Univerzitet u Nišu

Elektronski fakultet

Predmet: Prikupljanje i predobrada podataka za mašinsko učenje

Kvalitet podataka

Seminarski rad

Smer: Veštačka inteligencija i mašinsko učenje

Student:

Anđelija Mladenović, br. ind. 1625

Profesor:

Doc. dr Aleksandar Stanimirović

Niš, februar 2024. godine

Sadržaj

1. Uvod

Mnoge velike tehnološke promene odigrale su se u industriji informacionih tehnologija od početka 21. veka: došlo je do razvoja računarsta „u oblaku“ (engl. *Cloud Computing*), interneta stvari (engl. *Internet of Things* – IoT), društvenih mreža... Razvojem navedenih oblasti, povećavala se naša sposobnost prikupljanja podataka – ove oblasti najavile su dolazak „velikih podataka“ (engl. *Big Data*) i svega što je usledilo [1].

Naravno, sa prikupljanjem velikog broja podataka, došla je i želja i potreba za analizom istih. Istraživači, naučnici i ljudi u raznim industrijama brzo su shvatili da znanje dobijeno analizom ovih podataka donosi višestruku korist – od boljeg razumevanja potreba korisnika, poboljšanja kvaliteta usluga, pa sve do predviđanja i prevencije rizika [1].

Ipak, zaključci se ne mogu izvoditi tek tako – kvalitet zaključaka ogleda se u kvalitetu podataka na kojima su oni bazirani. Samo visoko kvalitetni podaci mogu rezultirati vrednim zaključcima [1]. Zbog toga, akvizicija podataka i njihova validacija predstavljaju značajne probleme. Kvalitetni podaci donose korist u vidu bolje informisanih i bržih odluka, smanjenja troškova i optimizacije određenih procesa [2].

Postavlja se pitanje – šta zapravo predstavlja kvalitet podataka? Odgovor na ovo pitanje zavisi od samog problema koji se tim podacima modeluje i često se definiše time koliko su podaci prikladni (engl. *data fitness*) za primenu u datoj oblasti. Kvalitet podataka zavisi od toga koliko su podaci potpuni, konzistentni, da li sadrže duplikate, i da li su tačni i vremenski relevantni za problem za koji se koriste [2].

Definisanje i provera kvaliteta podataka je težak zadatak, jer su podaci prikupljeni u jednom, a potrebno ih je primeniti u potpuno drugačijem kontekstu. Pored toga, provera kvaliteta podataka je određena samom oblašću na koju se podaci odnose, često nije objektivna i zahteva ljudski uticaj i znanja [2]. Kvalitetni podaci preduslov su za analizu i korišćenje prikupljenih podataka, i garancija vrednosti tih podataka [1]. Zbog toga, razvijen je širok spektar metoda specifičnih za domen (engl. *domain specific methods*) kako bi se izvršila provera i poboljšao kvalitet podataka [2].

2. Atributi i njihova podela

Skup podataka često se posmatra kao kolekcija objekata podataka. Druga imena za objekat podataka jesu rekord, tačka, vektor, obrazac, događaj, slučaj, uzorak (engl. *sample*), obzervacija ili entitet. Svaki objekat opisan je nizom atributa koji opisuju osnovne karakteristike tog objekta – na primer, atribut nekog predmeta može biti njegova masa, dok vremenski trenutak može biti atribut nekog događaja i opisuje kada se taj događaj odigrao. Druga imena za atribute jesu varijable, karakteristike, polja, osobine (engl. *feature*) ili dimenzije [3].

Podaci se obično predstavljaju i čuvaju u vidu tabele, gde objekti predstavljaju vrste (redove) tabele, a svako polje vrste, odnosno kolone, predstavljaju određeni atribut. Tabela 2.1 predstavlja primer ovakvog prikaza i skladištenja podataka. Svaka vrsta tabele predstavlja jednog studenta, dok svaka kolona sadrži samo određene podatke o studentima, kao što su, na primer, broj indeksa, prosečna ocena i sl. Presek određene vrste i određene kolone sadrži odgovarajući podatak o odgovarajućem studentu [3].

Broj indeksa	Smer	Prosečna ocena	...
...	...	....	...
1234	Računarstvo i informatika	7.80	...
1235	Veštačka inteligencija i mašinsko učenje	8.92	...
1236	Automatsko upravljanje	9.00	...
...	...	...	...

Tabela 2.1 – Podaci o studentima [3]

Atribut je svojstvo ili karakteristika objekta koje može da varira u zavisnosti od objekta ili u zavisnosti od vremenskog trenutka u kom je atribut zabeležen. Na primer, boja očiju varira od osobe do osobe, dok temperatura varira u vremenu. Ovde je takođe značajno primetiti da je boja očiju predstavljena imenom boje (simbolom) i ima mali broj mogućih vrednosti (npr. braon, zelene, plave, sive i sl.), dok je temperatura predstavljena numeričkom vrednošću i ima potencijalno neograničen broj mogućih vrednosti [3].

Na najosnovnijem nivou, atributi nisu samo brojevi i simboli. Ipak, radi preciznije analize karakteristika objekta, dodeljujemo im određene vrednosti. Da bismo ovo uradili na jasno definisan način, neophodna nam je merna skala [3].

Merna skala je pravilo (funkcija) koje asocira numeričku ili simboličku vrednost sa atributom nekog objekta. Proces merenja predstavlja primenu merne skale za preslikavanje atributa objekata u neku vrednost. Iako ovo može delovati apstraktno, svakodnevno obavljamo različita merenja. Na primer, koristimo vagu da izmerimo masu nečega, klasifikujemo ljude u muškarce i žene, ili prebrojavamo stolice u prostoriji da bi proverili da li ima dovoljno mesta za sedenje. U svim ovim slučajevima, „fizička vrednost“ atributa objekta se mapira na simboličku ili numeričku vrednost [3].

Bitno je primetiti da svojstva atributa ne moraju biti nužno ista kao svojstva vrednosti koje se koriste za njihovo merenje. Ovo znači da atributi mogu imati svojstva koja vrednosti iskorišćene za njihovo predstavljanje ne poseduju, ili obrnuto. Primer ovoga mogu predstavljati ID nekog zaposlenog i njegova starost. Oba atributa mogu se predstaviti prirodnim brojevima. Ipak, iako je savršeno smisleno računati prosečni broj godina svih zaposlenih, nema ni malo smisla govoriti o prosečnoj vrednosti identifikacionog broja. Ovo je proisteklo iz toga što je jedina svrha identifikacionog broja radnika da se osigura da su svi radnici različiti ljudi, te je stoga jedina validna operacija nad ID brojevima provera da li su dva broja jednaka. Međutim, ne postoji nikakva implikacija ovoga u tipu podataka koji je iskorišćen za predstavljanje ID broja. U slučaju godina, svojstva brojeva koji su iskorišćeni za predstavljanje starosti gotovo se u potpunosti slažu sa svojstvima atributa. Ipak, i ovde postoje izuzeci – starost mora biti u određenim granicama, dok brojevi nisu ograničeni [3].

Zbog svega navedenog, neophodno je postojanje različitih tipova atributa. Tip atributa treba da nam stavi do znanja koja svojstva atributa su istovremeno i svojstva vrednosti kojima se atribut predstavlja i obrnuto. Jednostavan način da se specificira tip atributa jeste da se identifikuju svojstva brojeva koja odgovaraju svojstvima atributa. Sledeća svojstva (operacije) brojeva se tipično koriste za opisivanje atribura [3]:

- Različitost (= , ≠)
- Uređenost (< , > , ≤ , ≥)
- Sabirljivost (+ , -)
- Pomnoživost (* , /)

Na osnovu ovih svojstava, može se izvršiti podela atributa na sledeće tipove [4]:

- Kategorički atributi
  - Nominalni
  - Binarni
  - Ordinalni
- Numerički atributi
  - Atributi zasnovani na intervalima
  - Atributi zasnovani na razmeri

2.1 Nominalni (kategorički) atributi

Nominalno znači „odnositi se na imena“ – vrednosti nominalnih atributa su simboli ili, bukvalno, imena stvari. Svaka vrednost predstavlja neku vrstu kategorije, koda ili stanja. Zbog ovoga, nominalni atrinuti često se nazivaju i kategorički atributi. Vrednosti ovih atributa nemaju nikakav smisleni poredak. U računarstvu, ovakvi atributi poznati su pod nazivom enumeracije [4].

Na primer, boja kose predstavlja nominalni atribut – može imati vrednosti „crna“, „smeđa“, „plava“ i dr. Iako smo rekli da su vrednosti nominalnih atributa simboli ili „imena stvari“, moguće je predstaviti ove simboje brojevima. U slučaju boje kose, možemo dodeliti kod 0 vrednosti „crna“, kod 1 vrednosti „smeđa“, kod 2 vrednosti „plava“ i tako dalje. Ipak, u ovoj situaciji, brojevi nisu namenjeni da se kvantitativno koriste, odnosno matematičke operacije na vrednostima nominalnih atributa su besmislene. Iako nominalni atributi mogu imati cele brojeve za svoje vrednosti, oni se tada ne mogu posmatrati kao numerički atributi jer ti brojevi nisu podložni bilo kakvoj manipulaciji niti kombinovanju [4].

