prikupljanja i održavanja, tako i koristi od velikih količina podataka.

Očigledno je da Big Data postavlja niz izazova kako korisnicima, tako i analitičarima u pogledu prikupljanja, skladištenja, deljenja, postavljanja upita, rukovanja privatnošću i slično. Često nije dovoljno imati samo specijalizovani softver za rad sa velikim podacima, već i specijalizovani hardver i/ili stotine ili hiljade servera.

Neophodna komponenta opreme svakog ko se bavi DS-om jesu alati (i veštine!) za *vizuelizaciju podataka*. Veliki ili ne tako veliki, podaci se često moraju vizuelizovati da bi imalo smisla da se koriste. Ne radi se samo o tome da slika vredi kao hiljadu reči, niti se radi o predstavljanju podataka u privlačnijem obliku nego što su to skupovi podataka / tabele. Prvenstveno se radi o tome da se krajnjim korisnicima omogući da shvate teške pojmove ili uoče nove obrasce, trendove i do tada nepoznate klastere u podacima. Takođe se radi o brzom razumevanju velikih količina podataka i efektivnom prenošenju značenja podataka. I poslednje, mada ne i najmanje bitno, radi se o eksperimentisanju sa različitim scenarijima budućih promena podataka.

Tehnologija vizuelizacije podataka danas je napredna i sofisticirana. Ona omogućuje istraživanje odgovarajućih vizuelnih prikaza podataka sa što više detalja, interakciju sa grafovima i grafikonima, promenu pogleda i fokusa i posmatranje animacija dinamike podataka. Alati za vizuelizaciju podataka omogućavaju kreativno i maštovito istraživanje podataka da bi se stekle nove predstave. 1D (linearne), 2D (planarne), 3D (trodimenzione), nD (višedimenzione), vremenske, stablolike/hijerarhijske i mrežne vizuelizacije samo su široke kategorije tipova vizuelizacije podataka; svaka kategorija uključuje veći broj specifičnih tipova vizuelizacije. Izbor odgovarajuće vizuelizacije za raspoložive podatke je vrlo intrigantan zadatak, može u velikoj meri da ubrza proces uočavanja zanimljivih obrazaca u moru podataka i, svakako, estetski je privlačan.

Veštačka inteligencija (*Atificial Intelligence*, *AI*) je visoko interdisciplinarna oblast nauke i prakse usredsređena na razvoj

sistema koji mogu da obavljaju zadatke koji obično zahtevaju ljudsku inteligenciju – predstavljanje znanja, zaključivanje o procesima i pojavama, obrada prirodnog jezika, prevođenje s jednog jezika na drugi, opažanje, obrada slika, prepoznavanje govora, rešavanje problema, odlučivanje, i tako dalje.

U poslednje vreme najpopularnije polje AI je *mašinsko učenje* (*Machine Learning*, *ML*). Wikipedia ga definiše kao "naučno proučavanje algoritama i statističkih modela koje računarski sistemi koriste da bi efektivno izvršili određeni zadatak bez upotrebe eksplicitnih uputstava, oslanjajući se, umesto toga, na patterne i zaključivanje". Tipični zadaci ML uključuju klasifikaciju (izgradnja modela koji nam sa većom ili manjom sigurnošću može reći kojoj klasi iz unapred određenog skupa klasa pripada opažanje – podatak), klasterovanje (identifikovanje više ili manje koherentnih grupa – klastera – podataka iz skupa svih podataka), regresija (predviđanje ishoda događaja ili vrednosti izlazne varijable na osnovu odnosa između ostalih varijabli u skupu podataka) i tako dalje. Praktične ML aplikacije grade model na osnovu uzorka podataka iz skupa podataka, poznatog kao "podaci za obučavanje", verifikuju tačnost modela na osnovu drugog uzorka podataka ("podaci za testiranje"), a zatim ga koriste u cilju predviđanja ili odlučivanja o novim ili ranije nepoznatim podacima. ML je najrelevantnije polje AI za DS.

Neuronske mreže (*Neural Networks*, *NN*) su među najpopularnijim ML tehnikama. Donekle se modeliraju prema ljudskom (životinjskom) mozgu i nervnom sistemu. Kada se neuronskoj mreži prikaže dovoljan broj primera ("podaci za obučavanje"), ona može da nauči kako da prepozna obrasce i zatim izvrši zadatke kao što su identifikacija, klasifikacija i klasteriovanje. Na primer, NN može da se obuči za prepoznavanje otisaka prstiju, rukopisnih znakova, predmeta (čak i onih u pokretu), ljudskih lica i još mnogo toga.

Tehnički, NN je organizovana kao veći broj malih procesnih jedinica (neurona) međusobno povezanih i grupisanih u slojeve. Više slojeva između ulaznog i izlaznog sloja može da omogući *duboko učenje*, pri čemu svaki sloj transformiše svoj ulaz (koji dolazi iz prethodnog sloja) u malo apstraktniji i složeniji prikaz. Na primer, NN za duboko učenje

obučena da prepozna ljudsku ruku može da ima jedan sloj koji od piksela apstrahuje ivice, zatim sledeći sloj za prepoznavanje različitih položaja ivica u odnosu na delove ruke, praćen slojem koji prepoznaje zglobove, zatim prste, dlan i na kraju ruku. Mreže dubokog učenja mogu zahtevati veliku količinu podataka za obučavanje i moćne računare, ali omogućavaju razvoj uzbudljivih aplikacija, kao što su automobili sa automatskim vozačem, automatsko prevođenje teksta s jednog na drugi jezik, automatsko generisanje teksta i rukopisa, otkrivanje raka dojke, predviđanje zemljotresa i mnogo, mnogo više.

Koncept vrlo sličan ML je *istraživanje podataka* (*data mining*). Alati istraživanja podataka tragaju za značenjem u velikim količinama podataka, pronalazeći patterne u naizgled nepovezanim podacima i sastavljajući ih tako da omoguće interpretaciju. To obično zahteva mnogo ljudske interakcije, dok ML sistemi, jednom obučeni, mogu da rade samostalno. Istraživanje podataka može se smatrati "hranom" za ML, jer ML često koristi skupove podataka formirane iz podataka dobijenih tehnikama istraživanja podataka.

Istraživanje teksta / *Istraživanje tekstualnih podataka* / *Analitika teksta* (*Text mining*) je proces ekstrakcije znanja iz nestrukturiranog teksta. To obično uključuje prvo strukturiranje ulaznog teksta, otkrivanje patterna u strukturiranim tekstualnim podacima i evaluaciju i interpretaciju rezultata. Intrigantne aplikacije i podoblasti istraživanja teksta uključuju kategorizaciju teksta, klasterovanje teksta, modeliranje tema, ekstrakciju koncepata/entiteta, ekstrakciju taksonomije, analizu osećanja, pravljenje rezimea dokumenata i tako dalje.

Istraživanje teksta preklapa se sa *obradom prirodnog jezika* (*natural language processing*, *NLP*), koja se fokusira na analizu velike količine prirodno-jezičkih podataka, prepoznavanje govora, razumevanje prirodnog jezika, prevođenje prirodnog jezika i generisanje teksta na prirodnom jeziku.

Analiza društvenih mreža (*Social network analysis*, *SNA*) koristi teoriju mreža/grafova i vizuelizaciju podataka za istraživanje društvenih struktura, interakcija i obrazaca. Čvorovi u vizuelizovanim strukturama

društvenih mreža (sociogrami) su ljudi (ili bilo koje druge vrste aktera ili stvari koje interaguju), a veze predstavljaju interakcije ili odnose među njima. SNA dobija sve veću popularnost u DS-u. Koristi se za analizu mreža društvenih medija, mreža prijateljstava i poznanstava, poslovnih mreža, prenosa bolesti, ljubavnih veza i tako dalje. Iako se koreni SNA nalaze u savremenoj sociologiji, ona se sve više koristi u drugim društvenim naukama (antropologiji, demografiji, studijama komunikacije, ekonomiji, geografiji, istoriji, organizacionim naukama, političkim naukama, socijalnoj psihologiji, razvojnim studijama, sociolingvistici), kao i u biologiji, informatici i računarskim naukama.

Programiranje i *softversko inženjerstvo* su u osnovi svih ovih tehnologija na steku DS tehnologija. Čak i ako cilj DS projekta nije razviti aplikaciju, neki programi su uvek uključeni u analizu podataka. Tako korišćenje programskog jezika poput Python-a ili R uvek predstavlja deo DS projekta. Oni koji se bave DS-om koriste i različite programske alate i biblioteke koje dolaze s ovim jezicima kako bi obavili čišćenje podataka, transformaciju, analizu i vizuelizaciju.

Postoje neka tipična pitanja u vezi sa programiranjem u DS-u: Da li bi onaj ko se bavi DS-om trebalo više da se bavi analizom, interpretacijom, matematikom, nego programiranjem? Šta je lose u korišćenju alata zasnovanog na GUI? Pa, postoje neke veštine koje nisu tako očigledne a stiču se zajedno sa programerskim veštinama i od suštinskog su značaja za DS. Alati zasnovani na GUI-u su korisni do određene tačke. Međutim, na kraju svi modeli zahtevaju neko podešavanje, sve analize zahtevaju određenu ponovljivost, pa samo stvarno programiranje omogućava pravo prilagođavanje modela. Slično tome, programiranje je od suštinske važnosti za razvoj računarskog mišljenja, što je u osnovi gotovo čitavog DS-a. Razumevanje načina na koji različiti algoritmi funkcionišu i koje su brojne opcije u prilagođavanju modela nemoguće je bez praktičnog iskustva u programiranju.

