

SVEUČILIŠTE U ZAGREBU  
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 5052

# **Primjena regresijske analize u edukacijskoj domeni**

Marin Krešo

Zagreb, lipanj 2017.

Zagreb, 8. ožujka 2017.

## ZAVRŠNI ZADATAK br. 5052

Pristupnik: **Marin Krešo (0036483331)**  
Studij: Računarstvo  
Modul: Računarska znanost

Zadatak: **Primjena regresijske analize u edukacijskoj domeni**

### Opis zadatka:

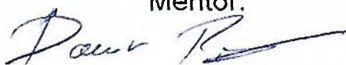
Znanstveno polje EDM (engl. Educational Data Mining - rudarenje podataka u edukaciji) bavi se primjenom metoda i tehnologija rudarenja podataka u nastavnom okruženju. To je relativno novo, interdisciplinarno polje koje se orijentira na analitiku specifičnih tipova podataka vezanih uz edukaciju. Cilj je bolje razumijevanje procesa usvajanja znanja, prepoznavanje svojstava okruženja u kojem se uči te dobivanje uvida u nastavne fenomene.

Vaš zadatak je proučiti metode regresijske analize i na studijskom primjeru prikazati slučaj uporabe i učinkovitost istih nad realnim podacima iz edukacijske domene.

Zadatak uručen pristupniku: 10. ožujka 2017.

Rok za predaju rada: 9. lipnja 2017.

Mentor:



Doc. dr. sc. Damir Pintar

Djelovođa:



Doc. dr. sc. Tomislav Hrkać

Predsjednik odbora za  
završni rad modula:



Prof. dr. sc. Siniša Srbljić

*Zahvaljujem se mentoru doc.dr.sc. Damiru Pintaru na pruženoj pomoći i savjetima koji su bili od velike pomoći pri izradi ovog završnog rada. Također sam mu zahvalan jer mi je omogućio da samostalno donosim većinu važnih odluka vezanih uz realizaciju ovog zadatka, zbog čega sam naučio nove načine pristupa i rješavanja problema. Zahvaljujem se i ZPM-u jer mi je pružio stvarne podatke za potrebe mog rada. Zahvaljujem se i svim autorima literature i algoritama koje sam koristio, jer su svoje radove i dostignuća učinili javno dostupnima za korištenje. Najviše se zahvaljujem mojoj obitelji jer me podržavaju i pomažu mi pri studiranju.*

# SADRŽAJ

<b>1. Edukacijska domena</b>	<b>1</b>
1.1. EDM . . . . .	1
1.2. Povijest EDM-a . . . . .	2
1.3. Proces istraživanja edukacijskih podataka . . . . .	2
1.3.1. Edukacijska okolina . . . . .	2
1.3.2. Pretprocesiranje . . . . .	2
1.3.3. Dubinska analiza podataka . . . . .	3
1.3.4. Interpretacija rezultata . . . . .	3
1.4. Metode EDM-a . . . . .	3
1.4.1. Predikcija . . . . .	4
1.4.2. Grupiranje(engl. clustering) . . . . .	4
1.4.3. Detekcija stršćih vrijednosti . . . . .	4
1.4.4. Analiza i obrada teksta . . . . .	5
1.5. Faze EDM-a . . . . .	5
<b>2. Strojno Učenje</b>	<b>6</b>
2.1. Zašto strojno učenje ? . . . . .	6
2.2. Koraci strojnog učenja . . . . .	7
2.3. Metode i tehnike strojnog učenja . . . . .	8
2.3.1. Linearna regresija . . . . .	8
2.3.2. Polinomijalna regresija . . . . .	12
2.3.3. Regresijska stabla i model-stabla . . . . .	12
2.3.4. Odabir varijabli . . . . .	13
<b>3. Prediktivna analiza podataka</b>	<b>15</b>
3.1. Treniranje i predikcija . . . . .	15
3.2. Prenaučenost . . . . .	15
3.3. Unakrsna provjera . . . . .	15

3.4. Odabir metode strojnog učenja . . . . .	17
<b>4. Studijski primjer</b>	<b>18</b>
4.1. Ciljevi istraživanja . . . . .	18
4.2. Odabir alata za analizu . . . . .	19
4.3. Prikupljanje podataka . . . . .	19
4.4. Pripremanje i pretprocesiranje podataka . . . . .	19
4.5. Analiza podataka i primjena regresijskih metoda . . . . .	20
4.5.1. Predviđanje ukupnog broja bodova pomoću ostvarenog rezultata na prvom međuispitu - korištene metode linearna i polinomijalna regresija . . . . .	21
4.5.2. Predviđanje broja bodova na završnoj provjeri pomoću prethodnih provjera - korištena metoda regresijsko stablo . . .	26
<b>Literatura</b>	<b>32</b>

# UVOD

Jedan od najvažnijih i najbrže rastućih trendova u informacijskim tehnologijama je ogroman porast podataka iz edukacijske domene. Važno je pristupiti takvim podacima s pažnjom kako se ne bi izgubili u njima. Problem je što unatoč postojanju toliko puno podataka, nemamo puno znanja izvedenog iz njih. Zato je posebno bitno imati sustavan i organiziran pristup takvim podacima. To postizemo odabirom modela i metoda pomoću kojih dolazimo do identificiranja bitnih i korisnih informacija iz takvih glomaznih skupova podataka.

U ovom radu ćemo se upoznati sa potrebnim disciplinama i tehnikama koje će nam bit od velike pomoći pri dubinskoj analizi. U prvom poglavlju ćemo se upoznati sa dubinskom analizom podataka u edukacijskoj domeni, prikazati ćemo kako pristupiti takvim podacima te koje su nam tehnike na raspolaganju. U drugom poglavlju ćemo predstaviti strojno učenje, objasniti koja je veza strojnog učenja i dubinske analize podataka te zašto nam je ono potrebno, pri čemu ćemo se orijentirati na regresijske metode (metode koje predviđaju numeričku varijablu). U idućem poglavlju ćemo se upoznati sa korisnim tehnikama pri prediktivnoj analizi podataka. Na kraju ćemo prikazati kako primjeniti regresijske metode nad stvarnim podacima iz edukacijske domene pri čemu ćemo koristiti programski jezik R.

# 1. Edukacijska domena

Dubinska analiza podataka (engl. Data Mining(DM)) je analiza velikih podatkovnih skupova s ciljem pronalaženja neočekivanih veza ili prikaza podataka koji su novi, korisni i donose nova znanja[3]. Prijašnje tehnike izvlačenja određenih informacija iz podataka su ovisile isključivo o statistici, sve dok se u 90-tim godinama 20. stoljeća nije pojavila disciplina dubinska analiza podataka koja koristeći mogućnosti suvremenih računala omogućuje pronalaženje i izvlačenje korisnih i primjenjivih modela i uzoraka u glomaznim skupovima podataka.

## 1.1. EDM

EDM (engl. Educational Data Mining - dubinska analiza u edukaciji) je znanstveno polje koje koristi razne algoritme (statističke, strojno učenje...) za istraživanje podataka iz edukacijske domene. S napretkom tehnologije i stvaranjem novih tehnologija poput "e-učenja", instrumentalnih edukacijskih softvera te kreiranjem golemih baza podataka sa informacijama o studentima dolazimo do golemih skladišta podataka iz edukacijske domene.[9] EDM smatramo dijetetom DM-a specificiranog na podatke iz edukacijske domene. Glavni cilj je analiziranje takvih tipova podataka kako bi izvukli informacije koje su korisne za bolje razumijevanje procesa usvajanja znanja, prepoznavanje svojstava okruženja u kojem se uči te dobivanje uvida u nastavne fenomene. Problem je što svaki poseban cilj ima specifične karakteristike koji zahtijevaju drugačije pristupe rješavanja problema dubinske analize. To nas dovodi do toga da ne možemo rješavati probleme iz EDM-a izravnim korištenjem tehnika DM-a(engl. Data Mining - dubinska analiza podataka). Posljedica toga je da moramo adaptirati klasične tehnike pri izvlačenju bitnih podataka iz edukacijske domene.[9]

## **1.2. Povijest EDM-a**

Znanstveno polje EDM se pojavilo kao nezavisno područje istraživanja u zadnjih par godina. Korijeni joj leže u seriji radionica koje su organizirane početkom 21. stoljeća. Prvu je radionicu, organiziranu 2005. pod nazivom "Educational Data Mining", slijedila serija novih radionica što je kulminiralo stvaranjem godišnje konferencije "International Conference on Educational Data Mining" organizirane od "International Working Group on Educational Data Mining". Kako je dolazilo do sve većeg interesa za EDM stvoren je i časopis "Journal of Education Data Mining", čija je namjena dijeljenje rezultata raznih istraživanja iz edukacijske domene.[9]

## **1.3. Proces istraživanja edukacijskih podataka**

Proces primjene metoda DM-a na podatke iz edukacijske domene možemo interpretirati iz više perspektiva.

Iz edukacijske i eksperimentalne perspektive, proces vidimo kao iterativan ciklus formiranja hipoteza, testiranja te pročišćavanja. U ovom procesu cilj nije samo izvlačenje znanja iz podataka nego i filtriranje tog znanja za dobivanje odluka koje nam govore kako modificirati edukacijsku okolinu da poboljšamo proces učenja kod studenata. Iz perspektive DM-a, proces možemo promatrati kao otkrivanje općenitog znanja i dubinske analize, iako postoji par bitnih razlika u svakom koraku.[9]

### **1.3.1. Edukacijska okolina**

Podaci iz edukacijske domene nam mogu dolaziti iz raznih izvora. Tu su uvijek podaci dobiveni tradicionalnim putem iz "učionica", zatim imamo podatke dobivene iz internetske verzije obrazovanja (e-učenje, online tečajevi...). Prikupljanje i integriranje takvih podataka je jako važan korak.