Zbog činjenice da se vrednosti nominalnih atributa ne mogu urediti ni u kakav smisleni poredak, nema smisla tražiti njihovu srednju ili medijalnu vrednost. Ono što ima smisla tražiti jeste najfrekventnija vrednost, odnosno mod ili modus (engl. *mode*). Ova vrednost može se koristiti kao mera centralne tendencije, o čemu će biti reči kasnije [4].

2.2 Binarni atributi

Binarni atributi predstavljaju podvrstu nominalnih atributa koji imaju samo dve kategorije ili stanja: 0, koje obično predstavlja odsustvo atributa, i 1, koja obično predstavlja prisustvo posmatranog atributa. U računarstvu, ovakvi atributi često se predstavljaju *boolean* vrednostima [4].

Primer ovakvih atributa može biti podatak da li pacijent puši ili ne, gde će stanje 0 predstavljati činjenicu da pacijent ne puši, dok će stanje 1 implicirati da je pacijent pušač [4].

Binarni atributi su simetrični ukoliko su oba njihova stanja podjednako značajna i imaju istu težinu, odnosno ne postoji preferenca koje stanje treba biti kodirano jedinicom, a koje nulom. Primer ovakvog atributa može biti pol osobe [4].

Nasuprot tome, binarni atribut je asimetričan ukoliko njegova dva stanja nisu podjednako značajna. Primer ovakvog atributa jeste ishod medicinskog testa na HIV. Prema konvenciji, najznačajniji ishod (stanje), koji je ujedno i ređi, kodira se jedinicom (1 = HIV pozitivan), dok se drugi ishod kodira nulom (0 = HIV negativan) [4].

2.3 Ordinalni atributi

Ordinalni atributi predstavljaju kategoričke atribute čije vrednosti se mogu urediti u smisleni poredak, odnosno između čijih vrednosti postoji neko rangiranje, pri čemu magnituda između dve uzastopne vrednosti atributa nije poznata [4].

Primer ovakvog atributa može biti veličina pića dostupna u nekom restoranu brze hrane. Ovaj atribut obično može imati tri vrednosti – „malo“, „srednje“ i „veliko“. Ove vrednosti mogu se poređati u smislenu sekvencu, koja odgovara povećanju veličine pića. Međutim, ne možemo iz vrednosti atributa zaključiti koliko je, na primer, „veliko“ piće veće od „srednjeg“ [4].

Ordinalni atributi mogu se dobiti i diskretizacijom numeričkih količina podelom njihovog opsega vrednosti u konačan broj uređenih kategorija. Centralna tendencija ovakvih podataka može se predstaviti modom i medijalnom vrednošću, ali srednja vrednost ne može biti definisana [4].

Bitno je primetiti da su nominalni, binarni i ordinalni atributi kvalitativni, odnosno da oni opisuju neku karakteristiku objekta bez davanja stvarne veličine ili količine. Vrednosti ovih atributa obično su reči koje predstavljaju kategorije, ili brojevi gde svaki broj predstavlja jednu kategoriju [4].

2.4 Numerički atributi

Numerički atributi su kvantitivni – to je merljiva količina nečega, predstavljena celim ili realnim brojem. Numerički atributi mogu biti zasnovani na intervalima ili na razmeri [4].

Atributi zasnovani na intervalima mere se skalom koja se sastoji od ravnomerno raspoređenih jedinica. Vrednosti ovakvih atributa mogu biti pozitvne, nula ili negativne. Zbog toga, pored rangiranja vrednosti, ovi atributi omogućavaju i kvantifikovanje razlike između nekih vrednosti. Primer ovakvog atributa može biti temperatura. Na primer, ukoliko nekoliko dana merimo temperaturu, možemo poređati objekte (dane kada su temperature merene) na osnovu ovog atributa. Pored toga, možemo videti razliku u temperaturi između, na primer, dva uzastopna dana. Još jedan primer ovakvog atributa su godine – 2002. godina je bila 8 godina pre 2010. godine [4].

Zbog toga što su atributi zasnovani na intervalima numerički, za njih je moguće izračunati srednju i medijalnu vrednost, mod i druge statistike [4].

Temperatura, bilo izražena u Celzijusovim ili Farenhajtovim stepenima, nema pravu nultu tačku, odnosno ni $0^{\circ}C$ ni $0^{\circ}F$ ne predstavlja odsustvo temperature. Iako možemo izračunati razliku između dve vrednosti temperature, ne možemo govoriti o vrednosti temperature kao o umnošku neke druge temperature. Bez prave nulte vrednosti, nema smisla reći da je 10 stepeni duplo više od 5. To znači da ne možemo zaista govoriti o razmerama. Slično ovome, ne postoji nulta tačka za kalendarse datume ni godine (nulta godina ne predstavlja početak vremena – čak i ne postoji!) [4].

Zbog navedenih ograničenja atributa zasnovanih na intervalima, postoje i atributi zasnovani na razmeri. Ovi atributi uvek poseduju nultu tačku. Zbog ovoga, moguće je govoriti o jednoj vrednosti kao o umnošku druge, ili o tome kako su te dve vrednosti u nekoj razmeri. Pored toga, ove vrednosti imaju određeni poredak, i može se izračunati razlika između dve vrednosti, kao i srednja, medijalna vrednost i mod ovih podataka [4].

Na primer, za razliku od Celzijusa i Farenhajta, Kelvinova temperaturna skala poseduje apsolutnu nulu – to je tačka u kojoj čestice od kojih se materija sastoji više ne poseduju kinetičku energiju. Drigi primeri atributa zasnovanih na razmeri su atributi koji nešto prebrojavaju – godine radnog iskustva zaposlenih, broj reči u tekstu, ali i atributi kao što su masa, visina, geografska širina i dužina, koordinate i slično [4].

2.5 Diskretni i kontinualni atributi

Iako smo do sada govorili o podeli atributa na nominalne i numeričke, ovo nije jedini način klasifikacije atributa. Postoji više načina podele atributa na tipove, i ove podele nisu međusobno isključive [4].

Algoritmi za klasifikaciju razvijeni u polju mašinskog učenja često govore o atributima kao o diskretnim i kontinualnim. Svaki od ovih tipova neohodno je obraditi na specifičan način. Diskretni atributi imaju konačan ili prebrojivo beskonačan broj vrednosti, koje mogu, ali ne moraju biti predstavljene celim brojevima. Atributi kao što su boja kose, veličina pića i da li je osoba pušač imaju konačan broj vrednosti, te su stoga diskretni. Treba primetiti da diskretni atributi mogu imati numeričke vrednosti, kao što su 0 i 1 za binarne atribute, ili vrednosti od 0 do 110 za godine. Atribut ima prebrojivo beskonačno vrednosti ako je broj njegovih mogućih vrednosti beskonačan, ali postoji 1-na-1 preslikavanje između vrednosti atributa i prirodnih brojeva. Na primer, atribut ID kupca ima prebrojivo beskonačno vrednosti. Broj kupaca može rasti do u beskonačnost, ali u realnosti, kupci su prebrojivi. Drugi ovakav primer su poštanski brojevi [4].

Ako atribut nije diskretan, onda je kontinualan. Iako se u literaturi često simultano koriste izrazi “numerički” i “kontinualni”, ovo može biti dosta zbunjujuće. Kontinualni podaci se obično predstavljaju podacima u pokretnom zarezu, te (ako se izuzmu ograničenja tehnologije i činjenica da se i za ove vrednosti koristi konačan broj cifara) ne postoji konačan broj mogućih vrednosti ovakvih atributa [4].

3. Statističke mere kvaliteta podataka

Da bi preprocesiranje podataka bilo uspešno, neophodno je imati opštu sliku o podacima sa kojima se radi. Statistički opis podataka može se upotrebiti za identifikaciju nekih svojstava podataka i za uočavanje šuma (engl. *outliers*) [4]. Svojstva podataka su značajna jer nam opisuju same podatke, ali i način na koji bi trebalo manipulisati tim podacima da bi se poboljšao njihov kvalitet i, samim tim, izvukao maksimum iz tih podataka. Poželjna svojstva unutar podataka zavise od same primene za koju se podaci koriste.

Neke od statističkih metoda kojima ćemo se baviti biće mere centralne tendencije, mere disperzije podataka i mere oblika distribucije podataka.

3.1 Mere centralne tendencije

Pretpostavimo da imamo neki atribut X, na primer godine radnog iskustva zaposlenih, i da je ovaj podatak izmeren za skup objekata. Neka je $x_1,\ x_2, \dots, x_N$ skup od N opservacija za atribut X. Ukoliko bismo grafički predstavili ove vrednosti, gde bi se nalazila većina vrednosti? Odgovor na ovo pitanje može da da mera centralne tendencije podataka. Najčešće korišćene mere centralne tendencije su srednja vrednost, medijalna vrednost, mod podataka i srednji opseg podataka.

