Primjena metoda analize i rudarenja podataka u edukaciji

Bašić, Tea

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Tea Bašić

PRIMJENA METODA ANALIZE I RUDARENJA PODATAKA U EDUKACIJI

Diplomski rad

Voditelj rada:
izv. prof. dr. sc. Saša Singer

Zagreb, srpanj, 2020.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Za mamu i Brunu!

Sadržaj

Uvod

Podatkovna znanost (eng. data science) interdisciplinarno je područje kojim podatkovni znanstvenici proširuju svoja znanja, učeći kako razmišljati na sistematičan način, integrirajući znanja iz različitih područja (vidi [4]). Zapravo, konkretna definicija podatkovne znanosti još ne postoji. Podatkovna znanost je relativno novi pojam. William S. Cleveland je prvi koristio termin Data science 2001. godine, u sklopu rada pod naslovom *„Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics“*. Tek godinu dana kasnije, Međunarodno vijeće za znanost zapravo je priznalo podatkovnu znanost i stvorilo odbor za to.

U ovom diplomskom radu će se objasniti što je to podatkovna znanost te kako ju se trenutno definira. Opisat će se što je sve potrebno da bi netko radio kao podatkovni znanstvenik (eng. data scientist), koja znanja i vještine su potrebne te koji se sve zadaci ubrajaju u ovo područje. Posebno će se objasniti kako podatkovni znanstvenici koriste Python i R u svom radu te koje su značajne razlike ako se koristi jedan ili drugi programski jezik. U drugom dijelu diplomskog rada će se prikazati na koji način se podatkovna znanost može ukomponirati u trenutni kurikulum informatike i matematike. Obradit će se nekoliko aktivnosti koje se mogu izvoditi na dodatnim satima ili regularnoj nastavi. Cilj je proučiti koliko je moguće upoznati djecu u osnovnoj i/ili srednjoj školi s jednim od popularnijih zanimanja budućnosti.

Poglavlje 1

Podatkovna znanost

1.1 Što je to *Data science*?

Mueller i Massaroni u knjizi *Python for Data Science* opisuju podatkovnu znanost na sljedeći način: Podatkovna znanost uključuje upotrebu naprednih matematičkih tehnika, statistika i velikih podataka. No, podatkovna znanost, također, uključuje pomaganje u donošenju odluka, stvaranje prijedloga opcija na temelju prethodnih izbora te razvijanje robota, odnosno kako roboti vide predmete (vidi [7], str. 26). Ljudi koriste podatkovnu znanost na toliko različitih načina da, radeći bilo što, mogu osjetiti učinak podatkovne znanosti na svoj život.

Kao što je navedeno u uvodu, podatkovnu znanost nije lako definirati. Najčešći način definiranja podatkovne znanosti je, zapravo, opisivanje što to podatkovni znanstvenik radi, no ponekad ni to nije u potpunosti jasno.

Zanimanje: Podatkovni znanstvenik

Različita okruženja u kojim podatkovni znanstvenik radi uvjetuju drugačiji opis posla. Zanimanje se može promatrati u akademskom okruženju te u industriji. U akademskom okruženju, O'Neil i Schutt (vidi [8]) definiraju ga kao znanstvenika koji posjeduje razne vještine, od društvenih znanosti do biologije, te koji radi s velikom količinom podataka i suočava se s problemima vezanim uz strukturu, veličinu, red i složenost prirode podataka. U isto vrijeme, rješava i probleme stvarnog svijeta.

U industriji se opis posla razlikuje po stažu i znanju koje podatkovni znanstvenik ima.

Glavni podatkovni znanstvenik, netko koga bi smatrali senior zaposlenikom[1], postavlja podatkovnu strategiju tvrtke, koja uključuje razne stvari: postavljanje infrastrukture za sakupljanje podataka, problem privatnosti kod prikupljanja podataka od korisnika, kako iskoristiti podatke za donošenje odluka i organizacija inženjeringa. U većini slučajeva, jedna osoba ne radi sav navedeni posao. Podatkovni znanstvenici se specijaliziraju za jedno područje te kao tim, uključujući inžinjere i analitičare, dolaze do potrebnih rezultata. A rezultati su najčešće analize i predviđanja poslovanja za tvrtku: u što uložiti, smanjiti budžet, predvidjeti rezultate kampanje i slično.

Podatkovni znanstvenik na nižoj poziciji će se uglavnom baviti izvlačenjem i interpretiranjem podataka, što zahtijeva alate i metode iz statistike i strojnog učenja. Najviše vremena provodi na procesu prikupljanja, čišćenja i mijenjanja podataka. Proces zahtijeva upornost, znanje statistike i vještine softverskog inženjeringa. Kada se dobije pristup podacima koji su pročišćeni, sljedeći korak je istraživačka analiza podataka, koja kombinira vizualizaciju i shvaćanje podataka. Cilj je pronaći uzorke, izraditi modele i algoritme koji proizlaze iz sirovih podataka.

Jasno je da, kao što je teško točno definirati podatkovnu znanost, nije jednostavno odrediti ni što to točno podatkovni znanstvenik radi. Opis posla se može razlikovati ne samo od tvrtke do tvrtke, već i kod individualnih odjela unutar iste tvrtke. No, svakako je bitno naglasiti da u baš svakoj industriji postoji otvorena pozicija za podatkovnog znanstvenika. Veliki razlog tome je što se višestruko povećao broj podataka koji se prikuplja (zahvaljujući IoT), a iz svih tih podataka se može doći do zanimljivih i često jako bitnih zaključaka vezanih uz profit pojedine tvrtke.

Kad se sagledaju razne definicije i opisi posla, sve se zapravo svodi na tri glavna procesa ili faze koje svaki podatkovni znanstvenik provodi, a to su: prikupljanje podataka, analiza podataka i prezentacija podataka.

1. **Prikupljanje podataka**
   Ovo je jedan od najvažnijih procesa, jer ako nema podataka za analizu, onda se ne može ni analizirati ni doći do bitnih zaključaka. Podaci se najčešće dobavljaju iz baza podataka i to u sirovom obliku. Ponekad nije najjednostavnije ni pronaći sve potrebne podatke, tako da ovaj proces može biti dugotrajan. Bitno je razumjeti podatkovnu domenu kako bi se podaci mogli lakše pregledavati i formulirati. Vještine modeliranja podataka, odnosno kako su podaci povezani i strukturirani, su ključne u ovom procesu.

[1] U IT industriji je česta podjela zaposlenika po znanju i stažu na junior, middle i senior.

2. **Analiza**
   Nakon što su podaci prikupljeni i postavljena je jasna struktura nad njima, može se početi obavljati analiza. Neke analize se izvode koristeći osnovne vještine statističkih alata, međutim, upotreba specijaliziranih matematičkih modela i algoritama može izvući jasnije i zanimljivije zaključke, koji se na prvi pregled jasno ne vide u podacima.

3. **Prezentacija**
   U analizi, podatkovnom znanstveniku podaci postaju posve jasni, te lako vidi neke povezanosti. Takve podatke ipak nije moguće predstaviti ostalim ljudima, najčešće voditeljima tvrtke. Većina ljudi ne razumije brojeve dobro i ne mogu primijetiti obrasce koje podatkovni znanstvenik vidi. Zato je važno pružiti grafički prikaz podataka i obrazaca, odnosno vizualizirati što brojevi znače i kako ih smisleno primijeniti.

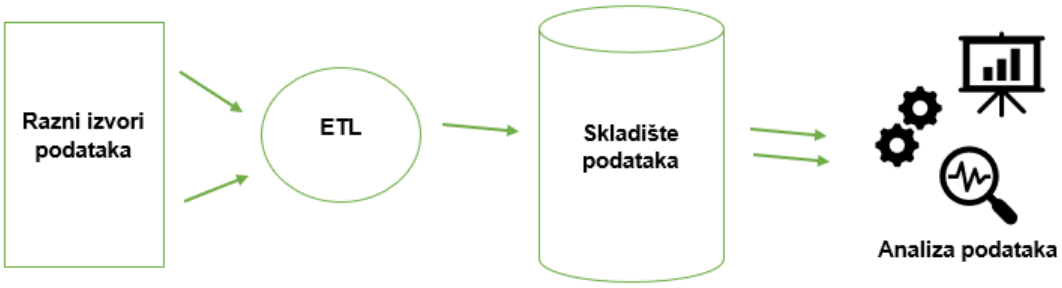
ETL

Proučavajući podatkovnu znanost često se spominje pojam ETL. To je vrsta integracije podataka koja se sastoji od tri koraka (eng. extract, transform, load) koji se koriste za kombiniranje podataka iz više izvora (vidi [4]). Često se koristi za izgradnju skladišta podataka. Tijekom ovog procesa podaci se izdvajaju iz izvornog sustava, pretvaraju u format koji se može analizirati i pohranjuju u skladište podataka ili drugi sustav. Može izdvojiti podatke iz bilo kojih izvora podataka, kao što su datoteke, bilo koje RDBMS / NoSQl baze podataka, web stranice ili korisničke aktivnosti u stvarnom vremenu, transformirati stečene podatke i zatim učitati pretvorene podatke u skladište podataka za poslovne svrhe, kao što su izvještavanje ili analitika.