### **1.3.2. Pretprocesiranje**

U edukacijskom kontekstu prirodno je da pretprocesiranje predstavlja važan i kompleksan zadatak. Zna se dogoditi da samo pretprocesiranje podataka traje više nego pola ukupnog vremena provedenog na rješavanju problema dubinske



analize. Često se događa podaci iz edukacijske domene nisu u primjerenom obliku (potrebne su transformacije podataka kako bismo istražili pitanja koja nas zanimaju). Također se postavlja i etičko pitanje. Važno je zaštititi osobne podatke studenata poput imena, telefonskog broja, JMBAG-a itd. To postizemo anonimiziranjem podataka, npr. pridjeljivanjem slučajnog broja koji predstavlja oznaku studenta.

### **1.3.3. Dubinska analiza podataka**

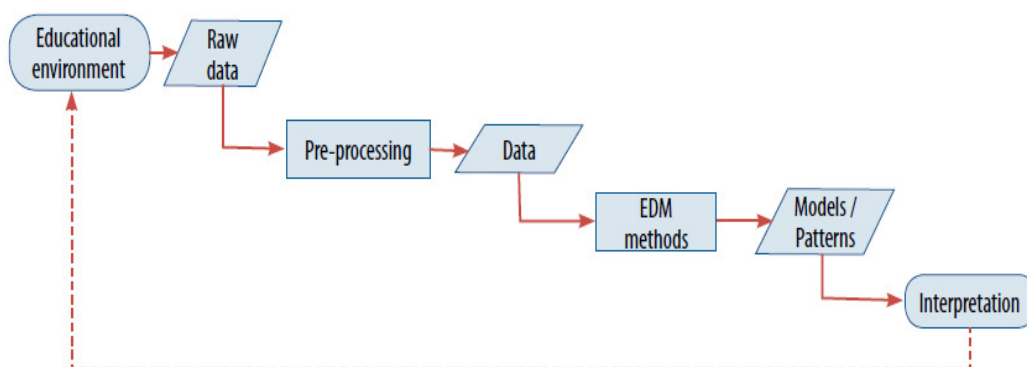
Većina DM tehnika poput klasifikacije, regresije, grupiranja su već uspješno primjenjene na podatke iz edukacijske domene. Često podaci iz edukacijske domene zahtijevaju poseban tretman. No, EDM je i dalje relativno novo područje istraživanja te možemo predvidjeti da će razvoj ovog znanstvenog polja rezultirati sa boljim razumijevanjem problema i prepreka pri analizi.[9]

### **1.3.4. Interpretacija rezultata**

Modeli dobiveni dubinskom analizom podataka trebaju biti razumljivi i korisni pri procesu stvaranja odluka. Često je bitnija interpretabilnost od preciznosti rezultata. Vizualizacija je jedna od korisnih tehnika kojima objašnjavamo i pokazujemo rezultate te ih tako činimo razumljivijim. Rezultati trebaju biti razumljivi i onima koji nisu stručnjaci u poljima statistike i dubinske analize. Ovisno o odabranoj metodi, model može biti tzv. "bijela kutija", tj. može nam pružiti informacije o tome kako računa izlaznu vrijednost iz čega se potencijalno mogu izvući korisne informacije o stvarima iz stvarnog svijeta koje analizirani podaci opisuju. Ponekad nam je ova interpretabilnost modela važnija od njegove prediktivne snage, tj. nije nam toliko bitna sama točnost predikcija koliko prikupljanje znanja o tome kako ulazne varijable utječu na izlaznu i u kojoj mjeri.

## **1.4. Metode EDM-a**

Postoji puno metoda u polju EDM-a. Neke su poznate kao univerzalne kroz različite tipove DM-a, poput metoda predikcije, grupiranja, detekcije stršćih vrijednosti te analize i obrade teksta.



**Slika 1.1:** Proces dubinske analize u edukacijskoj domeni[7]

#### 1.4.1. Predikcija

Glavni cilj predikcije je zaključivanje o jednoj varijabli (ciljna varijabla) pomoću jedne ili više drugih varijabli (regresori). Vrlo često u praksi imamo specifičan odnos među varijablama takav da jedne varijable možemo smatrati slučajnim reakcijama na neke druge nezavisne varijable. Metode predikcije možemo podijeliti s obzirom na to da li predviđaju kategorijsku varijablu (klasifikacijske metode) ili predviđaju numeričku varijablu (regresijske metode).

#### 1.4.2. Grupiranje(engl. clustering)

Cilj grupiranja je identificiranje grupe jedinki unutar populacije koje su dovoljno slične na način da su jedinke unutar iste grupe više slične jedne drugima nego jedinkama iz druge grupe. U EDM-u grupiranje možemo koristiti za pronalaženje sličnih materijala za predmete ili za grupiranje studenata po njihovim navikama učenja.

#### 1.4.3. Detekcija stršećih vrijednosti

Cilj detekcije stršećih vrijednosti je pronalazak zapažanja koja su značajno drugačije od vrijednosti u ostatku skupa podataka. Stršeća vrijednost može biti rezultat greške pri unosu podataka, ali i ne mora. Često se dolazi u napast izbaciti stršeće vrijednosti jer "kvare" istraživanje. No detaljnijim promatranjem i ispitivanjem stršećih vrijednosti možemo doći do novih korisnih informacija. U EDM-u stršeće vrijednosti nam mogu biti korisne za identificiranje studenata sa

problemima pri učenju, devijacije ili pristranosti pri odlukama učitelja odnosno profesora.

#### **1.4.4. Analiza i obrada teksta**

Glavni cilj metode analize i obrade teksta je pronalazak bitnih i korisnih informacija iz teksta. Tipični zadaci pri tome su kategorizacija i grupiranje teksta, te sažimanje teksta. U EDM-u ovu metodu koristimo pri analizi foruma, materijala za učenja i ostalih materijala koji su bitni za edukacijsku domenu.

### **1.5. Faze EDM-a**

Postupak EDM-a dijelimo u 4 faze[9] :

1. Prva faza EDM-a je otkrivanje dosad skrivenih veza u podacima. Tražimo konstantnu vezu između varijabli iz podataka koji dolaze iz edukacijske domene. Za otkrivanje veza koristimo poznate algoritme i metode poput regresije, klasifikacije, analize i obrade teksta i grupiranja
2. Otkrivene veze(pravila) moramo validirati kako bi izbjegli prenaučenosť (engl. overfit) podataka. Jedna od poznatijih metoda je unakrsna provjera (engl. cross-validation).
3. Otkrivene veze (pravila) koje smo uspješno validirali koristimo kako bi stvarali predikcije o budućim događajima u edukacijskoj okolini.
4. Predikcije koristimo kao potporu za stvaranje odluka koje će dovesti do korisnih stvari u edukacijskoj okolini.

## 2. Strojno Učenje

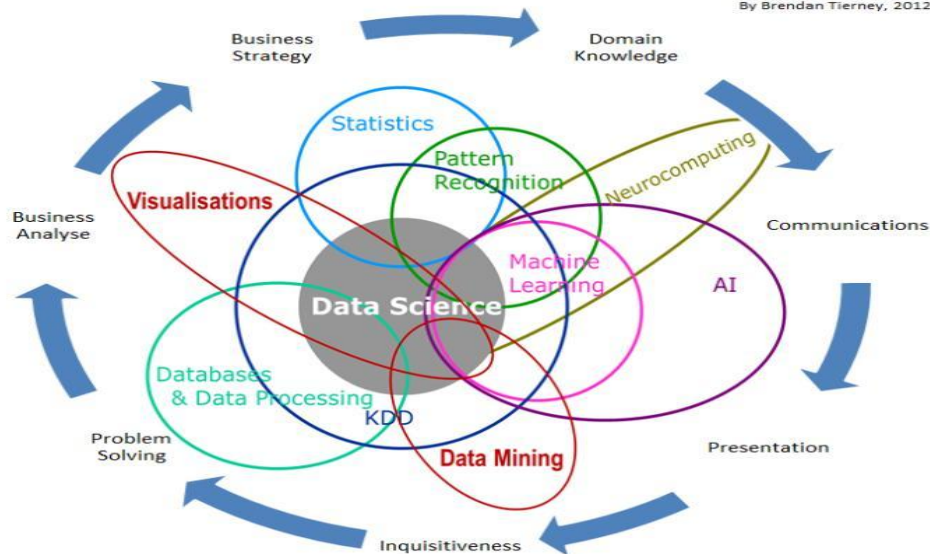
Strojno učenje (engl. Machine Learning) je polje računalne znanosti koje se bavi specifičnim načinom programiranja u kojem računalu ne dajemo eksplicitne instrukcije, već očekujemo da računalno samostalno dođe do određenih spoznaja na osnovu odabranih podatkovnih skupova i određene metode "učenja". Strojno učenje se često dijeli na tzv. "nadzirano učenje" (engl. supervised learning), gdje imamo jasno definirane ulaze i izlaze tj. ciljeve, te "nenadzirano učenje", gdje nemamo unaprijed definirane izlaze već očekujemo da će računalno analizirajući samo ulaze doći do nekih korisnih spoznaja o samim podacima.[8] Ovo polje korijene vuče iz okoline u kojoj dolazi do simultanog rapidnog rasta količine podataka, statističkih metoda te dodatne računalne snage. Porast podataka zahtijevao je dodatnu računalnu snagu, što je dovelo do razvoja statističkih metoda za analizu velikih skupova podataka. Ovo je rezultiralo ciklusom poboljšanja koja omogućuju kolekciju i analizu sve većih skupova podataka.

### 2.1. Zašto strojno učenje ?

Strojno učenje i DM su bliski rođaci. U mnogim stvarima se ova dva polja znanosti preklapaju no ipak postoje bitne razlike. Strojno učenje vežemo s proučavanjem, dizajniranjem i razvojem algoritama koji pružaju računalu sposobnost da uče (ne programiramo eksplicitno naredbe). DM vežemo s procesom izvlačenja znanja i zanimljivih uzoraka iz različitih skupova podataka. Dakle ključna razlika je što DM izvlači nova pravila iz dostupnih podataka, a strojno učenje podučava računalno da uči i razumije dana, već poznata pravila. No važna je veza između ova dva polja znanosti. Tijekom postupka DM-a koristimo algoritme strojnog učenja, primjenjujemo ih na probleme čija rješenja i karakteristike nas zanimaju.