Primene

Ne postoji praktično nijedna oblast gde DS ne može da se upotrebi, dokle god imamo odgovarajuću bazu podataka. Ono što sledi je samo mali sažetak primena DS.

Primene u biznisu

Ako su podaci novo pogonsko gorivo, tada je ovladavanje procesom njihovog pretvaranja u poslovne odluke ključ za oslobađanje njihovog potencijala. DS omogućava ispitivanje velike količine podataka radi otkrivanja skrivenih pravilnosti, korelacija i takođe daje odgovarajući uvid radi donošenja ispravnih poslovnih odluka.

Generalno, većina preduzeća ima nekoliko ciljeva prilikom usvajanja DS projekata. Iako je glavni cilj većine njih poboljšanje zadovoljstva kupaca, postoje i drugi ciljevi koji uključuju smanjenje troškova, bolje ciljani marketing i unapređenje postojećih proizvoda i poslovnih procesa.

Tehnike DS kao deo proizvoda / usluge

Kao što je ranije pomenuto, Netflix koristi analizu podataka za ciljano oglašavanje, kao i za slanje predloga za sledeće filmove koje bi pretplatnik trebalo da pogleda. Netflix to radi koristeći prethodne podatke o pretraživanju i gledanju korisnika i na taj način može davati predloge za sličan sadržaj. Kompanija takođe koristi rezultate analize podataka radi donošenja odluka o tome šta treba da kupuju, licenciraju i koji novi sadržaj da stvore kako bi udovoljili potrebama svojih pretplatnika.

Spotify, jedan od najvećih digitalnih muzičkih servisa, takođe koristi analizu podataka za kreiranje personalizovanih play lista za svakog pretplatnika. Sugestije koje daju ove liste su vrlo dobre, a evo i zašto. Analitičari podataka Spotify analiziraju sve izvođače, njihove stilove i kategorišu ih po glasnosti, mogućnosti za ples, energiji i još mnogo toga. Zatim se svi izvođači grupišu u klastere na osnovu ovih karakteristika i spremni su za pokretanje preporuka. Mehanizam

preporuka uzima u obzir omiljene umetnike pretplatnika i daje dodatne preporuke u pogledu umetnika koji se nalaze u istom klasteru.

Rolls Royce, drugi najveći svetski proizvođač sistema za pogon aviona i morskih brodova, ulaže mnogo u izgradnju Interneta inteligentnih uređaja i tehnologija za analizu podataka, kako na svojim objektima, tako i na svim vozilima/brodovima/avionima koje prave kako bi pružili vrhunske usluge. Oni su ugradili tehnologije za analizu podataka u svoj glavni proizvod – motore aviona koje koristi više od 500 civilnih i 150 vojnih kompanija/država u svetu. Njihovim motorima upravlja Engine Health System, sistem koji koristi do 100 parametara iz podataka sačuvanih u opisima letova i koji se analiziraju se nakon svakog leta. Plan za sledeću generaciju motora je da ovaj sistem nadgleda više od 5000 parametara i da bude povezan sa Cloud okrženjem, kako bi se problemi uočili i sprečili pre nego što se pojave.

Kompanija Buxtonco koristi analizu podataka da bi pomogla drugima da odaberu dobru lokaciju za maloprodaju. Ovo je posebno važno za restorane i prodavnice. Buxtonco to radi tako što traži gde kupci provode svoje vreme i šta rade na određenim lokacijama – nešto vrlo slično geofencing-u. Na taj način oni mogu odrediti mesto gde bi bilo najbolje otvoriti sledeću prodavnicu.

Gramener, kompanija za vizuelizaciju podataka i prediktivnu analitiku, implementirala je Web-bazirani AI program za Nisqually River Foundation (Vašington), organizaciju za zaštitu prirode. Ovaj program meri i prati vrste riba prisutnih u reci Niskuali tako što automatski analizira podatake iz video kamere i infracrvenih senzora u vodi.

Cargill je razvio mobilnu aplikaciju koja pomaže proizvođačima škampa da smanje gubitak svojih prinosa zbog bolesti. Aplikacija koristi analizu podataka za predviđanje biomase u ribnjacima (na osnovu temperature, pH vrednosti i količine hrane) i radi zajedno sa automatizovanim sistemom hranjenja škampa kako bi se postigli optimalni rezultati. Podaci se prenose iz aplikacije u Cloud, a zatim radnici mogu pristupiti sistemu uživo i videti performanse ribnjaka, što omogućava preduzimanje odgovarajućih mera i prediktivnu analitiku

koja im pomaže da bolje upravljaju zdravljem škampa i povećavaju prinose.

DS može poboljšati zadovoljstvo korisnika

U 2015. godini, Coca-Cola je počela da koristi analizu podataka za svoj program lojalnosti u okviru digitalnog upravljanja. To je omogućilo kompaniji da kreira i distribuira različite reklamne sadržaje za različite vrste publike: ljubitelje sporta, ljubitelje muzike itd. Kao posledica, došlo je do povećanja broja novih i do zadržavanja starih potrošača.

Optimizacija HR-a pomoću DS

Xerox je međunarodna kompanija za proizvodnju kancelarijskih uređaja i pružanje pratećih usluga. Godinama se kompanija borila sa visokim stopama osipanja zaposlenih u centrima za podršku, jer su ljudi napuštali posao bez obzira na napore u odeljenjima za ljudske resurse, bez obzira na dodatne pogodnosti, beneficije ili bonuse koje je kompanija uvela. Iskusni profesionalci, visoko obučeni, koji rade za pristojnu svotu novca, napustili su posao. Nakon primene analitike podataka nad podacima o zaposlenima, kompanija je otkrila da prethodno iskustvo kao operatora pozivnog centra uopšte nije bilo važno. Otkriveno je da su osobine zaposlenih ključni faktor. Imati članove tima sa duhom partnerstva i međusobnog poštovanja pokazalo se mnogo efikasnijim od "superstar-a" u timu, ili od organizovanja takmičenja u timu, ili nuđenja bonusa za dopunski rad. Te informacije dovele su do toga da je Xerox reorganizovao pristup tokom zapošljavanja i smanjio odliv osoblja za 20%, čime je kompanija dugoročno uštedela milione dolara.

DS u drugim domenima

PepsiCo, kao velika multinacionalna kompanija, oslanja se na efikasno upravljanje lancem snabdevanja. Unapredila je proces upravljanja lancem snabdevanja uz pomoć analitike podataka tako što je iskoristila izveštaje o prodaji i zalihama da bi predvidela potrebe za proizvodnjom i isporukom. Na ovaj način, kompanija obezbeđuje da trgovci imaju prave proizvode, u pravim količinama i u pravo vreme.

UOB banka iz Singapura koristi velike podatke da bi poboljšala upravljanje rizikom, tj. da bi smanjila vreme izračunavanja rizika. Ceo proces je trajao do 18 sati, ali sveden je na nekoliko minuta sa potencijalom da se izvrši analiza rizika u realnom vremenu.

Transport za London (TfL) je kompanija koja upravlja milionima autobusa, taksi vozila, vozova podzemne železnice, trajekata i semafora u Londonu. Oni opslužuju gotovo 10 miliona stanovnika Londona i svakodnevno prikupljaju ogromne količine podataka. Ovi podaci omogućavaju kompaniji da razume kako i kada putnici putuju, dužine ruta koje se koriste, koji su putevi i mostovi opterećeniji i koje su rute javnog prevoza suočene sa većim prometom. Svakodnevna analiza takvih podataka omogućava raspoređivanje više autobusa na više opterećenih ruta kako bi se smanjilo vreme provedeno na putu i poboljšao protok saobraćaja, prilagođavanje semafora na opterećenijim putevima kako bi im se povećao propusni kapacitet i izbegli zastoji u saobraćaju, kao i mnoge druge aktivnosti.

Primene u medicini i biološkim naukama

Analiza podataka u biološkim naukama pretrpela je revolucionarne promene u poslednjih 10 godina, a ove promene su najvidljivije u medicini i zdravstvu. Uvođenje novih i inovativnih sistema upravljanja informacijama dovelo je do značajnih promena u funkcionisanju savremenih zdravstvenih ustanova kao što su klinički centri i velike bolnice. Nove strategije i pristupi u analizi podataka takođe su omogućili veliki napredak u naučnoistraživačkom radu u biologiji i medicini.

Istraživanja i u biologiji i u medicini gotovo su nemoguća bez kompetentnog stručnjaka za analizu podataka. Često, nakon eksperimenata i kliničkih ispitivanja, postoji obilje sirovih podataka koje je potrebno obraditi i strukturirati. Doktorima medicine i biolozima često nedostaje profesionalno znanje iz statistike, tako da je svaka pomoć stručnjaka za analizu podataka dragocena. Statistički testovi kao što su t-test, hi-kvadrat, ANOVA, Pirsonova i Spirmanova korelacija su osnovni alati i u fundamentalnim i kliničkim medicinskim istraživanjima.