3.1.1 Srednja vrednost

Možda i najčešće korišćena mera centralne tendencije jeste srednja vrednost, ili drugačije, aritmetička sredina. Neka je $x_1, x_2, \dots, x_N$ skup od N opservacija za atribut X. Aritmetička sredina vrednosti ovog atributa može se izračunati po sledećoj formuli [4][5]:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N} \tag{3.1}$$

Na primer, ukoliko bismo za pet radnika imali broj godina radnog iskustva (3, 5, 2, 1 i 4 godine), aritmetička sredina njihovog radnog iskustva bila bi:

$$\bar{x} = \frac{3 + 5 + 2 + 1 + 4}{5} = \frac{15}{5} = 3\ godine$$

Za neke primene, može se dogoditi da nam nisu sve opservacija istog atributa podjednako važne, odnosno da svaka ima svoju težinu. Težina odražava značaj ili frekvenciju ponavljanja neke vrednosti. Ukoliko želimo da podaci imaju težine, srednju vrednost možemo izračunati po formuli:

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N} \tag{3.2}$$

gde $w_1,\ w_2, \ldots,\ w_N$ predstavljaju vrednosti težina za odgovarajući podatak. Ovako izračunata srednja vrednost naziva se težinska aritmetička sredina [4].

Iako je srednja vrednost jako korisna za opisivanje atributa skupa podataka, nije uvek najbolja metoda za merenje centralne tendencije. Značajan problem aritmetičke sredine leži u njenoj osetljivosti na ekstremne vrednosti [4]. Na primer, ukoliko bismo u prethodnom primeru sa računanjem srednje vrednosti radnog iskustva radnika imali 4 nova radnika koji u kompaniji rade po godinu dana, i jednog koji u kompaniji radi 8 godina, srednja vrednost bi bila:

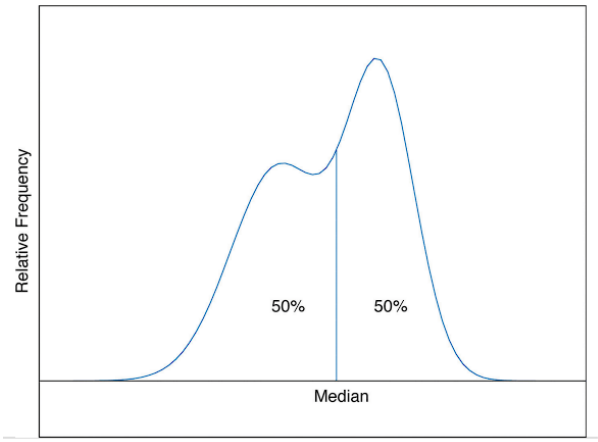
$$\bar{x} = \frac{1 + 1 + 1 + 1 + 8}{5} = \frac{12}{5} = 2.4\ godine$$

što ne bi oslikavalo realnu situaciju, a to je da skoro svi imaju samo godinu dana radnog iskustva.

Čak i mali broj ekstremnih vrednosti može jako nepovoljno uticati na aritmetičku sredinu. Jedan od načina da se izbegne ovakav efekat jeste računanje „odsečene“ srednje vrednosti – ovo se postiže time što se vrednosti opservacija atributa sortiraju, izbace se ekstremne vrednosti, i onda računa aritmetička sredina preostalih vrednosti podataka. Problem sa ovim pristupom je to što može doći do gubitka informacija, naročito ukoliko se izbaci veliki broj ekstremnih vrednosti [4].

3.1.2 Medijalna vrednost (medijana)

U situacijama kada je raspodela podataka asimetrična, ili onda kada ima mnogo vrednosti van očekivanih granica (engl. *outliers*), umesto aritmetičke sredine bolje je računati medijalnu vrednost (medijanu) podataka. Medijalna vrednost skupa podataka računa se tako što se podaci sortiraju od najmanjeg ka najvećem, a onda se za medijanu uzme vrednost koja se nalazi na sredini sortitanog niza. Ukoliko imamo paran broj vrednosti za koji tražimo medijanu, medijana se računa kao aritmetička sredina dve vrednosti koje se nalaze oko centra sortiranog niza. Ono što je značajno za medijanu jeste da tačno 50% podataka iz skupa ima vrednost manju ili jednaku medijalnoj, dok druga polovina skupa ima vrednost veću ili jednaku medijalnoj [4][5].



Slika 3.1 Vizuelni prikaz medijalne vrednosti [5]

U prethodnom primeru sa radnim iskustvom zaposlenih, medijana bi se računala na sledeći način:

$$1, 1, 1, 1, 8$$

Dakle, medijalna vrednost je 1, što mnogo bolje oslikava realno stanje u kompaniji.

Iako se medijalna vrednost generalno primenjuje na numeričke podatke, ovaj koncept može se primeniti i na ordinalne podatke: jedina razlika u ovom slučaju jeste da, ukoliko imamo paran broj podataka i ukoliko se oko „centra“ sortiranog niza nalaze dve različite vrednosti, medijana nije precizno definisana [4].

Računanje medijalne vrednosti može biti skup proces ukoliko imamo veliki broj podataka. Međutim, za numeričke vrednosti možemo lako aproksimirati medijalnu vrednost. Pretpostavimo da su podaci grupisani na osnovu intervala vrednosti atributa za koji računamo medijalnu vrednost, i da se za svaki interval zna frekvencija pojavljivanja vrednosti iz tog opsega. Na primer, radnici se po godinama staža mogu grupisati u juniore (do 2 godine iskustva), mediore (do 6 godina iskustva) i seniore (6+ godina iskustva), ili na bilo koji drugi pogodan način. Nazovimo interval koji sadrži medijanu medijalni interval. Možemo aproksimirati medijanu celog skupa podataka interpolacijom koristeći formulu:

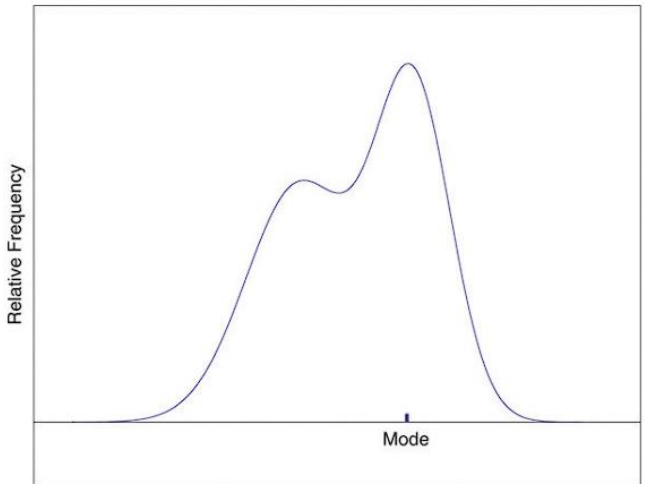
$$\tilde{x} = L_1 + \left(\frac{\frac{N}{2} - (\sum freq)_l}{freq_{\tilde{x}}} \right) w \tag{3.2}$$

gde L_1 predstavlja donju granicu medijalnog intervala, N je broj vrednosti u celom skupu podataka, $(\sum freq)_l$ je suma frekvencija pojavljivanja svih intervala sa vrednostima manjim od onih u medijalnom intervalu, $freq_{\tilde{x}}$ je frekvencija pojavljivanja vrednosti iz medijalnog intervala, i w je širina medijalnog intervala [4].

3.1.3 Mod (modus) podataka

U slučajevima kada nema mnogo smisla govoriti o podacima u razlomljenim vrednostima (1.43 automobila po glavi stanovnika i sl.), neretko se kao mera centralne tendencije koristi mod (modus) podataka. Mod skupa podataka je vrednost sa najvećom frekvencijom pojavljivanja u skupu. Ova vrednost može se odrediti i za kvalitativne i za kvantitativne atribute [4][5].

Može se dogoditi da nekoliko vrednosti ima istu frekvenciju pojavljivanja. Skupovi podataka sa jednom, dve ili tri modalne vrednosti nazivaju se redom unimodalni, bimodalni i trimodalni skupovi podataka. Generalno, skup podataka sa više od jedne modalne vrednosti naziva se multimodalni skup. U slučaju da se sve vrednosti u skupu podataka javljaju isti broj puta, mod ne postoji [4].



Slika 3.2 Vizuelni prikaz moda podataka

U ranije pomenutom primeru sa radnicima, mod podataka bi bio 1, zato što 4 od 5 radnika ima toliko godina staža.

Mod se koristi kao mera centarlne tendencije zbog toga što većina realnih skupova podataka ima više podataka koji su iz središta opsega vresnosti za taj podatak, a manje vrednosti pri krajevima opsega vrednosti. Vrednost sa najvećom frekvencijom pojavljivanja jako često se nalazi upravo u središtu ovog opsega [5].