Česta je greška pretpostaviti da je ETL neki alat kojeg podatkovni znanstvenici, analitičari i inženjeri koriste. To je, zapravo, koncept kretanja podataka koji se može uspostaviti uz pomoć različitih alata, kao što su Informatica, Tableau, programski jezici, itd. Analiza podataka se uglavnom vrti oko ETL-a. Ključni razlozi za korištenje ETL-a su:

- Vizualizacija cjelokupnog protoka podataka koji pomaže poslovanju i donošenju kritičkih poslovnih odluka.
- Transakcijske baze podataka ne mogu odgovoriti na složena poslovna pitanja na koja može odgovoriti ETL.

- ETL pruža metodu premještanja podataka iz različitih izvora u skladište podataka.
- Omogućuje automatsko ažuriranje skladišta podataka.
- Omogućuje izvođenje složenih transformacija.
- Pomaže kod migracije podataka, održavanja točnih formata i dosljednosti pojedinog sustava.
- ETL nudi dubok povijesni kontekst za posao.
- Unaprijed se definira na ciljanoj bazi podataka.



Slika 1.1: ETL proces

1.2 Veliki podaci (Big data)

Sve većom brzinom se stvara velika količina podataka (eng. Big data). Svijet u kojem živimo se može promatrati kao jedan veliki stroj za generiranje podataka. Takvu količinu podataka je teško spremati i analizirati u stvarnom vremenu, pogotovo ako ne prate neku uobičajenu strukturu. Često su veliki podaci različiti, nisu uređeni, nemaju strukturu i nisu nešto što bi se moglo pronaći u urednoj bazi podataka. Podatkovna znanost je ”alat” kojim se *Big data* transformira u korisne i čitljive podatke.

U knjizi EMC Education Services (vidi [4]), veliki podaci su definirani kao podaci čija veličina, distribucija, raznolikost i/ili pravovremenost zahtijeva korištenje novih tehnoloških arhitektura i analitika, da bi se omogućio uvid koji će otključati nove izvore poslovne vrijednosti. Za opisivanje karakteristika velikih podataka obično se koristi izraz 3V,

koji uključuje volumen, varijantnost i velocitet. No, sve više se, osim navedenih 3V, dodaju još tri karakteristike, tako da možemo govoriti o ukupno 6V.

- Volumen (eng. Volume)
  Količina se odnosi na velike količine podataka, koje se svake sekunde generiraju iz društvenih mreža, mobitela, automobila, kreditnih kartica, M2M senzora, fotografija, videa, itd. Ogromne količine podataka postale su toliko velike da se više ne mogu pohraniti i analizirati koristeći tradicionalnu tehnologiju baza podataka. Prikupljanje i analiza ovih podataka je inženjerski izazov velikih razmjera.

- Raznolikost (eng. Variety)
  Podaci dolaze u brojnim oblicima, od geografskih i prostornih podataka, do kolačića i sadržaja web stranica, od tweetova do vizualnih podataka, poput fotografija i videozapisa. Iako se često zanemaruju u usporedbi s javno objavljenim "volumenom", raznolikost podataka može predstavljati problem za većinu poduzeća.

- Brzina (eng. Velocity)
  Dok količina podataka raste, brzina stvaranja i korištenja podataka također se povećava. To znači da se obrada, pohrana i analiza moraju ubrzati. Prednost poslovanja leži u postojanju i djelovanju najsvježijih podataka, što znači primanje podataka i uvid što je prije moguće, a zatim djelovanje jednakom brzinom.

- Istinitost (eng. Veracity)
  Istinitost je u osiguravanju pouzdanosti i valjanosti uvida dobivenih iz podataka. Netočni podaci gotovo su bezvrijedni, u nekim slučajevima čak i štetni.

- Ranjivost (eng. Vulnerability)
  Širenje podataka učinilo je da se mnogi ljudi osjećaju izloženi i ranjivi na način na koji se njihovi podaci koriste. Nakon što istraže nešto više o prikupljanju podataka, uglavnom se osjećaj nelagode poveća, jer prije nisu bili ni svjesni koliko se podataka zapravo prikuplja i koristi.

- Vrijednost (eng. Value)
  Na najjednostavnijoj razini podaci nemaju neku vrijednost. Postaju korisni samo kad se izvuku uvidi koji su potreban za rješavanje određenog problema ili za zadovoljavanje određene potrebe. Jednom kada se to učini, podaci dobivaju vrijednost pomoću poslovnog utjecaja i vrijednosti potrošača koje ovaj uvid pruža.

1.3 Python

Python je viši programski jezik dizajniran tako da se lako čita i jednostavno implementira. Open-source je, što znači da se može slobodno koristiti, čak i za komercijalne aplikacije. Python se može pokretati na Mac, Windows i Unix sustavima, a također je prijenosan na Java i .NET virtualne strojeve. Python se smatra skriptnim jezikom, poput Ruby ili Perl, i često se koristi za stvaranje web aplikacija i dinamičkog web sadržaja. Podržani su i brojni programi dvodimenzionalnog i 3D prikaza, što korisnicima omogućuje stvaranje prilagođenih dodataka i proširenja s Python-om.

Mnogi podatkovni znanstvenici rado koriste Python, jer on pruža bogati izbor biblioteka, poput NumPy, SciPy, MatPlotLib, Pandas i Scikit-learning, koje značajno olakšavaju zadatke iz područja podatkovne znanosti. Osim toga, Python je precizan jezik koji olakšava upotrebu višestruke obrade na velikim skupovima podataka, time smanjujući vrijeme potrebno za njihovu analizu. Jedan od najpoznatijih IDE-a[2] je Anaconda, koja ima implementiran koncept Jupyter Notebook-a.

Biblioteke

Pandas je Python biblioteka za analizu podataka koja se koristi za sve, od uvoza podataka iz Excel proračunskih tablica, do obrade skupova za analizu vremenskih serija. Pandas stavlja gotovo svaki uobičajeni alat za pregled podataka na dohvat ruke. To znači da se osnovno čišćenje i napredne manipulacije mogu provesti s Pandas-ovim moćnim okvirima. Pandas je izgrađen na temelju NumPy-a, jedne od najranijih biblioteka za Python, vezane uz analizu podataka. Funkcije NumPy koriste se u Pandas za naprednu numeričku analizu.

Ako je potrebno nešto veća specijalnost, Python i to nudi. Neke od popularnih i često korištenih biblioteka:

- SciPy je znanstveni ekvivalent NumPy-a, nudi alate i tehnike za analizu znanstvenih podataka.
- Statsmodels se fokusira na alate za statističku analizu.
- Scilkit-Learn i PyBrain su biblioteke strojnog učenja koje pružaju module za izgradnju neuronskih mreža i obradu podataka unaprijed.

[2]IDE: integrated development environment

Istraživanje i analiza podataka	Pandas, NumPy, SciPy
Vizualizacija podataka	Matplotlib, Bokeh, d3py, ggplot, Plotly, prettyplotlib
Strojno učenje	Scikit-Learn, StatsModels, Shogun, PyLearn2, PyMC, PyBrain
Pohrana i formatiranje podataka	csvkit, PyTables, SQLite3
Duboko učenje	Keras, TensorFlow

Tablica 1.1: Najpopularnije Python biblioteke i API

Korištenje Pythona

U prvom poglavlju su navedene tri glavne faze kroz koje svaki podatkovni znanstvenik prolazi, obrađujući podatke. Python se koristi u svakoj od tih faza i to na sljedeći način.

- Prva faza

  Potrebno je dobiti potrebne podatke za analizu. Podaci nisu uvijek lako dostupni, pa je potrebno na što jednostavniji način doći do kvalitetnih podataka iz raznih izvora. U ovom poslu pomažu biblioteke Scrapy i BeautifulSoup.

- Druga faza

  Koristeći biblioteke kao što su Pandas i NumPy, brzo se mogu dobiti čitljivi i analizirani podaci. Koristeći navedene biblioteke, podaci se mogu i paralelno procesirati. Ako se radi o strojnom učenju, što je vrlo složena računalna tehnika koja uključuje matematičke alate poput vjerojatnosti, statistike, matrica i slično, onda se mogu koristiti biblioteke Scikit-Learn ili PyBrain.

