UNIVERZITET U NIŠU  
ELEKTRONSKI FAKULTET

TRANSFORMACIJA PODATAKA

Prikupljanje i predobrada podataka za mašinsko učenje

Student: Katarina Stanojković 1773

Mentor: prof. dr Aleksandar Stanimirović

Sadržaj

[1. Uvod 4](#_Toc180768403)

[1.1. Šta je transformacija podataka? 4](#_Toc180768404)

[1.2. Značaj transformacije podataka u mašinskom učenju 4](#_Toc180768405)

[1.3. Ciljevi i struktura rada 4](#_Toc180768406)

[2. Identifikacija problema u analizi podataka 5](#_Toc180768407)

[2.1. Različiti tipovi podataka i izazovi u radu sa njima 5](#_Toc180768408)

[2.2. Uticaj nepripremljenih podataka na performanse modela 5](#_Toc180768409)

[2.3. Klasifikacija problema u pripremi podataka 5](#_Toc180768410)

[2.4. Pregled mogućih rešenja i njihova primena 5](#_Toc180768411)

[5. Skaliranje podataka 6](#_Toc180768412)

[5.1. Standardizacija (Z-score) 6](#_Toc180768413)

[5.2. Min-Max skaliranje 7](#_Toc180768414)

[5.3. Max-Abs skaliranje 7](#_Toc180768415)

[5.4. Robust skaliranje 7](#_Toc180768416)

[6. Transformacije koje menjaju raspodelu vrednosti 9](#_Toc180768417)

[6.1. Logaritamska transformacija 9](#_Toc180768418)

[6.2. Box Cox transformacija 10](#_Toc180768419)

[6.3. Yeo-Johnson transformacija 11](#_Toc180768420)

[6.4. Kvantilna transformacija 12](#_Toc180768421)

[7. Enkodiranje kategoričkih podataka 13](#_Toc180768422)

[7.1. One Hot enkodiranje 13](#_Toc180768423)

[7.2. Dummy enkodiranje 14](#_Toc180768424)

[7.3. Label enkodiranje 14](#_Toc180768425)

[7.4. Ordinalno enkodiranje 14](#_Toc180768426)

[7.5. Binarno enkodiranje 14](#_Toc180768427)

[7.6. Count and Frequency enkodiranje 15](#_Toc180768428)

[7.7. Target enkodiranje 15](#_Toc180768429)

[7.8. Effect enkodiranje 15](#_Toc180768430)

[7.9. Feature Hashing enkodiranje 16](#_Toc180768431)

[8. Diskretizacija (binovanje) podataka 17](#_Toc180768432)

[8.1. Podela u intervale jednakih širina 17](#_Toc180768433)

[8.2. Podela na intervale sa jednakom frekvencijom 17](#_Toc180768434)

[8.3. Diskretizacija korišćenjem klasterizacije 17](#_Toc180768435)

[8.4. Diskretizacija korišćenjem stabla odlučivanja 18](#_Toc180768436)

[8.5. Chi Merge 18](#_Toc180768437)

[9. Rad sa outlier-ima 19](#_Toc180768438)

[9.1. Tehnike vizuelizacije za detekciju outlier-a 19](#_Toc180768439)

[9.2. Z-Score 21](#_Toc180768440)

[9.3. Interquartile Range (IQR) 21](#_Toc180768441)

[9.4. Percentile 22](#_Toc180768442)

[9.5. Dbscan 22](#_Toc180768443)

[9.6. Isolation Forest 23](#_Toc180768444)

[9.7. Tehnike rukovanja outlier-ima 23](#_Toc180768445)

[10. Konstrukcija atributa (Feature creation) 25](#_Toc180768446)

[10.1. Tipovi kreiranja karakteristika 25](#_Toc180768447)

[10.2. Interakcione karakteristike 25](#_Toc180768448)

[10.3. Polinomske karakteristike 25](#_Toc180768449)

[10.4. Vremenske karakteristike 26](#_Toc180768450)

[11. Zaključak 27](#_Toc180768451)

[12. Reference 28](#_Toc180768452)

# Uvod

U savremenom svetu, mašinsko učenje predstavlja ključni alat za analizu i interpretaciju velikih skupova podataka. Kvalitet ulaznih podataka ostaje presudan za uspeh procesa učenja, što naglašava značaj njihove transformacije. Ovaj rad razmatra različite tehnike transformacije podataka, njihov uticaj na performanse modela, kao i prednosti i nedostatke u različitim kontekstima. Cilj je pružiti sveobuhvatan pregled ovih tehnika kroz praktične primere i analize, te ponuditi preporuke za njihovu primenu u zavisnosti od specifičnih potreba i izazova.

## 1.1. Šta je transformacija podataka?

Transformacija podataka podrazumeva pretvaranje sirovih podataka u format prikladniji za analizu i modeliranje. U mašinskom učenju, ona omogućava efikasnije učenje iz podataka smanjenjem složenosti, poboljšanjem interpretabilnosti i eliminisanjem šuma. Neke od osnovnih tehnika transformacije uključuju skaliranje, normalizaciju, enkodiranje kategoričkih atributa, detekciju i obradu outlier-a.

## 1.2. Značaj transformacije podataka u mašinskom učenju

Neobrađeni podaci često sadrže nepravilnosti koje mogu otežati proces učenja i smanjiti performanse modela. Transformacija podataka rešava ove probleme i optimizuje ulazne podatke. Na primer, skaliranje omogućava modelima da bolje funkcionišu sa podacima različitih skala, dok enkodiranje kategoričkih podataka omogućava modelima poput regresije i neuronskih mreža da efikasno rade sa diskretnim vrednostima.

## 1.3. Ciljevi i struktura rada

Cilj rada je pružiti sveobuhvatan pregled tehnika transformacije podataka koje se koriste u mašinskom učenju. Rad je podeljen na nekoliko celina:

* Identifikacija problema u pripremi podataka
* Tehnike skaliranja i normalizacije
* Transformacije raspodele podataka
* Enkodiranje kategoričkih atributa
* Rad sa outlier-ima
* Konstrukcija novih atributa (Feature Creation)

# Identifikacija problema u analizi podataka

## Različiti tipovi podataka i izazovi u radu sa njima

Podaci u mašinskom učenju mogu biti numerički, kategorički, vremenski, tekstualni, slikovni, itd. Svaki tip donosi specifične izazove. Numerički podaci mogu imati različite skale i raspodele, dok kategorički podaci zahtevaju enkodiranje kako bi bili korisni za modele. Vremenski i tekstualni podaci zahtevaju dodatnu obradu za ekstrakciju relevantnih karakteristika, dok slikovni podaci često zahtevaju tehnike obrade slike za smanjenje dimenzionalnosti i izdvajanje ključnih atributa.

## Uticaj nepripremljenih podataka na performanse modela

Nepripremljeni podaci negativno utiču na performanse modela. Algoritmi osetljivi na skalu mogu favorizovati veće vrednosti, a outlieri mogu značajno izobličiti rezultate. Kategorički podaci koji nisu pravilno enkodirani mogu stvoriti lažne korelacije između atributa, što može dovesti do smanjenja tačnosti modela.

## Klasifikacija problema u pripremi podataka

Problemi u pripremi podataka mogu se klasifikovati u nekoliko glavnih kategorija:

1. **Problemi sa skalom i normalizacijom:** Uključuju podatke različitih jedinica i raspona, što može otežati učenje modela.
2. **Problemi sa raspodelom:** Nelinearne raspodele ili podaci koji nisu normalno raspoređeni mogu zahtevati transformacije kako bi se poboljšala performansa modela.
3. **Problemi sa kategoričkim podacima:** Uključuju podatke koji nisu numerički, već nominalni ili ordinalni, što zahteva specifične tehnike enkodiranja.
4. **Problemi sa outlier-ima (ekstremnim vrednostima):** Outlieri mogu značajno uticati na performanse modela, posebno kod algoritama koji su osetljivi na anomalije.

## Pregled mogućih rešenja i njihova primena

Rešenja za identifikovane probleme uključuju:

* **Skaliranje:** Tehnike poput Min-Max skaliranja, Z-Score standardizacije i robustnog skaliranja pomažu u normalizaciji podataka različitih skala.
* **Transformacije raspodele:** Logaritamska transformacija i Box-Cox transformacija omogućavaju prilagođavanje raspodele podataka kako bi se poboljšala normalnost i smanjila asimetrija.
* **Enkodiranje kategoričkih atributa:** Metode poput One-Hot enkodiranja, target enkodiranja i label enkodiranja omogućavaju konverziju kategoričkih podataka u numerički format pogodniji za modele.
* **Metode za rad sa outlier-ima:** Tehnike kao što su Z-Score detekcija, Isolation Forest i DBSCAN omogućavaju identifikaciju i tretman outliera, čime se poboljšava robusnost modela.