3.1.4 Srednji opseg podataka

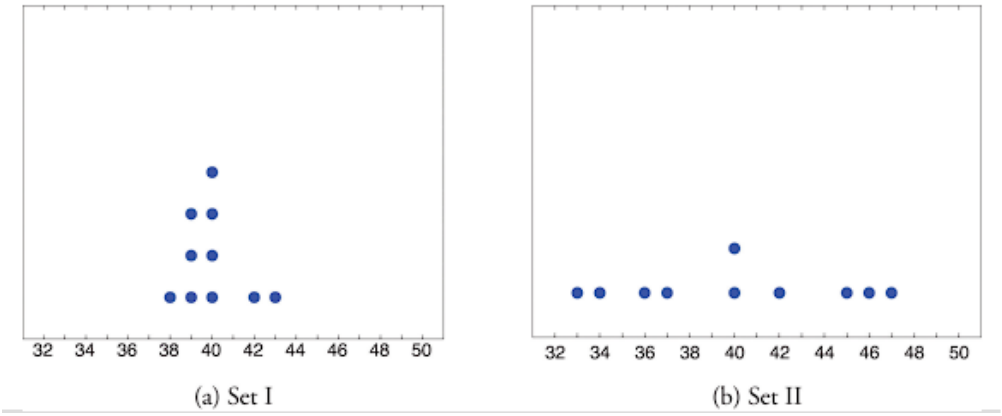
Srednji oseg podataka (engl. *midrange*) još jedna je mera centralne tendencije numeričkih podataka. Računa se kao prosek maksimalne i minimalne vrednosti atributa, odnosno [4]:

(3.3)

$$midrange = \frac{\max(x_1, x_2, \dots, x_N) + \min(x_1, x_2, \dots, x_N)}{2}$$

3.2 Mere disperzije podataka

Posmatrajmo dva seta podataka sa slike 3.3: oba skupa podataka imaju isti broj opservacija. I jednom i drugom su i srednja i medijalna vrednost i mod podataka 40. Ipak, pogled na grafički prikaz ovih podataka dovoljan je da se zaključi da se ova dva skupa podataka značajno razlikuju. Kod prvog seta podataka, opservacije su bliske centru i malo odstupaju od njega, dok podaci unutar drugog seta značajno variraju [5].



Slika 3.3 Dva seta podataka [5]

Kako bismo lakše razlikovali ova dva skupa podataka, kao što smo računali poziciju centra skupa, sada želimo da dodelimo skupovima podataka vrednosti takve da će one opisivati kako se podaci udaljavaju od centra, ili kako se klasterizuju oko njega. Za takvu vrstu opisa koriste se mere disperzije podataka – opseg, kvantili, varijansa, standardna devijacija, interkvartilni opseg i druge, o kojima će biti reči u ovom delu rada [4]. Ove mere imaju smisla za numeričke podatke.

3.2.1 Opseg

Neka je $x_1, x_2, \dots, x_N$ skup od N opservacija za neki numerički atribut X. Opseg skupa vrednosti predstavlja razliku između najveće i najmanje vrednosti u tom skupu, odnosno [4]:

$$range = \max(x_1, x_2, \dots, x_N) - \min(x_1, x_2, \dots, x_N) \quad (3.4)$$

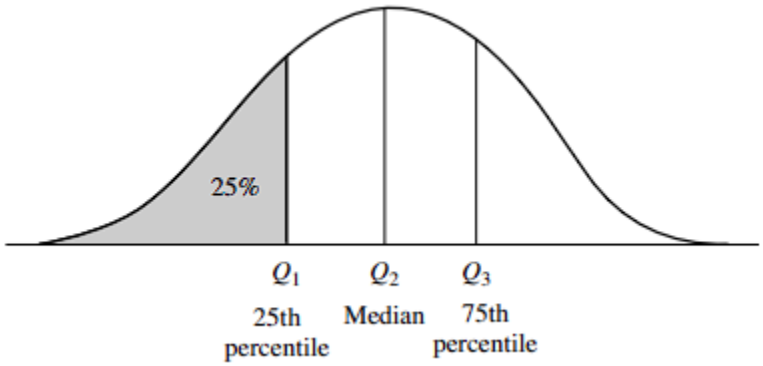
Opseg je mera disperzije podataka zato što ukazuje na veličinu intervala duž kog su vrednosti podataka raspoređene. Manji opseg znači manju varijaciju (manja disperzija) unutar podataka, dok veći opseg podataka indicira suprotno [5].

3.2.2 Kvantili i *outlier*-i

Pretpostavimo da su vrednosti atributa X sortirane u rastući poredak. Zatim, zamislimo da možemo izabrati određene tačke tako da podatke podelimo u određeni broj grupa iste veličine (sadrže isti broj opservacija). Ove tačke nazivaju se kvantili [4].

Kvantili su tačke na ravnomernom rastojanju unutar distribucije podataka, koje dele skup podataka na delove podjednakih veličina po broju podataka u njima. K-ti q-kvantil za neki skup podataka za zadatu raspodelu podataka predstavlja vrednost x takva da najviše k/q podataka ima vrednosti manju od x i najviše $(q - k)/q$ podataka ima vrednost veću od x, gde je k ceo broj takav da $0 < k < q$. Postoji $(q - 1)$ q-kvantila [4].

2-kvantil je tačka koja polovi distribuciju podataka, i ova tačka zapravo odgovara medijani. 4-kvantili su tri tačke koje dele raspodelu podataka na četiri jednaka dela: svaki deo predstavlja jednu četvrtinu ukupnog broja podataka. Ovi kvantili se češće nazivaju kvartili (engl. *quartiles*). 100-kvantili se češće nazivaju percentili (engl. *percentiles*); oni dele distribuciju podataka u 100 podskupova. Medijana, kvartili i percentili su najčešće korišćeni oblici kvantila [4]. Q_1 kvantil predsavlja medijalnu vrednost skupa podataka sa vrednostima manjim od Q_2, dok Q_3 predstavlja medijalnu vrednost skupa podataka koji sadrži vrednosti atributa veće od medijalne [5].



Slika 3.4 Raspodela podataka nekog atributa X, sa prikazanim kvartilima. Drugi kvartil odgovara medijani [4]

Kvartili daju indikaciju o centru distribucjue podataka, njenoj disperziji i obliku. Prvi kvartil, Q_1, je istovremeno i 25-ti percentil. On odseca najnižih 25% podataka, Treći, Q_3 kvartil, predstavlja i 75-ti percentil – odseca najnižih 75% podataka. Drugi kvartil je ujedno i 50%, i kao medijalna vrednost atributa predstavlja i centar distribucije atributa [4],

Rastojanje između prvog i trećeg kvartila je jednostavna mera disperzije koja daje opseg unutar kog se nalazi središnja polovina podataka. Ova distanca se naziva interkvartilni opseg (engl. *interquartile range* – IQR) i definiše se kao [4][5]:

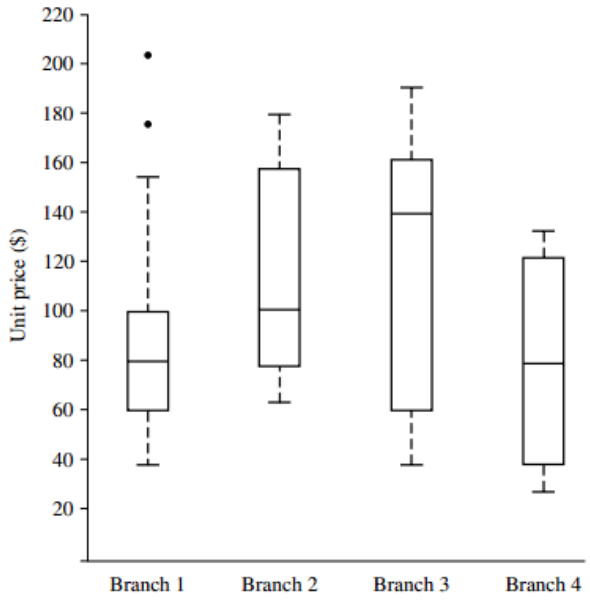
$$IQR = Q_3 - Q_1 \tag{3.5}$$

Pored vrednosti tri kvartila, dve ekstremne vrednosti, minimum vrednosti atributa X x_{min} i maksimalna vrednost atributa X x_{max} su takođe korisne za opisivanje celog skupa podataka. Zajedno, ovih pet brojeva naziva se petobrojni rezime (engl. *five-number summary*) i zapisuje se u obliku [5]:

$$\{ x_{min}, Q_1, Q_2, Q_3, x_{max} \} \tag{3.6}$$

Ovih pet brojeva koristi se za konstrukciju *boxplot* dijagrama. Ovi dijagrami predstavljaju popularan način vizuelizacije distribucije podataka. *Boxplot* inkorporira pet brojeva na sledeći način:

- Tipično, krajevi kutije su prvi i treći kvartil, tako da dužina kutije predstavlja interkvartilni opseg.
- Medijalna vrednost je označena linijom unutar kutije.
- Dve linije koje se nazivaju „brci“ (engl. *whiskers*) izvan kutije pružaju se do najmanje i najveće opservacije [4][5].



Slika 3.5 *Boxplot* cena proizvoda koji su se prodavali u 4 prodavnice tokom nekog vremenskog perioda [4]

Kada se radi sa velikim brojem opservacija nekog atributa, može biti korisno označiti vrednosti u podacima koje se potencijalno nalaze izvan nekog očekivanog opsega vrednosti (engl. *outliers*). Kako bi se ovo uradilo, neophodno je da se „brci“ protežu samo do onih vrednosti koje su u opsegu $1.5 \times IQR$. Sve anomalije koje su ostale van opsega označavaju se individualno tačkama, što se može videti na slici 3.5 [4][5].

Detektovanje i obrada *outlier*-a predstavlja jedan od najvažnijih koraka preprocesiranja podataka za algoritme mašinskog učenja, jer ovi podaci mogu negativno uticati na statističku analizu i proces treniranja ovih modela, što će rezultovati manjom tačnošću [6]. Nakon detekcije, postoji više načina na koje možemo obraditi *outlier*-e tako da manje utiču na konačne rezultate.