Najočigledniji primer vrednosti analize podataka u medicinskim istraživanjima ogleda se kroz uvođenje novih dijagnostičkih testova u oblastima kao što su interna medicina, neurologija, onkologija ili radiologija. Pre nego što se test može primeniti u kliničkoj praksi, treba utvrditi njegovu senzitivnost (sposobnost identifikacije osoba sa bolešću), specifičnost (sposobnost identifikacije osoba koje nemaju bolest) i razne druge karakteristike. Svaki pacijent je jedinstven, a iako test može dati izvanredne rezultate kod određene grupe pacijenata, u drugim okolnostima njegov učinak može biti loš. Prilikom evaluacije budućeg dijagnostičkog testa, istraživači se moraju baviti složenim i heterogenim podacima dobijenim pod različitim uslovima.

Veliki podaci u biomedicini, posebno oni koji se odnose na genomiku, epigenomiku i proteomiku, danas su sve popularniji među lekarima i istraživačima. Sa razvojem novih i inovativnih informacionih tehnologija postalo je moguće adekvatno organizovati, klasifikovati i transformisati te podatke i izvući informacije korisne za buduće biomedicinske aplikacije. Danas postoje brojne slobodno dostupne baze podataka sa ogromnim količinama potencijalno vrednih informacija koje čekaju da budu otkrivene i primenjene u kliničkoj praksi. Jedini razlog za odlaganje je nedostatak kompetentnih stručnjaka i naučnika za analizu podataka koji imaju odgovarajuću ekspertizu iz oblasti bioinformatike!

Projekat *Ljudski genom* koji je završen 2003. godine, predstavljao je jedan od najvećih svetskih naučnih napora namenjenih mapiranju svih ljudskih gena i proceni njihovih strukturnih i funkcionalnih karakteristika. Podaci o genskim sekvencama se sada čuvaju u mnogim bazama podataka, kao što su Nacionalni centar za biotehnološke informacije u SAD, Ensembl (European Bioinformatics Institute and the Wellcome Trust Sanger Institute database) i drugi. Tumačenje podataka o genomu je vrlo teško bez složenog softvera za analizu podataka i kvalifikovanog IT stručnjaka. Količina potencijalno korisnih informacija koje se mogu dobiti iz tih baza podataka je zaista ogromna, kao i potencijalna primena u budućem razvoju novih lekova i terapijskih strategija za mnoge bolesti. S druge strane, zbog nedostatka ljudskih, finansijskih i drugih resursa koji se odnose na

analizu podataka, stepen korišćenja ovih baza je relativno nizak. Ovde je prilika da budući DS stručnjak doprinese integrišući znanja iz područja informatike, biologije i medicine, kako bi pomogao budućem lečenju raznih bolesti koje se danas smatraju neizlečivim. A to je glavni cilj današnje medicine.

U velikim zdravstvenim ustanovama i sistemima, lekari su često suočeni sa ogromnom količinom podataka koji sadrže informacije o hiljadama bivših i sadašnjih pacijenata. Upravljanje ovim podacima od presudnog je značaja za pravilno funkcionisanje ovih institucija, a različite strategije i informacioni sistemi razvijeni su da bi se ovo postiglo. Podaci kao što su imena pacijenata, podaci iz raznih dokumenata, starost, pol, dijagnoza, zakazani pregledi, izabrani lekari itd. moraju se skladištiti i biti lako dostupni za budući klinički i istraživački rad. Takođe, ovi sistemi mogu da skladište obimne dijagnostičke datoteke, poput radiografskih slika (rendgenski snimci, snimci nastali kompjuterizovanom tomografijom, ultrazvuk, nuklearna magnetna rezonanca) koji se dalje mogu analizirati pomoću različitih kompjuterskih algoritama pre postavljanja definitivne dijagnoze. Konačno, pitanja poput zaštite privatnosti pacijenata od najvećeg su značaja u kliničkoj praksi i upravljanje sigurnošću podataka je presudna veština koju svaki analitičar treba da poseduje.

U Srbiji, tipičan primer velikog centralizovanog sistema za čuvanje i uređivanje podataka o pacijentima je *Integrisani zdravstveni informacioni sistem Republike Srbije*. Ovaj sistem omogućava efikasno zakazivanje i upravljanje specijalističkim pregledima, kao i elektronskim upućivanjem pacijenata različitim lekarima i medicinskim ustanovama. Kroz ovaj sistem pacijent može zatražiti i izabrati odgovarajuće vreme za sastanak sa svojim izabranim lekarom / lekarom opšte prakse. Takođe, lekari opšte prakse kroz ovaj sistem komuniciraju sa medicinskim ustanovama sekundarnog i tercijarnog nivoa, kao što su bolnice i klinički centri. Sve ovo značajno olakšava komunikaciju, štedi vreme i za pacijenta i za lekara, i može dovesti do pravovremene dijagnoze i lečenja. I naravno, stvaranje i održavanje takvog složenog sistema ne bi bilo moguće bez tima visoko kvalifikovanih stručnjaka za analizu podataka.

Osim velikih integrisanih zdravstvenih sistema na državnom nivou, mnoge velike bolnice i klinički centri imaju svoje lokalne informacione sisteme u kojima se čuvaju podaci vezani za zaposlene, pacijente, dijagnostičke postupke, projekte itd. Iako su lokalne, ove baze podataka imaju tendenciju da budu veoma velike i ponekad je teško njima upravljati. U prošlosti je bilo mnogo slučajeva u kojima je bolnica, zbog tehničkih poteškoća ili sigurnosnih propusta u vezi sa pristupom, uređivanjem i skladištenjem u bazi podataka, imala ozbiljne poremećaje u zdravstvenim uslugama. Primer bi mogao biti prekid u 2015. u bolnici Hillingdon u Londonu (opslužuje aerodrom Heathrow), gde zbog kvara na mreži medicinsko osoblje nije moglo pristupiti podacima potrebnim za lečenje pacijenata. Do pada sistema došlo je i u mnogim drugim bolnicama, poput one u Bolderu, Kolorado 2013. godine kao i kod nekih provajdera zdravstvenih usluga u Australiji. U jednom istraživanju koje je obuhvatilo 50 različitih zdravstvenih ustanova, u poslednje 3 godine 70% je prijavilo najmanje jedan pad IT sistema, što je uzrokovalo poremećaje veće od 8 sati. To je jedan od razloga zašto je kvalitetan stručnjak za analizu podatke veoma cenjen (a često i dobro plaćen) u mnogim svetskim bolnicama.

Drugo važno područje medicine u kome su neophodne sofisticirane veštine analitičara podataka je elektronsko generisanje recepta za lekove ili takozvani e-recept. Ova tehnologija omogućava medicinskom stručnjaku da elektronskim putem direktno pošalje u apoteku recept za izdavanje lekova ili drugu dozvolu za upotrebu leka. U nekim zemljama, poput Danske i Švedske, ovo je rutinska praksa. Druge zemlje, poput onih u jugoistočnoj Evropi, trenutno je pokreću ili su već počele da je primenjuju u svom zdravstvenom sistemu. Elektronsko propisivanje lekova, osim što štedi vreme i smanjuje birokratske probleme, značajno doprinosi ukupnoj efikasnosti lekara. Lekar sada može da dobije podatke i kompletnu listu aktivnih i odobrenih lekova zajedno sa zvaničnim preporukama i upozorenjima (npr. interakcije lekova, alergijske reakcije) i mudro izabere najbolji. Takođe, verovatnoća lekarske greške sada je znatno smanjena, kao i mogućnost zloupotrebe lekova. Upravljanje sistemom e-recepta kao novom i inovativnom tehnologijom zahteva mnogo iskusnih i

kvalifikovanih stručnjaka za analizu podataka, a mogućnosti dobijanja posla u ovoj oblasti definitivno će se povećati u narednim godinama.

Studijski program u iz oblasti analitike podataka može se osmisliti tako da pored ostalog obuhvata i kurseve vezane za analizu podataka iz medicine i bioloških nauka. U takvim programima, ovi kursevi pomažu sadašnjim i budućim profesionalcima u oblasti informacionih tehnologija da izgrade odgovarajuće iskustvo i ekspertizu, a multidisciplinarna priroda takvog studijskog programa doprinosi uspostavljanju adekvatne saradnje sa lekarima i biolozima, što uveliko koristi trenutnom znanju i praksi u ovim oblastima.

Primena u društvenim i humanističkim naukama

Nauka o društvenim podacima (engl. social data science - SDS) postaje sve popularnija. SDS iz korena menja oblast društvenih i bihejvioralnih nauka. Koristeći skupove digitalno generisanih podataka („veliki društveni podaci“) i inovativne analitičke tehnike, stručnjaci za analizu podataka (engl. data scientists) postaju sve sposobniji da se bave složenim društvenim problemima. Za SDS su matematika, statistika i računanje od suštinskog značaja, ali isto tako su važni i teorijski modeli, kao i metodološki i etički standardi razvijeni u okviru društvenih nauka. Neko bi sa izvesnom dozom humora mogao reći da je SDS društvena nauka „na steroidima“, ali SDS zapravo predstavlja društvene nauke pogodne za borbu sa globalnim izazovima 21. veka.