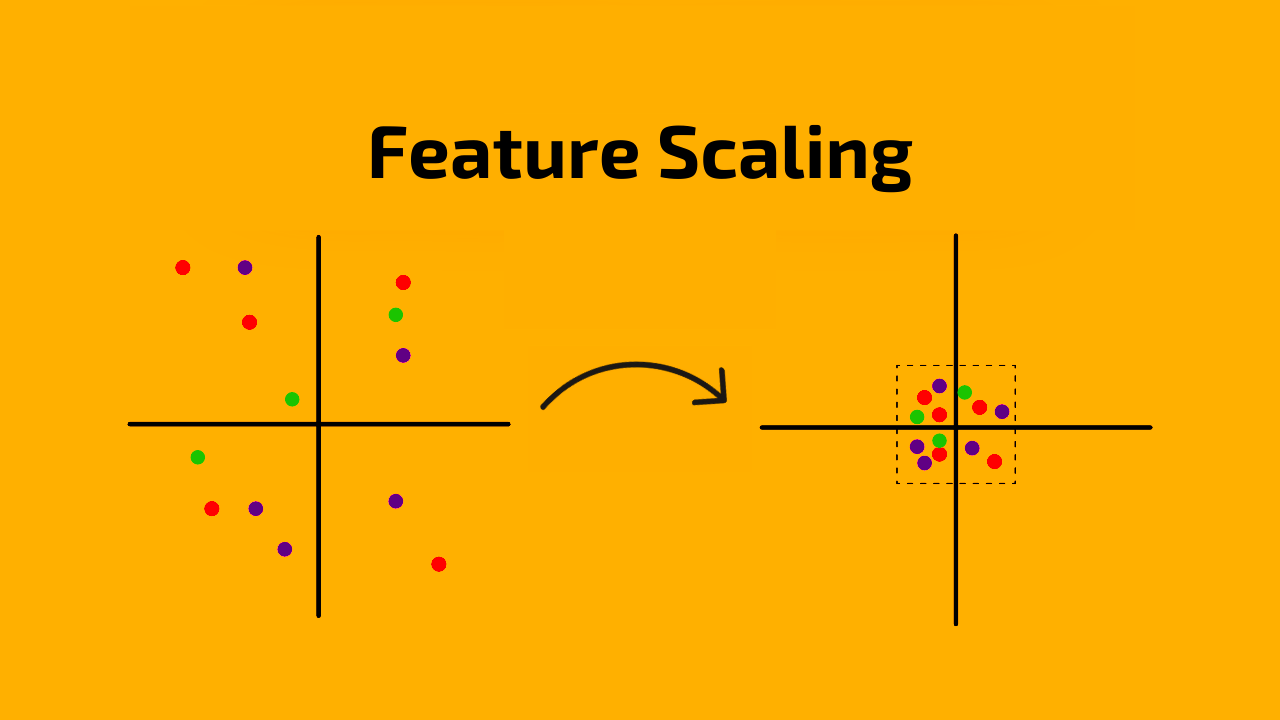
Ova rešenja omogućavaju efikasniju pripremu podataka, što direktno doprinosi poboljšanju performansi mašinskih modela i tačnosti njihovih predikcija.

# Skaliranje podataka

Skaliranje podataka je ključan korak u pripremi numeričkih atributa za mašinsko učenje, jer omogućava transformaciju vrednosti u uporedive opsege. Ova tehnika pomaže u minimiziranju uticaja različitih skala i ekstremnih vrednosti, što može značajno poboljšati performanse modela. Bez skaliranja, algoritmi često favorizuju varijable sa većim opsegom vrednosti, što dovodi do pristrasnih i nepreciznih predikcija.

Skaliranje podataka omogućava ravnotežan tretman svih atributa u modelu, čime se poboljšava efikasnost algoritama poput linearne regresije, k najbližih suseda (k-NN), i metoda zasnovanih na gradijentnom spustu.

U nastavku rada biće detaljno analizirane prednosti i nedostaci različitih tehnika skaliranja, sa ciljem identifikacije najpogodnijih metoda za različite tipove podataka i analitičke procese.



## Standardizacija (Z-score)

Standardizacija podataka, često poznata kao Z-score standardizacija, je tehnika koja se koristi za pretvaranje podataka tako da imaju srednju vrednost od 0 i standardnu devijaciju od 1. Ovaj postupak omogućava poređenje podataka različitih skala i smanjuje efekat različitih jedinica merenja u statističkim analizama i mašinskom učenju.

Formula za standardizaciju je:

gde je:

µ - srednja vrednost,  
σ – standardna devijacija

Rezultat ove transformacije je niz podataka gde su svi elementi izraženi u odnosu na odstupanje od srednje vrednosti u jedinicama standardne devijacije.

Ova tehnika je naročito korisna za podatke koji su normalno distribuirani, jer omogućava algoritmima da efikasnije rade sa podacima različitih skala, posebno kada se koriste modeli zasnovani na udaljenosti, poput linearne regresije ili k-najbližih suseda. Jedna od glavnih prednosti standardizacije je što eliminiše uticaj različitih jedinica merenja, omogućavajući konzistentnu analizu.

Međutim, Z-score standardizacija može biti osetljiva na preveliki broj outlier-a, jer ekstremne vrednosti značajno utiču na srednju vrednost i standardnu devijaciju, što može iskriviti rezultate. Ovu tehniku je pogodno koristiti kada skup podataka ima određeni broj outliera, koje nije pogodno izbaciti i koji nose informacije od značaja. Uprkos tome, ako su podaci približno normalno distribuirani, standardizacija je efikasan način da se obezbedi jednaka važnost za sve promenljive i smanji uticaj veličine promenljivih na krajnje rezultate analize.

## Min-Max skaliranje

Min-Max skaliranje je metoda transformacije podataka koja prilagođava vrednosti tako da budu unutar raspona između 0 i 1. Ova tehnika je korisna kada je potrebno da svi podaci budu na istoj skali, na primer, prilikom korišćenja algoritama poput K-Means klasterovanja. Proces Min-Max skaliranja se vrši oduzimanjem minimalne vrednosti od svake tačke podataka i deljenjem dobijenog rezultata sa opsegom vrednosti, čime se obezbeđuje da minimalna vrednost postane 0, a maksimalna 1, uz očuvanje relativnih odnosa među ostalim vrednostima.

Glavna prednost ove metode je to što omogućava lakšu interpretaciju podataka i osigurava da sve vrednosti budu unutar ograničenog raspona. Međutim, Min-Max skaliranje je veoma osetljivo na outliere, jer prisustvo ekstremnih vrednosti može značajno uticati na rezultat transformacije, dovodeći do kompresije većine podataka u uskom rasponu.

## Max-Abs skaliranje

Max-Abs skaliranje je tehnika koja transformiše numeričke vrednosti tako da one budu u opsegu od -1 do 1, koristeći maksimalnu apsolutnu vrednost kao referentnu tačku za skaliranje. Ova metoda je korisna za očuvanje pozitivnih i negativnih vrednosti u podacima, jer ne menja znakove vrednosti, već samo smanjuje opseg na osnovu maksimalne apsolutne vrednosti.

Jedna od glavnih prednosti Max-Abs skaliranja je njegova efikasnost i brzina, što ga čini pogodnim za rad sa velikim skupovima podataka. Još jedna značajna prednost je očuvanje retkosti podataka, što znači da ovaj pristup ne utiče negativno na podatke koji imaju mnogo nultih vrednosti, što je važno prilikom rada sa retkim matricama, kao što su one koje se koriste u obradi teksta ili recommended sistemima.

Ipak, kao i druge metode skaliranja, Max-Abs skaliranje je osetljivo na outliere. Ekstremne vrednosti mogu značajno uticati na referentnu maksimalnu apsolutnu vrednost, što dovodi do nepravilnog skaliranja ostatka podataka. Zbog toga je važno biti oprezan kada se ova metoda koristi na podacima sa izraženim outlier-ima.

## Robust skaliranje

Robust skaliranje je metoda koja transformiše podatke korišćenjem statistika koje su otporne na outliere. Umesto da koristi prosečnu vrednost i standardnu devijaciju, ova tehnika oduzima medijanu i skalira podatke koristeći opseg između prvog i trećeg kvartila, poznat kao interkvartilni opseg (IQR). To znači da se podaci skaliraju u rasponu između 25. i 75. percentila, čime se minimizira uticaj ekstremnih vrednosti.

Jedna od glavnih prednosti Robust skaliranja je to što je manje osetljivo na outliere u poređenju sa metodama koje koriste prosečnu vrednost i standardnu devijaciju, jer su medijana i interkvartilni opseg otporniji na ekstreme.

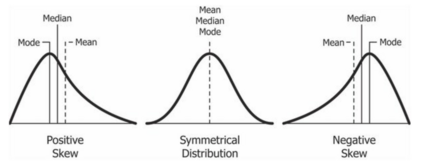
Robust skaliranje je naročito korisno kada podaci sadrže outliere i kada je potrebno da se očuva stabilnost transformacije, ali je važno proceniti veličinu i broj ekstremnih vrednosti u podacima pre nego što se ova tehnika primeni.

# Transformacije koje menjaju raspodelu vrednosti

**Asimetrija raspodele (Skewness)** je mera koja opisuje odstupanje distribucije neke nasumične varijable od simetrične raspodele, kao što je normalna distribucija. U normalnoj distribuciji, podaci su simetrično raspoređeni oko srednje vrednosti. Ako je kriva pomerena udesno ili ulevo, kažemo da je raspodela asimetrična (skewed). Asimetrija može biti pozitivna ili negativna, zavisno od toga na koju stranu je „rep“ krive izdužen. Poznavanje stepena i tipa asimetrije važno je jer pomaže u pravilnoj analizi i interpretaciji raspodela podataka, naročito u ekonomiji, finansijama i mašinskom učenju.

**Tipovi asimetrije:**

* **Pozitivna asimetrija (right-skewed):** Podaci su nagnuti ulevo, a rep distribucije ide udesno. U ovoj situaciji, srednja vrednost je veća od medijane.
* **Negativna asimetrija (left-skewed):** Podaci su nagnuti udesno, a rep distribucije ide ulevo. U ovoj situaciji, srednja vrednost je manja od medijane.



## Logaritamska transformacija

**Logaritamska transformacija** je matematička operacija koja se koristi za pretvaranje podataka sa ciljem da se poboljša njihova raspodela, posebno kada podaci pokazuju značajnu asimetriju ili ekstremne vrednosti. Logaritamska transformacija podrazumeva zamenu svake vrednosti u skupu podataka sa njenim logaritmom. Najčešće korišćena logaritamska funkcija je prirodni logaritam, koji koristi osnovu „e“ (približno 2.71828). Postoje i druge osnove, kao što su 10 i 2, ali prirodni logaritam je najzastupljeniji.