Najjednostavniji način obrade jeste potpuno izbacivanje podataka koji su van očekivanog opsega. Iako najjednostavnija, nije naročito poželjno koristiti ovu metodu, jer njeno korišćenje može dovesti do gubitka informacija. Češće korišćene metoda podrazumeva postavljanje svih vrednosti većih od 90-tog percentila na njegovu vrednost, odnosno svih vrednosti manjih od 10-og percentila na vrednost 10-og percentila; druga često korišćena metoda podrazumeva zamenu vrednosti *outlier*-a medijalnom ili srednom vrednošću tog atributa [6].

3.2.3 Varijansa i standardna devijacija

Varijansa i standardna devijacija su mere disperzije podataka. Pomoću njih možemo utvrditi koliko su podaci „raširenи“. Mala standardna devijacija znači da je većina podataka jako bliska aritmetičkoj sredini podataka, dok visoka vrednost standardne devijacije označava da se vrednosti podataka nalaze u nekom širem opsegu [4].

Varijansa N opservacija $x_1,\ x_2, \dots, x_N$ nekog atributa X računa se kao:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{3.7}$$

gde $\bar{x}$ predstavlja aritmetičku sredinu vrednosti svih opservacija atributa. Standardna devijacija σ računa se kao kvadratni koren varijanse [4][5].

Osnovna svojsta standardne devijacije kao mere disperzije podataka su:

- Standardna devijacija meri rasprostranjenost podataka oko njihove aritmetičke sredine i treba je razmatrati samo onda kada se aritmetička sredina koristi kao mera centralne tendencije [4].
- Standardna devijacija je jednaka nuli samo onda kada svi podaci imaju istu vrednost, inače je standardna devijacija veća od nule [4].

Računanje standardne devijacije i varijanse skalabilno je na velike skupove podataka [4].

3.3 Mere oblika distribucije podataka

Ranije opisane mere ne preciziraju u potpunosti izgled distribucije podataka. Mere centralne tendencije nam govore o koncentraciji opservacija oko centra distribucije, dok mere disperzije daju ideju o „raspršenosti“ opservacija. Možemo naići na dve distribucije podataka koje se po prirodi i kompoziciji značajno razlikuju, ali imaju istu centralnu tendenciju i disperziju podataka. Zbog svega ovoga, neophodno je uvesti i mere koje će opisivati samu raspodelu podataka. Te mere su:

- Iskošenost (engl. *skewness*)
- Kurtozis (engl. *kurtosis*)

Centralna tendencija, disperzija, iskošenost i kurtozis dovoljni su da u potpunosti odrede neku raspodelu podataka [7]. Ove mere mogu se primeniti na numeričke podatke.

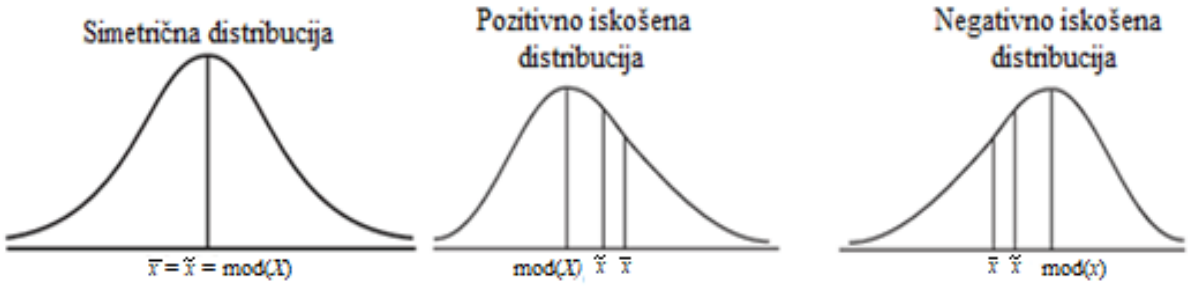
Ove mere se takođe mogu iskoristiti da opišu koliko je raspodela naših podataka bliska normalnoj raspodeli. Važno je da sličnost između ove dve raspodele bude što veća, zbog toga što brojni algoritmi mašinskog učenja, kao što su na primer linearna i drugi oblici regresije, podrazumevaju da je raspodela podaka s kojima rade približna normalnoj raspodeli [8]. Zbog toga, postoje tehnike koje transformišu podatke tako da njihovu raspodelu približe normalnoj.

3.3.1 Iskošenost

Bukvalno značenje ove mere je „nedostatak simetrije“. Posmatranjem iskošenosti možemo dobiti ideju o obliku raspodele podataka koju opisujemo. Ova mera omogućava da odredimo prirodu koncentracije opservacija ka višem ili nižem opsegu iz intervala svih vrednosti tog atributa [7].

Za distribuciju podataka kažemo da je iskošena ako:

- Kriva koja grafički predstavlja distribuciju nije simetrična i oblika zvona, već je rastegnuta više na jednu stranu. Drugim rečima, ima duži „rep“ (engl. *tail*) sa jedne nego sa druge strane. Ukoliko je rep duži sa desne strane, onda je distribucija pozitivno iskošena; u suprotnom, distribucija je negativno iskošena. Ovo je prikazano na slici 3.6, pri čemu $\bar{x}$ predstavlja aritmetičku sredinu, $mod(X)$ mod podataka, a $\tilde{x}$ medijalnu vrednost atributa X koji se posmatra.[7].
- Aritmetička sredina, medijalna vrednost i mod atributa nisu u istoj tački [7].
- Kvartili Q_1 i Q_2 nisu na jednakom rastojanju od medijane [7].



Slika 3.6 Primeri iskošenosti [7]

Neke od apsolutnih mera iskošenosti su:

$$Sk = \bar{x} - \tilde{x} \tag{3.8}$$

$$Sk = \bar{x} - mod(X) \tag{3.9}$$

$$Sk = (Q_3 - \tilde{x}) - (\tilde{x} - Q_1) = Q_3 + Q_1 - 2\tilde{x} \tag{3.10}$$

Međutim, ovo su apsolutne mere iskošenosti i iskošenost u ovakvom obliku nema mnogo praktične primene iz sledećih razloga:

- Pošto apsolutne mere iskošenosti uključuju i merne jedinice, ne mogu se iskorititi za upoređivanje dve raspodele podataka koje su u različitim jedinicama mere.
- Čak i ako raspodele imaju iste jedinice, apsolutne mere iskošenosti se ne preporučuju iz razloga što se može naići na različite raspodele podataka koje imaju gotovo istu meru iskošenosti, ali koje se značajno razlikuju po merama centralne tendencije i disperzije [7].

Zbog toga, za upoređivanje dve ili više distribucija potrebno je izračunati njihove relativne mere iskošenosti, takođe poznate i kao koeficijenti iskošenosti, koji su čisti brojevi nezavisni od mernih jedinica [7]. Često korišćeni koeficijenti iskošenosti su:

- Karl Person koeficijent (engl. *Karl Pearson coefficient*): Ovaj koeficijent računa se po sledećoj formuli:

$$Sk = \frac{\bar{x} - mod(X)}{\sigma} \tag{3.11}$$

Često, zbog loše definisanosti moda raspodele, umesto formule (3.11) koristi se formula:

$$Sk = \frac{3(\bar{x} - \tilde{x})}{\sigma} \tag{3.12}$$

- Boulijev koeficijent (engl. *Bowley's coefficient*): ovaj koeficijent je baziran na kvartilima i računa se kao:

$$Sk = \frac{Q_3 + Q_1 - 2\tilde{x}}{Q_3 - Q_1} \tag{3.13}$$

Ovaj koeficijent poznat je i pod nazivom kvartilni koeficijent iskošenosti i posebno je koristan u situacijama kada je mod podataka loše definisan, postoje *outlier*-i u podacima i slično. Loša strana ovog koeficijenta je što ignoriše 50% podataka (onih koji su u blizini minimuma i maksimuma opsega vrednosti) [7].

- Kelijev koeficijent (engl. *Kelly's measure*): predstavlja poboljšanje Boulijevog koeficijenta, i računa se kao:

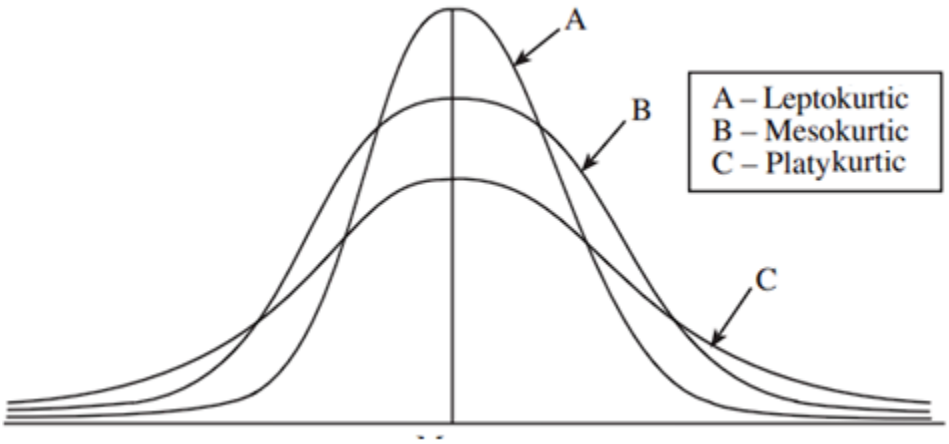
$$Sk = P_{90} + P_{10} - 2P_{50} \tag{3.14}$$

gde je P_x oznaka za x-ti percentil [7].