- Treća faza

  U ovoj fazi moramo vizualizirati ili grafički prikazati podatke. Najbolji način da se to učini je predstavljanjem podataka u obliku grafova, grafičkih pita i drugih formata. Za obavljanje ove funkcije koriste se Python biblioteke Seaborn i Matplotlib.

Glavni razlozi zašto koristiti Python su sljedeći:

- **Snažan i jednostavan za korištenje**

  Python se smatra jezikom početnika i svaki student ili znanstvenik sa samo osnovnim znanjem može početi raditi u njemu. Vrijeme provedeno za uklanjanje pogrešaka i za različita ograničenja softverskog inženjeringa je, također, minimalno. U usporedbi s drugim programskim jezicima, kao što su C, Java i C#, vrijeme za implementaciju koda je kratko, što pomaže programerima i softverskim inženjerima da provedu više vremena radeći na algoritmima.

- **Biblioteke**

  Python pruža veliku bazu biblioteka vezanih uz analizu podataka, umjetnu inteligenciju i strojno učenje. Neke od najpopularnijih biblioteka uključuju Scikit Learn, TensorFlow, Seaborn, Pytorch, Matplotlib i mnoge druge.

- **Skalabilnost**

  U usporedbi s drugim programskim jezicima, poput Java i R, Python se pokazao kao visoko skalabilan i brži jezik. Pruža fleksibilnost za rješavanje problema koji se ne mogu riješiti s drugim programskim jezicima. Mnoge tvrtke ga koriste za razvoj brzih aplikacija i alata svih vrsta.

- **Vizualizacija i grafika**

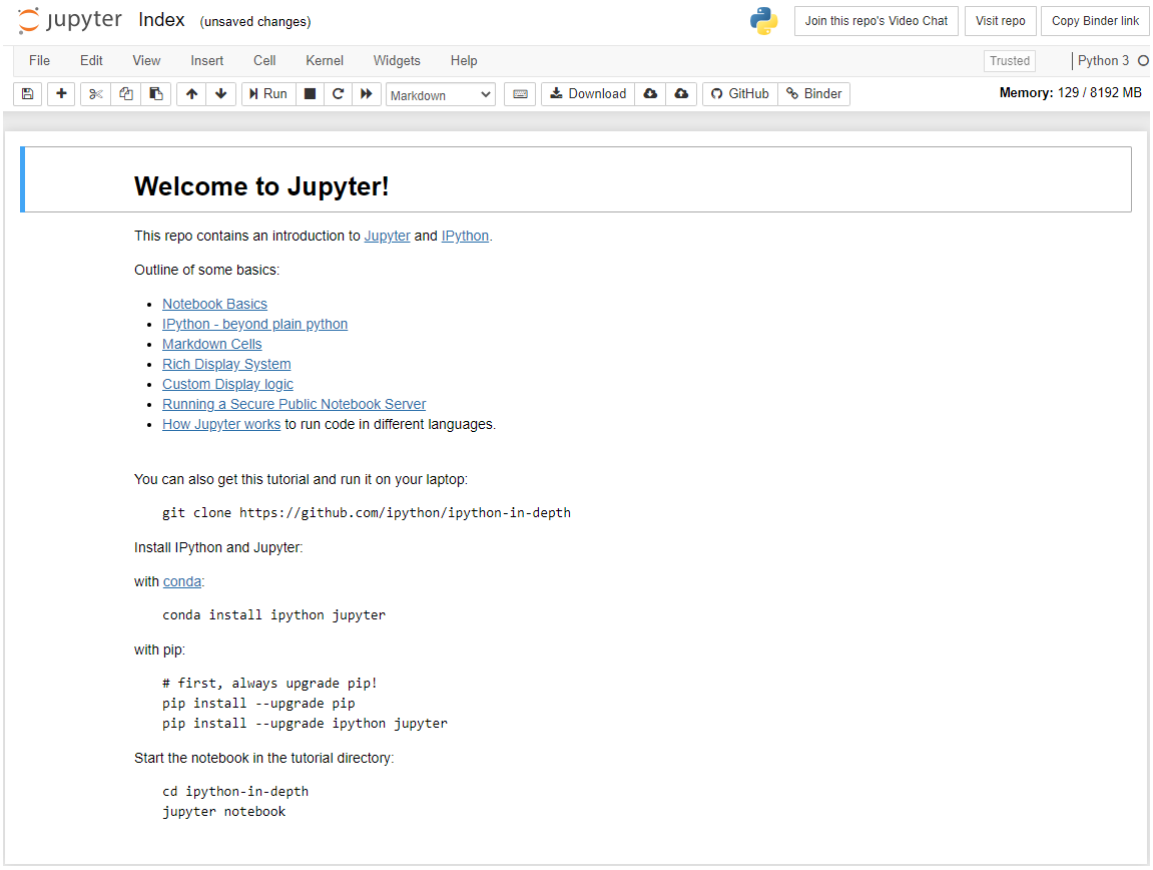
  Na Pythonu su dostupne različite mogućnosti vizualizacije. Njegova biblioteka Matplotlib pruža snažne temelje oko kojih se grade druge biblioteke, poput ggplot, pandas plotting, pytorch i druge.

Jupyter Notebook

Jupyter Notebook je web-aplikacija otvorenog koda, koja omogućuje stvaranje i razmjenu dokumenata koji sadrže kôd uživo, jednadžbe, vizualizacije i narativni tekst. Upotrebe uključuju: čišćenje i transformaciju podataka, numeričku simulaciju, statističko modeliranje, vizualizaciju podataka, strojno učenje i još mnogo toga.

Jupyter Notebook-u se može pristupiti online ili se može preuzeti program na računalo. Ako se instalira na računalo, korisnici se najčešće odluče instalirati Anacondu. Anaconda je najpopularnija Python distribucija za podatkovnu znanost i dolazi s unaprijed instaliranim svim najpopularnijim bibliotekama i alatima. Python biblioteke prisutne u Anacondi

uključuju NumPy, Pandas i Matplotlib, a čitav popis je veći od 1000 biblioteka. Važno je napomenuti da Jupyter notebook podržava razne programske jezike, uključujući i R.



Slika 1.2: Juptyer Notebook početna stranica, pristup preko web preglednika

1.4 R

R je programski jezik i okruženje za statističko računanje i grafiku. To je projekt GNU-a sličan jeziku S, kojeg su u Bell Laboratories (ranije AT&T, sada Lucent Technologies) razvili John Chambers i njegovi kolege. R se može smatrati različitim načinom realizacije jezika S. Postoje neke važne razlike, ali većina koda napisanog za S radi nepromijenjeno u R (vidi [5].

R pruža široku paletu statističkih tehnika (linearno i nelinearno modeliranje, klasična statistička ispitivanja, analiza vremenskih serija, klasifikacija, klasteriranje) i grafičkih tehnika, te je lako proširiv. Jezik S je često izbor kod istraživanja u statističkoj metodologiji, a R pruža put otvorenog koda za sudjelovanje u toj aktivnosti.

Jedna od prednosti R-a je lakoća s kojom se mogu proizvesti dobro dizajnirani grafovi, uključujući matematičke simbole i formule po potrebi. U mnogim aspektima Python i R dijele iste vrste funkcionalnosti, ali ih primjenjuju na različite načine. Ovisno o izvoru kojeg gledate, Python i R imaju približno isti broj zagovornika, a neki ljudi Python i R koriste naizmjenično (ili ponekad u tandemu). Za razliku od Pythona, R nudi svoje okruženje, tako da vam ne treba proizvod treće strane, kao što je Anaconda. No, R se ne miješa s drugim jezicima s lakoćom koju pruža Python.

R okruženje

R je integrirani paket softverske opreme za obradu podataka, proračun i grafički prikaz. Bitni dijelovi tog paketa su:

- Učinkovito sredstvo za obradu i pohranu podataka.
- Paket operatora za proračun na nizovima, posebno matrica.
- Velika, koherentna, integrirana zbirka posrednih alata za analizu podataka.
- Grafički uređaji za analizu podataka i prikaz na zaslonu ili na papirnoj kopiji.
- Dobro razvijen, jednostavan i učinkovit programski jezik, koji uključuje uvjetovanja, petlje, rekurzivne funkcije i mogućnosti ulaza i izlaza.

R ima značajne probleme s upravljanjem memorijom i radnim performansama, što je bitno za istraživače koji rade s vrlo velikim skupovima podataka. Ali mogućnosti jezika

nadmašuju problem. Paketi, poput dplyr i data.table, napisani su kako bi se R mogao nositi s izazovima rudarstva ogromnih skupova podataka.

Matematičari i statističari koji nemaju znanja iz programiranja imaju tendenciju koristiti R, u usporedbi s ostalim programskim jezicima. Jedan od glavnih razloga zbog kojeg je R bio popularan izvan akademske zajednice, je taj što ga je neprogramerima relativno lako učiti, ali zadržava velik dio snage namjenskog programskog jezika. Inženjerima, znanstvenicima i statističarima je relativno lako odabrati jezik, jer koristi pojmove općeg podrijetla koji su im poznati iz matematike i statistike.