Kakvi se podaci koriste u SDS-u? Uglavnom digitalni društveni podaci. Danas milijarde ljudi širom sveta koriste pametne telefone, različite elektronske uređaje i Internet za komunikaciju, učenje, poslovanje, pronalaženje informacija, planiranje odmora, deljenje sadržaja na društvenim mrežama, trgovinu, plaćanje računa i poreza. Svaki put kada posetimo neku Web lokaciju, kupujemo online, pretražujemo medicinsku dijagnozu, nadgledamo svoje zdravlje pomoću medicinskih aplikacija, slušamo muziku na YouTube-u, gledamo hit-seriju na Netflx-u ili komentarišemo postove naših prijatelja na Facebook-u mi, kao korisnici, stvaramo potencijalno dragocene podatke. Sve ove svakodnevne aktivnosti ostavljaju iza nas digitalne

tragove na Internetu. Na ovaj način kolektivno stvaramo džinovski društveni "selfie". Međutim, naši digitalni otisci predstavljaju ogromnu količinu neuređenih odnosno nestrukturiranih podataka, tako da su nam potrebni "stručnjaci za podatke" kako bismo mogli da vidimo i razumemo ovu veliku sliku društva. Naučnici koji se bave analizom masovnih društvenih podataka koriste matematiku, različite tehnike statističkog modeliranja i programiranja kako bi prikupili, organizovali, očistili, vizuelizovali, analizirali i tumačili ove podatke.

Uz malu pomoć SDS-a, masovni društveni podaci mogu se dobro iskoristiti. Analiza masovnih podataka može pomoći kreatorima javnih politika u praćenju, razumevanju i rešavanju društvenih problema. Na primer, SDS se uspešno koristi za istraživanje govora mržnje i prikrivanje diskriminacije na Facebook-u, za istraživanje trendova u političkoj komunikaciji tokom predsedničkih izbora, za mapiranje političkog pejzaža na Twitter-u, za istraživanje korupcije visokog nivoa u javnim nabavkama u različitim zemljama. Tehnike SDS primenjene su u otkrivanju lažnih vesti, ali i za istraživanje profila preminulih korisnika Facebook-a (čiji će broj uskoro prevazići broj živih) otvarajući etičko pitanje čuvanja i korišćenja digitalnih podataka. Uspešna u razumevanju i pronalaženju rešenja za složena društvena pitanja, SDS posebno vodi računa o etičkim aspektima upotrebe velikih (društvenih) podataka.

Da li su ljudi bolje raspoloženi ujutro ili popodne? Ponedeljkom ili sredom? A šta je sa vikendima? Iako starije socio-psihološke studije daju izvestan uvid u obrasce ljudskih afektivnih ritmova, njihova otkrića imaju tendenciju da budu nepouzdana (usled malih i nereprezentativnih uzoraka) i zasnovana na retrospektivnim samoizveštavanju o ponašanju, koje je podložno različitim greškama (npr. loše pamćenje).

S obzirom na nedostatke ranijih istraživanja, dva sociologa sa Univerziteta Kornel – Skot Golder i Majkl Masi (2011), odlučili su se za upotrebu tehnika SDS za istraživanje promene ritma raspoloženja pojedinaca. U okviru svoje istraživačke studije, Golder i Masi su analizirali približno pola milijarde javnih tvitova, od 2.4 miliona korisnika Twittter-a koji žive u 84 zemlje širom planete. Nešto potpuno

nezamislivo do pre samo nekoliko godina, kada društveni naučnici nisu imali alate za praćenje ponašanja velike populacije u realnom vremenu. Istraživanje je pokazalo da ljudi imaju bolje raspoloženje kad se probude i da im se raspoloženje pogoršava kao dan odmiče. Odgovor na postavljeno pitanje je, dakle – ljudi su vikendom srećniji, ali jutarnji vrhunac "dobrog raspoloženja" kasni 2 sata u odnosu na radnu nedelju, jer se ljudi vikendom bude kasnije.

Još jedan dobar primer primene analize velikih podataka generisanih na društvenim mrežama jeste razvoj forenzike društvenih mreža fokusirano na praćenje sajber kriminala i aktivnosti kao što su krađa identiteta, elektronsko nasilje, seksualno uznemiravanje i govor mržnje. Naučnici sa Univerziteta u Najrobiju koristili su podatke prikupljene sa Twitter-a kako bi predvideli govor mržnje. Korišćenjem ključnih reči govora mržnje, poput "ubiti", "silovati", "ubistvo", "napad", "kidnapovati", "pucati", "pištolj", "zločin" itd. napravili su izbor od približno 3 miliona tvitova koji su zatim klasifikovani primenom Naive Bayes Classifier-a u 3 kategorije: pozitivan, neutralan i negativan govor mržnje / onlajn zlostavljanje. Model se pokazao uspešnim u otkrivanju i klasifikaciji govora mržnje na Twitter-u, koji je uglavnom bio etnički zasnovan. Ova studija pokazala je kako se podaci Twitter-a mogu prikupiti i sačuvati u bazi podataka za forenzičku analizu, obezbeđujući pritom i to da budu priznati kao dokaz pred sudom.

Za naučnike koji se bave društvenim podacima, postizanje društvenog dobra je od velikog značaja. Već postoji mnogo primera uspešne upotrebe digitalnih društvenih podataka u borbi protiv ozbiljnih zdravstvenih, klimatskih i socijalnih pitanja.

Tuberkuloza (TBC) je jedan od vodećih globalnih uzroka smrti koji odnese do 2 miliona života svake godine. Pošto populacija Indije čini gotovo trećinu ukupnog broja umrlih od TBC-a, indijska vlada je odlučila da ovu bolest iskoreni do 2025. godine uz pomoć novih tehnologija i nauke o podacima. Zajednička inicijativa vlasti, Svetske zdravstvene organizacije i glavnih mobilnih operatera rezultirala je uspešnim projektom mapiranja rizičnih geografskih područja i otkrivanja obrazaca širenja bolesti analizom velikih skupova podataka. Opseg, preciznost i neposredan pristup podacima omogućili su

identifikaciju područja sa niskom stopom TBC-a, ali koja su intenzivno povezana sa oblastima sa visokim stopama TBC-a. Statistička analiza prikupljenih podataka pružila je dragocen uvid u obrasce širenja bolesti, pokazujući da je redovno kretanje stanovništva važniji faktor širenja TBC-a od prostorne blizine između visokih i niskih TBC regiona. Ovaj nalaz je bio izuzetno važan u oblikovanju i primeni mera prevencije i dijagnoze u ovim oblastima i smanjenju broja novih TBC infekcija u zemlji.

Na sličan način korišćeni su veliki setovi podataka u Kolumbiji za sa ciljem smanjenja negativnih uticaja klimatskih promena, u Brazilu za predviđanje zagađenja vazduha, a u Turskoj i Japanu u kreiranju politika koje imaju za cilj smanjenje negativnih uticaja prirodnih katastrofa.

Naučnici u društvenim naukama sve su više zainteresovani za velike skupove podataka koje generišu javne institucije. Podaci zdravstvenog sistema, javnog upravljanja, policije, obrazovnog sistema, podaci o popisu stanovništva, sve se više koriste za poboljšanje performansi institucija. Cilj je pametno upravljanje. Na primer, da bi se adekvatno upravljalo rizicima u gradu, analizom podataka o kriminalnim aktivnostima (vreme, mesto, vrsta krivičnog dela), prestupnicima (pol, starost, biografija itd.) itd., gradske strukture pomažu u adekvatnoj distribuciji policijskih patrola, kao u razvoju preventivnih mera u rizičnim oblastima. Podaci o obrazovnim tranzicijama, generisani u obrazovnom sistemu, mogu nam dati odgovore u kojoj je meri prisutna nejednakost u pristupu obrazovanju i, prema tome, tržištu rada. Analiza velikih setova digitalno generisanih podataka, između ostalog, pruža nam mogućnost efikasnijeg i jednostavnijeg transporta u velikim gradovima.