# Data Science Is Multidisciplinary

By Brendan Tierney, 2012

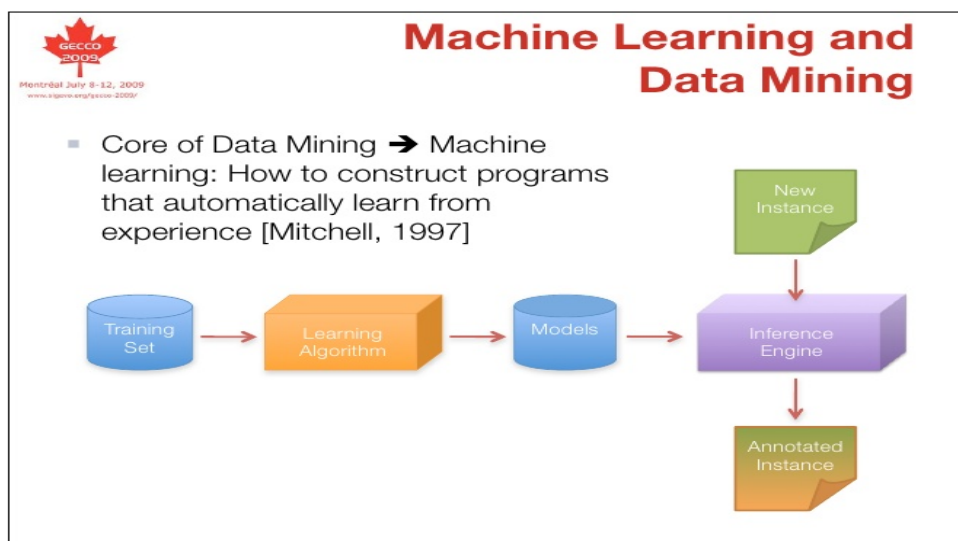


**Slika 2.1:** isprepletenost strojnog učenja, dubinske analize podataka te drugih znanstvenih polja[5]

## 2.2. Koraci strojnog učenja

Svaki zadatak strojnog učenja možemo razlomiti na seriju koraka. Jedna od dobrih podjela je ova[6] :

1. Prikupljanje podataka: Bilo da su podaci zapisani na papiru, u tekstualnim datotekama, spremljeni u jednu od SQL baza podataka, podatke moramo skupiti i reorganizirati u elektronički oblik pogodan za alat u kojem ćemo analizirati podatke.
2. Istraživanje i pripremanje podataka: Kvaliteta bilo kojeg projekta u kojem koristimo strojno učenje uvelike ovisi o kvaliteti podataka koje koristimo. Ovaj dio je dosta ovisan o ljudskoj intervenciji. Većinu vremena potrebnog za rješavanje problema provedemo u upoznavanju podataka.
3. Treniranje modela nad podacima: Kad su podaci spremni za analizu, trebali bi imati predosjećaj što možemo očekivati od podataka. Odabirom primjerenog algoritma strojnog učenja gradimo model koji reprezentira podatke.
4. Evaluiranje performanse modela: Zato što svaki model strojnog učenja rezultira pristranim rješenjem, jako je važno evaluirati koliko je dobro algoritam "naučio" iz iskustva. Ovisno o tipu modela, jedno od rješenja je



**Slika 2.2:** Dubinskoj analizi podataka je potrebno strojno učenje[1]

provjeravanje preciznosti modela nad test podacima (npr. pri gradnji linearnog modela jedno od popularnih rješenja je unakrsna provjera)

5. Poboljšanje performanse modela: Ako je potrebna veća preciznost, postaju nužne naprednije metode koje će poboljšati model. Ponekad će biti nužno promijeniti model.

## 2.3. Metode i tehnike strojnog učenja

Ovdje između ostalog govorimo o metodama (algoritmima) strojnog učenja. No mnoge metode poput linearne regresije nam dolaze iz statistike, odnosno statističkog učenja. Strojno učenje je na neki način "posudilo" i preradilo dane algoritme. Vrlo često je linija između strojnog i statističkog učenja "zamagljena". Orijentirani ćemo se na regresijske metode, odnosno metode koje predviđaju numeričke kontinuirane vrijednosti.

### 2.3.1. Linearna regresija

Linearna regresija je metoda nadziranog strojnog učenja za predviđanje ciljne numeričke varijable uz pomoć linearne funkcije jedne ili više ulaznih varijabli. Na ovaj način stvaranje prediktivnog modela svodi se na postupak određivanja koeficijenta smjera (engl. slope) i odsjeka (engl. intercept) koji će tvoriti jednostavnu

formulu za izračun ciljne varijable uz pomoć ulaznih parametara. Budući da se ova metoda svodi na pogađanje navedenih parametara, metoda linearne regresije spada u tzv. "parametarske metode" strojnog učenja[8]. Parametarske metode sadrže iduće korake[4] :

1. Stvaramo pretpostavku o obliku funkcije  $f$  (funkcije koja predstavlja model linearne regresije). Ako nam je pretpostavka da je funkcija  $f$  linearna, problem računanja koeficijenata modela smo si bitno pojednostavili.
2. Nakon odabira modela, trebamo postupak koji će iskoristiti trening skup za treniranje modela. Jedna od najpoznatijih funkcija za računanje koeficijenata modela je metoda najmanjih kvadrata.

Ovisno o broju ulaznih varijabli razlikujemo dva tipa linearne regresije:

1. Jednostavnu linearnu regresiju

$$y = \beta_0 + \beta_1 x_1$$

2. I višestruku linearnu regresiju

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots$$

U praksi veza nije deterministička tj. postoji još niz faktora koji utječu na reakciju. Pretpostavka modela (na primjeru jednostavne linearne regresije) je :

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

$\epsilon$  nam predstavlja slučajnu varijablu sa konstantnom varijancom (znamo ju kao rezidualna varijanca).[4] Linearna regresija korisna je u raznim istraživačkim i praktičnim situacijama, a daje odgovore na nekoliko bitnih pitanja:

- Postoji li veza između ulazne varijable (ili više ulaznih varijabli) - regresora, i izlazne varijable (reakcije)?
- Koliko je jaka ta veza?
- Koje ulazne varijable najviše utječu na izlaznu varijablu i koliko je jak taj efekt?
- Možemo li predvidjeti izlaz za neke nove vrijednosti ulaznih varijabli i s kojom točnošću?

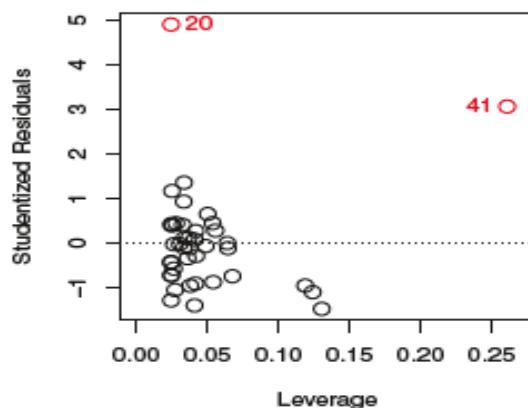
Nama je prava vrijednost parametara  $\beta_0$  i  $\beta_1$  nepoznata, kao i pogreška  $\epsilon_i$ . Bitni pojmovi vezani uz regresiju :

- Postoji puno načina kako procijeniti koeficijente linearne regresije. Jedna od najpopularnijih metoda je metoda najmanjih kvadrata. Metoda najmanjih kvadrata minimizira sumu kvadrata vertikalnih odstojanja od pravca.
- Rezidual nam predstavlja razliku između stvarne vrijednosti ciljne varijable i vrijednosti izračunate modelom. Standardna greška reziduala je mjera koja procjena koliko (prosječno!) model "promašuje" kod svojih predviđanja ciljne varijable. Iznos "prihvatljive" greške razlikovati će se ovisno o konkretnom scenariju
- Pearsonov koeficijent korelacije nam služi za opis snage linearnog odnosa između dviju varijabli. Ova mjera nam je jako korisna kod jednostavne linearne regresije, ali ima i prednosti kod višestruke linearne regresije (procjena nezavisnosti između regresora)
- Koeficijent determinacije (engl. R-squared) je jedna od najkorisnijih mjera. Ona se definira kao "količina varijabilnosti koja je objašnjena modelom", odnosno ako nam se promijeni iznos ciljne varijable, koliko zasluga pri tome ima promjena iznosa u jednom ili više regresora.[8] Važno je napomenuti da je koeficijent determinacije jedan od važnijih kriterija za ocjenu kvalitete linearnog modela te je kao takav često sadržan u opisu rezultata modela.
- Razlikujemo stršeće vrijednosti (engl. outliers) te ekstremne vrijednosti (engl. high leverage points). Stršeće vrijednosti su one koje imaju veliki rezidual, odnosno linearni model ih ne aproksimira precizno. Ekstremne vrijednosti su vrijednosti regresorske varijable koje odskakuju od ostatka vrijednosti regresorske varijable.