R je najpopularniji izbor za podatkovne znanstvenike. Slijedi nekoliko ključnih razloga zašto:

- R je pouzdan i koristan u akademskoj zajednici dugi niz godina. Tradicionalno, R se koristio u istraživačke svrhe, jer je pružio različite statističke alate za analizu. S napretkom u podatkovnoj znanosti i potrebom za analizom podataka, R je postao popularan izbor i u industriji.
- R je idealan alat kada je u pitanju obrada podataka. Omogućuje upotrebu nekoliko unaprijed obrađenih paketa, što olakšava slaganje podataka. To je jedan od glavnih razloga zašto je R preferiran u zajednici podatkovnih znanstvenika.
- R pruža svoj poznati ggplot2 paket, koji je najpoznatiji po svojim vizualizacijama. Ggplot2 pruža estetske vizualizacije koje se bave svim podacima. Nadalje, ggplot2 pruža stupanj interaktivnosti korisnicima kako bi jasnije razumjeli podatke ugrađene u vizualizaciju.
- R sadrži pakete za strojno učenje za razne operacije. Bilo da se radi o poticanju, izgradnji slučajnih šuma ili regresiji i klasifikaciji, strojno učenje je podržano širokom paletom paketa.

Za početak rada s R-om potrebno je instalirati Rstudio IDE i sljedeće popularne pakete:

- dplyr, plyr i data.table za jednostavno upravljanje paketima,
- stringr za manipuliranje nizovima,
- zoo koji radi s redovitim i nepravilnim vremenskim serijama,
- ggvis, lattice i ggplot2 za vizualizaciju podataka i za strojno učenje.

1.5 Python vs R

Svaki od navedenih programskih jezika ima svojih prednosti i nedostataka. Iako je trenutno R popularniji izbor kod podatkovnih znanstvenika, Python ubrzano dobiva na popularnosti i uskoro bi mogao uzeti vodstvo. Najjednostavniji način po kojem izabrati jezik je proučiti što je točno potrebno za analizu, odnosno što je to bitno u analizi.

Kada koristiti R

R se uglavnom koristi kada zadatak analize podataka zahtijeva samostalno računanje ili analizu na pojedinim poslužiteljima. Odličan je za istraživački rad, a prikladan je za gotovo svaku vrstu analiza podataka, zbog ogromnog broja paketa i lako koristivih testova, koji često pružaju potrebne alate za brzo pokretanje i rad. R čak može biti dio rješenja velikih podataka.

Kada koristiti Python

Python se koristi kada zadatke za analizu podataka treba integrirati s web aplikacijama ili ako je statistički kôd potrebno ugraditi u produkcijsku bazu podataka. Budući da se radi o programskom jeziku, izvrstan je alat za implementaciju algoritama u produkciju.

	Prednosti	**Nedostaci**
R	Vizualizacija R okruženje Nije bitno znati programirati	Upravljanje memorijom Radne performanse Može biti kompliciran za naučiti ako ne postoji predznanje u statistici
Python	Juptyer Notebook Višenamjenski jezik Vizualizacija	Prevelik izbor biblioteka (osjećaj preplavljenosti) Potrebno predznanje u programiranju

Tablica 1.2: Nedostaci i prednosti programskih jezika R i Python

Poglavlje 2

Data science u edukaciji

Podatkovna znanost se već koristi u edukaciji na način da pomaže definirati probleme, analizirati podatke i pretpostaviti teze kako poboljšati trenutnu nastavu. No, u ovom dijelu, nije naglasak istražiti kako podatkovna znanost može pomoći edukaciji, nego kako upoznati učenike s njom.

Razloga za to je mnogo, a jedan od najvažnijih je to što su upravo podaci okarakterizirani kao ulje 21. stoljeća. Odnosno, podaci su svuda oko nas i bitno ih je znati razaznati, pročitati i analizirati. Upravo je i to jedan od razloga što su u novom kuriukulumu iz matematike, već od prvog razreda osnovne škole, u ishodima uključeni i prikazi podataka:

MAT OŠ E.1.1. Služi se podatcima i prikazuje ih piktogramima i jednostavnim tablicama.	Određuje skup prema nekome svojstvu. Prebrojava članove skupa. Uspoređuje skupove. Prikazuje iste matematičke pojmove na različite načine (crtež, skup, piktogram i jednostavna tablica). Čita i tumači podatke prikazane piktogramima i jednostavnim tablicama. Prošireni sadržaji: Prikazivanje podataka različitih nastavnih predmeta. Korelacija s Hrvatskim jezikom, Prirodom i društvom, međupredmetnim temama Učiti kako učiti i Poduzetništvo.	Čita i prikazuje podatke piktogramima.

Sadržaj: Čitanje, tumačenje i prikazivanje podataka. Piktogrami i jednostavne tablice.
Prošireni sadržaj: Prikazivanje podataka različitih nastavnih predmeta.

Slika 2.1: Isječak iz kurikuluma iz predmeta Matematika (vidi [2])

Cilj navedenih aktivnosti je osvijestiti učenika što je to podatkovna znanost i neki najvažniji pojmovi vezani uz nju, kao što su veliki podaci, analiza i vizualizacija.

2.1 Informatika

U informatici se podatkovna znanost može ukomponirati već u osnovnoj školi. Sve aktivnosti koje uključuju analiziranje podataka iz Excela, već mogu pripadati aktivnostima vezanima uz podatkovnu znanost. Uz Excel, učenici po novom kurikulumu uče i Python pa se mogu odraditi jednostavne aktivnosti koje uključuju Jupyter Notebook. Osim aktivnosti koje se mogu provesti na satu, podatkovna znanost je izvrstan izvor tema koje učenici mogu obraditi u sklopu svojih projektnih zadataka.

Prva aktivnost se može provesti na dodatnoj nastavi u osnovnoj školi, ali i srednjoj. Prije provedbe aktivnosti bitno je učenicima predstaviti Jupyter Notebook i pomoći učenicima da se povežu na `https://colab.research.google.com/`, koristeći google korisničke račune. To omogućuje zajednički rad na jednom dokumentu, odnosno omogućuje pristup učenicima već definiranim dokumentima koje je nastavnik/učitelj pripremio.

Aktivnost #1 – Ruke i prsti

Cilj aktivnosti: Upoznati učenike s osnovama Jupyter Notebook-a, osnovnim okvirima i prikazom podataka.

Nastavni oblik: Diferencirana nastava u obliku individualnog rada

Nastavna metoda: Metoda demonstracije, Metoda rada s tekstom

Potrebni materijali: pristup `https://colab.research.google.com`

Pristup digitalnom listiću na adresi: `https://colab.research.google.com/drive/1lBYSLFc5OHglPF-LpLDxLQ2do0MC1Lvr?usp=sharing`

Tijek aktivnosti:

Na početku učitelj/ica, odnosno nastavnik/ica kratko objašnjava što je to Jupyter Notebook i pomaže učenicima pristupiti colab.research.google.com. Na Padletu[1] ili nekom drugom alatu, objavljuje link na dokument koji je kreirao/la ranije. Dokument je izrađen u Jupyter Notebooku i koncipiran je kao nastavni listić.

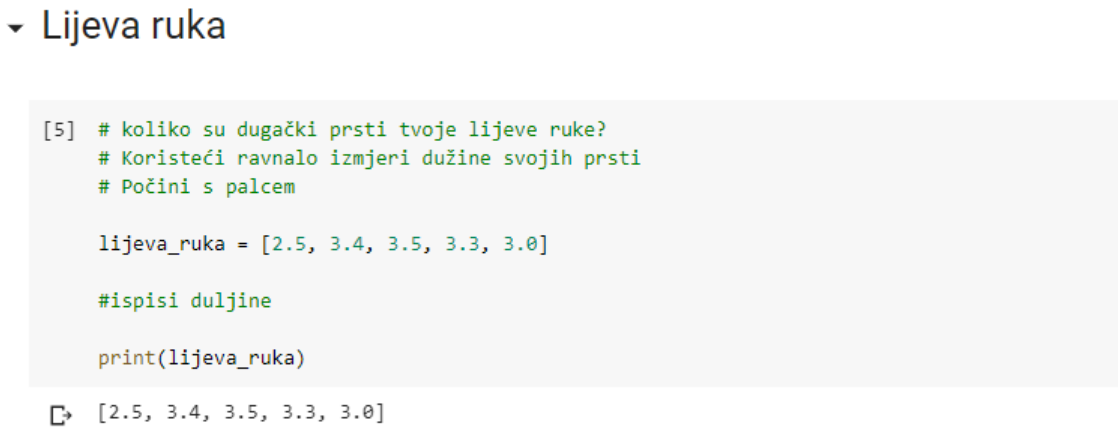
[1]Padlet je internetska virtualna oglasna ploča na sigurnoj lokaciji, na kojoj učenici i nastavnici mogu surađivati, komunicirati, dijeliti veze i slike.