Primer uspešne primene novih tehnika analize velikih skupova podataka je grad London. Tradicionalne metode analize nisu mogle tačno da mapiraju maršrute i načine korišćenja različitih vrsta prevoza (metro, autobus, taksi i drugo) u ovom gradu. U 2017. godini sproveden je četvoronedeljni pilot projekat sa ciljem da se bolje shvati kako se građani kreću podzemnim sistemom u Londonu. Na osnovu podataka prikupljenih putem iBus sistema, Oyster kartice i mobilnih

aplikacija napravljenih u tu svrhu, bilo je moguće precizno pratiti kretanje putnika. Prema zvaničnicima kompanije TFL (Transport for London) koristi od ovih analiza su brojne:

- "omogućavanje osoblju da bolje informiše kupce o najboljem načinu da izbegnu zastoje ili nepotrebne gužve
- pomoć klijentima u planiranju rute koja im najviše odgovara - bilo na osnovu vremena putovanja, gužve ili pešačke udaljenosti
- omogućavanje veće sofisticiranosti u pružanju informacija u realnom vremenu klijentima dok putuju po Londonu
- pomoć prilikom određivaja prioriteta u investiranju, sa ciljem poboljšanja usluga obezbeđujući maksimalnu korist građanima
- pružanje boljeg uvida u karakterisitke klijenata – ko se i kada kreće transposrtnom mrežom – koji bi mogli povećati komercijalni prihod kompanija koje se reklamiraju ili iznajmljuju maloprodajne jedinice u transportnoj mreži"

SDS integriše znanja društvenih nauka i analitičke tehnike DS. DS se brzo razvija u dinamičnom okruženju koje je obeleženo sinergijom akademskog rada, praktičnih politika i poslovne primene. SDS posebnu pažnju posvećuje etičkim aspektima upravljanja i obrade podataka: digitalnim nejednakostima i isključenosti; otvorenosti podataka, vlasništvu i pristupu; privatnosti; socijalnim uzrocima i posledicama algoritamskih pristrasnosti itd.

Ersamus+ projekat: Napredna analiza podataka u biznisu – ADA

Ako bacite pogled na popularnu Web stranicu datascience community (http://datascience.community/), videćete listu stotina (ako ne i hiljada) koledža i boot kampova koji nude treninge i diplome iz oblasti DS širom sveta. Većina izlistnih koledža nudi DS MSc diplome. DS hype je ogroman, a zaraza se proširila i u Srbiji. Industrija i usluge u Srbiji

postepeno postaju svesni potrebe za DS-om, jer potražnja za DS specijalistima raste kada je u pitanju obvavljanje regularnih poslovnih praksi. Istovremeno, studenti su postali svesni novih trendova.

Ova dva talasa dobila su svoj odraz u evropskom projektu Napredna analiza podataka u biznisu (ADA), http://www.ada.ac.rs/, koji finansira Evropska Komisija u sklopu Erazmus plus+ projekta (evidencioni broj EACEA 598829-EPP-1-2018-1-RS-EPPKA2-CBHE-JP). Glavni cilj ADA projekta je razvijanje master studijskih programa u oblastima DS i analitike podataka, na različitim univerzitetima u Srbiji, koji bi bili usklađeni sa sličnim programima na evropskim univerzitetima.

Ovaj glavni cilj može se podeliti na nekoliko užih, međusobno povezanih ciljeva:

- Podizanje nivoa kvalifikacija stručnjaka iz oblasti DS / analitike podataka na onaj nivo koji se zahteva od stručnjaka u ovoj oblasti iz Evropske Unije. Pojedini nastavnici na univerzitetima u Srbiji već su dostigli potreban nivo kvalifikacija i ekspertize, ali najčešće u specifičnim užim oblastima DS, pa smo tako suočeni sa situacijom da samo nekolicina nastavnika poseduje celovita znanja iz DS i istovremeno ima predavačka iskustva u okviru studijskih programa koji se izvode u ovoj oblasti. Kako bi se ova situacija promenila, uz pomoć projektnih partnera iz Evropske Unije, ADA projekat organizuje trening programe za nastavnike, letnje škole i studijske posete relevantnim istraživačkim laboratorijama i centrima na univerzitetima u Evropskoj Uniji, kao i diskusione sesije sa međunarodnim ekspertima i studentima.
- Pored toga, važan cilj se odnosi i na podizanje sveti o značaju DS među potencijalnim studentima. Širu sliku o DS

neophodno je predstaviti i temeljno objasniti studentima, i a istovremeno i ukazati na potencijalne društvene implikacije ove oblasti istraživanja (kao što su promenjene mogućnosti zaposlenja, promene na tržištu rada, novi pravni i ekonomski aspekti DS, i sl.).

- Privreda, usluge i javni sektor u Srbiji takođe moraju imati važnu ulogu u definisanju potreba za novom radnom snagom i kvalifikovanim stručnjacima iz oblasti analitike podataka, ali i u procesu definisanja privrednih problema u okviru kojih diplomci iz oblasti DS treba da budu obučeni, kao i u obezbeđivanju prakse za studente i master kandidate.
- Neophodno je inicirati promene na univerzitetima u Srbiji. Nije samo reč o uspostavljanju nekolicine studijskih programa, već i o promenama koje bi dovele do uspostavljanja obrazovnog kontinuuma od osnovnog do doktorskog nivoa studija, a koje bi bile u skladu sa savremenim tendencijama u visokom obrazovanju na globalnom nivou i koje bi uzimale u obzir celovit pristup obrazovanju. Razvijanje i implementacija ADA master programa podstiče i neke druge programe osnovnih akademskih studija ili pojedinačne kurseve da prilagode svoje pristupe nastavi tako da budu fokusirani na rešavanje problema i na taj način omoguće i studentima da steknu potrebne kvalifikacije nakon završetka master programa u oblasti DS i analitike podataka.
- Takođe, prepoznaje se jasna potreba za uvođenjem i primenom novih analitičkih alata u nastavi, kao što su složena pretraga velikih baza podataka i primena modela predviđanja, ali i alata za vizuelizaciju podataka, softvera za prikupljanje poslovnih podataka i pravljenje izveštaja na osnovu njih, samouslužnih analitičkih platformi i platformi za prikupljanje i analizu velikih količina podataka.
- Jedan od ciljeva projekta odnosi se i na razvijanje saradnje i fleksibilnosti. Kako bi se izbegla situacija da ADA master programi ostanu "samo nekolicina studijskih programa" iz oblasti DS, cilj je uspostavljanje intenzivne koordinacije aktivnosti i saradnje između nastavnika sa 27 različitih fakulteta na svim relevantnim univerzitetima u Srbiji i sa svim relevantnim akreditacionim telima, kako bi se podigao stepen

interdisciplinarnosti studija i njihove efikasnosti, postigao visok kvalitet i, u krajnjoj liniji, olakšalo iniciranje i razvijanje novih studijskih programa.

- ADA programi su međunarodno orijentisani, akreditovani za nastavu na srpskom i engleskom jeziku, sa naučnim osobljem koje nije isključivo vezano za univerzitete u Srbiji, već i za univerzitete u Evropskoj Uniji, što nije uobičajena situacija u visokom obrazovanju u Srbiji.
- Cilj projekta je i postepeno uvođenje programa celoživotnog učenja iz oblasti DS / analitike podataka za stručnjake koji dolaze iz različitih oblasti ekspertize. Svi univerziteti u Srbiji koji učestvuju u ovom projektu planiraju uspostavljanje programa celoživotnog učenja i organizaciju seminara za stručnjake koji su zaposleni u različitim sektorima (kompanijama, javnom sektoru, zdravstvu, istraživačkim institucijama, itd.). Sastanci koje organizuje ADA projektni konzorcijum imaju za cilj da pomognu učesnicima u programu i zainteresovanim stranama da naprave izbor odgovarajućih oblasti i tema koje bi bile pokrivene programima celoživotnog učenja.
- Jedan od ključnih ciljeva ADA master programa je da izvrše uticaj na visoko obrazovanje, kako na nacionalnom, tako i na regionalnom nivou. Cilj nije da ostanu "samo nekolicina studijskih programa" iz oblasti DS.
- Spajanje delova slagalice u koherentnu celinu zahteva koordinisani pristup koji uzima u obzir nekoliko ključnih aktera i zainteresovanih strana (univerzitete, privredu, javni sektor, medije, Ministarstvo prosvete, nauke i tehnološkog razvoja, akreditaciona tela, partnere iz Evropske Unije, itd.).

Partneri ADA projekta su:

- Univerzitet u Novom Sadu (koordinator projekta za Srbiju)
- Univerzitet u Beogradu (Srbija)
- Univerzitet u Kragujevcu (Srbija)
- Univerzitet u Nišu (Srbija)
- Bečki univerzitet za ekonomiju i poslovanje (Beč, Austrija)
- Škola postdiplomskih studija informatike i matematičkog inženjeringa (Serži, Francuska)

- Univerzitet u Rimu Tor Vergata (Rim, Italija)
- Centar za istraživanje i tehnologiju Grčke (Solun, Grčka)
- Srpska asocijacija menadžera (Beograd, Srbija)

ADA studijski programi

Neke od polaznih tačaka u razvoju ADA studijskih programa na univerzitetima u Srbiji uključuju usaglašeni skup veština svršenih studenata ovih programa, uvid u brojne master programe u DS / analitici podataka širom sveta, kao i opšte smernice za izradu nastavnih programa ovakvih studijskih programa, koje objavljuju nadležne institucije.

Iz perspektive studenata ("Zašto bi trebalo da se prijavim za ovaj studijski program?"), najvažnija poenta je da je skup veština svršenih studenata ADA studija osmišljen tako da odgovara onome što kolokvijalno nazivamo stručnjakom DS trećeg talasa. To uključuje iskustvo sa aktuelnim paketom statističkih alata i algoritmima za analizu podataka, veštinu softverskog inženjerstva, neophodne "tranverzalne" veštine (komunikacija, rešavanje problema, timski rad, istraživački stav, preduzetnički duh,...) i poslovni način razmišljanja – zapošljavanje u oblasti DS-a da bi se kreirala poslovna vrednost, a ne samo gradili modeli, davanje prioriteta radu i osećaju kada se treba zaustaviti, praćenje toka novca u organizaciji i eksperimentisanje i stvaranje inovativne kulture podataka kako bi se postigle stvarne promene.