Logaritamska transformacija je korisna u raznim situacijama, kao što su:

* **Smanjenje uticaja ekstremnih vrednosti:** Ekstremne vrednosti mogu imati veliki uticaj na rezultate analize. Logaritamska transformacija kompresuje ove vrednosti, smanjujući njihov uticaj.
* **Korekcija asimetrije:** Kada su podaci pozitivno ili negativno nagnuti, logaritamska transformacija može pomoći u približavanju raspodele normalnoj raspodeli.
* **Linearizacija odnosa:** U mnogim modelima, poput linearne regresije, pretpostavlja se da je odnos između varijabli linearan. Ako to nije slučaj, primena logaritamske transformacije može pomoći u linearizaciji.
* **Stabilizacija varijanse:** Ukoliko varijansa nije konstantna, logaritamska transformacija može pomoći u stabilizaciji varijanse.

Da bi se uspešno primenila logaritamska transformacija, sve vrednosti u skupu podataka moraju biti pozitivne, jer logaritam nije definisan za nule ili negativne brojeve. Ako su prisutne nule ili negativne vrednosti, potrebno je dodati konstantu pre nego što se primeni transformacija.

**Interpretacija koeficijenata kod logaritamski transformisanih modela:**

Logaritamska transformacija je korisna za poboljšanje modela, ali ona menja način na koji interpretiramo koeficijente regresije. Dok su u standardnim (netransformisanim) modelima koeficijenti lako razumljivi — kao promena zavisne varijable za svaku jedinicu promene nezavisne varijable — logaritamski transformisani modeli zahtevaju pažljiviju interpretaciju.

* **Log — Level regresija**: U **log-level regresiji**, zavisna varijabla je logaritamski transformisana, dok su nezavisne varijable ostale u svom originalnom obliku. Koeficijenti u ovom modelu se tumače kao procentualna promena zavisne varijable za svaku promenu od jedne jedinice u nezavisnoj varijabli. Dakle, umesto apsolutne promene, rezultat koeficijenata ukazuje na relativne promene u zavisnoj varijabli.
* **Level — Log regresija**: U **level-log regresiji**, nezavisna varijabla je logaritamski transformisana, dok zavisna varijabla ostaje u originalnom obliku. Koeficijenti u ovom modelu se tumače kao apsolutne promene zavisne varijable koje odgovaraju relativnim promenama u nezavisnoj varijabli. Drugim rečima, svaka procentualna promena nezavisne varijable dovodi do određene apsolutne promene u zavisnoj varijabli.
* **Log — Log regresija**: U **log-log regresiji**, i zavisna i nezavisna varijabla su logaritamski transformisane. Koeficijenti u ovom modelu se tumače kao elastičnost — to jest, procentualna promena zavisne varijable kao odgovor na procentualnu promenu nezavisne varijable. Ovaj tip modela najčešće se koristi kada su i zavisna i nezavisna varijabla bolje opisane kroz relativne promene nego apsolutne vrednosti.

Logaritamska transformacija je moćan alat u statističkoj analizi i modeliranju podataka, posebno kada podaci pokazuju asimetriju ili ekstremne vrednosti. Korišćenje ove transformacije ne samo da poboljšava raspodelu podataka, već omogućava linearniju interpretaciju odnosa između varijabli, stabilizuje varijansu i smanjuje uticaj ekstremnih vrednosti. Iako njena primena zahteva pažljiviju interpretaciju koeficijenata u regresijskim modelima, ona omogućava dublje i preciznije razumevanje odnosa među podacima, čime postaje ključna u mnogim analitičkim i naučnim disciplinama.

## Box Cox transformacija

Box-Cox transformacija je tehnika koja se koristi za transformaciju podataka sa ciljem da se približe normalnoj raspodeli, što je posebno korisno u modelima koji pretpostavljaju normalnost grešaka. Box-Cox transformacija se primenjuje na ciljnu varijablu, a najčešće koristi parametar 𝞴 koji se podešava da bi se postigla što bolja aproksimacija normalne raspodele.

Box-Cox transformacija je korisna u različitim situacijama, kao što su:

* **Približavanje normalnoj raspodeli:** Transformacija podataka prema normalnoj raspodeli omogućava korišćenje statističkih tehnika koje zahtevaju normalnost grešaka.
* **Povećanje prediktivne moći modela:** Smanjenjem šuma u podacima, ova transformacija može poboljšati preciznost modela.
* **Uklanjanje asimetrije:** Box-Cox transformacija pomaže da se smanji asimetrija u podacima, što vodi ka simetričnijoj raspodeli.

Izbor optimalne vrednosti za 𝞴 je ključan za postizanje normalnosti podataka. Vrednost 𝞴 se obično bira tako da minimizuje odstupanje od normalne raspodele, a može se automatski odrediti pomoću statističkih alata, poput funkcije boxcox iz biblioteke SciPy.

gde je:

T(Y) – transformisana vrednost ciljne varijable  
Y – originalna vrednost ciljne varijable  
𝞴 – parametar koji kontriliše oblik transformacije

Ako je 𝞴 =1, podaci ostaju gotovo nepromenjeni. Kada je 𝞴 =0, transformacija se svodi na logaritamsku transformaciju. Optimalna vrednost 𝞴 omogućava minimizaciju odstupanja od normalne raspodele, što se postiže evaluacijom različitih vrednosti 𝞴 i izborom one koja daje najbolje rezultate.

Box-Cox transformacija je moćna statistička tehnika koja omogućava efikasno prilagođavanje podataka normalnoj raspodeli, čime se poboljšava upotrebljivost statističkih modela. Posebno je korisna u smanjenju asimetrije i šuma u podacima, što doprinosi preciznijim prognozama i validnijim analizama. Ključni aspekt ove transformacije je pažljiv odabir parametra 𝞴 koji određuje intenzitet transformacije i omogućava optimalno usklađivanje podataka sa normalnom raspodelom. Iako se izbor 𝞴 može automatizovati, pravilna implementacija zahteva razumevanje njenog uticaja na rezultate modela i interpretaciju podataka.

## Yeo-Johnson transformacija

Yeo-Johnson transformacija je statistička tehnika, slična Box-Cox transformaciji, ali proširena za rad sa dataset-ovima koji sadrže i pozitivne i negativne vrednosti.

Yeo-Johnson transformacija je korisna u raznim situacijama, kao što su:

* **Obrada negativnih i pozitivnih vrednosti:** Za razliku od Box-Cox transformacije, Yeo-Johnson može da radi sa podacima koji uključuju negativne vrednosti, čineći je fleksibilnijom.
* **Približavanje** **normalnoj raspodeli:** Pomaže u transformaciji podataka ka normalnijoj raspodeli, što je važno za primenu mnogih statističkih metoda.
* **Smanjenje asimetrije:** Kao i kod Box-Cox transformacije, smanjuje asimetriju podataka i pomaže u stabilizaciji varijanse.

Formula za Yeo-Johnson transformaciju zavisi od toga da li su vrednosti pozitivne ili negativne:

Izbor parametra 𝞴 je ključan i sličan kao kod Box-Cox transformacije. Optimalna vrednost 𝞴 se određuje putem maksimum-likelihood procene, a njen cilj je da se minimizuje odstupanje od normalne raspodele. Kada su vrednosti striktno pozitivne, Yeo-Johnson transformacija se ponaša kao Box-Cox transformacija.

[dodaj neku referencu za ovu transformaciju]

## Kvantilna transformacija

Kvantilna transformacija, poznata i kao **Rank transformacija**, je tehnika koja mapira vrednosti promenljive na različitu raspodelu, najčešće uniformnu ili normalnu. Ova transformacija koristi rangiranje podataka kako bi se očuvala relativna pozicija svake vrednosti, a istovremeno se postiže ravnomerno raspoređivanje vrednosti unutar nove raspodele. Kvantilna funkcija (poznata i kao percent-point funkcija, PPF) predstavlja inverznu funkciju kumulativne distribucije verovatnoće (CDF), što omogućava mapiranje vrednosti u odnosu na željenu raspodelu.

Kvantilna transformacija je korisna u različitim situacijama, kao što su:

* **Otpornost na ekstremne vrednosti:** Zbog rangiranja podataka, kvantilna transformacija je robusna prema ekstremnim vrednostima i može ublažiti njihov uticaj na analizu.
* **Očuvanje ranga:** Rang vrednosti ostaje očuvan čak i nakon transformacije, što omogućava stabilnost podataka u smislu relativnog poređenja.
* **Promena distribucije:** Kvantilna transformacija može promeniti oblik raspodele podataka, omogućavajući prebacivanje sa originalne raspodele na uniformnu ili normalnu raspodelu.

Formula za kvantilnu transformaciju oslanja se na PPF funkciju, koja za zadatu verovatnoću vraća odgovarajuću vrednost, dok CDF funkcija radi obrnuto — vraća verovatnoću za zadatu vrednost. Primena ove transformacije na podatke omogućava prilagođavanje različitim modelima mašinskog učenja, posebno kada su podaci različitih skala ili sadrže izražene outliere.

Kvantilna transformacija je efikasna metoda za unapređenje prediktivnih modela, naročito u situacijama kada je potrebna ravnomernija raspodela podataka, otpornost na outliere, i kada podaci dolaze iz različitih skala ili raspodela.

# Enkodiranje kategoričkih podataka

Enkodiranje kategoričkih podataka je proces pretvaranja varijabli koje sadrže kategoričke vrednosti u numeričke karakteristike. Ovaj korak u preprocesiranju podataka je esencijalan za većinu zadataka u mašinskom učenju, jer omogućava algoritmima da adekvatno obrađuju i interpretiraju kategoričke podatke.