Bitno je napomenuti da navedeni koeficijenti nisu međusobno uporedivi, odnosno dva koeficijenta mogu se porediti samo ako su računata na isti način [7].

3.3.2 Kurtozis

Do sada smo govorili o tri mere: centralnoj tendenciji, disperziji i iskošenosti kako bismo ispitali karakteristike distribucije podataka. Ipak, čak i ako znamo sve tri mere ne možemo u potpunosti okarakterisati raspodelu. Primer ovoga može se videti na slici 3.7 [7].



Slika 3.7 Različite raspodele sa istom centralnom tendencijom, disperzijom i iskošenošću [7]

Sve tri krive su simetrične oko svoje srednje vrednosti i imaju isti oseg vrednosti. Zbog toga, da bismo u potpunosti opisali ove krive, moramo ispitati njihov kurtozis, odnosno konveksnost krive. Dok nam iskošenost pomaže da odredimo sa koje strane raspodela ima „rep", kurtozis nam daje informaciju o obliku i prirodi „grbe" (središnjeg dela) raspodele podataka. Drugim rečima, kurtozis razmatra koliko je kriva izbočena ili spljoštena [7].

Kurtozis se računa po sledećoj formuli [7]:

$$kurtosis = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^4}{N\sigma^4} \tag{3.15}$$

gde je N broj opservacija koje ima atribut koji posmatramo, x_i vrednost i-te opservacije, $\overline{x}$ je aritmetička sredina vrednosti atributa i σ je standardna devijacija podataka.

Pošto normalna kriva ima kurtozis 3, uvedena je mera koja označava razliku između kurtozisa posmatrane krive i kurtozisa normalne raspodele (engl. *excess kurtosis*) [7]:

$$excess\ kurtosis = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^4}{N\sigma^4} - 3 \tag{3.16}$$

Kriva označena sa B na slici 3.7 nije ni previše izbočena, niti previše spljoštena. Ovakve krive se nazivaju normalne krive i oblik njihove „grbe“ smatra se standardnim.. Za ovakve krive se kaže da imaju normalan kurtozis (koji je jednak 3) i nazivaju se mezokurtične krive (engl. *mesokurtic*). Krive tipa A, koje su više izbočene, nazivaju se leptokurtične (engl. *leptokutic*) i one imaju kurtozis veći od 3, odnosno pozitivan *excess kurtosis*. Sa druge strane, krive tipa C, koje su spljoštene u odnosu na normalnu krivu zovu se platikurtične (engl. *platykurtic*) i one imaju kurtozis manji od 3, odnosno negativan *excess kurtosis* [7].

Kao što smo ranije pomenuli, postoje metode koje treba da utiču na raspodelu podataka tako da je što više približe normalnoj raspodeli. Naravno, retko se događa da je rezultat ovih transformacija baš normalna raspodela, ali jeste raspodela približnija normalnoj od originalne. U ove svrhe često se koristi logaritamska transformacija. Ona podatke menja vrednošću njihovog logaritma za neku određenu osnovu. Naravno, pošto logaritam nije definisan za vrednost 0, neophodno je ovu vrednost zameniti nekom malom vrednošću koja je približna nuli. Pored ove, postoje i brojne druge metode, pri čemu od slučaja zavisi koja će se bolje pokazati [9].

4. Mere sličnosti podataka

Često imamo potrebu za upoređivanjem nekih objekata od interesa. Na primer, neka *online* prodavnica bi možda želela da grupiše svoje kupce na osnovu njihovih karakteristika (atributa) zarad targetiranog marketinga [4].

U ovom odeljku, bavićemo se merama sličnosti i različitosti podataka, koje se jednim imenom nazivaju mere bliskost (engl. *measures of proximity*). Sličnost i različitost su povezani. Mera sličnosti dva objekta *i* i *j* obično vraća vrednost 0 ako su objekti različiti. Ukoliko su objekti identični, vrednost sličnosti je 1. Mera različitosti radi suprotno: 0 označava iste objekte, a 1 znači da su potpuno različiti [4]. Ove mere su značajne jer je često neophodno porediti neke podatke, a od ishoda tog poređenja može zavisiti da li su podaci zadovoljavajućeg kvaliteta ili ne.

4.1 Strukture podataka za merenje sličnosti i različitosti podataka

Do sada smo se bavili merama centralne tendencije, disperzijom i raširenošću opservacija nekog atributa X. Naši objekti su u tom slučaju bili jednodimenzionalni, odnosno opisivao ih je samo jedan atribut. U ovom poglavlju, govorićemo o objektima opisanim pomoću više atributa. Zbog toga uvodimo sledeću notaciju: pretpostavimo da imamo n objekata, i da je svaki od objekata opisan sa p atributa. Te objekte predstavićemo kao $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, gde je x_{ij} vrednost j-tog atributa i-tog objekta u skupu objekata. Ovakvi objekti se drugačije nazivaju uzorci podataka (engl. *data samples*) ili vektori osobina (engl. *feature vectors*) [4].

Većina algoritama za klasterizaciju ili za pronalaženje najbližeg suseda (engl. *nearest-neighbour algorihtm*) rade sa nekom od sledećih struktura podataka:

- Matrica podataka (engl. *data matrix*): Ova struktura smešta n objekata u formi relacione tabele, odnosno $n \times p$ matrice, gde je p broj atributa objekata. Svaki red matrice predstavlja jedan objekat. Ovakva jedna matrica prikazana je na slici 4.1, pri čemu je indeks f iskorišćen za indeksiranje kroz atribute objekata [4].

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}.$$

Slika 4.1 Primer matrice podataka [4]

- Matrica različitosti (engl. *dissimilarity matrix*): ova struktura sadrži kolekciju „razdaljina“ između objekata, za svaki par n objekata. Najčešće je predstavljena $n \times n$ matricom poput one na slici 4.2, pri čemu $d(i, j)$ predstavlja meru različitosti između objekata i i j. Obično je $d(i, j)$ nenegativan broj blizak nuli kada su objekti i i j veoma slični, odnosno „blizu“ jedan drugom, i koji se povećava što su ova dva objekta više različita. Treba takođe primetiti da važi $d(i, i) = 0$, odnosno $d(i, j) = d(j, i)$ – ovo znači da je matrica simetrična [4]. Mere različitosti, odnosno sličnosti, biće obrađenje u nastavku poglavlja.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Slika 4.2 Primer matrice različitosti [4]

Mere sličnosti često se mogu izraziti kao funkcija mere različitosti. Veza između ove dve mere obično se predstavlja u obliku:

$$sim(i,j) = 1 - d(i,j) \tag{4.1}$$

gde $sim(i,j)$ predstavlja sličnost između objekata i i j [4].

4.2 Mere sličnosti i različitosti nominalnih atributa

Kao što je razmatrano u ranijim poglavljima, nominalni atributi su oni koji mogu imati određeni broj vrednosti: na primer, atribut „boja“ može imati vrednosti „crvena“, „zelena“, „plava“.

Neka atribut ima M mogućih vrednosti. Ove vrednosti mogu biti predstavljene slovima tj. rečima, simbolima ili skupom prirodnih brojeva, kao što su 1, ..., M. Treba obratiti pažnju na činjenicu da su u ovom slučaju brojevi iskorišćeni samo za reprezentaciju podataka, i između njih ne postoji nikakav poredak. Mera različitosti između dva objekta i i j čiji su svi atributi nominalni predstavlja se u obliku:

$$d(i,j) = \frac{p - m}{p} \tag{4.2}$$

gde je m broj poklapanja (broj atributa koji i u podatku i i u podatku j imaju istu vrednost), a p predstavlja ukupan broj atributa koji opisuju objekte. Takođe, mogu se uključiti težine kako bi se dao veći značaj atributima koji imaju više mogućih vrednosti [4].

Nasuprot ovome, mera sličnosti između podataka može se izračunati kao:

$$sim(i,j) = 1 - d(i,j) = \frac{m}{p} \tag{4.3}$$

Sličnost između objekata predstavljenih pomoću nominalnih atributa može biti izračunata korišćenjem alternativnog načina enkodiranja. Nominalni atributi mogu se enkodirati korišćenjem asimetričnih binarnih atributa na način da se za svako od M stanja atributa kreira novi binarni atribut. Za dati objekat, samo onaj binarni atribut koji predstavlja stanje odgovarajućeg atributa biće postavljen na 1, dok će svi ostali binarni atributi koji reprezentuju vrednosti tog nominalnog atributa imati vrednost 0 [4]. Način za računanje sličnosti između ovako predstavljenih nominalnih podataka biće predstavljen u nastavku.

4.3 Mere sličnosti i različitosti binarnih atributa

Kao što je ranije razmatrano, binarni atributi mogu imati samo dve vrednosti: 0 i 1, gde 0 predstavlja odsustvo tog atributa, a 1 njegovo prisustvo u objektu. Tretiranje binarnih podataka kao numeričkih može dovesti do pogrešnih zaključaka, te su zbog toga neophodne posebne metode za računanje različitosti između ovih podataka [4].

