Универзитет у Нишу
Електронски факултет

Предмет: Прикупљање и предобрада података за машинско учење

Квалитет података

Семинарски рад

Студент:

Радомир Стајковић

Ментор:

Доц. др Александар Станимировић

Ниш, 2024. године

Садржај

1. Увод у квалитет података

Квалитет података постао је кључно питање за јавне институције, приватне компаније и ширу јавност, јер се све више ослањају на доношење одлука заснованих на подацима. Технолошки напредак довео је до развоја софтверских и хардверских система који генеришу огромне количине података, који служе као основа за модерне софтверске апликације у различитим доменима. Ово ослањање на податке доноси и могућности и изазове, посебно за организације које пружају информације ради подршке доношењу одлука у јавном и приватном сектору.

Појава нових извора података, иновативне употребе постојећих података, напредне методе анализе и интеграција података из више извора нуде потенцијал за свеобухватније увиде. Међутим, квалитет ових података мора бити пажљиво евалуиран како би се избегли погрешни закључци и лоше одлуке. Подаци ниског квалитета могу довести до искривљених резултата, док подаци високог квалитета побољшавају доношење одлука, помажу у идентификацији образаца унутар скупова података и повећавају укупну поузданост и кредибилитет њихове употребе.

Јасно документовано разумевање снага и слабости података осигурава да се слабости решавају, омогућава одговарајућу употребу података и јача поверење у њихову примену. Сви подаци показују снаге и слабости кроз више мера квалитета, што често укључује компромис између свеобухватности, тачности и благовремености. Обезбеђивање прикладности за употребу у различитим контекстима—било за развој нових података, спровођење анализа или подршку доношењу одлука—захтева темељно разумевање ових компромиса.

Процес обезбеђивања високог квалитета података почиње прикупљањем и предобрадом података, који укључују кораке као што су агрегација, валидација, анализа дистрибуција и процена односа између појединачних података. Ова фаза предобраде је кључна за побољшање квалитета података јер помаже у идентификацији и решавању аномалија и грешака које могу угрозити скуп података. Само пажљивом провером и предобрадом подаци се могу сматрати погодним за даљу анализу.

Разумевање мера квалитета података и имплементација робусних процеса за обезбеђивање квалитета података су стога кључни кораци за коришћење података у циљу поузданог, тачног и увида богатог доношења одлука. Овај рад ће даље истражити ове мере и размотрити методе за унапређење квалитета података како би се максимизирала вредност и употребљивост података у различитим доменима.

2. Мере квалитета података

Квалитет података односи се на различите елементе или компоненте података, као што су променљиве и поља података, као и на целокупан скуп података. Овај концепт примењује се на постојеће методе као што су традиционални дизајн анкета, али и на нове примене већ успостављених и нових метода које се користе за креирање скупова података, као што су вештачка интелигенција и машинско учење. Дефиниција квалитета података односи се на податке произведене из различитих врста извора, укључујући податке прикупљене у ненаменске статистичке сврхе, као што су подаци из административних записа и сензора.

Квалитет података је такође релевантан за интегрисане производе података (нпр. повезани подаци, моделирани подаци), где интегрисани подаци могу да садрже статистичке податке, нестатистичке податке или комбинацију различитих извора података. Иако се термин "интегрисани подаци" често користи, ова врста података може се описати и другим терминима, укључујући комбиноване податке, податке из више извора и повезане податке (када је то применљиво).

Мере квалитета података, као што су тачност, потпуност, конзистентност, ажурност, валидност, релевантност, приступачност и јасноћа, примењују се на све ове различите типове и изворе података. Ове мере служе као основе за процену колико су подаци погодни за употребу у различитим контекстима, било за анализу, моделирање или доношење одлука.

Мере квалитета у контексту различитих типова података:

1. **Традиционални извори података:** За традиционалне изворе података, као што су анкете и административни записи, мере квалитета се често фокусирају на тачност и потпуност, осигуравајући да подаци тачно одражавају оно што се мери и да не недостају критичне информације. Конзистентност је такође важна, јер подаци морају бити усаглашени између различитих истраживања и извора.
2. **Нови извори и методе:** Код података добијених помоћу вештачке интелигенције и машинског учења, мере квалитета се морају проширити да би обухватиле тачност алгоритма, валидност резултата и способност модела да прецизно генерализује на непознате податке. Овде је посебно важна транспарентност у методологији и јасноћа у погледу резултата.
3. **Интегрисани и комбиновани подаци:** Код интегрисаних података, мере квалитета као што су конзистентност и ажурност играју кључну улогу, јер подаци из различитих извора морају бити усклађени и ажурирани како би осигурали њихову тачност и релевантност. Поред тога, релевантност података постаје критична, јер различити извори података могу садржати различите нивое детаља и корисности.
4. **Подаци из сензора и IoT уређаја:** За податке прикупљене сензорима и IoT уређајима, мере као што су тачност и ажурност су од суштинског значаја. Будући да ови подаци могу бити подложни шуму или неочекиваним променама у окружењу, потребно је осигурати њихову валидност и доследност кроз процедуре филтрирања и валидације.

Мере квалитета података могу се поделити у две категорије: објективне и субјективне метрике. Објективне мере се односе на мере које се могу јасно дефинисати и мерити на основу доступних података, док се субјективне мере односе на перцепцију корисника о квалитету података и могу се оценити кроз анкете и корисничка искуства. Испод су детаљно објашњене најчешће мере у обе категорије.