**Šta su kategorički podaci?**

Kategorički podaci predstavljaju promenljive čije vrednosti dolaze iz određenih grupa kategorija ili oznaka. Vrednosti ovih promenljivih mogu biti predstavljene stringovima ili brojevima, ali u osnovi predstavljaju klase ili oznake koje se ne mogu numerički obraditi bez odgovarajuće transformacije. Ove varijable mogu biti podeljene u dve grupe:

* **Nominalni podaci:** Varijable koje nemaju inherentan redosled ili rangiranje. Na primer, boje automobila kao što su crvena, plava ili zelena.
* **Ordinalni podaci:** Varijable koje imaju prirodan redosled ili rangiranje. Na primer, nivo obrazovanja (osnovna škola, srednja škola, fakultet).

**Zašto je enkodiranje važno?**

Mašinsko učenje zahteva numeričke ulazne podatke za procesiranje i prepoznavanje obrazaca. Enkodiranje kategoričkih atributa omogućava modelima da koriste ove podatke za treniranje i predikciju. Ključni razlozi za enkodiranje kategoričkih podataka uključuju:

* **Kompatibilnost sa modelima:** Većina algoritama zahteva numeričke vrednosti, pa je konverzija kategoričkih podataka neophodna za njihovu obradu.
* **Prepoznavanje obrazaca:** Efikasno enkodiranje pomaže modelima da prepoznaju ključne obrasce i odnose unutar podataka.
* **Prevencija pristrasnosti:** Enkodiranje kategoričkih podataka osigurava ravnotežu između različitih kategorija, sprečavajući pristrasnost ka određenim vrednostima.
* **Smanjenje dimenzionalnosti:** Određene tehnike enkodiranja, kao što su Target enkodiranje ili Frequency enkodiranje, mogu pomoći u smanjenju složenosti i dimenzionalnosti podataka.

Izbor odgovarajuće tehnike enkodiranja zavisi od prirode podataka, broja kategorija, kao i specifičnih zahteva algoritma koji se koristi. Efikasno enkodiranje poboljšava performanse modela, optimizuje prepoznavanje obrazaca, i smanjuje mogućnost grešaka u predikcijama izazvanih pogrešnim obradama kategorijskih vrednosti.

## One Hot enkodiranje

**One-hot enkodiranje** je tehnika za konvertovanje kategoričkih varijabli u numerički format, posebno pogodna za **nominalne kategorije** bez prirodnog poretka. Ova metoda funkcioniše tako što za svaku jedinstvenu kategoriju kreira novu binarnu kolonu. Svaka kolona predstavlja jednu kategoriju, a vrednosti u kolonama su 1 ako podatak pripada toj kategoriji, ili 0 ako ne pripada. Ova tehnika dobro podnosi nedostajuće kategorije, jer kreira sve nule u binarnim kolonama kada kategorija nije prisutna.

Iako je ova tehnika efikasna, može dovesti do problema poput **prokletstva dimenzionalnosti**, gde se broj kolona drastično povećava za varijable sa mnogo kategorija, ili **multikolinearnosti**, jer kreirane kolone mogu biti korelisane.

## Dummy enkodiranje

**Dummy enkodiranje** je varijacija one-hot enkodiranja koja se koristi za konvertovanje kategoričkih varijabli u numerički format, ali sa ključnom razlikom – umesto kreiranja binarne kolone za svaku kategoriju, dummy encoding kreira **(N-1)** binarnih kolona za **N** kategorija.

Kao i one-hot enkodiranje, dummy enkodiranje je pogodno za rad sa nominalnim podacima, ali smanjuje broj novih kolona, čineći ga efikasnijim u slučajevima gde postoji veliki broj kategorija.

Međutim, slični izazovi kao kod one-hot enkodiranja ostaju prisutni, poput **prokletstva dimenzionalnosti** i **retkosti podataka**, gde veliki broj binarnih kolona može rezultirati matricama sa mnogo nula, što utiče na efikasnost skladištenja i performanse modela.

## Label enkodiranje

**Label enkodiranje** je tehnika koja se koristi za konvertovanje kategoričkih podataka u numeričke vrednosti dodeljivanjem jedinstvene, celobrojne vrednosti svakoj kategoriji unutar karakteristike. Svaka kategorija dobija numeričku oznaku, obično zasnovanu na abecednom poretku ili redosledu pojavljivanja u skupu podataka.

Ova metoda je posebno pogodna za **ordinalne podatke**, gde kategorije imaju prirodan redosled. Međutim, kod nominalnih podataka, gde kategorije nemaju prirodan redosled, label enkodiranje može uvesti neželjen redosled među kategorijama, što može navesti model da pretpostavi rangiranje koje ne postoji.

## Ordinalno enkodiranje

**Ordinalno enkodiranje** je tehnika slična label enkodiranju, ali se koristi isključivo kada kategorije imaju **inherentni poredak** ili rang. Za razliku od label enkodiranja, ordinalno enkodiranje eksplicitno koristi prirodni redosled kategorija prilikom dodeljivanja numeričkih vrednosti, što ga čini pogodnim za karakteristike poput nivoa obrazovanja, ocena zadovoljstva ili rangova u takmičenju.

Glavna razlika između **label** i **ordinalnog enkodiranja** leži u primeni. Label enkodiranje može biti pogrešno primenjeno na nominalne podatke, gde ordinalno enkodiranje ne bi bilo prikladno, dok je ordinalno enkodiranje posebno dizajnirano da očuva redosled među kategorijama. Ova metoda smanjuje dimenzionalnost skupa podataka u poređenju sa one-hot encoding-om, što ga čini efikasnijim za podatke sa jasnim poretkom.

Međutim, kao i kod label enkodiranja, jedan od izazova ordinalnog enkodiranja je da neki algoritmi mašinskog učenja mogu **pretpostaviti linearan odnos** između dodeljenih numeričkih vrednosti, što nije uvek prikladno za sve ordinalne karakteristike.

## Binarno enkodiranje

**Binarno enkodiranje** je tehnika koja se koristi za konvertovanje kategoričkih varijabli u numerički format, pri čemu se kategorije predstavljaju pomoću binarnih cifara. Ova metoda kombinuje prednosti one-hot i label enkodiranja, ali smanjuje dimenzionalnost, što je posebno korisno kod varijabli sa visokom kardinalnošću, odnosno velikim brojem jedinstvenih kategorija.

Jedna od ključnih prednosti binarnog enkodiranja je smanjenje dimenzionalnosti, jer broj kreiranih kolona raste sporije u odnosu na one-hot encoding. Takođe je efikasnije u smislu memorije i relativno jednostavno za implementaciju i interpretaciju. Međutim, kod varijabli sa ekstremno visokom kardinalnošću, kompleksnost može i dalje biti izazov, a takođe je potrebno obratiti pažnju na način rukovanja sa nedostajućim vrednostima tokom procesa enkodiranja.

Binarno enkodiranje predstavlja efikasan način za rad sa kategoričkim podacima u situacijama gde je broj jedinstvenih kategorija velik, pružajući balans između preciznosti i dimenzionalnosti.

## Count and Frequency enkodiranje

**Count** (**frequency) enkodiranje** je tehnika za pretvaranje kategoričkih varijabli u numerički format, dodeljivanjem numeričke vrednosti svakoj kategoriji na osnovu njene učestalosti u skupu podataka. Ova metoda dodeljuje veću vrednost kategorijama koje se pojavljuju češće, a manju vrednost onima koje se pojavljuju ređe.

Count enkodiranje je korisno u situacijama gde je frekvencija kategorija značajan faktor, kao što je analiza kupovnih obrazaca u segmentaciji korisnika. Ova tehnika takođe smanjuje dimenzionalnost u poređenju sa one-hot encoding-om, što je posebno korisno za varijable sa visokom kardinalnošću.

Ipak, izazov sa ovom metodom je mogućnost gubitka specifičnih informacija o kategorijama, jer kategorije sa istom frekvencijom dobijaju istu vrednost. Takođe, nije pogodna za ordinalne podatke gde je redosled kategorija važan, jer frekvencija ne odražava njihov poredak.

## Target enkodiranje

**Target enkodiranje**, poznato i kao **Mean enkodiranje**, je napredna tehnika za enkodiranje kategoričkih varijabli, koja se posebno koristi kod varijabli sa visokom kardinalnošću. Ova metoda funkcioniše tako što svakoj kategoriji dodeljuje numeričku vrednost na osnovu prosečne vrednosti ciljne promenljive unutar te kategorije, čime se uspostavlja direktna veza između kategoričke varijable i ciljne varijable.

Target enkodiranje je posebno korisno u klasifikacionim problemima jer uspešno zadržava informacije o tome koliko je verovatno da će neka kategorija izazvati određenu vrednost ciljne promenljive. Takođe, smanjuje dimenzionalnost skupa podataka u poređenju sa one-hot enkodiranjem, čime se efikasno obrađuju varijable sa mnogo kategorija, a zadržava se važno ponašanje specifično za kategorije.

Međutim, target enkodiranje dolazi sa izazovima, uključujući mogućnost overfittinga i curenja podataka. Target enkodiranje je snažna tehnika kada je potrebno očuvati odnose između kategorijskih varijabli i ciljne promenljive, ali zahteva pažljivu primenu kako bi se osigurala pravilna generalizacija modela.

## Effect enkodiranje

**Effect enkodiranje**, poznato i kao **deviation encoding** ili **sum encoding**, je tehnika za enkodiranje kategoričkih varijabli. Ova metoda je slična dummy enkodiranju, ali sa ključnom razlikom – umesto binarnih vrednosti (0 i 1), koristi tri vrednosti: 1, 0, i -1. Ova tehnika je posebno korisna kod linearnog modeliranja jer efikasno rešava problem multikolinearnosti i omogućava lakše tumačenje koeficijenata u modelima.