Kada je reč o sadržaju studijskih programa DS na master nivou, treba imati na umu da se oni razlikuju prema onome na čemu je naglasak na tim programima. Mnogi od njih su programi poslovne analitike u kojima programom i sadržajem kurseva dominiraju aplikacije DS u poslovanju. Postoje i programi u kojima su naglašeni kvantitativni i algoritamski aspekti analize podataka, kao i oni u kojima komponente DS-a orijentisane na tehnologiju vode implementaciju programa. Stručni profil nastavnog osoblja može da pruži prednost jednom ili drugom domenu primene, ili može da rezultira nastavnim planom i programom sa više kurseva vezanih za oblasti Big Data, ili statističke

analize, ili prikupljanja podataka, bezbednosti i transformacije podataka. Brojni su i primeri studijskih programa koji uključuju izborne predmete iz srodnih oblasti, kao što su teorija informacija, obrada signala, teorija uzorkovanja (sampling) i slično.

Projektni tim ADA oslanjao se na smernice za nastavni plan i program DS koje su objavila udruženja poput Association for Computing Machinery (ACM) i Park City Math Institute (PCMI). Tim je istraživao i specifične implementacije ovih smernica u različitim master studijskim programima i pokušao da se što više približi implementaciji tih uzora.

Univerzitet u Novom Sadu: Napredna analitika podataka u poslovanju

Studijski master program ADA iz oblasti DS, razvijen na Univerzitetu u Novom Sadu, jeste kompletan program koji se u potpunosti izvodi na engleskom jeziku, uz učešće predavača sa nekoliko evropskih univerziteta, sa ciljem da pripremi stručnjake zainteresovane za upravljanje podacima i tehnike analize. Program traje dve godine i ima 120 ECTS kredita. Svi kursevi su jednosemestralni. Ovaj studijski program kombinuje dubinsko teorijsko razumevanje prikupljanja podataka i arhitekture sistema podataka sa više poslovnih veština, kao što su analiza podataka i vizuelizacija. Uglavnom je orijentisan na analitiku poslovnih podataka, pa obuhvata sve važne aspekte matematike i statistike, poslovne primene i IT-a. Studijski program ispunjava visoke standarde kvaliteta obrazovnog sistema u Srbiji, a u skladu je i sa zahtevima vrhunskog znanja iz oblasti DS i ekonomije u evropskim i svetskim razmerama. Jedan od najvažnijih ciljeva studijskog programa je edukacija stručnjaka, koji pored značajnog teorijskog znanja, imaju i dovoljno iskustva u primeni analitike podataka na stvarne probleme savremenog poslovanja.

Svrha ovog studijskog programa je postizanje sledećih opštih ishoda učenja: savladavanje kompetencija u razumevanju velikih količina podataka, njihove pripreme, modeliranja, evaluacije i implementacije rešenja u poslovanju primenom programiranja, statistike, mašinskog učenja, manipulacije i vizuelizacije podataka. Predviđeni ishodi učenja

bazirani su na iskustvima nekoliko vodećih evropskih univerziteta i kompanija koje primenjuju DS u poslovanju. Rezultati su u skladu i sa potrebama naše ekonomije, tržišta rada i šire zajednice.

Studijski program osmišljen je tako da obezbedi sticanje kompetencija koje su definisane u saradnji sa kompanijama iz Srbije i regiona, ali i da bude kompatibilan sa sličnim master studijama širom Evrope, u cilju privlačenja stranih studenata. Implementacijom tako definisanog studijskog programa, poslovni analitičari stiču vrhunsku kombinaciju znanja iz ekonomije, računarskih nauka i kvantitativnih metoda, prihvatljivu u evropskom i svetskom kontekstu.

Program nudi obavezne predmete u prva dva semestra, izborne predmete u trećem, kao i rad na master tezi u četvrtom semestru. Pre početka prvog semestra nudi se jedan pripremni kurs, *Kampus podataka*. Kursevi u prvom semestru uglavnom se odnose na računarske osnove savremene analitike podataka (*Osnove velikih podataka* i *Mašinsko učenje*) i specifične teme i alate koji pomažu analitičarima podataka u radu na praktičnim problemima (*Analitika društvenih medija*, *Upravljanje, skladištenje i vizuelizacija velikih podataka*, *R za nauku o podacima*). Obvezni predmeti u drugom semestru fokusirani su na kvantitativno modeliranje i statističku analizu (*Vremenske serije*), zajedno sa praktičnom primenom nauke o podacima u poslovanju (*Poslovni slučajevi*).

Treći semestar nudi različite izborne predmete koji omogućavaju različite putanje učenja (od *Analitike podataka u finansijama*, *Lanaca snabdevanja*, *Marketinga* ili *Menadžmenta*, do *Dubokog učenja* ili *Akademskog pisanja*). Četvrti semestar rezervisan je za praktičan rad i za rad na master tezi.

Univerzitet u Beogradu: Napredna analitika podataka

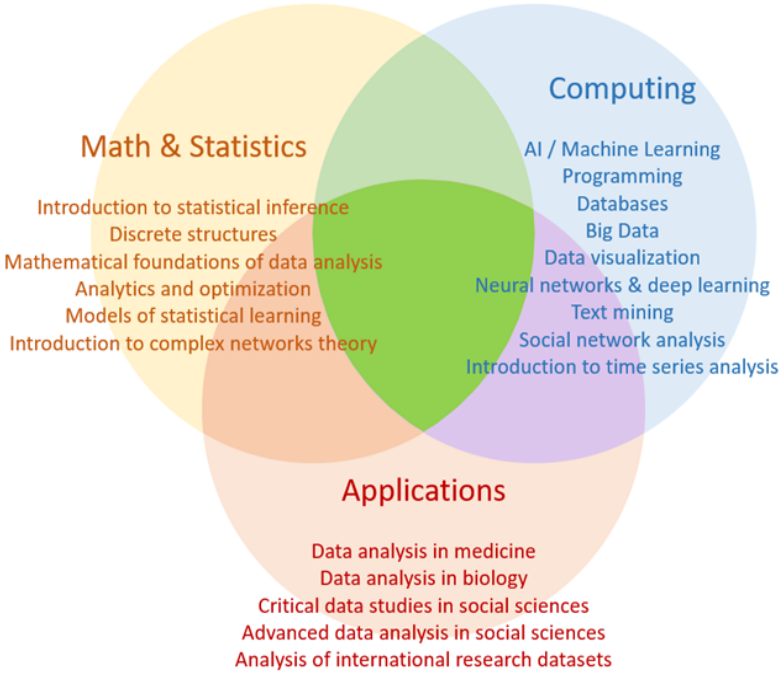
Studijski ADA master program iz oblasti DS-a razvijen na Univerzitetu u Beogradu, *Napredna analiza podataka*, polazi od opšteg razumevanja šta je DS, kao što je objašnjeno u uvodnom delu ove

brošure. U skladu s tim, nudi tri različite grupe kurseva – Matematika i statistika, Računarstvo i Primene. Program ima snažnu tehnološku komponentu, implementiranu u kursevima vezanim za računarstvo. Kada je reč o primenama DS, program trenutno cilja prvenstveno na oblasti nauka o životu (life sciences), društvenih nauka i naučnog izračunavanja (scientific computing). Buduće verzije programa mogu da ponude i kurseve za analizu poslovnih podataka, analizu velikih količina podataka u domenu energije i tako dalje. Mada je program prvenstveno namenjen studentima koji su već završili osnovne studije u nekoj od kvantitativnih disciplina, otvoren je i za studente koji su pohađali druge discipline.

Predmeti koji se nude u ovom studijskom programu razvijeni su oko tri glavna stuba savremene analitike podataka: matematičkih/statističkih osnova, tehnoloških osnova i primena. Kao sastavni deo u nekoliko kurseva, posebna pažnja se posvećuje etičkim aspektima upravljanja i obrade podataka (digitalne nejednakosti i isključenost; otvorenost podataka, vlasništvo i pristup; privatnost; socijalni uzroci i posledice algoritamskih pristrasnosti (biases); itd.).

Program uključuje i obaveznu praksu (glavni projekat / praktikum) za studente kako bi stekli praktično iskustvo u radu na projektima za analitiku podataka, kao i obavezni master rad. Praktičnu orijentaciju karakterišu i mnogi kursevi navedeni na gornjoj slici, pošto se mnogo sati iz jezgra programa zapravo odnosi na sate praktičnih vežbi koje su usmerene na praktične projekte.

Ovo je trosemestralnii studijski program sa 90 kredita (ECTS). Svi kursevi su jednosemestralni. Program nema modula, ali bogata ponuda izbornih predmeta omogućava studentima da odaberu one predmete koji ih vode ka unapređenju znanja u odabranim disciplinama kvantitativnih nauka.