Linearna regresija pretpostavlja dva bitna svojstva u podacima :

1. Pretpostavka nezavisnosti između regresora (pretpostavka je bitna za višestruku regresiju). Ova pretpostavka nam govori da je efekt promjene regresorske varijable utječe na ciljnu varijablu nezavisno od ostalih regresorskih varijabli. No to često nije slučaj, tj. ulazne varijable nisu samo kolinearne sa ciljem, nego i između sebe. Pretpostavka da se jedan ulazni parametar mijenja dok njemu kolinearan ostaje fiksiran je nerealan, što se odražava u podatkovnom skupu a samim time i u porastu "nesigurnosti"

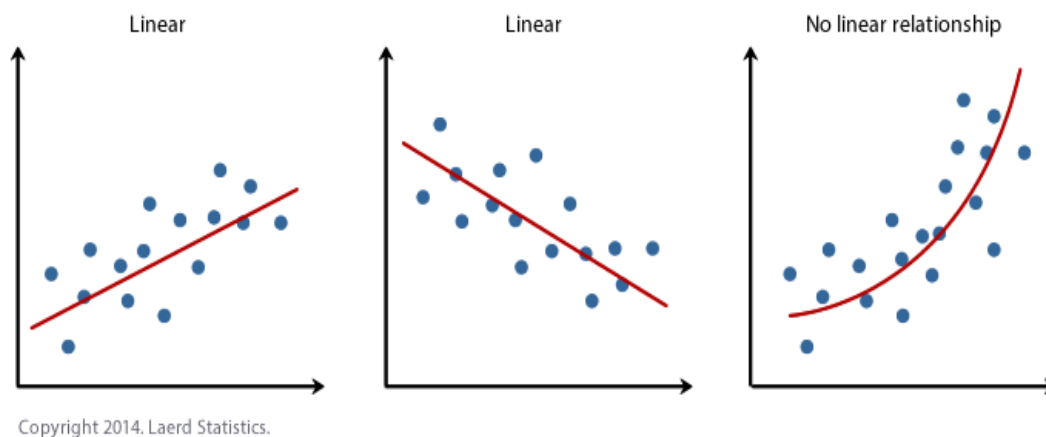




**Slika 2.3:** Graf reziduala, točka 20 je stršeća vrijednost, dok je točka 41 posebno opasna jer i stršeća i ekstremna vrijednost[4]

linearnog modela. Zbog toga u konačnom rezultatu modela možemo dobiti veće p-vrijednosti ulaznih varijabli, tj. one mogu biti tretirane kao irelevantne, iako su zapravo snažno linearno povezane s ciljem. No postoji još jedna zanimljiva pojava - multikolinearnost. Naime može se dogoditi da nema izravne veze između dvije varijable, ali se kolinearnost očituje tek u kombinaciji tri i više varijabli. Kolinearnost varijabli ne mora nužno utjecati na prediktivnu moć modela, ali unosi potencijalno veliku nesigurnost u modelu smislu da sve kolinearne prediktore izbacimo iz modela kao irelevantne. Pitanje je što učiniti kada naletimo na ovaj problem? Možemo izbaciti jednu od problematičnih varijabli iz para problematičnih varijabli. Jedno od rješenja je stvaranje novog regresora koji je umnožak para problematičnih varijabli. Ovo nam je korisno rješenje jer promjena u jednom regresoru više nije konstantna nego ovisi i o drugom (problematičnom) regresoru.

2. Pretpostavka linearnosti modela nam govori da je reakcija ciljne varijable konstantna na promjenu regresorske varijable, neovisno o trenutačnoj vrijednosti regresorske varijable. Ali u nekim slučajevima ovo nije istina, tj. nekad postoji nelinearna veza između ciljne i regresorske varijable. Jedno od jednostavnijih rješenja ovog problema je upotreba polinomijalne regresije.



**Slika 2.4:** Lijeva i srednja slika pokazuju linearnu vezu između podataka, dok desna slika predstavlja vezu koja nije linearna.[4]

### 2.3.2. Polinomijalna regresija

Polinomijalna regresija je oblik regresijske analize u kojoj je veza između ciljne varijable i regresora modelirana kao n-ti stupanj polinoma regresorske varijable.[4] Glavni razlog korištenja polinomijalne regresije je "ugađanje" (engl. fit) nelinearne veze između ciljne i regresorske varijable. Standardni model višestruke linearne regresije

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots$$

mijenjamo sa modelom

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_3^3 \dots$$

Možemo primjetiti da koeficijente uz prediktore računamo na isti način kao i kod višestruke linearne regresije, odnosno metodom najmanjih kvadrata. Polinomijalna regresija je specijalni slučaj višestruke linearne regresije.

### 2.3.3. Regresijska stabla i model-stabla

Upoznali smo se sa linearnom regresijom. Tu je nezavisna varijabla Y modelirana kao linearna funkcija nezavisnih varijabli. Linearna regresija je globalni model sa jednom formulom koja sadrži cijeli skup podataka. Međutim, kada podaci imaju više značajki koje međusobno djeluju na nelinearan način, tada je sastavljanje modela iznimno teško i komplicirano za objašnjenje rezultata. Da bi se olakšao način modeliranja podataka koriste se particije koje dijele podatke u manje dijelove s kojima se radi i takav se način modeliranja zove rekurzivna particija. Kažemo da

takvim načinima modeliranja gradimo regresijsko stablo odlučivanja. Svaki list (krajnji čvor u stablu) predstavlja jednu ćeliju u particiji. Kako bi saznali u kojoj particiji se određeni podatak nalazi, moramo krenuti od vršnog čvora u stablu te se pratiti vrijednosti koje odgovaraju našim podacima, postupak ponavljamo dok ne dođemo do čvora koji predstavlja list[6]. Takav način ima određene prednosti poput :

- Lagano saznajemo vrijednost za novi podatak (pregledavanjem stabla i praćenjem odgovarajućih grana), te ih mogu interpretirati i laici.
- Jednostavno saznajemo koje varijable su bitne pri predviđanju (one koje se nalaze u stablu), naime algoritmi za gradnju ovakvih stabla automatski rade odabir varijabli.
- Regresijska stabla (i klasifikacijska) su laka i jednostavna za objašnjavanje drugim ljudima.

Imamo dvije vrste stabla odlučivanja za numeričke podatke :

- Regresijsko stablo (engl. regression tree) uvedeno je 1980. kao dio CART (engl. Classification and Regression Tree) algoritma. Unatoč imenu, regresijska stabla ne koriste linearnu regresiju, nego stvaraju predviđanja na osnovu prosječne vrijednosti podataka koji stignu do određenog lista.
- Model-stabla(engl. model tree) su uvedena par godina poslije regresijskih stabla. Manje su poznata, ali unatoč tome smatrana su "snažnijom" metodom od regresijskih stabla. Stabla gradimo na sličan način kao i stabla kod metode regresijskih stabla, s razlikom da gradimo model linearne regresije za podatke koje stignu do određenog lista. Ovo čini model-stabla težima za interpretaciju ali zato imaju preciznija predviđanja. Jedan od najpoznatijih algoritama za izgradnju model-stabla je **M5** algoritam.

#### 2.3.4. Odabir varijabli

Odabir varijabli (engl. variable selection) jedan je od ključnih izazova s kojima se suočavamo u izradi prediktivnih modela, ne samo kod linearne regresije već i općenito. Kolinearnost varijabli ne mora nužno utjecati na prediktivnu moć modela, ali unosi potencijalno veliku nesigurnost u modelu smislu da sve kolinearne prediktore izbacimo iz modela kao irelevantne. To bi se mogao pokazati kao velik problem kada imamo više potencijalnih prediktora i pokušavamo odabrati relevantni podskup. Mogući kriterij za odluku koje varijablu odabrati za

ugrađivanje u model tako može biti utjecaj na povećanje zajedničke "R-kvadrat" mjere, smanjenje standardne greške reziduala ili p-vrijednost koeficijenta za tu ulaznu varijablu. Pored ovih "standardnih" kriterija postoje i razni drugi, kao npr. popularni AIC (engl. Akaike information criterion) koji procjenjuje informativnost modela uz penaliziranje većeg broj varijabli. Varijable možemo odabirati ručno, no puno je lakše taj posao ostaviti računalu. Statistički alati često imaju ugrađene algoritme koji na osnovu zadanog kriterija izgrađuju prediktivni model iterativnim odabirom varijabli.[4][8] Najčešće strategije izgradnje modela su:

- Odabir unatrag : Krećemo od modela koji sadrži sve varijable te na osnovu odabranog kriterija izbacujemo "nepodobne varijable". Kad više nemamo varijabli za izbaciti, odabrane su varijable koje model treba sadržavati.
- Odabir unaprijed: Krećemo od praznog modela te na osnovu odabranog kriterija dodajemo varijable u model. Kada nemamo više varijabli za dodati, sve dodane varijable su one koje model treba sadržavati.
- Razne hibridne metode

## 3. Prediktivna analiza podataka

Prediktivna analiza podataka jest vrsta statističke i dubinske analize podataka koja koristi metode nadziranog strojnog učenja kako bi na osnovu povijesnih podataka omogućila predviđanje određenih varijabli za buduće slučajeve. Varijabla koju pogađamo može biti kategorijska (nebrojčana) ili numerička (metrička). Ako pogađamo kategorijsku varijablu najčešće koristimo klasifikacijske metode, dok za pogađanje numeričke varijable najčešće koristimo regresijske metode.[8]

### 3.1. Treniranje i predikcija

Nadzirani model treba najprije trenirati (naučiti) na označenim primjerima, a nakon toga se može koristiti za predikciju (klasifikaciju ili regresiju) na dotad neviđenim primjerima. Jedna od uobičajenih podjela je da 70% podataka stavljamo u skup za treniranje, a ostatak u skup za testiranje.

### 3.2. Prenaučenost

Naš glavni cilj je da trenirani model dobro generalizira. Ako je model presložen, previše će se prilagoditi podacima na kojima je treniran, a davat će loše predikcije nad neviđenim podacima (prenaučenost). Prenaučen model je beskoristan jer loše generalizira! Prenaučenost je jedan od glavnih problema strojnog učenja. Do prenaučenosti dolazi jer nastojimo modelirati šum u podacima. Šum je neobjašnjiv po definiciji, te pokušaj objašnjenja šuma rezultira pogrešnim zaključcima koji se neće dobro slagati sa novim podacima.