**Effect enkodiranje** funkcioniše tako što svakoj kategoriji unutar varijable dodeljuje binarne vrednosti, ali koristi tri različite vrednosti: 1, 0 i -1. Prvi korak je identifikacija kategoričke varijable koja treba biti enkodirana, nakon čega se kreiraju binarne kolone za sve kategorije osim jedne, koja služi kao bazna kategorija. Kategorijama se dodeljuju vrednosti 1 za prisustvo određene kategorije, 0 za odsustvo, dok se za redove koji bi inače imali sve 0 (u dummy enkodiranju) koristi vrednost -1 za baznu kategoriju, što omogućava bolju interpretaciju i analizu modela.

Glavna prednost effect enkodiranja je to što efikasnije rešava problem **multikolinearnosti**, koji se javlja u linearnim modelima, u poređenju sa dummy enkodiranjem. Korišćenjem vrednosti -1 za baznu kategoriju, smanjuje se redundantnost informacija, što vodi stabilnijim modelima. Takođe, koeficijenti u linearnim modelima postaju lakši za tumačenje, jer predstavljaju odstupanja svake kategorije u odnosu na ukupnu prosečnu vrednost, što omogućava intuitivnije zaključke u analizi rezultata.

## Feature Hashing enkodiranje

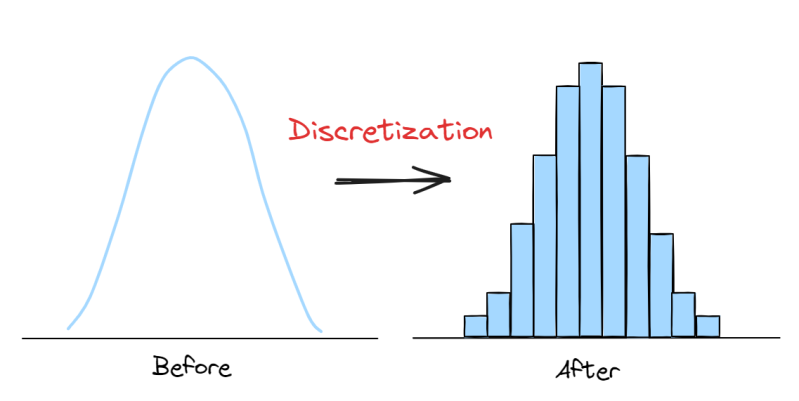
**Feature hashing**, poznato i kao **hashing trik**, je tehnika za enkodiranje kategoričkih varijabli koja omogućava efikasnu obradu podataka sa visokom kardinalnošću. Funkcioniše tako što se na kategoričke podatke primenjuje hash funkcija koja svaku kategoriju mapira na fiksan broj numeričkih kolona. Ova hash funkcija distribuira kategorije kroz više kolona, a svaka kategorija doprinosi vrednostima u više kolona istovremeno. Tako se smanjuje dimenzionalnost skupa podataka, dok se zadržava numerička reprezentacija kategorijskih varijabli.

Ova tehnika omogućava smanjenje dimenzionalnosti skupa podataka, što je od velike koristi kada su memorijski i računarski resursi ograničeni. Međutim, potencijalni problem kod feature hashinga su **hash kolizije**, gde različite kategorije mogu biti mapirane na iste kolone, što može dovesti do gubitka specifičnosti podataka. Ipak, čak i uz ovu mogućnost, feature hashing nudi jednostavno i efikasno rešenje za rad sa velikim skupovima podataka.

# Diskretizacija (binovanje) podataka

**Diskretizacija** (poznata i kao **binovanje**) je proces pretvaranja kontinualnih varijabli u diskretne grupisanjem vrednosti u susedne intervale (binove). Ovaj postupak se često koristi u pripremi podataka za modele mašinskog učenja, kao i u rudarenju podataka. Diskretizacija omogućava jednostavniju obradu i analizu podataka, poboljšava performanse algoritama poput stabala odlučivanja i Naive Bayes-a, i olakšava tumačenje rezultata. Takođe, može smanjiti uticaj outlier-a grupisanjem ekstremnih vrednosti u odgovarajuće intervale, čime doprinosi stabilnijim modelima.

Glavni cilj diskretizacije je podela kontinualnih vrednosti na što manji broj intervala bez značajnog gubitka informacija, čime se ubrzava proces obrade i treniranja modela. Iako može dovesti do gubitka detalja, pravilno primenjena diskretizacija smanjuje kompleksnost podataka i može rezultirati preciznijim modelima.



## Podela u intervale jednakih širina

**Diskretizacija na osnovu intervala jednake širine** podrazumeva podelu opsega kontinuiranih vrednosti varijable na **k** intervala iste veličine. Ova metoda je jednostavna za implementaciju i tumačenje, jer ravnomerno deli podatke prema opsegu vrednosti.

Međutim, nedostatak ove metode je što ne uzima u obzir distribuciju podataka, pa u slučajevima kada podaci imaju **asimetričnu distribuciju** ili sadrže outliere, može stvoriti **prazne ili retke binove**, što može smanjiti kvalitet analize. Diskretizacija jednake širine ne menja osnovnu distribuciju podataka, pa varijable koje su inicijalno asimetrične ostaju takve i nakon diskretizacije.

## Podela na intervale sa jednakom frekvencijom

**Diskretizacija na osnovu jednake frekvencije** podrazumeva podelu kontinuirane varijable u intervale tako da svaki interval sadrži približno isti broj instanci. Širina intervala se određuje na osnovu kvantila, čime se postiže ravnomerna raspodela podataka unutar intervala, što je naročito korisno kod **asimetrično raspodeljenih** varijabli.

Prednost ove metode je što kreira **uravnotežene intervale** koji mogu bolje da upravljaju podacima sa outlierima ili značajnom asimetrijom. Međutim, nedostatak je što može doći do **izobličavanja distribucije** podataka i stvaranja **nepravilnih širina intervala**, što može otežati tumačenje i analizu podataka.

## Diskretizacija korišćenjem klasterizacije

Diskretizacija uz pomoć **klasterizacije** koristi algoritme poput **K-Means** kako bi kontinuirane varijable podelila u grupe ili klastere na osnovu sličnosti između posmatranja. Ova metoda formira intervale (binove) tako što grupiše slične vrednosti u klastere, pri čemu podaci unutar svakog klastera imaju visoku međusobnu sličnost, dok su podaci iz različitih klastera što različitiji.

Na primer, algoritam **K-Means** deli podatke na **k** klastera, pri čemu svaki klaster predstavlja grupu sličnih vrednosti varijable. Svaki klaster se može opisati ključnim karakteristikama kao što su **centroid** (središnja tačka klastera) i **dijametar** (razdaljina unutar klastera). Prilikom diskretizacije na ovaj način, kontinuirane vrednosti varijable bivaju zamenjene oznakom odgovarajućeg klastera.

Diskretizacija korišćenjem klasterizacije je posebno korisna kada se traži dublja struktura unutar podataka, jer algoritmi poput **K-Means** mogu otkriti prirodne grupe u podacima. Međutim, izbor broja klastera (**k**) je ključni parametar i može značajno uticati na rezultat diskretizacije. Evaluacija klastera se često vrši pomoću metrika kao što su **inercija**, koja meri kompaktnost klastera, i **Dunn-ov indeks**, koji ocenjuje udaljenost između klastera i homogenost unutar klastera.

## Diskretizacija korišćenjem stabla odlučivanja

**Diskretizacija pomoću stabla odlučivanja** je metoda koja automatski deli kontinuirane varijable u diskretne intervale tokom procesa učenja modela. Ova tehnika funkcioniše tako što stablo odlučivanja analizira sve moguće vrednosti varijable i odabira **tačke preseka** koje maksimalno poboljšavaju razdvajanje klasa, koristeći metrike kao što su **entropija** ili **Gini impurity**. Proces se ponavlja na svakom čvoru stabla dok se ne postigne zadati kriterijum zaustavljanja. Na taj način, stablo odlučivanja prirodno pronalazi optimalne tačke za diskretizaciju, kreirajući intervale koji najbolje razdvajaju podatke u koherentne grupe.

Glavna prednost ove metode je što se **optimalna tačka preseka** automatski identifikuje na osnovu strukture podataka, bez potrebe za ručnim određivanjem broja binova. Nakon što stablo odlučivanja pronađe tačke preseka, predikcije stabla se koriste kao oznake za diskretne grupe (binove), što čini ovaj pristup posebno korisnim za klasifikacione zadatke.

## Chi Merge

**Chi-merge** je **nadgledana metoda** diskretizacije koja koristi klasičnu statističku tehniku za grupisanje vrednosti kontinuiranih varijabli. Ova metoda radi tako što prvo sortira vrednosti varijable u rastućem redosledu i grupiše ih u intervale koji sadrže identične vrednosti. Nakon toga, **Chi-merge** iterativno spaja susedne intervale na osnovu rezultata **chi-square testa** (χ2), koji procenjuje sličnost distribucije klasa unutar tih intervala. Intervali se spajaju dok se ne postigne unapred definisani kriterijum zaustavljanja, a koristi se isključivo u situacijama kada je ciljna varijabla diskretna.

Ovaj **bottom-up pristup** spaja intervale sa niskim χ2 vrednostima, jer niska vrednost ovog testa ukazuje na to da su distribucije klasa u susednim intervalima slične. Proces se ponavlja dok ne ostane odgovarajući broj intervala, koji je određen stop-kriterijumom. Chi-merge je posebno koristan za zadatke klasifikacije jer koristi informacije o ciljnim klasama za optimizaciju intervala, osiguravajući da se različite klase efikasno razdvoje unutar diskretizovanih intervala.

# Rad sa outlier-ima

**Outlieri** predstavljaju tačke u skupu podataka koje značajno odstupaju od većine. Oni mogu biti izuzetno visoke ili niske vrednosti u poređenju sa ostatkom podataka i često iskrivljuju distribuciju podataka. Outlieri nastaju iz različitih razloga, poput grešaka u unosu podataka, nepravilnih merenja ili stvarnih, ali retkih pojava. Prepoznavanje i tretiranje outlier-a je ključan korak u obradi podataka, jer mogu značajno uticati na performanse modela mašinskog učenja i iskriviti rezultate analize.