U ovom studijskom programu budući studenti mogu da izaberu veći broj putanja učenja. Pažljivo osmišljeni preduslovi za mnoge kurseve omogućavaju ovu raznolikost. U suštini, mnogi kursevi ponuđeni u prvom semestru odnose se na savladavanje verovatnoće, statistike i drugih matematičkih osnova DS-a (diferencijalni i integralni račun, linearna algebra, diskretne strukture i slično). Zatim u drugom semestru dolaze kursevi koji se odnose na računarske osnove savremene analitike podataka (*Programiranje*, *Baze podataka*, *Big Data analitika*) i različite teme iz veštačke inteligencije (AI) neophodne za naprednu analitiku podataka (*Mašinsko učenje*, *Neuronske mreže i duboko učenje*). Nekoliko kurseva u drugom semestru pokriva i određene teme i alate koji pomažu analitičarima podataka u radu na praktičnim problemima (*Vizuelizacija podataka*, *Text mining* i *Analiza društvenih mreža*). Treći semestar je rezervisan za kurseve orijentisane na primene, sa mnogo praktičnog rada, za prakse i za rad na master tezi. Uz pomoć drugih partnera u projektu, brojne institucije

obezbeđuju prakse za studente, a praksa je osmišljena kao osnova za master rad/tezu.

Odbranom master rada kandidat dobija akademsko zvanje *Master napredne analize podataka*. Studenti koji su završili master program *Napredna analiza podataka* na Univerzitetu u Beogradu postaju kompetentni za:

- samostalni rad na analizi skupova podataka različite složenosti u odabranim domenima, uz naprednu upotrebu aktuelnih alata i tehnologija za analizu podataka
- pripremu, modifikaciju, prilagođavanje i kombinovanje skupova podataka za analizu, iz neobrađenih podataka dobijenih iz različitih aplikacija i drugih izvora
- uključivanje u različite interdisciplinarne radne timove, gde se očekuju veštine analize podataka u različitim disciplinama i savladavanje aktuelnih alata i tehnologija za analizu podataka, ne samo u rešavanju rutinskih praktičnih problema, već i u nestandardnim situacijama gde je potrebna kreativnost i istraživački pristup
- rad sa velikim skupovima podataka

Predmetne kompetencije svršenih studenata uključuju:

- sposobnost razumevanja i analize različitih skupova podataka iz perspektive matematičkih osnova napredne analize podataka (linearna algebra, diferencijalni i integralni račun, diskretna matematika, višedimenziona geometrija, optimizacija, itd.)
- ovladavanje statističkim osnovama napredne analize podataka (sumarizacija podataka, testiranje hipoteza, varijansa podataka, korelacija podataka, verovatnoća i funkcija raspodele verovatnoće, primena deskriptivnih i inferencijalnih statistika na skupove podataka, itd.)
- programiranje na savremenim (state-of-the-art) programskim jezicima za analizu podataka
- veštine korišćenja naprednih i adekvatnih tehnika vizuelizacije podataka, kao i aktuelnih softverskih alata i tehnologija koje omogućavaju kreiranje bogatih vizuelizacija

ADA - perspektiva poslodavaca

Jedan od partnera ADA projekta je i Udruženje menadžera Srbije (Serbian Association of Managers, SAM). Udruženje obuhvata više od 400 članova. Oni su menadžeri uspešnih kompanija koje ukupno zapošljavaju više od 70.000 ljudi u Srbiji. U proleće 2019. SAM je sproveo anketu među svojim menadžerima i stručnjacima iz oblasti analize podataka. Cilj ankete bio je da se identifikuju potrebe industrije u Srbiji u kontekstu DS-a, kao i veštine analitičara podataka koje su potrebne srpskim kompanijama, imajući u vidu postojeću organizacionu strukturu u tim kompanijama. Analiza potreba koja je rezultat ankete pokazala je zanimljive rezultate.

Šta kompanije i institucije u Srbiji mogu da dobiju od ADA studijskih programa? Očigledan odgovor na ovo pitanje je: obrazovane stručnjake iz oblasti analize podataka koji su diplomirali na ADA studijskim programima. Ali ima još puno toga. Zanimljivo je da su menadžeri koji su učestvovali u anketi pokazali samo prosečno razumevanje šta je tačno posao stručnjaka za analizu podataka, i to više u srednjim i velikim preduzećima i u onim koji svoje poslovanje obavljaju u inostranstvu ili imaju strane vlasnike. Tek svaka peta kompanija zapošljava istraživače koji se bave podacima, a ispitanici uglavnom ne znaju da li će biti zaposleni u narednoj godini. Istraživači koji rade u kompanijama prilično su raštrkani unutar organizacija i rade u različitim sektorima: konsalting, prodaja, upravljanje programom i prodajom, menadžment, operacije, poslovna inteligencija, cene, analitika, upravljanje poslovnim podacima, tehnologija, kontrola, finansije, marketing, strateško upravljanje, upravljanje segmentima,... Čini se da nema univerzalnog sektora / funkcije tamo gde rade[1].

Takođe je primetna razlika između velikih i malih kompanija: u manjim kompanijama generalni direktor je taj koji je uglavnom zadužen za

[1] U sličnoj studiji u SAD (Big Data and AI Executive Survey 2019, NewVantage Partners) koje je sprovedeno na 65 of Fortune 1000 kompanija, dve funkcije su se jasno izdvojile: Chief Data Analytics Officer i Chief Information Officer.

upravljanje podacima i njihovo korišćenje za poboljšanje poslovnog odlučivanja, dok su u većim kompanijama funkcije mnogo raznovrsnije: većina funkcija je povezana sa finansijama i upravljanjem. Ovo jasno ukazuje na potrebu za dodatnom obukom menadžera o tome kako stručnjak za podatke može da poboljša posao i šta tačno treba da radi. A ovo opravdava ideju ADA projekta u smislu organizovanja programa celoživotnog učenja u DS / analitici za profesionalce iz različitih oblasti.

Šta menadžeri misle o DS / naprednoj analitici podataka i u kojoj meri se ona koristi u njihovim kompanijama? Skoro dve trećine menadžera koji su učestvovali u anketi (59%) tvrdi da, prema njihovim saznanjima, njihove kompanije do sada nisu sprovele nijedan DS projekat. One koji jesu, za to su koristile unutrašnje resurse. Veoma malo kompanija (samo 2 u ukupnom uzorku) je angažovalo stručnjake van kompanije da pomognu u implementaciji DS projekata. Logično je da veće kompanije i kompanije u stranom vlasništvu češće primenjuju DS projekte.

Koji su nedostaci otkriveni? Menadžeri koji su učestvovali u anketi procenjuju da je upotreba napredne analitike u njihovim kompanijama na prilično niskom nivou. Svaka četvrta kompanija još uvek nije ni započela sa primenom napredne analitike, što je posebno zabrinjavajuće, jer SAM okuplja najuspešnije kompanije u Srbiji (i zato se može očekivati da će, ako se posmatra ekonomija u celini, aplikacija biti znatno manja). Opet, veće kompanije, kompanije u stranom vlasništvu i kompanije koje trguju u inostranstvu pokazuju znatno veći stepen razvoja u ovom pogledu.

Koje su barijere i prepreke u prihvatanju DS-a? Osnovna barijera za primenu napredne analitike je nerazumevanje DS-a i njegove primene u poslu. Manja preduzeća su ograničena i nedostatkom finansijskih sredstava, dok se veće kompanije bore sa organizacionim poteškoćama.

Koji su prioriteti u uvođenju i intenzivnijoj primeni napredne analitike podataka? Istraživanje SAM pokazalo je da programski jezici, tehnologije, Cloud platforme ili alati koji su važni za poziciju DS

stručnjaka uključuju: SQL, alate za vizuelizaciju (PowerBI, Tableau, GGPlot, Plotly, Qlik), Python i / ili R i Cloud platformu (AWS, Google Cloud, Private Cloud). Alati i tehnike koje će biti potrebne da stručnjak iz analitike podataka u budućnosti savršeno savlada su: Hadoop, Scala, Hive, Spark, TensorFlow, NoSQL i Natural Language Processing (NLP). Poznavanje SAS-a, SPSS-a i Excel-a je poželjno jer se ovi alati danas široko koriste u kompanijama, ali DS naučnici ne moraju biti eksperti da bi ih koristili. Na pitanje koje druge veštine (pored tehničkih veština) ovi stručnjaci treba da jasno pokažu, ispitanici (DS stručnjaci koji su učestvovali u anketi) izdvojili su strateško upravljanje, lance vrednosti, poslovne procese (jer se poslovni procesi menjaju sa naukom), osnove finansija (ROI, šta je ravnoteža i ravnoteža uspeha), kao i transverzalne veštine (veštine prezentacije, komunikacijske veštine,…).

Istraživanje SAM je takođe pokazalo da postoji prilično jasan uvid kad je u pitanju korišćenje DS-a u preduzećima u Srbiji. Ovi nalazi mogu biti od velike koristi kako za same kompanije, tako i za potencijalne studente ADA studijskih programa.