### 3.3. Unakrsna provjera

Kako bismo izbjegli prenaučenost, moramo ispitati koliko dobro će model raditi nad neviđenim podacima (metoda unakrsne provjere). Budući da neviđeni

primjeri nisu dostupni, dio primjera koje imamo izdvajamo da "glume" nevidene primjere. Podatke dijelimo na skup za treniranje i skup za testiranje (tipično 70% i 30%). Model treniramo na skupu za treniranje, a zatim pomoću tog modela radimo predikciju na skupu za testiranje i na tom skupu računamo točnost. Dolazimo do problema kad imamo više modela i trebamo procijeniti koji je najbolji. Za ovo su nam kod linearnih modela jako korisne statistike MSE (engl. mean squared error - usrednjena kvadratna greška),

$$\frac{1}{n} \sum_{t=1}^n e_t^2$$

te MAE (engl. mean absolute error - usrednjena apsolutna greška)

$$\frac{1}{n} \sum_{t=1}^n |e_t|$$

pri čemu  $e$  predstavlja razliku između stvarne vrijednosti i vrijednosti izračunate linearnim modelom, ta je varijabla poznata pod imenom rezidual. Imamo MSE od skupa za treniranje, no on nam je praktički beskoristan. Često je MSE od skupa za treniranje potpuno drugačija nego MSE skupa za testiranje (zbog prenaučivosti skupa za treniranje). Potrebno je provjeriti koliko model dobro radi na neviđenim podacima. Zato smo i odvajali podatke u dva skupa. Odabir koji je linearni model bolji možemo napraviti preko skupa za testiranje koji nam je dostupan. No treba paziti na jednu važnu činjenicu. Ako trebamo procijeniti koji je model bolji, odnosno koji preciznije aproksimira zapažanja ne smijemo to raditi preko skupa za testiranje. Namjena skupa za testiranje je prikaz rezultata, te podaci iz tog skupa niti u jednom trenutku ne smiju biti uključeni prilikom treniranja modela i odabira koji je model najbolji. Za ovaj problem je razvijeno više tehnika :

1. Podjela podataka na tri skupa. Skup za testiranje, skup za validaciju te skup za treniranje (koristimo omjer 40%-30%-30%). Skup za validaciju služi za odabir najboljeg modela među ponuđenima. Ova metoda ima više mana. MSE može biti jako nestabilna mjera. Naime jako ovisi o tome kako su zapažanja raspoređena po skupovima. Također smo osiromašili skup za treniranje te su modeli istrenirani nad manjim brojem podataka, što dovodi do manje preciznosti modela.
2. Tehnika LOOCV (engl. Leave-One-Out-Cross-Validation - izostavljanje jednog podatka prilikom treniranja) rješava gore navedene probleme. Tehnika stvara  $n$  modela (pri čemu je  $n$  broj zapažanja u skupu za treniranje)

pri čemu uvijek izostavljamo jedno zapažanje iz skupa za treniranje te nam to zapažanje predstavlja "skup" za validaciju. Zatim usrednjujemo MSE svih modela te nam usrednjena MSE vrijednost predstavlja grešku treniranog modela. No problem je u prevelikoj složenosti ove metode, pogotovo ako imamo velik broj zapažanja. Problem je još veći kad radimo odabir između više modela, te svaki model stvaramo  $n$  puta, što vrlo često ispada vremenski zahtjevno. Također moramo paziti na činjenicu da ne uključimo sve podatke u skup za treniranje. Uvijek trebamo imati primjere neviđene prilikom treniranja i namještanja modela, ti primjeri ne smiju nikako sudjelovati u tom procesu. Njih koristimo za prikaz rezultata.[4]

3. Tehnika k-fold Cross-Validation popravljala gore navedeni problem. Zapažanja iz skupa za treniranje dijelimo u  $k$  grupa (pri čemu je  $k$  najčešće 5 ili 10), pri čemu se svako zapažanje nalazi u točno jednoj grupi. Zatim koristimo LOOCV, ali na razini grupa. Ova tehnika je vremenski puno manje zahtjevnja nego tehnika LOOCV.[4]

### 3.4. Odabir metode strojnog učenja

Proces odabira metode strojnog učenja (tj. njezina implementacija u odabranom programskom jeziku ili analitičkom alatu) uključuje pronalažanje odgovarajućih karakteristika skupa podataka koji ćemo trenirati. Odabir metode strojnog učenja je ovisan o tipu podataka koje analiziramo te to trebamo imati na umu pri skupljanju, istraživanju i čišćenju podatkovnog skupa

## 4. Studijski primjer

Proces primjene regresijskih metoda nad podacima zna biti vrlo kompleksan. Jako je bitno dobro formulirati pitanja te očistiti i prilagoditi podatke u skladu s pitanjima. Također je vrlo važno ispravno interpretirati rezultate. Izvođenje krivih zaključaka je jedna od najvećih grešaka. Ovdje ćemo prikazati cjelokupan proces analize podataka uz primjenu regresijskih metoda.

### 4.1. Ciljevi istraživanja

Jedan od najbitnijih koraka je postaviti ciljeve istraživanja. Pošto prikazujemo primjenu regresijskih metoda nad podacima iz edukacijske domene, ciljeve ćemo uskladiti s tim preduvjetima.

Želimo poboljšati rezultate i proces učenja studenata prve godine na Fakultetu elektrotehnike i računarstva pri kontinuiranoj provjeri. Htjeli bi identificirati "kritične" studente, odnosno studente kojima prijeti pad predmeta kako bi im mogli pružiti potrebnu pomoć. Također bi htjeli dati studentima povratnu informaciju o tome kako su napisali neku od početnih provjera znanja. Kako to učiniti? Jedan od načina je pokušati predvidjeti broj bodova studenata. Ako bi to mogli te kad bi i studenti i profesori imali uvid u takve podatke, mogli bi shvatiti da li su studenti na dobrom putu ili trebaju pojačati rad na predmetu kako bi ga uspješno položili. Možemo istražiti rezultate studenata prijašnjih godina. Tako bi trenutni studenti nakon neke od provjere znanja imali informaciju kako su prolazili kolege studenti koji su imali rezultate slične njima. Ovo istraživanje će se orijentirati na dva pitanja :

1. Možemo li predvidjeti ukupan broj bodova preko rezultata ostvarenog na prvoj velikoj provjeri, odnosno preko prvog međuispita?
2. Možemo li predvidjeti rezultate završne provjere znanja preko rezultata ostvarenih na ranijim provjerama?



## 4.2. Odabir alata za analizu

Danas postoji mnogo statističkih alata. Neki od najpoznatijih su Microsoft Excel, SPSS, Matlab i programski jezik R. Koristiti ćemo programski jezik R zbog više razloga :

- R je potpuno otvorena i besplatna platforma.
- Velika baza korisnika koji pridonose konstantnom razvoju. Koristi se sistem paketa pomoću kojih korisnici na lak način mogu uzimati samo potrebne pakete.
- Ugrađena podrška za skupove podataka. Naime, R ima fantastične mehanizme za stvaranje struktura podataka.
- Ugrađena podrška za nedostajuće vrijednosti ("NA" vrijednosti)
- Kvalitetna podrška za vizualizaciju podataka (paket "ggplot2")

## 4.3. Prikupljanje podataka

Zahvaljući ZPM-u (Zavod za primjenjenu matematiku) imamo podatke studenata sa predmeta "Matematika 1" iz akademske godine 2015/2016. Pregledom podataka vidimo da su podaci anonimizirani te sadrže rezultate svih (4) provjera znanja na spomenutom predmetu te identifikacijsku oznaku grupe (Slika 4.2).

```
osnovni<-read.csv("Matematika1_2014-15.csv",sep=";")
colnames(osnovni)[1] <- "ID_GRUPE"
#dodavanje ukupnog broja bodova
osnovni$ukupno<-osnovni$KPZ1+osnovni$KPZ2+osnovni$MI+osnovni$ZI
osnovni<-filter(osnovni,!is.na(MI) & !is.na(ZI) &
                !is.na(KPZ1) &!is.na(KPZ2))
```

Slika 4.1: Učitavanje i priprema podataka

## 4.4. Pripremanje i pretprocesiranje podataka

Uvidom u podatke ustanovljeno je da ne sadrže informaciju o ukupnom broju bodova. No to je lako rješivo. Ukupan broj bodova je suma rezultata svih provjera znanja. Također je ustanovljeno da podaci nekih studenata sadrže "NA" vrijednosti (nemamo informaciju o rezultatima na nekoj od provjera znanja). To je

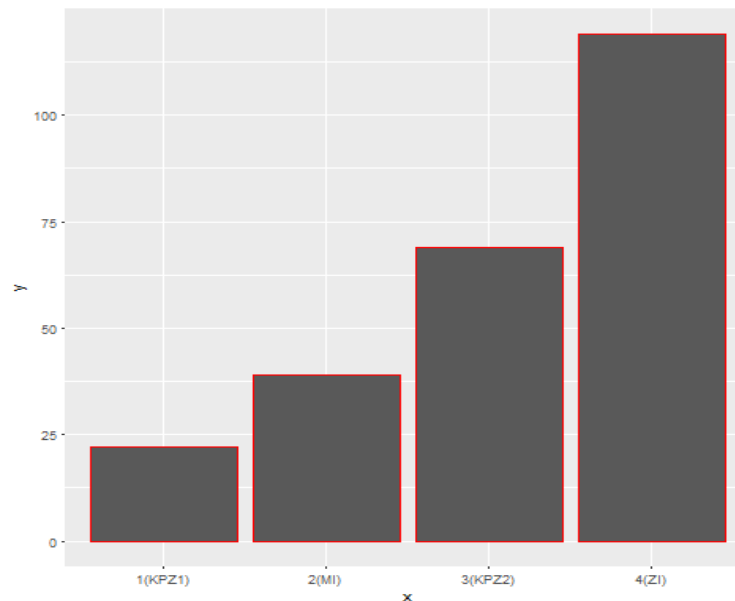
	ID_GRUPĚ	KPZ1	KPZ2	MI	ZI
1	ID915	10	10	28	31
2	ID188	5	NA	8	0
3	ID175	7	0	13	25
4	ID188	8	8	22	16
5	ID357	7	10	14	24
6	ID592	10	10	30	37
7	ID251	6	10	11	23
8	ID251	7	10	19	26
9	ID251	10	10	15	28
10	ID662	9	10	17	22
11	ID915	7	9	12	23
12	ID662	5	10	14	24
13	ID357	10	10	27	NA
14	ID662	8	5	4	6
15	ID251	7	10	11	18
16	ID251	5	4	7	20
17	ID188	5	7	13	14
18	ID636	8	9	16	20
19	ID662	9	NA	12	NA
20	ID357	10	9	13	26

**Slika 4.2:** Uvid u podatke

jedan od čestih izazova pri analizi podataka. Što učiniti sa takvim vrijednostima ? Možemo reći da "NA" vrijednost predstavlja nula bodova na provjeri znanja. Tako i u stvarnosti vrednujemo rezultate studenata koji nisu izašli na neku od provjera znanja. No dolazimo do jednog velikog problema. Da li želimo da nam npr. studenti koji nisu izašli na međuispit budu u istraživanju prvog pitanja (pogotovo ako ima puno takvih studenata) ? Jako je bitno sve vizualizirati. Grafom izostanaka(Slika 4.3) po provjeravama vidimo da što dalje ispiti idu, imamo sve više onih koji nisu pristupili ispitu. Zato ćemo takve studente (studente koji nisu izašli na neku od provjera) izbaciti iz istraživanja. Na slici 4.1 vidimo dio koda koji obavlja učitavanje i pripremu podataka te izbacuje studente koji nisu bili prisutni na svim provjerama.