Kao ekstremne vrednosti, outlieri mogu izobličiti **srednju vrednost** (prosek), povlačeći je ka sebi i dajući pogrešan utisak o centralnoj tendenciji podataka. Takođe, mogu značajno povećati **standardnu devijaciju**, što vodi ka većoj varijabilnosti podataka nego što to zapravo jeste. Efektivna detekcija outliera je važna jer omogućava uklanjanje tačaka koje su stvarno izuzetne, kako bi modeli bolje generalizovali i postigli bolje performanse na test podacima.

Tipovi odstupanja:

1. **Globalna odstupanja**: Tačke koje značajno odstupaju od celokupnog skupa podataka.
   * **Primer**: U skupu podataka o godinama studenata, vrednost od 150 godina bila bi globalno odstupanje.
2. **Lokalna odstupanja**: Tačke koje značajno odstupaju od okoline u okviru određenog dela skupa podataka.
   * **Primer**: Kuća sa neobično niskom cenom u luksuznom naselju može biti lokalno odstupanje unutar tog naselja.
3. **Univarijantna odstupanja**: Odstupanja koja se primećuju u jednoj promenljivoj.
   * **Primer**: Rezultat testa koji značajno odstupa od proseka ostalih rezultata na tom testu.
4. **Multivarijantna odstupanja**: Odstupanja koja postaju očigledna samo kada se razmatra zajednička distribucija dve ili više promenljivih.
   * **Primer**: Pojedinac sa izuzetno visokim prihodima u odnosu na nivo obrazovanja može biti multivarijantno odstupanje.

Uzroci odstupanja: (mozda moze da se brise)

1. **Greške u unosu podataka**: Ljudske greške pri unosu podataka.
2. **Greške u merenju**: Netacna očitavanja instrumenata.
3. **Eksperimentalne greške**: Greške u planiranju ili izvođenju eksperimenata.
4. **Namerno**: Lažna odstupanja napravljena radi testiranja metoda detekcije.
5. **Prirodna**: Noviteti u podacima koji nisu greške, već su retke pojave.

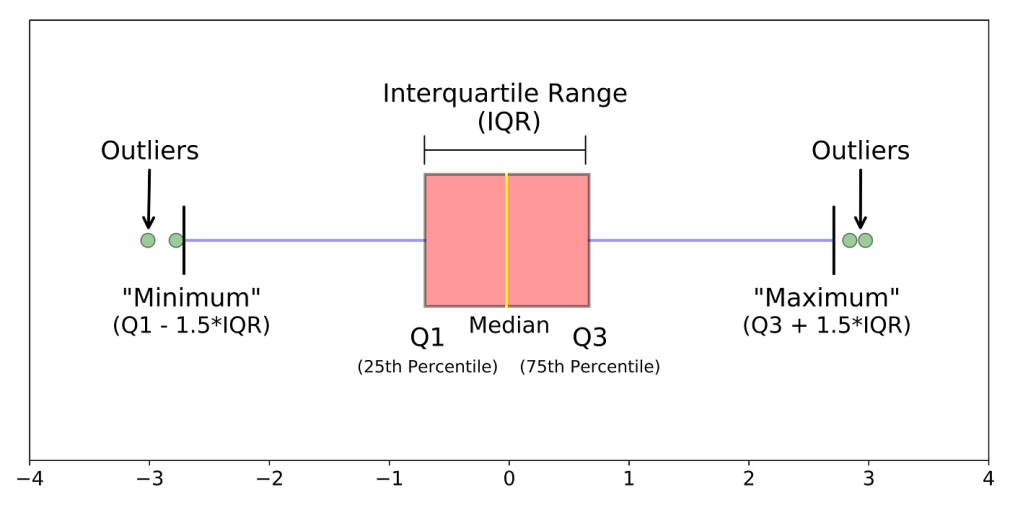
Vizuelne tehnike poput **scatter plotova**, **box plotova** i **histograma** mogu biti korisne za prepoznavanje outliera. Ipak, konačna odluka o tome da li treba ukloniti outliere zavisi od domen ekspertize i konteksta podataka, jer ono što izgleda kao odstupanje u nekim situacijama može zapravo predstavljati važne trendove ili obrasce u podacima.

## Tehnike vizuelizacije za detekciju outlier-a

Vizuelizacija podataka je izuzetno korisna tehnika za identifikaciju outliera, jer omogućava jednostavno uočavanje nepravilnosti i ekstremnih vrednosti unutar skupa podataka. Kroz različite grafičke prikaze, možemo brzo prepoznati tačke koje značajno odstupaju od opšteg trenda, što olakšava analizu i unapređuje kvalitet modela. Među najefikasnijim tehnikama za detekciju outliera su **box plot**, **scatter plot** i **histogram**.

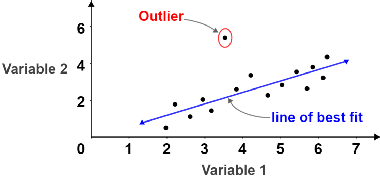
1. **Box Plot**

Box plot (poznat i kao "box-and-whisker" dijagram) prikazuje raspodelu podataka koristeći medijanu, kvartile i potencijalne outliere. Centralna polovina podataka nalazi se unutar interkvartilnog raspona (IQR), dok se podaci koji leže van 1.5 puta IQR od gornjeg ili donjeg kvartila smatraju mogućim outlierima i prikazuju kao pojedinačne tačke. Box plot je posebno koristan kada je potrebno uporediti distribucije između različitih grupa ili kategorija, kao i pri analizi jedne promenljive.



1. **Scatter Plot**

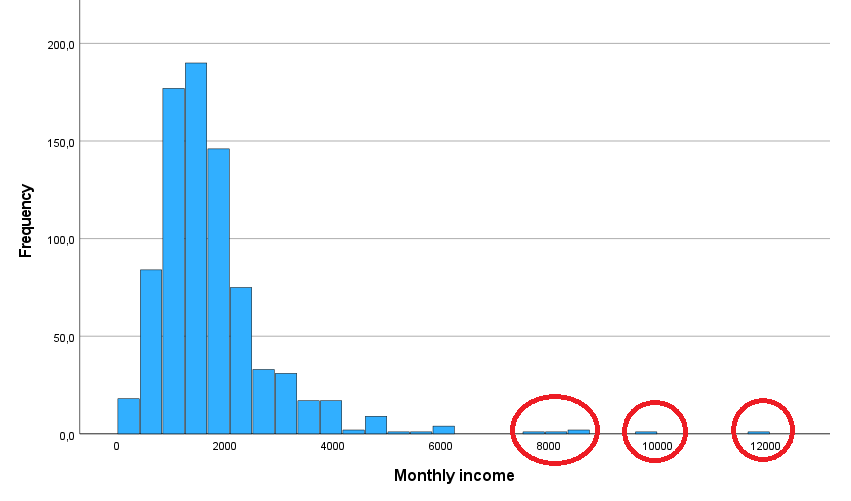
Scatter plot prikazuje pojedinačne tačke podataka u dvodimenzionalnom prostoru, što omogućava lako prepoznavanje odstupanja od opšteg obrasca ili odnosa između dve promenljive. Outlier-i se lako identifikuju kao tačke koje značajno odstupaju od trenda. Ova tehnika je idealna za analizu odnosa između dve kontinuirane promenljive i za otkrivanje tačaka koje se ne uklapaju u očekivanu relaciju.



1. **Histogram**

Histogram vizualizuje frekvencijsku distribuciju jedne promenljive tako što razbija podatke na binove. Ovaj grafički prikaz omogućava prepoznavanje oblasti sa ekstremnim vrednostima ili retkim podacima, što može ukazivati na prisustvo outliera. Histogram je koristan za analizu ukupne raspodele jedne promenljive i za identifikaciju tačaka koje se nalaze van tipičnih vrednosti.

Korišćenjem ovih tehnika vizuelizacije, možemo intuitivno prepoznati outliere, što pomaže u preciznijem procesu čišćenja podataka. Odabirom odgovarajuće metode, možemo efikasno identifikovati i tretirati odstupanja, čime se poboljšava tačnost i pouzdanost analiza i modela.



## Z-Score

Z-score (ili standardizovani skor) je mera koja pokazuje koliko je standardnih devijacija neki podatak udaljen od proseka skupa podataka. Ovaj metod je posebno koristan za detekciju outliera kada podaci prate normalnu distribuciju.

Z-score se računa pomoću jednostavne formule: oduzme se srednja vrednost skupa podataka od posmatrane vrednosti, a zatim se taj rezultat podeli sa standardnom devijacijom skupa podataka.

gde je x vrednost podatka, μ prosečna vrednost, a σ standardna devijacija. Rezultat pokazuje koliko je određeni podatak udaljen od proseka izraženo u standardnim devijacijama. Obično se podaci sa z-score-om većim od 3 (ili manjim od -3) smatraju odstupanjima.

Z-score je posebno koristan kada se radi sa podacima koji su približno normalno distribuirani, jer omogućava lako prepoznavanje ekstremnih vrednosti. Postavljanjem praga, poput z=3, možemo klasifikovati podatke koji leže izvan ovog raspona kao outliere. Ovaj metod je jednostavan za implementaciju i tumačenje, a pruža efikasno rešenje za identifikaciju outliera u standardnim distribucijama.

Međutim, Z-score ima svoja ograničenja. Kao parametarski metod, njegova efikasnost zavisi od toga da li podaci slede normalnu distribuciju. U slučajevima kada distribucija nije normalna, Z-score može biti manje efikasan, pa je potrebno primeniti druge metode ili transformacije podataka kako bi se detektovali outlieri.