Zadatak koji učenici imaju je izmjeriti prste na lijevoj i desnoj ruci, unesti izmjerene podatke i napraviti jednostavnu analizu.

Prolazeći kroz "nastavni listić", prolaze kroz sve navedene korake.

Slika 2.2: Nastavni listić – 1. dio

Prvo što vide je dio gdje se unosi biblioteka pandas. Učitelj/ica ovdje može objasniti što je to biblioteka, ako učenici nisu upoznati, i dodatno pojasniti što sadrži pandas.



Slika 2.3: Nastavni listić – 2. dio

Zatim unose podatke za lijevu ruku. U komentarima su napisana pitanja na koja moraju dobiti odgovore. U slučaju da se aktivnost obrađuje u srednjoj školi, nastavnik može pustiti učenike da samostalno odrade cijeli listić, dok bi u osnovnoj školi bilo poželjno prolaziti s učenicima svaki korak i objasniti što koji dio koda radi.

```
[ ]  # pretvori u podatkovni okvir (eng. dataframe)

     lijeva_ruka_df = pandas.DataFrame(lijeva_ruka, columns=['Duljina'])

     lijeva_ruka_df
```

	Duljina
0	2.5
1	3.4
2	3.5
3	3.3
4	3.0

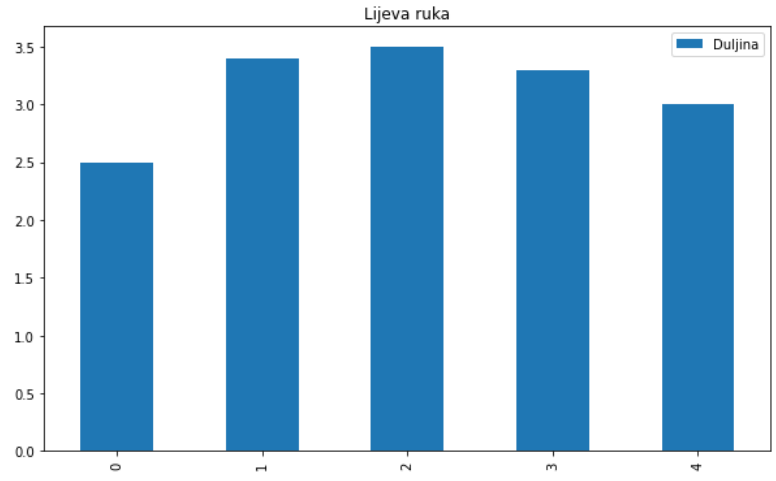
Slika 2.4: Nastavni listić – 3. dio

U sljedećem koraku se upoznaju s izrazom podatkovni okvir (eng. dataframe), odnosno kreiraju tablicu podataka za upisane duljine prstiju. Podatkovni okvir je dvodimenzionalna podatkovna struktura. Podaci u tablicama su poredani u redove i stupce.

```
# prikažimo graf duljine tvojih prsti

lijeva_ruka_df.plot(kind="bar", title="Lijeva ruka", figsize=(10,6))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9deb8dd2b0>
```



Slika 2.5: Nastavni listić – 4. dio

Na kraju crtaju graf, odnosno vizualiziraju podatke. Tu se može provesti diskusija s učenicima o tome što mogu pročitati iz grafa. Isti postupak ponavljaju za obje ruke.

Nakon što su unijeli potrebne podatke, cilj je s učenicima provesti i neku jednostavnu analizu. Potrebno je izračunati najveću, najmanju i prosječnu dužinu. Kod izračuna prosjeka vidljivo je da se koristi mean(), odnosno računa se medijan. Ako učenici nisu upoznati s navedenim izrazom može se zadržati na izračunavanju prosjeka.

Analizirajmo podatke

```
[11] # Koja je najveća dužina prsta na tvojoj lijevoj ruci?

     lijeva_ruka_df.max()
```

```
length    3.5
dtype: float64
```

```
[12] # Koja je najmanja dužina prsta na tvojoj desnoj ruci?

     desna_ruka_df.min()
```

```
length    2.4
dtype: float64
```

```
[13] # Koja je prosječna dužina prsti na lijevoj ruci?

     lijeva_ruka_df.mean()
```

```
length    3.14
dtype: float64
```

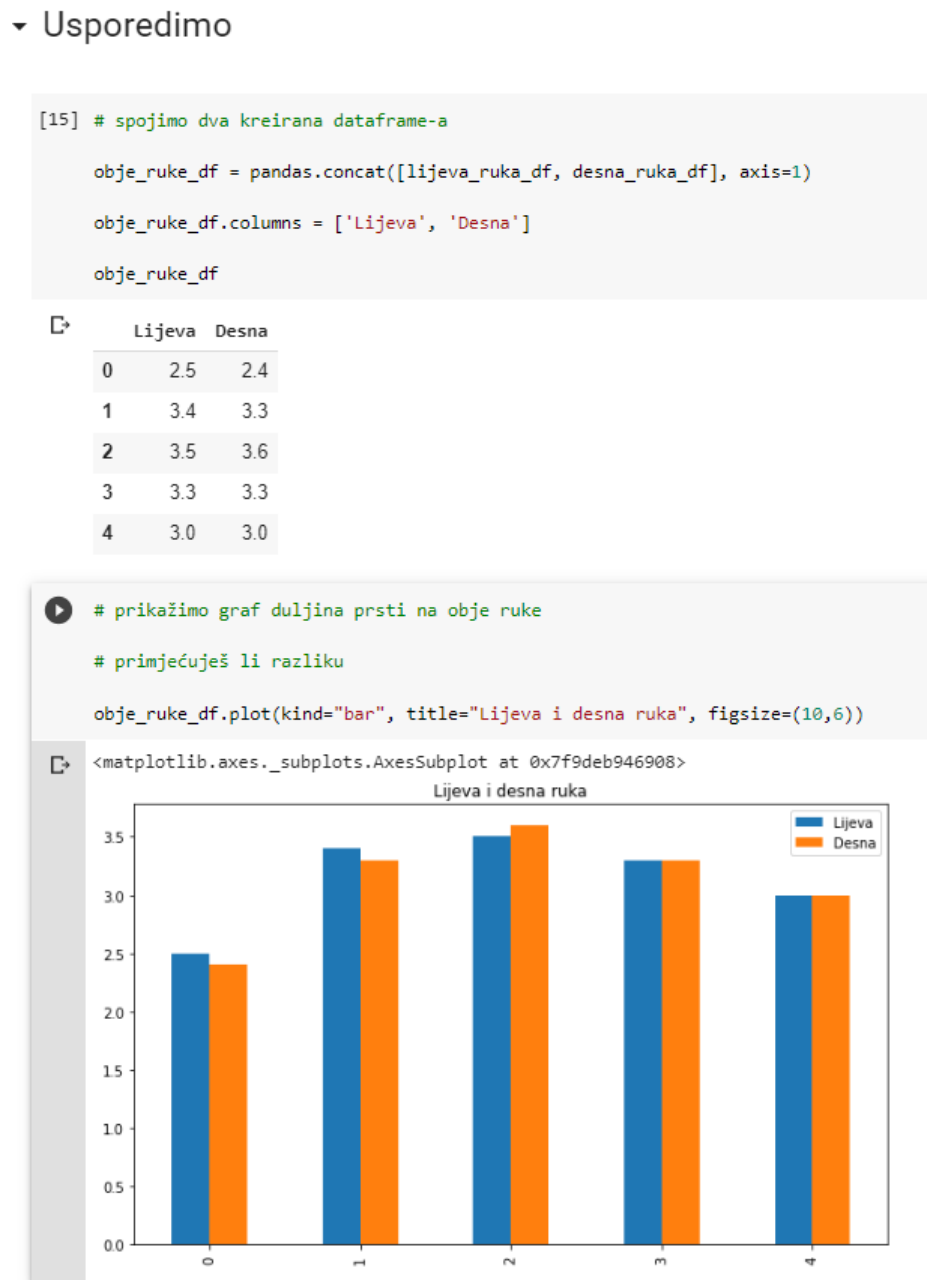
```
[14] # Koja je prosječna dužina prsti na desnoj ruci?


     desna_ruka_df.mean()
```

```
length    3.12
dtype: float64
```

Slika 2.6: Nastavni listić – 5. dio

Na kraju moraju usporediti podatke lijeve i desne ruke. Bitno je na kraju aktivnosti s učenicima provesti diskusiju. Proći kroz sve grafove, analizirati što primjećuju iz njih te odgovoriti na sva pitanja ako učenicima nešto nije jasno.