Jedan pristup za računanje različitosti između dva objekta predstavljena binarnim atributima jeste izračunavanje matrice slučaja. Ukoliko smatramo da svi atributi imaju iste težine, dobićemo 2×2 matricu slučaja (engl. *contigency matrix*) prikazanu na slici 4.3, gde q predstavlja broj atributa jednakih 1 u oba objekta i i j. r je broj atributa jednakih 1 u objektu i, ali koji su 0 u objektu j. s je broj atributa koji imaju vrednost 0 u objektu i, ali vrednost 1 u objektu j, i konačno, broj t je broj atributa koji imaju vrednost 0 u oba objekta. Ukupan broj atributa je p, gde je $p = q + r + s + t$ [4].

		Objekat j		
		1	0	sum
Objekat i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

Slika 4.3 Primer matrice slučaja [4]

Podsetimo se da su kod simetričnih binarnih atributa oba stanja podjednako važna, dok je kod asimetričnih jedno stanje važnije od drugog. Različitost između objekata bazirana na simetričnim binarnim atributima naziva se simetrična binarna različitost. Ako su objekti i i j opisani simetričnim binarnim atributima, onda se njihova različitost može predstaviti kao [4]:

$$d(i,j) = \frac{r+s}{q+r+s+t} \tag{4.4}$$

Mera različitosti bazirana na dva podatka predstavljena asimetričnim binarnim vrednostima naziva se i asimetrična binarna različitost. Kod ove različitosti, broj t, odnosno broj negativnih poklapanja, smatra se nebitnim, te se ignoriše u izračunavanju [4]:

$$d(i,j) = \frac{r+s}{q+r+s} \tag{4.5}$$

Komplementarno, možemo izračunati meru sličnosti pomoću različitosti između dva podatka. Na primer, asimetrična binarna sličnost između dva objekta može se izračunati kao:

$$sim(i,j) = \frac{q}{q+r+s} = 1 - d(i,j) \tag{4.6}$$

Koeficijent sličnosti iz jednačine (4.6) naziva se i Žakardov koeficijent (engl. *Jaccard coefficient*) [4].

Kada se u podacima jave i simetrični i asimetrični binarni atributi, mogu se primeniti tehnike za mešane atribute koje će biti obrađene kasnije [4].

4.4 Mere sličnosti i različitosti numeričkih podataka: Minkovski distanca

Kada su u pitanju numerički podaci, cilj je izračunati distancu, odnosno različitost između njih. U nekim slučajevima, podaci se normalizuju pre preračunavanja distanci između njih. Ovo podrazumeva transformaciju podataka tako da se sve vrednosti nalaze unutar nekog manjeg opsega, kao što su opsezi [-1, 1] ili [0, 1]. Normalizacija podataka za cilj ima davanje istih težina svim podacima. Ovaj pristup može biti, ali nije nužno koristan za svaku primenu podataka [4].

Najpopularnija mera udaljenosti između numeričkih podataka jeste euklidska distanca i ona predstavlja dužinu prave linije koja povezuje ova dva objekta u nekom p-dimenzionalnom prostoru. Neka su $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ i $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ dva objekta predstavljena sa po p numeričkih atributa. Euklidska distanca između ova dva objekta računa se kao [4]:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \tag{4.7}$$

Druga često korišćena distanca jeste Menhetn distanca, poznata kao i blok (engl. *city block*) distanca, nazvana tako jer podseća na udaljenost između dva bloka u nekom gradu (npr. dva bloka dole i tri desno – ukupno 5 blokova). Ova distanca definisana je kao [4]:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}| \tag{4.8}$$

Minkovski distanca predstavlja generalizaciju euklidske i Menhetn distance. Definisana je kao:

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h} \tag{4.9}$$

gde je h realan broj takav da $h \geq 1$. Ovakva distanca se u literaturi često naziva i L_p distanca, pri čemu je zbog konzistentnosti u nomenklaturi u formuli (4.9) p zamenjeno sa h. Minkovski distanca jednaka je Menhetn distanci za $h = 1$ (L_1 norma), i euklidskoj distanci za $h = 2$ (L_2 norma) [4].

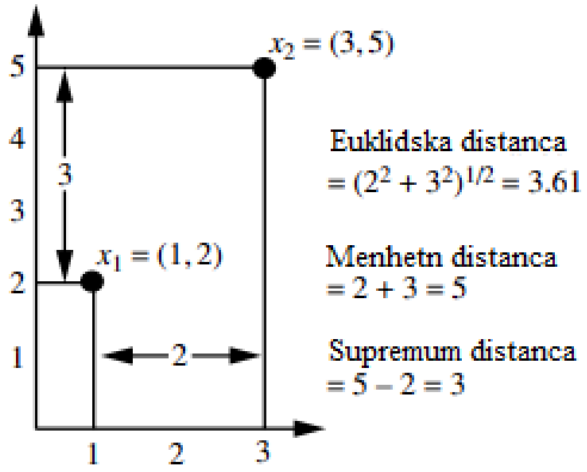
Supremum distanca, takođe poznata i kao L_{max}, L_∞ ili Čebiševljeva norma, je generalizacija Minkovski distance za $h \to \infty$. Ova norma definisana je kao [4]:

$$d(i,j) = \lim_{h\to\infty}\left(\sum_{f=1}^{p}|x_{if} - x_{jf}|^{h}\right)^{\frac{1}{h}} = \max_{f\in[1,p]}|x_{if} - x_{jf}| \qquad (4.10)$$

Sve norme (distance) moraju da ispunjavaju sledeće uslove:

- $d(i,j) \geq 0$: Distanca je nenegativan broj.
- $d(i,i) = 0$: Distanca između objekta i njega samog je 0.
- $d(i,j) = d(j,i)$: Distanca je simetrična funkcija.
- $d(i,j) \leq d(i,k) + d(k,j)$: Za distance važi nejednakost trougla [4].

Primer računanja distanci prikazan je na slici 4.4.



Slika 4.4 Primer računanja različitih distanci [4]

4.5 Mere sličnosti i različitosti ordinalnih atributa

Vrednosti ordinalnih atributa imaju određeni redosled, ali razlika između dve uzastopne vrednosti nije poznata. Ordinalni atributi mogu se dobiti i diskretizacijom numeričkih atributa podelom raspona vrednosti tih atributa na konačan broj intervala [4].

Neka je broj M broj stanja koje ordinalni atribut može da ima. Ova uređena stanja definišu rangiranje 1, ..., M_f [4].

Ordinalni atributi tretiraju se na sličan način kao i numerički atributi kada se računa različitost između objekata. Neka je f atribut iz skupa ordinalnih atributa koji opisuju n objekata. Računanje različitosti između objekata u odnosu na atribut f uključuje sledeće korake:

1. Neka je vrednost atributa f za i-ti objekat x_{if}, i neka f ima M_f uređenih stanja koja se mogu predstaviti pomoću svog ranga kao 1, ..., M_f. Zameniti svaku vrednost x_{if} njenim odgovarajućim rangom $r_{if} \in \{1, \dots, M_f\}$.
2. Pošto svaki od ordinalnih atributa može imati različiti broj stanja, često je neophodno mapirati raspon rangova svakog od atributa na interval [0.0, 1.0] kako bi svaki od atributa imao istu težinu. Ovu normalizaciju podataka vršimo zamenom ranga r_{if} atributa f i-tog objekta vrednošću:
$$z_{if} = \frac{r_{if} - 1}{M_f - 1} \tag{4.11}$$
3. Različitost između podataka se sada može izračunati pomoću bilo koje od metrika opisanih prilikom analize različitosti numeričkih atributa, koristeći vrednost z_{if} za predstavljanje vrednosti atributa f i-tog objekta [4].

Sličnost između objekata opisanih ordinalnim atributima može se dobiti pomoću njihove različitosti kao [4]:

$$sim(i, j) = 1 - d(i, j) \tag{4.12}$$

4.6 Mere sličnosti i različitosti atributa mešovitog tipa

Do sada smo razmatrali kako uporediti dva objekta koja su opisana samo atributima istog tipa, gde tip može biti nominalni, simetrični binarni, asimetrični binarni, numerički ili ordinalni. Ipak, u mnogim realnim skupovima podataka, objekti su opisani kombinacijom ovih tipova podataka. U opštem slučaju, skup podataka može sadržati sve ove tipove atributa [4].

Jedan pristup za upoređivanje objekata opisanih atributima mešovitog tipa jeste grupisanje atributa istog tipa, i primena neke od tehnika za dobijanje podataka (engl. *data mining techniques*), kao što je na primer klasterizacija, na svaku od dobijenih grupa atributa. Ovo je moguće izvesti ukoliko će svaka od ovih analiza proizvesti kompatibilne rezultate, što je malo verovatno sa realnim skupovima podataka [4].

Češće korišćeni pristup je procesiranje svih atributa zajedno. Jedna ovakva tehnika kombinuje različite atribute u jednu matricu različitosti, skaliranjem svih značajnih podataka na raspon [0.0, 1.0] [4].