## 4.5. Analiza podataka i primjena regresijskih metoda

Eksploratorna analiza podataka (engl. EDA - exploratory data analysis) proces je analize podatkovnog skupa s ciljem upoznavanja s podacima i donošenjem



**Slika 4.3:** Stupčasti graf izostanaka po provjerama

**Tablica 4.1:** Sažetak 5 brojeva i aritmetička sredina

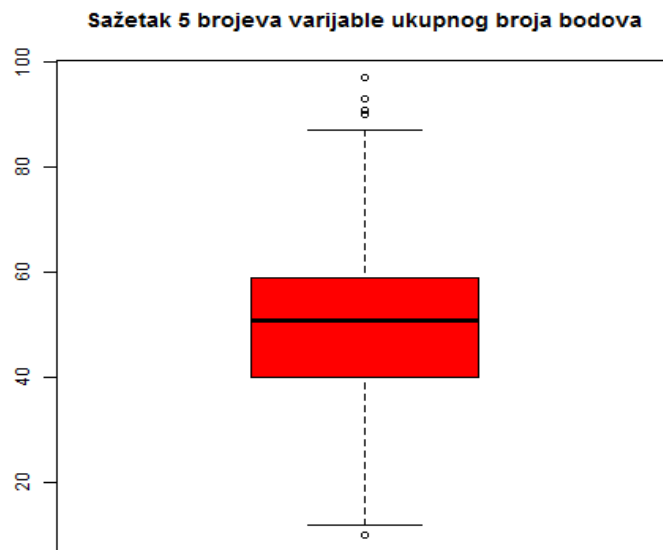
minimum	1. kvartil	medijan	3. kvartil	maksimum	ar. sredina
2.0	40.0	51.0	59.0	97.0	49.9

određenih zaključaka. Jedna od okosnica je vizualizacija podataka. Podatke smo uredili, sad nam je cilj izvući bitne informacije za naša pitanja te potražiti postoji li veza između varijabli. Analizu moramo orijentirati ka informacijama koje će nam pomoći pri analizi. Naše ciljne varijable su ukupan broj bodova (prvo pitanje) i broj bodova ostvaren na završnom ispitu (drugo pitanje). Obje varijable su numeričke, što ih čini pogodnim za uporabu regresijskih metoda.

#### 4.5.1. Predviđanje ukupnog broja bodova pomoću ostvarenog rezultata na prvom međuispitu - korištene metode linearna i polinomijalna regresija

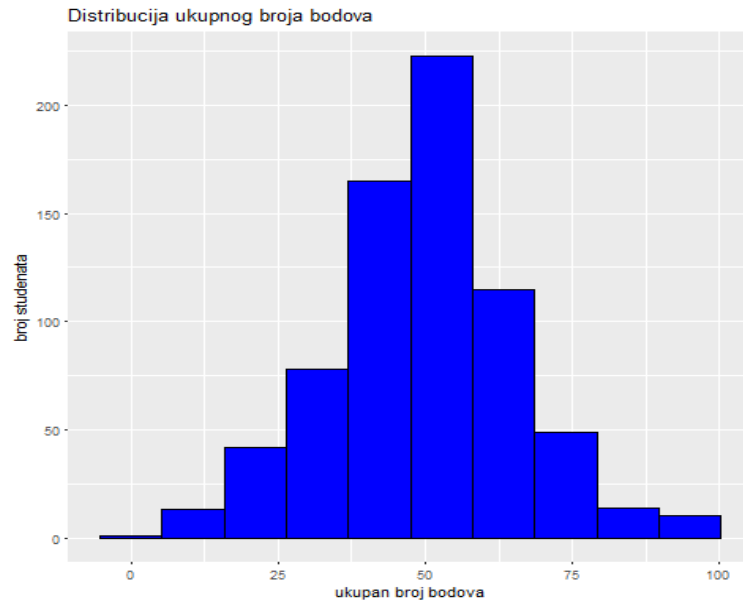
##### EDA

Ukupan broj bodova predstavlja zbroj bodova ostvaren na svim kontinuiranim provjerama. Uobičajeno je numeričku varijablu opisati uporabom "sažetka 5 brojeva" te aritmetičke sredine (Tablica 4.1) Jedna od najpopularnijih metoda za vizualizaciju "sažetka 5 brojeva" je tzv. boxplot dijagram (Slika 4.4) .



**Slika 4.4:** Boxplot graf varijable ukupnog broja bodova

Jedan od najpopularnijih grafova je histogram. Pomoću njega možemo vidjeti distribuciju numeričke varijable. To nam je jako bitno jer puno statističkih metoda i metoda strojnog učenja zahtijeva da varijable budu normalno distribuirane. Npr. linearna regresija pretpostavlja normalnu distribuciju ciljne varijable. No, ipak mnoge od ovih metoda su robustne, odnosno davati će dobre procjene i za varijable koje nisu normalno distribuirane. Vidimo da je distribucija varijable ukupnog broja bodova relativno normalno distribuirana (Slika 4.5). Nije "savršeno" simetrična te ima duže repove, no ovo smatramo sasvim zadovoljavajućom distribucijom. Kod analize dvije numeričke varijable često nas zanima pojava tzv. kolinearnosti, tj. linearne zavisnosti među varijablama. Jedna od popularnih metoda je "Pearsonov koeficijent korelacije". Kao rezultat dobivamo vrijednost između -1 i 1, što je rezultat po apsolutnom iznosu bliži jedinici, to su varijable jače linearno korelirane. Korelacija između ukupnog broja bodova i ostvarenog rezultata na međuispitu iznosi 0.843, što pokazuje veliku korelaciju između ove dvije varijable, te možemo pretpostaviti snažnu linearnu vezu. Da je to istina možemo se uvjeriti tzv. točkastim grafom (engl. scatterplot) između ove dvije varijable (Slika 4.6.b). Zadovoljeni su svi preduvjeti za linearnu regresiju. No, što ako veza između ove dvije varijable nije u potpunosti linearna, odnosno što ako bi vezu mogli bolje izraziti nekom krivuljom (Slika 4.6.a). Metoda koja nam pomaže tu je metoda polinomijalne regresije. Pitanje je kako izabrati koja

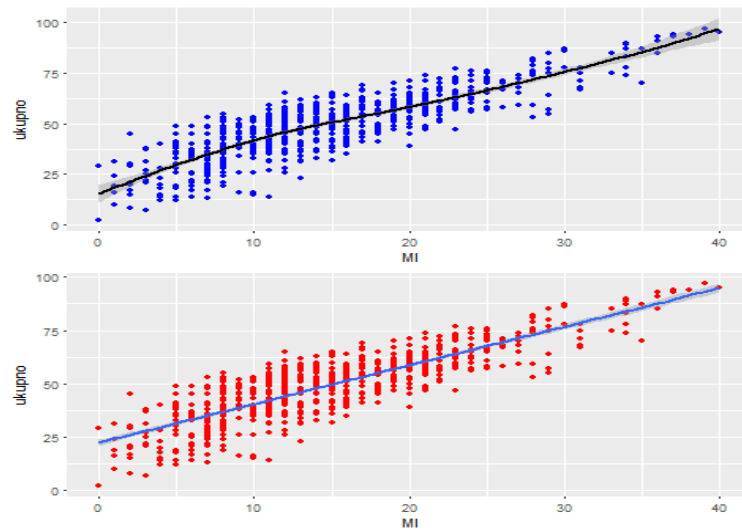


**Slika 4.5:** Distribucija ukupnog broja bodova

je metoda bolja, odnosno koja metoda će bolje generalizirati nad novim podacima (uvijek trebamo paziti na prenaučnost). Slika 4.7 prikazuje dio koda kojim provodimo eksploratornu analizu.

### Izgradnja modela

Uveli smo pojmove podjele podataka na trening i test skupove te pojam unakrsne provjere. Ovdje ćemo iskoristiti hibridan pristup. Podijeliti ćemo podatke na trening i test skup, te ćemo na skupu za treniranje provesti metodu unakrsne provjere (verzija k-fold CV, pri čemu parametar k iznosi 10) kako bi mogli odrediti koji model nam je bolji, a boljim modelom smatramo onaj koji ima manju usrednjenu grešku (jedan od tipičnih odabira za grešku je MSE, no možemo i druge koristiti). Kako znati koji će polinom pri polinomijalnoj regresiji biti najprecizniji? Jednostavno ćemo isprobati polinome do neke granice te ih uspoređivati međusobno te sa modelom linearne regresije (možemo reći i polinomom 1. stupnja), pri čemu ćemo ispitivati polinome do 5. stupnja. Još jedan način odabira najboljeg modela je upotrebom metode ANOVA (engl. Analysis of Variance - analize varijance), koja nam govori koji model najbolje objašnjava varijabilnost podataka. Usporedba usrednjenih grešaka unakrsnom provjerom te ANOVA metoda (Slika 4.12) se oboje slažu da je model polinomijalne regresije s polinomom 3. stupnja najbolji, te ćemo s tim modelom nastaviti analizu. Sad kad smo odabrali model treniramo ga nad cijelim skupom za treniranje. Na slici 4.10 vidimo dio koda koji



**Slika 4.6:** a) vezu opisujemo polinomom (gornja slika) , b) vezu opisujemo linearnom funkcijom (donja slika)

gradi modele i vrši unakrsnu provjeru za odabir najboljeg modela.