Slika 2.7: Nastavni listić – 6. dio

Druga aktivnost je primjerena za srednju školu. Može se provesti na dodatnoj nastavi ili u drugom razredu prirodoslovno-matematičke gimnazije, u sklopu cjeline gdje se obrađuju biblioteke.

Aktivnost #2 – Google dionice

Cilj aktivnosti: Koristeći biblioteke kreirati jednostavnu web-aplikaciju.

Nastavni oblik: Diferencirana nastava u obliku individualnog rada

Nastavna metoda: Metoda demonstracije, Metoda rada s tekstom

Potrebni materijali: nastavni listić

Tijek aktivnosti:

Na početku aktivnosti nastavnik/ica postavlja pitanja učenicima:

- Jeste čuli do sada za *Data science*?
- Možete li pretpostaviti što podatkovni znanstvenik radi?

Navedenim pitanjima potiče se diskusija o podatkovnoj znanosti i budi znatiželja među učenicima. Zatim učenici, preko Padleta ili nekog drugog alata, preuzimaju dvije datoteke. Jedna je običan *pdf*, dok je druga program napisan u Pythonu.

Učenici će, prolazeći kroz *pdf*, naučiti što napisani program radi, pokrenuti program i time kreirati svoju prvu web-aplikaciju vezanu uz podatkovnu znanost.

Aplikacija prikazuje cijene dionica i njihovu količinu. Za izradu takve aplikacije potrebno je negdje preuzeti podatke. Koristeći biblioteku yfinance dolazi se do potrebnih podataka iz Yahoo! Finance[2]. Podaci su spremljeni u kreirane podatkovne okvire i, koristeći biblioteku streamlit, podaci se prikazuju grafički.

Prvo pitanje na koje učenici moraju odgovoriti iz nastavnog listića je definiranje biblioteka yfinance i streamlit. Odnosno, zadatak im je istražiti što točno te biblioteke rade. Tim se potiče samostalni rad i uče koristiti pretraživanje interneta na koristan način.

[2]Yahoo! Financije su medijsko vlasništvo koje je dio Yahoo!. Pruža financijske vijesti, podatke i komentare, uključujući kotacije dionica, priopćenja za javnost, financijska izvješća i originalni sadržaj.

Sljedeći korak im je da otvore preuzeti program i promotre kôd. Zatim, kroz detaljan opis kôda, samostalno prođu kroz svaku liniju.

```
1  import yfinance as yf
2  import streamlit as st
3
4  st.write("""
5  # Moja prva web aplikacija
6  Prikazane su krajnje cijene i koli ina dionica od Googlea!
7  """)
8
9  #definraj simbol za dionice
10 tickerSymbol = 'GOOGL'
11 #preuzmi podatke
12 tickerData = yf.Ticker(tickerSymbol)
13 #kreiraj dataframe
14 tickerDf = tickerData.history(period='1d', start='2010-5-31', end='
       2020-5-31')
15
16
17 st.line_chart(tickerDf.Close)
18 st.line_chart(tickerDf.Volume)
```

Listing 2.1: Programski kôd web-aplikacije

Linije 1–2:

U prve dvije linije kôda se uvoze biblioteke yfinance i stremalit, te im se zadaju nadimci yf i st.

Linije 4–7:

Koristeći st.write() funkciju ispisuje se tekst, u ovom slučaju radi se o naslovu aplikacije.

Linije 9–14:

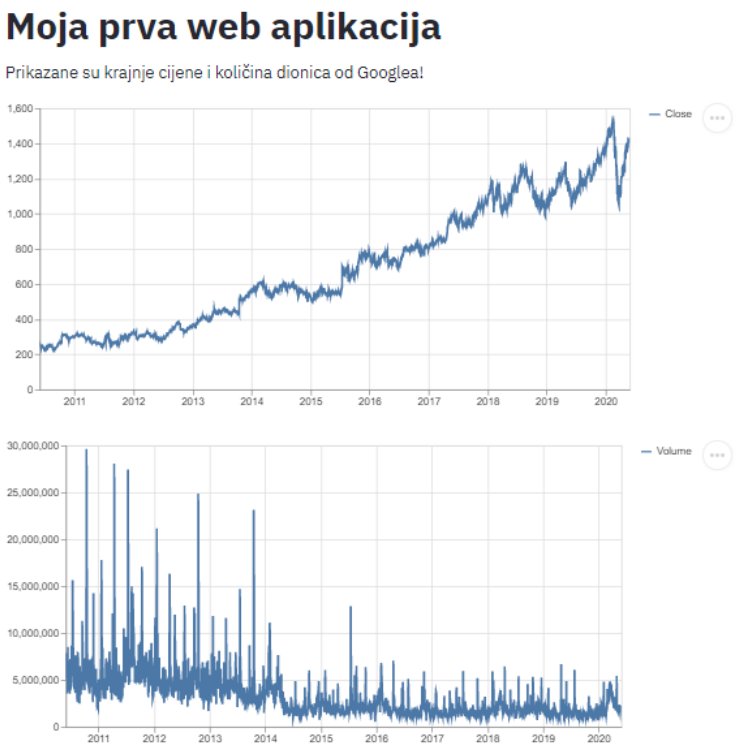
Koristi yfinance biblioteku za preuzimanje povijesnih tržišnih podataka s Yahoo! Financije.
Linija 10 – Definira simbol kao GOOGL. To je ime firme za koju želimo gledati podatke.
Linija 12 – Stvara varijablu tickerData korištenjem funkcije yf.Ticker() koja pristupa podacima o dionicama.
Linija 14 – Stvara tickerDf podatkovni okvir i definira datumski raspon (od 31. svibnja 2010. do 31. svibnja 2020.) i vremensko razdoblje (1 dan).

Linije 17–18:

Koristi funkciju st.line_chart() za crtanje linijskog grafikona, koristeći cijenu iz podatkovnog okvira tickerDf, kako je definirano u retku 14.

Nakon što su samostalno proučili kôd, nastavnik provodi diskusiju s učenicima. Prolaze kroz sve nedoumice i pitanja, ako im nešto nije bilo jasno. Na kraju učenici pokreću aplikaciju, prateći upute ispisane na listiću:

1. Otvori cmd ili PowerShell.
2. Upiši naredbu: streamlit run ime-aplikacije.py.
3. Ako se automatski ne otvori, pristupi sljedećoj adresi: `http://localhost:8501`.



Slika 2.8: Web-aplikacija

Na kraju aktivnosti se proučavaju podaci zajedno s učenicima. Bitno je donesti neke zaključke, kako bi učenici shvatili korisnost. Aktivnost se dodatno može proširiti tako da se aplikacija dodatno uredi. Navedeno se može ostaviti i kao domaća zadaća.

2.2 Matematika

Podatkovna znanost je usko vezana uz statistiku. U kurikulumu matematike (vidi [2]) jedna od 5 glavnih domena se zove Podaci, statistika i vjerojatnost. Navedeno je da se ta domena bavi prikupljanjem, razvrstavanjem, obradom, analizom i prikazivanjem podataka u odgovarajućemu obliku. Podatke dane grafičkim ili nekim drugim prikazom treba znati očitati te ih ispravno protumačiti i upotrijebiti. Sve se to postiže koristeći se jezikom statistike. Ona podrazumijeva uporabu matematičkoga aparata kojim se računaju mjere srednje vrijednosti, mjere raspršenja, mjere položaja i korelacije podataka.

Već od prvog razreda osnovne škole uvodi se piktogram te se podaci prikazuju u jednostavnim tablicama. U sljedećoj tablici su navedeni svi ishodi relevatni za podatkovnu znanost od 5. razreda osnovne škole do kraja školovanja.

MAT OŠ E.5.1 Barata podacima prikazanim na različite načine.	Tumači prikaz podataka tablicama, slikama, listama te različitim grafovima i dijagramima.
MAT OŠ E.6.1 Prikazuje podatke tablično te linijskim i stupčastim dijagramom frekvencija.	Određuje frekvencije razvrstanih podataka potrebne za grafički prikaz. Prikupljene podatke prikazuje linijskim dijagramom frekvencija.
MAT OŠ E.7.1 Organizira i analizira podatke prikazane dijagramom relativnih frekvencija.	Određuje relativne frekvencije razvrstanih podataka potrebne za grafički prikaz. Prikupljene podatke prikazuje stupčastim dijagramom relativnih frekvencija i tumači prikaz.
MAT OŠ E.8.2 Interpretira podatke povezane s novcem te na osnovi toga donosi odluke.	Opisuje pojam kamate na štednju i kamate na kredit na primjeru iz stvarnoga života. Uspoređuje i tumači kamate na stambeni i gotovinski kredit.
MAT SŠ E.1.2 Barata podatcima prikazanima na različite načine.	Prikazuje podatke tablično, stupčastim dijagramom, histogramom, dijagramom stablo – list, linijskim dijagramom. Određuje srednje vrijednosti: mod, medijan, donji i gornji kvartil te standardnu devijaciju. Crta brkatu kutiju.