Pretpostavimo da su objekti opisani pomoću p atributa mešovitog tipa. Različitost između dva ovakva objekta definiše se kao:

$$d(i, j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}} \tag{4.13}$$

gde je indikator $\delta_{ij}^{(f)}$= 0 ako je vrednost x_{if} ili x_{jf} nedostajuća, ili ako je $x_{if} = x_{jf} = 0$ i atribut f je binaran i asimetričan; u ostalim slučajevima, vrednost indikatora $\delta_{ij}^{(f)}$= 1. Doprinos atributa f različitosti između objekata i i j $d_{ij}^{(f)}$ računa se na osnovu tipa atributa:

- Ako je f numerički atribut:

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}} \tag{4.14}$$

  gde je h indeks pomoću kog se indeksiraju svi objekti koji poseduju vrednost za atribut f.
- Ako je f nominalni ili binarni atribut:

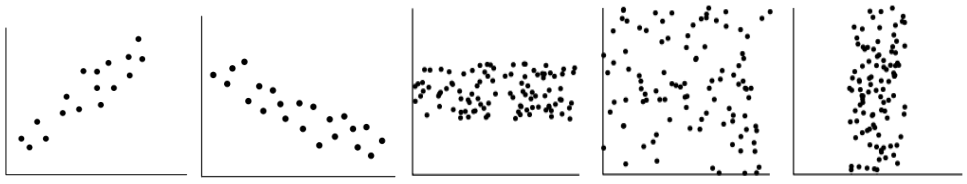
$$d_{ij}^{(f)} = \begin{cases} 0, & x_{if} = x_{jf} \\ 1, & x_{if} \neq x_{jf} \end{cases} \tag{4.15}$$

- Ako je f ordinalni atribut: preračunati rangove r_{if} i njihove vrednosti z_{if}. Tretirati z_{if} vrednosti kao numeričke [4].

Ovi koraci su identični onome što smo već videli za svaki od ovih individualnih tipova podataka. Jedina razlika je za numeričke atribute, gde normalizujemo razliku tako da se njena vrednost mapira na opseg [0.0, 1.0]. S toga, zaključujemo da se sličnost između objekata u skupu podataka može izračunati i kada atributi koji ih opisuju nisu svi istog tipa [4].

4.7 Korelacija i kovarijansa

Dva atributa *X* i *Y* su korelisani ukoliko su međusobno na neki način zavisni. Ukoliko postoji, korelacija između atributa može biti pozitivna i negativna. Vizuelno, korelacija se može uočiti prikazom podataka pomoću *scatter plot*-ova, pri čemu su osama ovih grafika predstavljene vrednosti atributa čija se korelacija ispituje. Primeri ovakvih vizuelizacija dati su na slici 4.5 [4].



Slika 4.5 *Scatter plot*-ovi za ispitivanje korelacije: a) pozitivna korelacija; b) negativna korelacija; c) d) e) nema korelacije [4]

Dva atributa su pozitivno korelisana ukoliko rastom vrednosti jednog atributa rastu i vrednosti drugog. Suprotno, atributi su negativno korelisani ukoliko je rastom vrednosti jednog uslovljeno opadanje vrednosti drugog atributa. Korelacija se može predstaviti pravom: ukoliko je njen koeficijent pravca pozitivan, i korelacija je pozitivna. Ukoliko je koeficijent pravca negativan, i korelacija je negativna, a ukoliko je prava paralelna sa nekom od osa, korelacija ne postoji [4].

Da bismo definisali korelaciju između atributa, neophodno je definisati i kovarijansu. Ukoliko posmatramo dva numerička atributa X i Y, i N objekata opisanih ovim atributima, pri čemu je $X = (x_1, x_2, \dots, x_N)$ i je $Y = (y_1, y_2, \dots, y_N)$, njihova kovarijansa može se izračunati kao:

$$Cov(X, Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{N} \tag{4.16}$$

U ovom izrazu, kada su atributi pozitivno korelisani, znači da oba rastu istovremeno, što dalje znači da, ukoliko je vrednost atributa X veća od njegove srednje vrednosti $\bar{X}$, verovatno je da je i vrednost Y veća od njegove srednje vrednosti $\bar{Y}$ – odavde sledi da će izraz (4.16) biti pozitivan. Na sličan način možemo zaključiti da će, u slučaju negativne korelacije između podataka, kovarijansa biti negativna, a da će biti jednak ili približno jednak nuli ukoliko atributi nisu korelisani [4].

Nakon što smo definisali kovarijansku, možemo definisati i koeficijent korelacije ili Pirsonov indeks (engl. *Pearson correlation coefficient*):

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{N\sigma_X \sigma_Y} \tag{4.17}$$

gde je σ_{XY} kovarijansa, a σ_X i σ_Y su standardne devijacije atributa X i Y. Koeficijent korelacije može imati vrednosti u opsegu [-1, 1], pri čemu korelacija približna nuli znači da su podaci slabo ili nisu uopšte korelisani. Vrednost veća od 0 ukazuje na pozitivnu korelisanost, pri čemu vrednost 1 ukazuje na potpunu korelaciju. Vrednost manja od 0 znači negativnu korelaciju, a vrednost -1 znači potpunu negativnu korelaciju [4].

Za računanje korelacije između dva nominalna atributa koristi se χ^2 mera. Neka je c broj mogućih vrednosti atributa X, a r broj mogućih vrednosti atributa Y. Najpre kreiramo matricu slučaja (engl. *contigency matrix*) sa c kolona i r vrsta, pri čemu su vrednosti ćelija jednake broju pojavljivanja parova (x_i, y_j). χ^2 mera definiše se kao:

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{4.18}$$

pri čemu je o_{ij} frekvencija pojavljivanja parova (x_i, y_j), a e_{ij} je očekivana frekvencija, koja se računa kao proizvod broja pojavljivanja vrednosti x_i i broja pojavljivanja vrednosti y_j podeljen ukupnim brojem opservacija, odnosno sa N [4].

χ^2 mera je zapravo statistički test hipoteze da atributi X i Y nisu korelisani. Ima $(r-1) * (c-1)$ stepen slobode, a rezultat se upoređuje sa referentnom χ^2 vrednošću za odgovarajući broj

stepena slobode. Ukoliko se hipoteza obori, odnosno ukoliko je dobijena vrednost veća od referentne, atributi su korelisani [4].

Atributi korelisani s drugim atributima se često odbacuju (odnosno samo jedan od njih se koristi u analizi) zbog toga što se slično ponašaju i imaće sličan uticaj na konačne rezultate, pa je korišćenje više ovakvih atributa redundantno. Uklanjanje korelisanih atributa štedi vreme i prostor pri izvršenju kompleksnijih algoritama. Takođe, rad sa manjim brojem atributa olakšava analizu i razumevanje rezultata [10].

5. Zaključak

U ovom radu bavili smo se načinom opisivanja objekata pomoću atributa, tipovima atributa, načinom njihovog opisivanja, i konačno načinom na koji se oni upoređuju. Sve ovo bilo je neophodno kako bismo imali odgovarajuće alate i tehnike za opisivanje i ispitaivanje kvaliteta podataka, i samim tim za ekstrakciju informacija i sticanje znanja iz njih.

Sve prethodno opisane tehnike služe da olakšaju rad sa podacima. Kada znamo sa kakvim tačno podacima radimo, kako su ti podaci opisani u kakvom su oni odnosu, možemo te podatke obraditi na odgovarajući način i maksimalno ih iskoristiti za svrhe za koje su nam potrebni.

Literatura

[1] L. Cai, Y. Yhou, „The Challenges of Data Quality and Data Quality Assessment in the Big Data Era“, *Data Science Journal*, vol. 14: 2, pp. 1-10, 2015

[2] V. N. Gudivada, A. Apon, J. Ding, „Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations”, *International Journal on Advances in Software*, vol. 10, no. 1 & 2, 2017

[3] P. Tan, M. Steinbach, V. Kumar, „Introduction to Data Mining“, *Pearson*, 2005

[4] J. Han, M. Kamber, J. Pei, „Data Mining: Concepts and Techniques – Third Edition“, *Morgan Kaufmann*, 2012

[5] D. S. Shafer, Z. Zhang, „Beggining Statistics“, *Creative Commons*, 2012

[6] P. Bansal, "Detecting and Treating Outliers: Treating the Odd One Out," *Analytics Vidhya*, May 19, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/ [Accessed: May 6, 2024]

[7] S. C. Gupta, I. Gupta, „Business Statistics“, *Himalaya Publishing House*, 2008

[8] P. Bansal, "Normal Distribution: An Ultimate Guide," *Analytics Vidhya*, May 11, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/normal-distribution-an-ultimate-guide/ [Accessed: May 6, 2024]

[9] P. Bansal, "How to Transform Features into Normal (Gaussian) Distribution," *Analytics Vidhya*, May 25, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/how-to-transform-features-into-normal-gaussian-distribution/ [Accessed: May 6, 2024]

[10] "Remove Correlated Attributes," in *RapidMiner Documentation. RapidMiner*, [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/blending/attributes/selection/remove_correlated_attributes.html#:~:text=Correlated%20attributes%20are%20usually%20removed,of%20calculation%20of%20complex%20algorithms [Accessed: May 6, 2024]