## Interquartile Range (IQR)

Interquartile Range (IQR) je mera raspodele podataka koja se koristi za identifikaciju outliera. IQR predstavlja razliku između prvog kvartila (Q1) i trećeg kvartila (Q3) u skupu podataka. Q1 označava 25. percentil, što znači da je 25% podataka ispod ove vrednosti, dok Q3 označava 75. percentil, pri čemu se 75% podataka nalazi ispod te vrednosti. IQR se računa kao razlika između Q3 i Q1, odnosno:

Outlieri se identifikuju kao tačke koje leže izvan granica definisanih na sledeći način: svaki podatak manji od ili veći od se smatra potencijalnim outlierom. Ovaj pristup je robustan prema ekstremnim vrednostima i efikasan za podatke koji ne prate normalnu distribuciju, kao i za podatke sa asimetričnim raspodelama.

IQR je posebno koristan kada podaci nisu normalno distribuirani ili kada postoje odstupanja u distribuciji, jer pruža pouzdan način za detekciju outliera bez prevelikog uticaja ekstremnih vrednosti na samu metodu.

## Percentile

Percentili predstavljaju relativnu poziciju podatka unutar raspodele, što ih čini korisnim za detekciju outliera u podacima sa širokim rasponom vrednosti. Umesto oslanjanja na standardizovane metode poput Z-skor metode ili IQR, percentile omogućavaju fleksibilnost prilikom definisanja pragova za outliere na osnovu specifičnih potreba skupa podataka. Na primer, vrednosti ispod 1. percentila ili iznad 99. percentila često se smatraju outlierima, jer predstavljaju ekstremne vrednosti koje odstupaju od većine podataka.

Korišćenjem ekstremnih percentila, možemo preciznije detektovati outliere u slučajevima kada su podaci široko distribuirani. Ovo je posebno korisno kada želimo uhvatiti specifičan procenat ekstremnih vrednosti u raspodeli, bez nepotrebnog eliminisanja velikog broja tačaka koje ne moraju biti pravi outlieri. Ovaj pristup je prilagodljiv i može se primeniti u različitim domenima, u zavisnosti od toga koliko strogo želimo definisati pragove za outliere.

## Dbscan

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) je metoda klasterizacije zasnovana na gustini, koja se koristi za detekciju outliera i grupisanje podataka u nepodeljenim raspodelama, naročito u višedimenzionalnim prostorima. Za razliku od drugih tehnika, DBSCAN definiše klastere na osnovu lokalne gustine podataka, gde su tačke unutar klastera gustinski povezane, dok outlieri nemaju dovoljnu gustinu i ostaju izolovani. Algoritam identifikuje osnovne tačke (core points), granične tačke (border points), i outliere, u zavisnosti od broja tačaka unutar zadatog radijusa (ɛ) i minimalnog broja tačaka (MinPts).

DBSCAN formira klastere tako što svaka osnovna tačka generiše klaster sa svim tačkama koje su povezane putem gustine. Tačke se povezuju ako postoji put od osnovnih tačaka, dok granične tačke ne mogu generisati nove klastere, ali su deo postojećih klastera. Tačke koje nisu povezane sa nijednim klasterom su outlieri i dodeljuju se posebnoj klasi (-1), što ih čini lako prepoznatljivim.

Prednosti DBSCAN metode uključuju njenu sposobnost da identifikuje klastere nepoznatog oblika i veličine, kao i da automatski detektuje outliere, što je korisno kod višedimenzionalnih skupova podataka. Algoritam je intuitivan za korišćenje i pruža mogućnost analize bez potrebe za definisanjem unapred broja klastera. Međutim, izazovi ove metode uključuju osetljivost na parametre, kao što su eps i MinPts, koji moraju biti pažljivo kalibrisani kako bi se postigli optimalni rezultati. Takođe, podaci moraju biti skalirani pre primene algoritma, a kalibracija mora biti ponovljena za svaki novi set podataka.

DBSCAN je efikasan alat za detekciju outliera u situacijama kada klasične metode zasnovane na pretpostavkama o raspodeli podataka nisu odgovarajuće, naročito u kompleksnim višedimenzionalnim prostorima.

## Isolation Forest

Za razliku od klasičnih pristupa, Isolation Forest funkcioniše na principu da su outlieri retki i udaljeni od ostatka podataka, što ih čini lakšima za izolaciju. Ova metoda koristi ansambl binarnih stabala odlučivanja, pri čemu se svako stablo gradi slučajnim odabirom promenljive i slučajnom vrednošću preseka između maksimalnih i minimalnih vrednosti promenljive. Proces se ponavlja za sve podatke u skupu za obuku, a rezultati različitih stabala se kombinuju kako bi se kreirala šuma.

Tokom predikcije, algoritam računa "dužinu puta" za svaku instancu – broj preseka potrebnih da se tačka izoluje u stablu. Očekivano, outlieri će imati kraće putanje jer se nalaze dalje od gusto naseljenih regiona podataka. Na osnovu dužine puta, algoritam izračunava "outlier skor" za svaku tačku, pri čemu skor bliži 1 ukazuje na veći stepen odstupanja, dok skor bliži 0 označava normalnost.

Prednosti Isolation Forest-a uključuju njegovu sposobnost da detektuje outliere bez potrebe za skaliranjem podataka i robustnost metode sa minimalnim brojem parametara. Ova tehnika je efikasna u slučajevima kada distribucija vrednosti nije unapred poznata.

Međutim, Isolation Forest može biti komplikovan za vizualizaciju, a ako nije pravilno optimizovan, može postati računski zahtevan. Uprkos tome, predstavlja moćan alat za otkrivanje outliera u raznovrsnim aplikacijama.

## Tehnike rukovanja outlier-ima

Postoji nekoliko strategija za rukovanje outlier-ima u zavisnosti od prirode podataka i uticaja koji ti outlieri imaju na analizu ili performanse modela. Ove tehnike mogu pomoći da se smanji negativan uticaj outliera, dok se zadržava što više informacija iz podataka. U nastavku su najčešće korišćene tehnike:

1. **Uklanjanje outliera (Trimming)**  
   Uklanjanje outliera je direktan pristup koji se koristi kada su odstupanja identifikovana kao šum ili greška. Ovo može poboljšati preciznost modela, ali može dovesti i do gubitka značajnih informacija. Triming se vrši tako što se uklanjaju samo ekstremne vrednosti iz gornjih i donjih percentila.
2. **Kraćenje (Capping)**  
   Kraćenje podrazumeva primenu granica na podatke, gde se vrednosti iznad ili ispod određenih pragova zamenjuju najvišim ili najnižim dozvoljenim vrednostima. Na ovaj način, outlieri se zadržavaju u skupu podataka, ali njihov uticaj na analizu ili modeliranje se smanjuje. Ova tehnika je korisna kada se želi očuvati struktura podataka, ali bez značajnog uticaja ekstremnih vrednosti.
3. **Transformacija**  
   Transformacije podataka kao što su logaritamska, kvadratna ili Box-Cox transformacija mogu pomoći u smanjenju uticaja outliera prilagođavanjem skale podataka. Transformacija pomaže da se podaci prilagode modelima, ali može promeniti njihovu interpretaciju.
4. **Imputacija**  
   Imputacija podrazumeva zamenu outliera sa odgovarajućim vrednostima, kao što su medijana, prosek ili mod. Ako se outlieri smatraju greškama, ovaj pristup može biti efikasan u vraćanju stabilnosti skupa podataka. Takođe se mogu koristiti napredne metode kao što su imputacija na osnovu najbližih suseda (K-nearest neighbors) ili regresiona imputacija.
5. **Vinzorizacija**  
   Vinzorizacija je tehnika koja uključuje postavljanje ekstremnih vrednosti na određene pragove, slično capping-u, ali zadržava outlieri u obliku modifikovanih vrednosti. Na primer, vrednosti iznad 95. percentila mogu biti postavljene na vrednost 95. percentila, a vrednosti ispod 5. percentila na vrednost 5. percentila. Ova metoda smanjuje uticaj outliera bez njihovog potpunog uklanjanja.
6. **Korišćenje robusnih modela**  
   Određeni modeli su manje osetljivi na outliere, kao što su random forest ili modeli koji koriste median-based loss funkcije. Korišćenje robusnih modela može prirodno ublažiti uticaj outliera bez potrebe za njihovim eksplicitnim uklanjanjem ili transformacijom podataka.
7. **Podela podataka (Partitioning)**  
   Podela podataka na setove sa i bez outliera omogućava zasebnu analizu ili modeliranje. Ova tehnika omogućava upoređivanje rezultata i uvida u to kako prisustvo outliera utiče na performanse modela.

Razumevanje prirode outliera i njihovog uticaja na podatke ključno je za izbor odgovarajuće metode, jer pogrešna obrada može dovesti do gubitka informacija ili iskrivljenih rezultata.

# Konstrukcija atributa (Feature creation)

Kreiranje karakteristika je ključni proces u mašinskom učenju, kojim se generišu nove karakteristike na osnovu postojećih podataka ili domenskog znanja. Ovaj korak je suštinski deo inženjeringa karakteristika, jer poboljšava sposobnost modela da prepozna obrasce i odnose u podacima, čime se značajno unapređuju performanse modela.

Proces kreiranja novih karakteristika podrazumeva razumevanje kako se različite varijable međusobno odnose, kao i uvođenje novih podataka koji dodatno opisuju posmatrani problem. Kreiranje karakteristika je korisno i za složenije modele, poput neuronskih mreža, ali i za jednostavnije pristupe kao što su linearna regresija.