Tablica 2.1: Ishodi iz domene Podaci, statistika i vjerojatnost

Prva aktivnost je primjerena za osnovnu školu. Može se obraditi na dodatnoj nastavi ili u regularnoj nastavi, kada se obrađuju različiti prikazi podataka. Aktivnost je moguće provesti u običnoj i u računalnoj učionici. Ako se aktivnost provodi u običnoj učionici, potrebno je imati barem jedno računalo (nastavničko) kako bi se mogli prikazati podaci.

Aktivnost #1 – M&M's

Cilj aktivnosti: Provesti prikupljanje podataka te ih prikazati na različite načine.

Nastavni oblik: Diferencirana nastava u obliku grupnog rada

Nastavna metoda: Metoda demonstracije, Metoda rada s tekstom

Potrebni materijali: PC računalo, nekoliko vrećica bonbona

Tijek aktivnosti:

Na početku aktivnosti potrebno je podijeliti učenike na nekoliko grupa, i to tako da u svakoj grupi bude najviše petero učenika. Ako se aktivnost provodi u računalnoj učionici, svaka grupa radi na jednom računalu i rade u proračunskoj tablici. U nastavku će biti opisano kako bi se aktivnost izvela u običnoj učionici.

Svaka grupa dobiva vrećicu šarenih bonbona, na primjer M&M's. Prva uputa koju učenici dobivaju je da prebroje koliko su dobili bonbona od svake boje i to zapišu na papir. Nakon što je svaka grupa prikupila podatke za svoju vrećicu bonbona, jedan od učenika iz svake grupe čita dobivene rezultate, dok ih nastavnik upisuje u proračunsku tablicu na nastavničkom računalu. Učenici prate putem projektora. Na kraju tablica može izgledati kao što je prikazano na slici 2.9.

Boja bonbona:	Grupa 1:	Grupa 2:	Grupa 3:	Grupa 4:	Ukupno:
	15	16	12	15	58
	17	17	14	13	61
	16	15	18	16	65
	12	13	13	17	55
	15	15	15	13	58
	18	17	19	18	72
Ukupno:	93	93	91	92	**369**

Slika 2.9: Primjer tablice

Sljedeći korak je odgovoriti na pitanja. Svaka grupa dobiva jedno pitanje i kroz diskusiju moraju doći do zaključka na koji način bi došli do odgovora, odnosno koji grafički prikaz bi im najviše olakšao pronalazak odgovora. Bitno je učenike usmjeriti da odgovore na pitanja koristeći grafičke prikaze. Pitanja su:

1. Koja boja se pojavljuje najviše?
2. Koja boja se pojavljuje najmanje?
3. Varira li totalni broj bonbona u vrećicama?
4. Varira li broj svake boje u vrećicama?

Prolazeći kroz svako pitanje po redu, nastavnik prikazuje podatke pomoću grafikona koje su učenici predložili.

1. *Koja boja se pojavljuje najviše?*

   Učenici tu mogu predložiti stupičasti ili tortni prikaz. Nastavnik postavlja podpitanja vezana uz nepredloženi prikaz kako bi učenici osvijestili i drugu mogućnost. Nakon kreiranja grafikona odgovaraju na postavljeno pitanje.

2. *Koja boja se pojavljuje najmanje?*

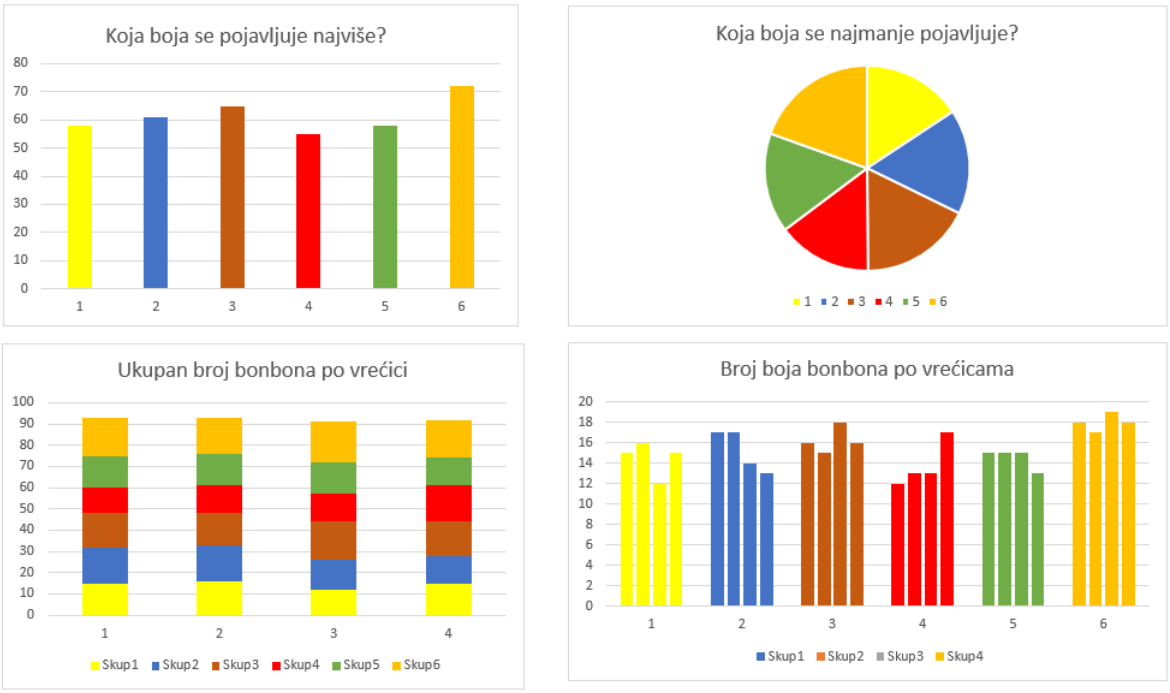
   Kao i u prvom pitanju, učenici mogu predložiti stupičasti ili tortni prikaz. U ovom slučaju nastavnik neka sam predloži prikaz drugačiji nego kod prvog pitanja. Nastavnik navodi učenike da odluče preko kojeg prikaza im je lakše odgovoriti na prvo i drugo pitanje.

3. *Varira li totalni broj bonbona u vrećicama?*

   Učenici kod ovog pitanja mogu postaviti pitanje zašto uopće to prikazati grafički, kada je u tablici vidljiv ukupan broj bonbona po vrećici. Bitno je objasniti učenicima da su podaci na kojima oni trenutno rade izuzetno maleni, neka zamisle kako bi izračunali traženo ako imaju 10 000 različitih vrećica i preko 100 boja. Podaci se prikazuju sa složenim stupičastim grafikonom i učenici, promatrajući graf, dolaze do zaključka.

4. *Varira li broj svake boje u vrećicama?*

   Četvrto pitanje se zapravo može analizirati na sličan način kao i treće pitanje. Ovdje je najbolje koristiti grupirani stupičasti grafikon.



Slika 2.10: Grafikoni kao odgovori na pitanja

Na kraju aktivnosti važno je provesti diskusiju s učenicima. Pitanja koja se mogu postaviti su:

- Je li stupičasti ili torta dijagram bio korisniji u odgovoru koje se boje najčešće i najmanje pojavljuju? Zašto? Što je učinilo jednu vrstu grafikona korisnijom od druge?
- Koliko smo sigurni u odgovore koji se dobiju iz vizualizacija? Postoje li okolnosti zbog kojih smo mogli donijeti pogrešne zaključke?
- Koje bi dodatne informacije ili podaci mogli biti korisni u poboljšanju točnosti zaključaka?
- Postoje li načini za poboljšanje vizualizacija? Postoji li bolji način prikazivanja podataka, koji bi olakšao donošenje zaključaka?

Druga aktivnost se može provesti u srednjoj školi ili na dodatnoj nastavi u osnovnoj školi. Kod ove aktivnosti je bitno da učenici imaju pristup grafičkom kalkulatoru kako bi mogli prikazati podatke.

Aktivnost #2 – Broj stranica

Cilj aktivnosti: Prikazati podatke grafički i naučiti kako koristiti predikcije.