Kvalitetu modela provjeravamo :

- provjerom normalnosti reziduala (Slika 4.8). Mada nisu savršeni, možemo reći da je sasvim zadovoljavajuća razina normalnosti reziduala. Također vidimo da ne postoji očit uzorak pri rezidualima što nam je i potrebno (Slika 4.9)
- koeficijentom determinacije koji iznosi 0.7149, što je dobar pokazatelj linearne veze između ciljne i regresorske varijable.

```
summary(osnovni$ukupno)

boxplot(trening$ukupno,col="red")

ggplot(osnovni,aes(x=ukupno))+geom_histogram(bins=10,
                                              fill="blue",color="black")
+labs(title="Distribucija ukupnog broja bodova",
      x="ukupan broj bodova",y="broj studenata")

cor(osnovni$MI,osnovni$ukupno)
```

**Slika 4.7:** Eksploratorna analiza podataka

## Predikcija

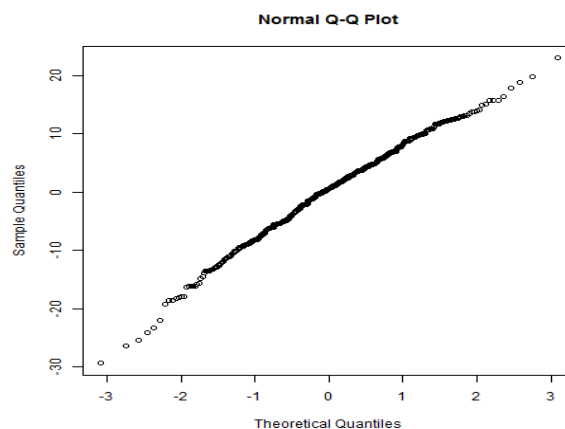
Važno nam je kako model radi s novim podacima (što kod nas predstavljaju podaci koje smo odvojili u skup za testiranje), odnosno da li ispravno predviđa vrijednosti. Kako interpretirati rezultate predikcije ? Možemo to učiniti kroz par mogućnosti :

- Predstavljanjem greške MAE, koja nam govori koliko u prosjeku model promašuje broj bodova
- Možemo stvoriti interval predikcije, te gledati koliko često stvarne vrijednosti upadaju u interval koji smo predvidjeli.
- Prikazom histograma grešaka

Slika 4.13 prikazuje dio koda kojim predviđamo nove vrijednosti.

Nakon obavljene predikcije, dobili smo iduće rezultate :

- MAE nam iznosi 6.15, dakle predviđena vrijednost u prosjeku je drugačija od stvarne vrijednosti za otprilike 6 bodova
- Od 213 zapažanja iz skupa za treniranje, od 207 je stvarna vrijednost unutar intervala predikcije.
- Histogram greške (Slika 4.11) je desno nakrivljen, unimodalan, te nam pokazuje da je većina grešaka manjeg iznosa.
- Usporedbom sažetka 5 brojeva te aritmetičke sredine predviđenih vrijednosti i stvarnih vrijednosti (Tablica 4.2) vidimo da vrijednosti između prvog i trećeg kvartila predviđamo s velikom preciznošću. Problem su ekstremno male (pogotovo) i ekstremno velike vrijednosti, kod kojih su veći



**Slika 4.8:** prikaz normalnosti reziduala sa kvantil-kvantil grafom

**Tablica 4.2:** prvi red predstavlja predviđene vrijednosti, drugi stvarne

minimum	1. kvartil	medijan	3. kvartil	maksimum	ar. sredina
15.88	43.36	50.56	58.57	97.45	51.04
2.0	42.0	52.0	59.0	95.0	51.23

**Tablica 4.3:** Sažetak 5 brojeva i aritmetička sredina varijable bodova na završnom ispitu

minimum	1. kvartil	medijan	3. kvartil	maksimum	ar. sredina
0	14	20	24	38	19.24

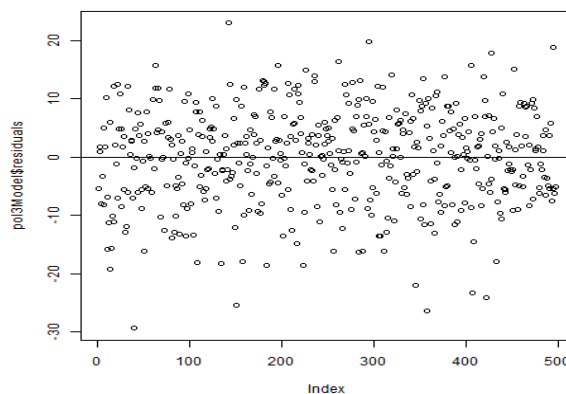
promašaji.

#### 4.5.2. Predviđanje broja bodova na završnoj provjeri pomoću prethodnih provjera - korištena metoda regresijsko stablo

##### EDA

Prije završne provjere imamo tri provjere : međuispit i dvije kratke provjere znanja. Pošto opet predviđamo numeričku varijablu pratimo postupak prve analize:

- Sažetak 5 brojeva i aritmetička sredina varijable (Tablica 4.3)
- Distribucija broja bodova na završnoj provjeri(Slika 4.14), koja je sasvim zadovoljavajuća u vidu normalnosti



**Slika 4.9:** graf reziduala



## Izgradnja modela

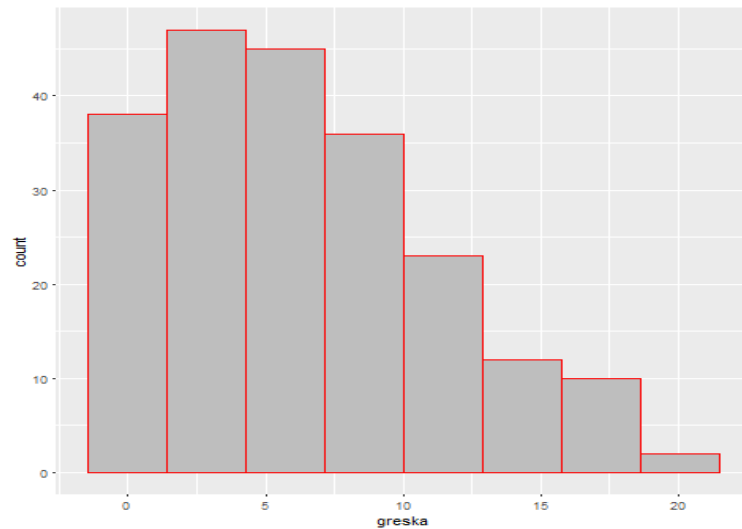
Iako su tradicionalne regresijske metode (poput linearne regresije) redovito prvi izbor za predviđanje numeričke varijable, u nekim slučajevima su stabla odlučivanja za numeričke podatke bolji izbor. Naprimjer, za zadatke koji sadrže puno prediktora ili prediktore između kojih je kompleksna, nelinearna veza (iako ovdje nemamo takav slučaj, zanemarit ćemo to u svrhu primjene regresijskog stabla). Za analizu ovog problema ćemo izgraditi jedno takvo regresijsko stablo. Regresijsko stablo ima prednost puno lakšeg vizualiziranja rezultata, odnosno puno je lakše objasniti običnom čovjeku (čak jednostavnije i od linearne regresije). Podatke dijelimo na skup za treniranje i skup za testiranje (omjer 70%-30%). Izgradnjom stabla imamo lagani način za pratiti predviđanje broja bodova sa završnog jednostavnim praćenjem grana koje odgovaraju vrijednostima istraživnog podatka (Slika 4.15). Ovdje možemo vidjeti jednu prednost ove metode, a to je automatski odabir varijabli. Naime, primjećujemo kako nijedan čvor ne dijeli stablo ovisno o rezultatu ostvarenom na prvoj kratkoj provjeri znanja (1. KPZ). Ako bi ušli dublje u način slaganja provjera na ovom predmetu shvatili bi da to i ima smisla. Jedan od razloga zašto je međuispit bitan za predviđanje bodova na završnom ispitu je svakako činjenica da međuispit pokazuje kako se student sprema za velike provjere. Dok kod druge kratke provjere znanje je svakako bitna spoznaja da je na završnom ispitu u velikoj mjeri zastupljeno gradivo koje se provjeravalo na drugoj kratkoj provjeri znanja. Slika 4.16 prikazuje izgradnju regresijskog stabla.

Rezultate predviđanja prikazujemo na sličan način kao i kod prve analize :

- MAE iznosi 5.321

```
linModel<-glm(data=trening,ukupno~MI)
pol2Model<-glm(data=trening,ukupno~poly(MI,2))
pol3Model<-glm(data=trening,ukupno~poly(MI,3))
pol4Model<-glm(data=trening,ukupno~poly(MI,4))
pol5Model<-glm(data=trening,ukupno~poly(MI,5))
cvErrors<-vector()
set.seed(1234)
cvErrors[1]<-cv.glm(trening,linModel,k=10)$delta[1]
cvErrors[2]<-cv.glm(trening,pol2Model,k=10)$delta[1]
cvErrors[3]<-cv.glm(trening,pol3Model,k=10)$delta[1]
cvErrors[4]<-cv.glm(trening,pol4Model,k=10)$delta[1]
cvErrors[5]<-cv.glm(trening,pol5Model,k=10)$delta[1]
cvErrors
```

**Slika 4.10:** Izgradnja modela i unakrsna provjera



**Slika 4.11:** histogram pogreške između stvarne i predviđene vrijednosti

#### Analysis of Variance Table

```

Model 1: ukupno ~ MI
Model 2: ukupno ~ poly(MI, 2)
Model 3: ukupno ~ poly(MI, 3)
Model 4: ukupno ~ poly(MI, 4)
Model 5: ukupno ~ poly(MI, 5)

```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	495	33341				
2	494	33076	1	265.40	3.9957	0.04617 *
3	493	32714	1	361.45	5.4416	0.02007 *
4	492	32659	1	55.44	0.8346	0.36138
5	491	32613	1	45.15	0.6798	0.41007

**Slika 4.12:** Ispis metode anova, koja uspoređuje RSS (engl- residual sum of squares - suma grešaka reziduala), te javlja da li je smanjenje RSS-a dovoljno kako bi preferirali kompleksniji model

- Histogram greške (Slika 4.17) je desno nagnut, unimodalan, te nam pokazuje da je veći broj malih grešaka.
- Usporedbom sažetka 5 brojeva te aritmetičke sredine (Tablica 4.4) predviđenih vrijednosti i stvarnih vrijednosti uporabom metode regresijskog stabla dobro predviđamo vrijednosti oko medijana.