## Tipovi kreiranja karakteristika

* **Specifične za domen**: Kreiranje karakteristika zasnovano na znanju o određenoj oblasti je možda najmoćniji način da se poveća prediktivna moć modela. Na primer, u finansijskim podacima, kreiranje varijable koja opisuje odnos prihoda i rashoda može dati važne uvide o solventnosti subjekta. Ovakve karakteristike proizlaze iz iskustva i specifičnog znanja o domenima kao što su poslovna pravila ili industrijski standardi.
* **Zasnovane na podacima**: Ovaj pristup uključuje kreiranje karakteristika na osnovu uočenih obrazaca u podacima. Na primer, agregacije (npr. prosečne vrednosti, maksimumi, minimumi) mogu obuhvatiti statistički značajne informacije o podacima. Takođe, kreiranje interakcijskih karakteristika, koje kombinuju dve ili više promenljivih, može omogućiti modelu da bolje prepozna odnose između tih promenljivih.
* **Sintetičke**: Sintetičke karakteristike su one koje se generišu kombinovanjem postojećih karakteristika. Na primer, spajanjem više sličnih karakteristika u jedinstvenu sintetičku karakteristiku možemo sažeti informacije u jednostavniji oblik, što pomaže u smanjenju složenosti modela.

## Interakcione karakteristike

Interakcione karakteristike se kreiraju kombinovanjem dve ili više postojećih karakteristika kako bi se obuhvatile međusobne interakcije između njih. Na primer, interakcija između varijabli kao što su godine i prihod može se ispitati kreiranjem nove karakteristike koja modeluje kako se efekat prihoda menja u zavisnosti od starosti. Kombinovanjem ovih promenljivih model može uočiti složenije odnose koje pojedinačne karakteristike ne bi mogle samostalno predstaviti.

Interakcione karakteristike su posebno korisne kod modela koji ne mogu lako da uoče nelinearne odnose između varijabli. Ovo je često slučaj kod linearnih modela. Kombinovanjem dve karakteristike množenjem ili nekim složenijim operacijama može se omogućiti modelu da bolje generalizuje podatke.

## Polinomske karakteristike

Polinomske karakteristike su proširenje interakcionih karakteristika, gde se vrednosti karakteristika podižu na različite stepene (npr. kvadrati, kubovi). Ovo je naročito korisno za modele kao što su linearna regresija, koji ne mogu direktno uhvatiti nelinearne odnose. Na primer, kreiranje karakteristika kao što su x2 ili x3 može modelu pomoći da prepozna zakrivljenost u podacima.

Polinomske karakteristike često donose značajna poboljšanja u modeliranju složenih relacija, ali se mora biti oprezan da se ne pretera sa brojem polinoma, kako bi se izbeglo pretreniranje modela na trening podacima. Treba odabrati odgovarajući stepen polinoma koji omogućava optimalnu generalizaciju.

## Vremenske karakteristike

Vremenske karakteristike se kreiraju iz podataka koji uključuju datume ili vremenske oznake. Na primer, moguće je izdvojiti informacije kao što su dan u nedelji, mesec ili godina. Ove karakteristike su od izuzetne važnosti u modelima koji analiziraju vremenske serije, kao što su predikcija prodaje, analiza tržišnih trendova ili vremenski rasporedi.

Sezonski obrasci, kao što su promene prodaje tokom vikenda ili praznika, mogu biti ključni za pravljenje boljih prognoza. Ove informacije se mogu kodirati kao nove karakteristike koje pomažu modelu da uhvati sezonalnost ili trendove kroz vreme, što značajno poboljšava performanse prediktivnih modela.

# Zaključak

Transformacija podataka predstavlja ključni korak u procesu pripreme podataka za mašinsko učenje. Kroz različite tehnike skaliranja, enkodiranja, diskretizacije i rukovanja odstupanjima, unapređuje se tačnost, robusnost i interpretabilnost modela. Primena odgovarajućih metoda obrade omogućava modelima da pravilno generalizuju podatke, čime se značajno povećava kvalitet predikcija i analiza.

Pravilna primena transformacija, kao što su standardizacija i normalizacija, osigurava ravnotežu između različitih atributa, dok enkodiranje kategoričkih podataka omogućava uspešnu integraciju ne-numeričkih vrednosti u modele mašinskog učenja. Takođe, tehnike za detekciju i obradu outliera, poput Z-skora, interkvartilnog raspona (IQR) i klasterizacije, omogućavaju efikasno prepoznavanje i minimiziranje negativnog uticaja ekstremnih vrednosti.

Poseban deo rada posvećen je kreiranju novih karakteristika (feature creation), koje mogu značajno poboljšati performanse modela. Interakcione, polinomske i vremenske karakteristike omogućavaju modelima da uhvate složene odnose u podacima, čineći ih fleksibilnijim i prilagodljivijim različitim tipovima zadataka.

Celokupan proces transformacije podataka zahteva pažljiv izbor metoda, uzimajući u obzir prirodu podataka i ciljeve analize. Ovaj rad pruža sveobuhvatan pregled najvažnijih tehnika, ukazujući na njihovu primenu i prednosti, uz naglasak na praktične implikacije u mašinskom učenju. Korišćenje ovih tehnika doprinosi kreiranju pouzdanijih, robusnijih i interpretabilnijih modela, čime se unapređuje kvalitet mašinskog učenja i analize podataka.

# Reference

Bala, P. C. (2022, July 5). *How to detect outliers in machine learning – 4 methods for outlier detection*. Freecodecamp.org. https://www.freecodecamp.org/news/how-to-detect-outliers-in-machine-learning/

Brown, I. (2024, February 8). *Handling outliers in ML: Best practices for robust data preprocessing*. Linkedin.com. https://www.linkedin.com/pulse/handling-outliers-ml-best-practices-robust-data-iain-brown-ph-d--mwf6e/

Coşgun, H. H. (2023, August 5). *Which data scaling technique should I use ?* Medium. https://medium.com/@hhuseyincosgun/which-data-scaling-technique-should-i-use-a1615292061e

Data Cleaning. (2023, March 9). *What are the pros and cons of different scaling methods for data normalization?* Linkedin.com; www.linkedin.com. https://www.linkedin.com/advice/1/what-pros-cons-different-scaling-methods-data-normalization

de la Calle, J. E. (2023, May 9). *Best tips and tricks: When and why to use logarithmic transformations in statistical analysis*. Medium. https://juandelacalle.medium.com/best-tips-and-tricks-when-and-why-to-use-logarithmic-transformations-in-statistical-analysis-9f1d72e83cfc

*Feature engineering A-Z*. (n.d.). Feature Engineering A-Z. Retrieved September 2, 2024, from https://feaz-book.com/numeric-maxabs

Galli, S. (2022, July 4). *Data discretization in machine learning*. Train in Data’s Blog; Train in Data. https://www.blog.trainindata.com/data-discretization-in-machine-learning/

Htoon, K. S. (2020, February 29). *Log transformation: Purpose and interpretation - Kyaw saw htoon*. Medium. https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9

Jarapala, K. N. (2023, March 13). *Categorical data Encoding techniques - AI skunks - medium*. AI Skunks. https://medium.com/aiskunks/categorical-data-encoding-techniques-d6296697a40f

Kiptoon, D. (2023, August 18). *Understanding feature engineering in machine learning*. Medium. https://medium.com/@jdkiptoon/understanding-feature-engineering-in-machine-learning-59fc343a29c9

Mahmood, H. (2024, July 23). Categorical data encoding: 7 effective techniques. *Data Science Dojo*. https://datasciencedojo.com/blog/categorical-data-encoding/

nikhilbhoi9739 Follow Improve. (2023, March 20). *What is feature engineering?* GeeksforGeeks. https://www.geeksforgeeks.org/what-is-feature-engineering/

*Numerical data: Normalization*. (n.d.). Google for Developers. Retrieved September 4, 2024, from https://developers.google.com/machine-learning/crash-course/numerical-data/normalization

Omardonia. (2023, March 9). *Data scaling and normalization: A guide for data scientists*. Generative AI. https://generativeai.pub/data-scaling-and-normalization-a-guide-for-data-scientists-d6f9fdfa7b2d

Patel, D. (2022, June 7). *Data discretization - CodeX - medium*. CodeX. https://medium.com/codex/data-discretization-b5faa2b77f06

Plummer, A. (2022, September 16). *Box-Cox transformation and target variable: A guide*. Built In. https://builtin.com/data-science/box-cox-transformation-target-variable

Santoyo, S. (2017, September 12). *A brief overview of outlier detection techniques*. Towards Data Science. https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561

Scaling and normalization: Preparing data for analysis. (2024, January 7). *Dataheadhunters.com*. https://dataheadhunters.com/academy/scaling-and-normalization-preparing-data-for-analysis/

Sidhikha, A. (2024, January 13). *Outliers detection*. Medium. https://medium.com/@ayeshasidhikha188/outliers-detection-9b39ede4eb20

Singh, H. (2024, May 26). *Polynomial regression: Exploring non-linear relationships*. DEV Community. https://dev.to/harsimranjit\_singh\_0133dc/polynomial-regression-exploring-non-linear-relationships-49nk

*StandardScaler, MinMaxScaler and RobustScaler techniques - ML*. (2020, July 15). GeeksforGeeks. https://www.geeksforgeeks.org/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/

Syam, S. S. (2024, March 10). *Understanding and handling outliers in data analysis*. Medium. https://medium.com/@heysan/understanding-and-handling-outliers-in-data-analysis-727a768650fe

Syed, A. H. (2023, April 20). *Dealing with outliers in data science: Techniques and best practices*. Medium. https://syedabis98.medium.com/dealing-with-outliers-in-data-science-techniques-and-best-practices-a08172643b7a

Taylor, S. (n.d.). *Skewness*. Corporate Finance Institute. Retrieved September 14, 2024, from https://corporatefinanceinstitute.com/resources/data-science/skewness/

*What are the advantages and disadvantages of equal-width and equal-frequency binning methods?* (n.d.). Linkedin.com. Retrieved September 9, 2024, from https://www.linkedin.com/advice/1/what-advantages-disadvantages-equal-width

(N.d.-b). Kantschants.com. Retrieved September 14, 2024, from https://kantschants.com/complete-guide-to-encoding-categorical-features#heading-advantages