Dominantne industrije i procesi. Poslovni procesi u kojima se već koriste novi pristupi i alati za analizu podataka uključuju: uvid u profile potrošača i preciznije ciljanje (češće kod srednjih i velikih kompanija), finansijsko planiranje i analize (češće kod malih preduzeća), cenovna politika i profitabilnost. Srednje i male kompanije danas su usredsređene na finansijsko planiranje i analize, dok se veće kompanije fokusiraju na stvaranje uvida u potrošače i preciznije ciljanje. Sektori koji su najčešće odgovorni za optimizaciju upravljanja podacima i poboljšanje poslovnog odlučivanja su finansije (dominantno), menadžment i poslovna inteligencija.

Raspodela odgovornosti. Stručnjaci koji su učestvovali u istraživanju ne slažu se oko toga da li i u kojoj meri bi DS stručnjak trebalo da bude stručnjak za domene, kao i da li će biti dobar predavač. Postoji šire i uže tumačenje obima rada DS stručnjaka (dve ili jednea funkcija), kao i toga da li je potrebno da se oni bave redovnim izveštavanjem ili ne. Sa druge strane, eksperti su se složili oko toga šta nije posao DS stručnjaka:

- da bude stručnjak za programiranje ili ekspert za programski jezik koji nije povezan sa njihovim radom
- da bude administrator baze podataka
- da bude finansijski ekspert
- da radi *sve* što je vezano za IT

S obzirom na to da opis posla DS stručnjaka nije u potpunosti poznat menadžerima, eksperti su definisali šta DS stručnjak treba da radi u kompaniji:

- tehnički aspekt
    - priprema i čišćenje podataka
    - razumevanje baza podataka
    - rad sa naprednim alatima
    - analiza, modeliranje i manipulacija podacima
    - pretvaranje podataka u informacije
    - vizuelizacija podataka
- poslovni aspekt
    - komunikacija sa ljudima na drugim pozicijama / funkcijama u kompaniji
    - razumevanje i učestvovanje u definisanju poslovnih procesa i podataka koje ti procesi stvaraju

Većina lekcija naučenih kroz odgovore dobijene u SAM istraživanju povezana je sa mogućnostima zapošljavanja diplomaca ADA studijskih programa.

Prednosti za studente i diplomce. Znanje i veštine koje nude ADA studijski programi i koji su danas sve značajniji u industriji uključuju:

- analiza podataka, statistika i algebra, vizualizacija podataka, deskriptivna analiza (transformacija podataka u informacije)
- otvorenost za saradnju sa demenskim ekspertima
- rad sa naprednim alatima za analizu podataka
- razumevanje podataka i njihove vrednosti za preduzeća
- sposobnost prepoznavanja poslovnih problema i modeliranje podataka u skladu sa tim
- poznavanje osnova strategije i finansija
- komunikacija s drugim poslovnim sektorima
- praksa

Prednosti za kompanije i institucije. Menadžeri koji su učestvovali u SAM istraživanju veruju da je činjenica da ADA studijski programi stvaraju diplomce sa čvrstim tehničkim i komunikacijskim veštinama, dobrim poznavanjem procesa u industriji i različitim uslugama, kao i dobrim poslovnim znanjem. Predstavnici kompanija trenutno više cene komunikacione i vizuelne veštine u poređenju sa procesima koji postepeno generišu kompletno poslovno rešenje (verovatno zbog nerazumevanja spektra tehnika koje DS stručnjak može da koristi).

ADA - Pratite nas!

Web lokacija ADA projekta: http://www.ada.ac.rs/

Facebook: https://www.facebook.com/ADA-Advance-Data-Analitics-in-Business-2319108561657462/

Instagram: https://www.instagram.com/advanceddataanalitics/

Twitter: https://tvitter.com/AdaAdvanced

Literatura

Study.EU: Your gateway to universities in Europe, Why you should study a Masters in Data Science: 3 reasons, https://www.study.eu/article/why-you-should-study-a-masters-in-data-science-3-reasons

Deepsense.ai, Why do we need more data scientists and why should you become one?, https://deepsense.ai/why-do-we-need-more-data-scientists-and-why-should-you-become-one/

Mentionlytics, 5 Real-World Examples of How Brands are Using Big Data Analytics, https://www.mentionlytics.com/blog/5-real-world-examples-of-how-brands-are-using-big-data-analytics

CIO, 10 data analytics success stories: An inside look, https://www.cio.com/article/3221621/6-data-analytics-success-stories-an-inside-look.html

University of Wisconsin Data Science Program, What Do Data Scientists Do?, https://datasciencedegree.wisconsin.edu/data-science/what-do-data-scientists-do/

Liip Data Science Stack, Tools in the Data Science Stack, http://datasciencestack.liip.ch/

Itsvit.com, Real-Life Business Success Stories Based on Big Data, https://itsvit.com/big-data/8-real-life-business-success-stories-based-big-data-part-1/

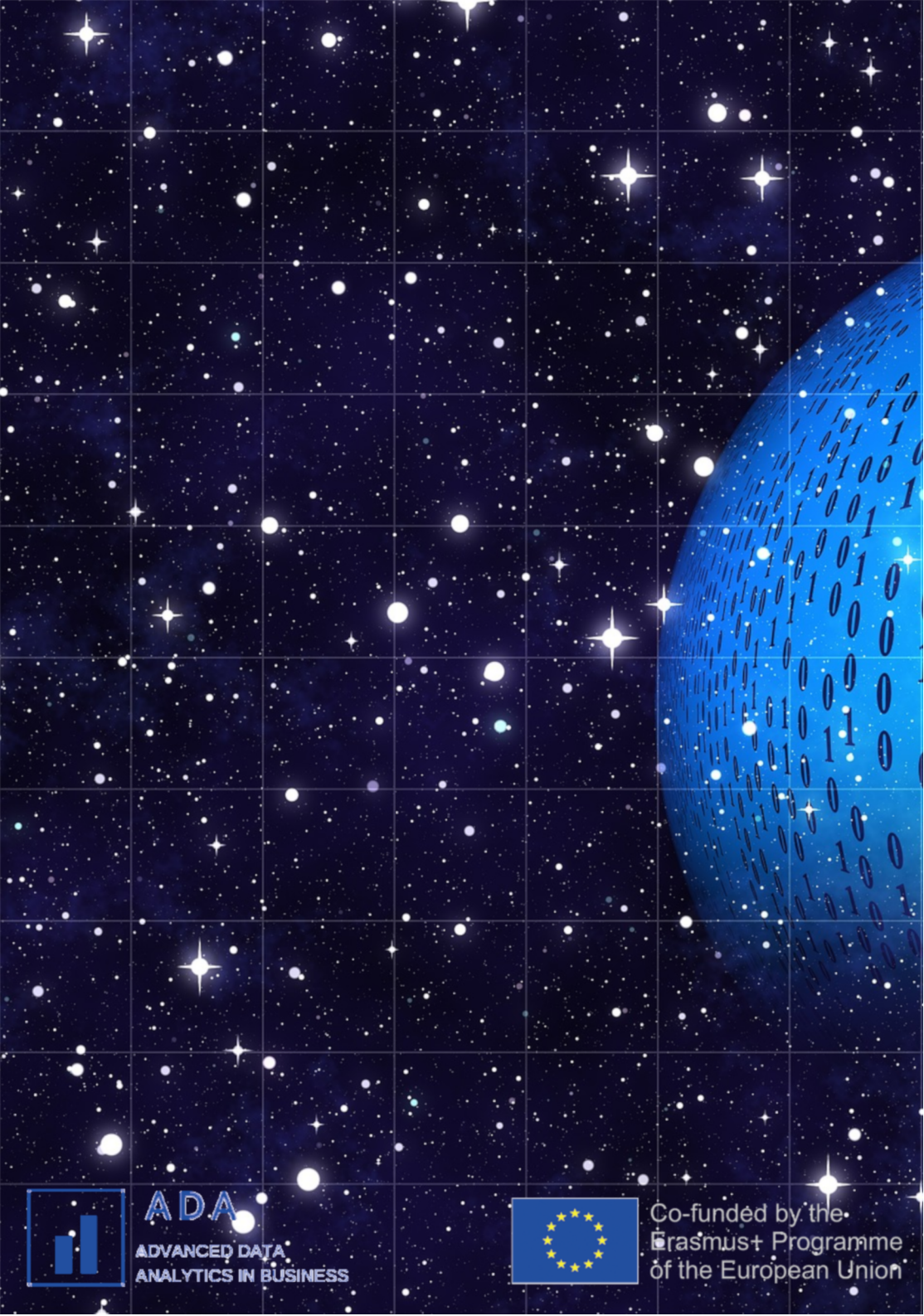
Data Science Community Website: http://datascience.community/

Data Science Graduate Programs, Data Science Careers, https://www.datasciencegraduateprograms.com/careers/

KDNuggets.com, Education in Analytics, Big Data, Data Mining, Data Science, Machine Learning, https://www.kdnuggets.com/education/index.html

Towards Data Science, The Third Wave Data Scientist, https://towardsdatascience.com/the-third-wave-data-scientist-1421df7433c9

Park City Math Institute (PCMI), Curriculum Guidelines for Undergraduate Programs in Data Science, https://arxiv.org/pdf/1801.06814.pdf


ADA
ADVANCED DATA
ANALYTICS IN BUSINESS
Co-funded by the
Erasmus+ Programme
of the European Union