**Tablica 4.4:** Usporedba stvarnih i predviđenih vrijednosti kroz sažetak 5 brojeva

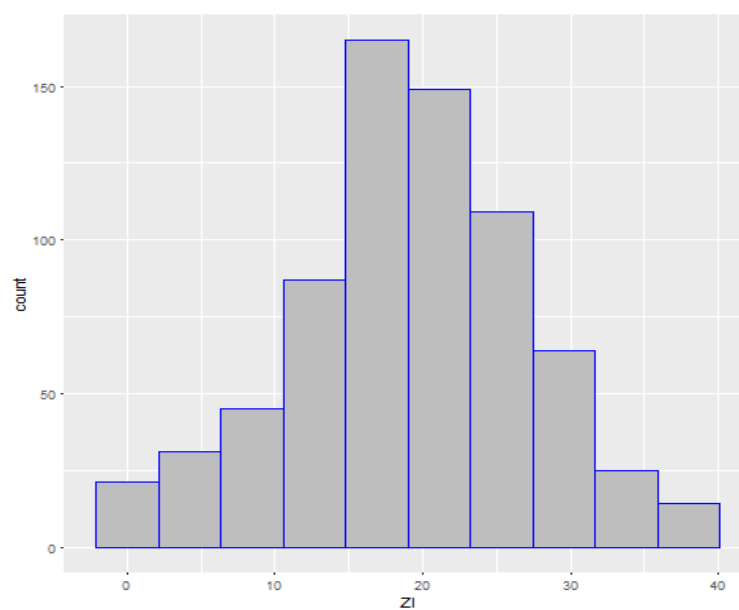
minimum	1. kvartil	medijan	3. kvartil	maksimum	ar. sredina
0	14	20	25	37	19.5
6.143	17.92	20.94	20.94	32.11	19.8

```

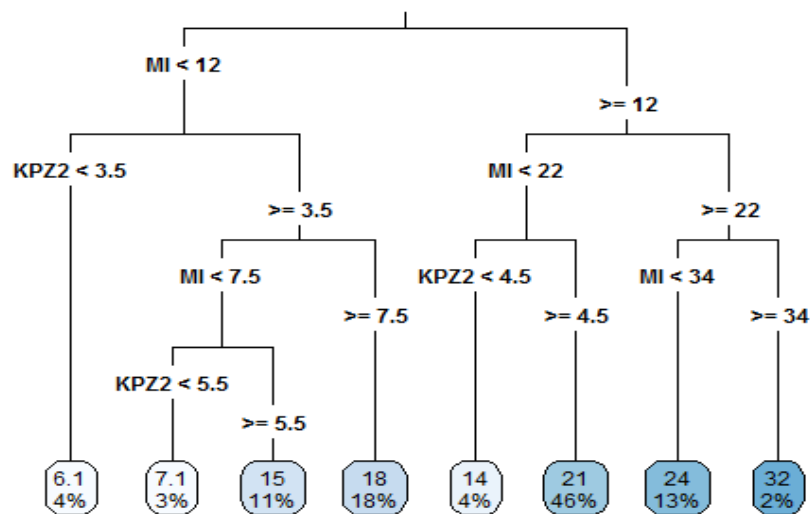
predictedValues<-predict (pol3Model ,data.frame(MI=test$MI),
interval ="prediction")
predictedValues<-data.frame(predictedValues)
predictedValues$stvarni<-test$ukupno
mae(predictedValues$stvarni,predictedValues$fit)
test[,c(4,6)]
predictedValues$stvarni<-test$ukupno
predictedValues
mae(predictedValues$stvarni,predictedValues$fit)

```

**Slika 4.13:** Predviđanje novih vrijednosti



**Slika 4.14:** Distribucija ukupnog broja bodova



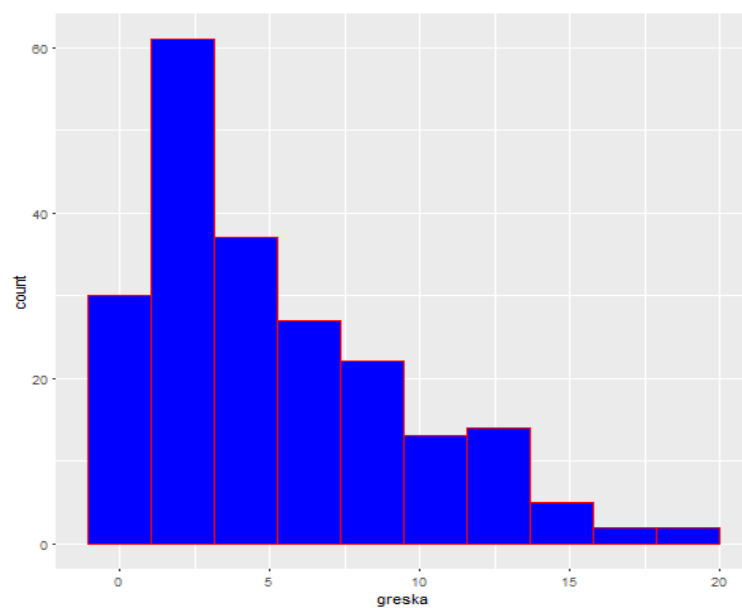
Slika 4.15: Regresijsko stablo

```

rTree<-rpart(data = trening, ZI~MI+KPZ1+KPZ2,method="anova")
rpart.plot(rTree,type=3,fallen.leaves = T)

```

Slika 4.16: Izgradnja regresijskog stabla



Slika 4.17: Distribucija greške

# ZAKLJUČAK

Znanstveno polje dubinske analize podataka koji dolaze iz edukacijske domene (EDM) je u velikom zamahu. Iako bi neki rekli da smo ušli u "BigData" eru, to nije istina. Podaci su već dulje vremena tu, zapravo smo sa napretkom tehnologije došli do tehnika za lakše izvlačenje i analizu podataka. Strojno učenje i DM su razvili mnoge alate i tehnike, koje u EDM-u uz prilagodbu iskoristavamo kako bi riješili probleme koje pred nas stavljaju takvi ogromni skupovi podataka. EDM je još uvijek relativno malo znanstveno polje, ali je sve više znanstvenika i inženjera koji su zainteresirani za EDM.

U ovom radu je prikazano kako pravilno primjeniti metode strojnog učenja (regresijske) nad velikim skupom podataka iz edukacijske domene za predviđanje novih vrijednosti (numeričkih). To nije jednostavan proces, jer moramo pravilno pripremiti, obraditi i istražiti podatke kako bi analiza i evaluacija rješenja problema bila provedena na ispravan način. Rezultati koje smo dobili nisu "savršeni", ali to i ne trebaju biti, svakako možemo reći da su korisni i da nam mogu pomoći u procesu boljeg razumijevanja kako rješavati probleme i poboljšati proces učenja kod studenata.

# LITERATURA

- [1] Jaume Bacardit. Large scale data mining using genetics-based machine learning. <https://www.slideshare.net/xllora/large-scale-data-mining-using-geneticsbased-machine-learning>, 2013.
- [2] David M Diez. *OpenIntro Statistics*. 2016.
- [3] David J. Hand. *Principles of Data Mining*. 2001.
- [4] Gareth James. *An Introduction To Statistical Learning*. 2013.
- [5] Alex Jones. Data science skills and business problems. <http://www.kdnuggets.com/2014/06/data-science-skills-business-problems.html>, 2014.
- [6] Brett Lantz. *Machine Learning with R*. 2013.
- [7] Calvet Liñán. Educational data mining and learning analytics: differences, similarities, and time evolution. <http://rusc.uoc.edu/rusc/ca/index.php/rusc/article/view/v12n3-calvet-juan/2746.html>, 2015.
- [8] Damir Pintar. *Programirajmo u R-u*. 2017.
- [9] Cristobal Romero. Data mining in education, 2013.

## Primjena regresijske analize u edukacijskoj domeni

### Sažetak

U ovom radu opisana su i objašnjena ključna znanstvena polja te cjelokupan proces za dubinsku analizu podataka nad podacima iz edukacijske domene. Objašnjene su i veoma bitne regresijske metode te kako ih pravilno upotrijebiti za rješavanje danih problema, također su i objašnjene bitne tehnike koje nam pomažu pri primjeni takvih metoda

Osim teorijskog dijela, u sklopu rada provedena je i konkretna analiza u kojoj su primjenjene prethodno opisane metode, tehnike i znanja. Na kraju su prezentirani rezultati predviđanja novih podataka dobivenih korištenjem regresijskih metoda.

**Ključne riječi:** linearna regresija, regresijsko stablo, strojno učenje, dubinska analiza podataka iz edukacijske domene

## Application of Regression Analysis in the Educational Domain

### Abstract

This work describes and explains important science fields and whole process of educational data mining. Also, it explains very important regression methods and how to apply them properly on given problems, with addition of explaining useful techniques during application of such methods.

In addition to the theoretical part of this paper, a concrete analysis was carried out in which were applied previously mentioned methods and techniques. At the end are presented the analysis results of predicting new values with regression methods.

**Keywords:** linear regression, regression trees, machine learning, educational data mining