Nastavni oblik: Diferencirana nastava u obliku grupnog rada

Nastavna metoda: Metoda demonstracije, Metoda rada s tekstom

Potrebni materijali: grafički kalkulator

Tijek aktivnosti:

Na početku aktivnosti potrebno je podijeliti učenike u parove. Svaki par dobiva upute ispisane na papiru:

- Koristite knjigu s više od 100 stranica i neka prvi učenik odabere broj.
- Drugi učenik pokušava otvoriti knjigu što je moguće bliže stranici koja odgovara odabranom broju.
- Zabilježite razliku broja stranica do odabranog broja.
- Provedite 7 pokusa i promatrajte uzorke u podacima.
- Predvidite rezultat za sljedećih 7 pokusa i napravite popis podataka prvih ispitivanja i predviđenih pokusa.
- Prikažite grafički oba podatka.

Cilj aktivnosti je da svaki igrač bilježi broj stranice, svoja nagađanja i njihovu pogrešku ili broj stranica za koliko je udaljen od zadanog broja. Nakon sedam igara, učenici analiziraju svoje podatke kako bi odgovorili na dva glavna pitanja:

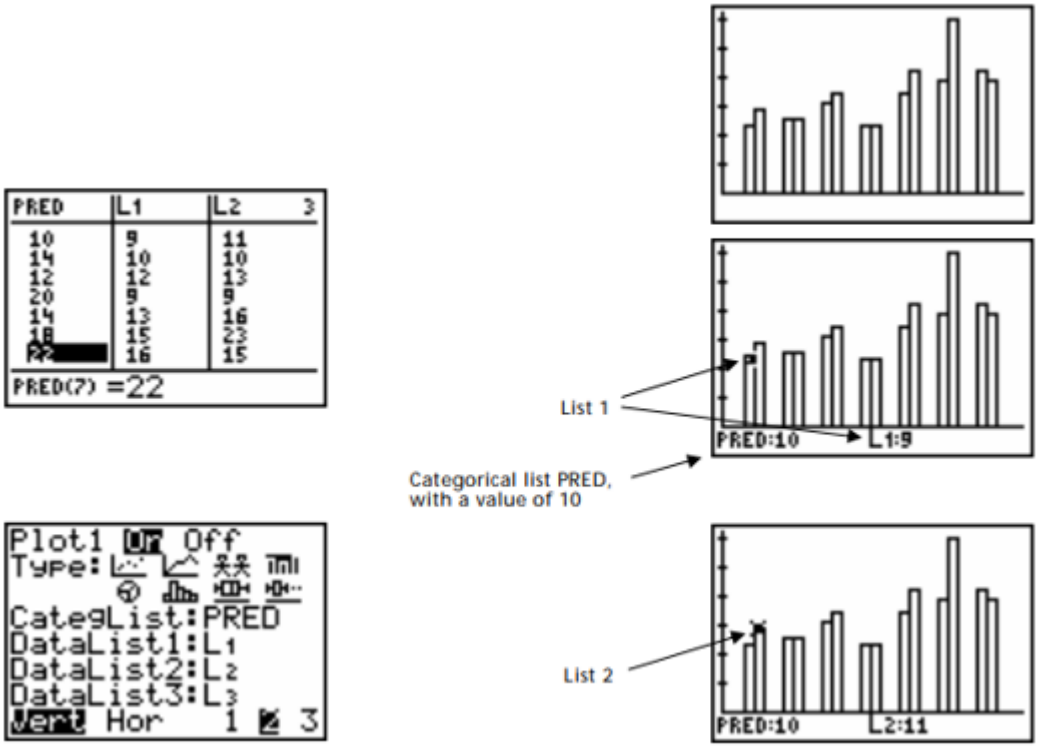
1. Poboljšava li se učenik u pogađanju?
2. Kako rezultati to podržavaju?

Kada učenici prođu kroz sve korake, nastavnik postavlja dodatna pitanja:

- Koju sliku možete opisati iz rezultata, bilo u tablici ili na grafovima?
- Poboljšavate li se? Koristeći statistiku i grafikone, opravdavajte bilo kakve tvrdnje o tome poboljšavate li se ili ne. Pravite li više grešaka kada je navedeni broj veći? Kako vaši podaci podržavaju vaš odgovor?
- Koje trendove primjećujete?
- Možete li predvidjeti koje ćete poboljšanje postići u sljedećih pet igara?

Na zadnja dva pitanja učenici bi trebali koristiti grafički kalkulator, odnosno putem njega mogu vidjeti neke trendove i grafički prikazati podatke. Upute za izradu kako napraviti dvostruki stupičasti grafikon na kalkulatoru TI-73:

1. U List editoru, umetnite novu listu imena PRED, u koju će se upisivati predviđanja.
2. Unesite predviđanja u listu za svih 7 testova. Unesite razlike u broju stranica između predviđanja i stvarnog suđenja za prvu igru u L1 i razlike za drugu igru u L2.
3. Pristupite izborniku [PLOT] i izaberite 1:Plot1.
4. Pritisnite tipku desno kako biste izabrali ON. Postavite svojstva kalkulatura kako su prikazana na slici 2.11.
5. Pristisnite GRAPH kako bi se prikazao graf.
6. Pristisnite TRACE.
7. Promatrajte vrijednost popisa PRED, prikazana je u donjem lijevom kutu zaslona.
8. Pritisnite tipku za desno, za usporedbu te vrijednosti s vrijednostima u L1 i L2.



Slika 2.11: Prikaz grafova na kalkulatoru T1-73 (vidi [1], str. 66)

Učenici će imati razne pristupe za odgovore na dva postavljena pitanja: 1) Poboljšavate li se? i 2) Kako rezultati to podržavaju? No, treba ih podsjetiti da moraju koristiti statistike i grafikone dobivene iz svojih igara, kako bi podržali svoje odgovore na pitanja.

Ova aktivnost samo je jedan od primjera postavljanja zanimljivih pitanja koja uključuju učenike u prikupljanje i analizu podataka. Učenike treba poticati na generiranje drugih pitanja koja bi mogla biti zanimljiva, a koja bi mogla dovesti do prikupljanja podataka, koristeći grafikone za interpretaciju podataka i analizu rezultata, kako bi odgovorili na postavljena pitanja.

Bibliografija

[1] *Integrating Handheld Technology Into the Elementary Mathematics Classroom*. Texas Instruments, 2002.

[2] *Kurikulum nastavnog predmeta Matematika*. Ministarstvo znanosti i obrazovanja, 2019.

[3] *Kurikulum nastavnog predmeta Informatika*. Ministarstvo znanosti i obrazovanja, 2020.

[4] EMC Education Services: *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley, 2015.

[5] Gardener, Mark: *Beginning R: The statistical programming language*. John Wiley & Sons, 2012.

[6] Grus, Joel: *Data science from scratch: first principles with python*. O'Reilly Media, 2019.

[7] Mueller, John Paul i Luca Massaron: *Python for data science for dummies*. John Wiley & Sons, 2019.

[8] O'Neill, Cathy i Rachel Schutt: *Doing data science: Straight talk from the frontline*. O'Reilly Media, Inc., 2013.

[9] Srikant, Shashank i Varun Aggarwal: *Introducing data science to school kids*. U *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, stranice 561–566, 2017.

[10] Voulgaris, Zacharias: *Data scientist: the definitive guide to becoming a data scientist*. Technics Publications, 2014.

Sažetak

Ovaj diplomski rad ima dva dijela. U prvom dijelu, opisano je što je znanost o podacima (eng. Data Science), gdje se koristi i zašto, s posebnim naglaskom na tzv. dubinsku analizu podataka. Također, detaljno je objašnjeno korištenje programskih jezika Python i R za tu svrhu, te analizirano koje su prednosti, nedostaci i razlike u korištenju jednog ili drugog. Drugi dio rada posvećen je korištenju metoda analize i rudarenja podataka u edukaciji. Konkretno, na koji način se ove metode mogu koristiti za poboljšanje nastave (posebno, nastave informatike) te prijedlog na koji način se "Data Science" može ukomponirati u trenutni kurikulum iz informatike i matematike.

Summary

This thesis has two parts. In the first part, it is described what Data science is, where it is used and why, with special emphasis on the so-called in-depth data analysis. Also, the use of programming languages Python and R is explained, and the advantages, disadvantages and differences in using one or the other are analyzed. The second part of the paper is dedicated to the use of data analysis and mining methods in education. In particular, how can these methods be used to improve teaching (in particular, teaching computer science) and a proposal on how ”Data Science” can be incorporated into the current curriculum in computer science and mathematics.

Životopis

Rođena sam 02.11.1992. u Zagrebu. Pohađala sam opću gimnaziju u Bjelovaru, koju sam završila 2011. Iste godine sam upisala preddiplomski studij na Prirodoslovno-matematičkom fakultetu, gdje sam 2017. stekla titulu sveučilišne prvostupnice edukacije matematike. Nakon godinu dana pauze, upisala sam diplomski studij Matematika i informatika, nastavnički smjer. Kroz cijelu svoju studensku karijeru sam radila razne poslove, primarno u IT industriji.