2.1 Објективне мере

Објективне мере квалитета података односе се на оне аспекте података који се могу прецизно измерити и који су независни од корисникове перцепције. Ове мере су кључне за обезбеђивање поузданости и тачности података, као и за њихову применљивост у различитим контекстима анализе

2.1.1 Тачност

Тачност података је кључни аспект у области машинског учења, где се мери колико подаци представљају стварне објекте и догађаје које описују. На пример, у системима машинског учења који користе ГПС податке за навигацију, тачност података одређује да ли ће модел правилно упутити корисника до жељене дестинације или ће га одвести на погрешно место.

Подаци коришћени за обуку модела машинског учења морају бити прецизни да би модел могао да учи тачно. На пример, ако се модел обучава на подацима о сликама са ознакама, било какве грешке у ознакама могу довести до погрешних закључака и смањити тачност модела. Актуелност и релевантност податка је веома важна за проблем који се решава. Старе или застареле информације могу утицати на способност модела да правилно предвиди тренутне трендове или ситуације.

Тачност података се разликује од комплетности података, која се односи на обухват и свеобухват података. У контексту машинсог учења, тачност података није исто што и квалитет података, иако се оба преклапају. Тачност осигурава да информације буду без грешака, док квалитет укључује и оцену корисности и вредности података за изградњу и обуку модела.

2.1.2 Комплетност

Комплетност података односи се на степен у којем подаци покривају све потребне информације за одређени контекст или сврху. Она је један од кључних аспеката квалитета података и одређује да ли су сви релевантни подаци доступни и укључени у анализу. Ако неке информације недостају, подаци се сматрају непотпуним. На пример, ако база података садржи само делимичне податке о продаји, без података о враћању производа, анализа може бити нетачна.

Комплетност података је важна јер недостајући или непотпуни подаци могу довести до погрешних закључака и одлука. У многим областима као што су истраживање, анализа пословних перформанси и управљање пројектима, осигурање потпуног обухвата података је кључно за тачне и ефикасне резултате.

2.1.3 Интегритет

Интегритет података подразумева осигурање да су подаци тачни, комплетни и доследни у свим фазама њиховог животног циклуса. Очување интегритета података укључује заштиту података од губитака, пропуштања и коруптивних утицаја, што је кључно за поуздане и ефикасне резултате у машинском учењу.

Чисти и поуздани подаци су неопходни за успешно обуку и процену модела машинског учења. Како обим података расте и како се подаци користе за доношење важних одлука, као што су предвиђања и класификације, максимизовање интегритета података постаје све важније.

Да би се осигурао интегритет података, примењују се различите методе укључујући проверу грешака, процедуре валидације и мере безбедности као што су енкрипција, контрола приступа и прављење резервних копија. Циљ ових мера је да обезбеде да аналитика података буде заснована на поузданим информацијама и да се осетљиве информације заштите од неовлашћеног приступа или злоупотребе.

Интегритет података није ограничен на један алат или платформу; уместо тога, то је свеобухватан приступ који укључује примену техника и политика за очување квалитета података током целокупног процеса машинског учења. Овај приступ захтева координисану употребу технолошке инфраструктуре, политика и процедура како би подаци остали поуздани и корисни за изградњу и тестирање модела

2.1.4 Конзистентност

Конзистентност података односи се на стање у којем су све копије или инстанце података идентичне у свим системима и базама података. Конзистентност помаже у осигурању да подаци буду тачни, ажурни и коерентни широм различитих база података, апликација и платформи. Она игра кључну улогу у обезбеђивању да корисници података могу веровати информацијама до којима имају приступ.

Постоји неколико начина да се осигура конзистентност података, укључујући:

- **Примену правила валидације података:** Ова правила помажу у идентификовању и исправљању неправилности и грешака у подацима пре него што буду унети у систем.
- **Коришћење техника стандардизације података:** Стандардизација осигурава да подаци буду представљени у истом формату или структури, чиме се олакшава њихова употреба и упоредивост.
- **Примена процеса синхронизације података:** Ови процеси осигуравају да се све верзије података у различитим системима ажурирају и синхронизују.

Конзистентност података је есенцијална из више разлога. Она помаже у осигуравању да корисници имају приступ тачним и актуелним информацијама, што им омогућава доношење информисаних одлука. Додатно, конзистентни подаци помажу предузећима да поједноставе своје операције, смање грешке и побољшају укупну ефикасност.

Док и конзистентност података и интегритет података имају за циљ одржавање тачних, поузданих и висококвалитетних података, они се разликују у свом основном фокусу:

- **Конзистентност података** се пре свега бави осигурањем да подаци остану униформни у свим системима и током целокупног животног циклуса. То значи да све копије или инстанце података морају бити идентичне, чиме се осигурава да корисници добијају исте информације без обзира на то где или како приступају подацима.
- **Интегритет података** се фокусира на очување тачности, поузданости и безгрешне природе података док се уносе, складиште и преузимају. То укључује заштиту података од грешака, корупције и губитака како би се осигурало да информације остану веродостојне и без икаквих недостатака.

2.1.5 Приступачност

Иако су подаци доступни, то не значи да су увек употребљиви. Овде је важна приступачност података, која се односи на степен до којег су подаци лако доступни и употребљиви. Приступачност података подразумева процес који чини доступне податке погодним за коришћење, без обзира на ниво искуства или стручности корисника. Међутим, с обзиром на то да подаци постоје у различитим форматима и типовима, изазов је учинити их потпуно приступачним, јер у свом тренутном стању можда неће бити од велике користи.

Да би подаци били корисни, потребно их је очистити, преформатирати и стандардизовати пре него што се интегришу са другим доступним подацима и учине спремним за употребу. Како се количина података континуирано повећава, приступачност података је сталан процес који је потребно спроводити како би се подаци наставили користити за доношење одлука заснованих на информацијама.

2.1.6 Кохерентност

Кохерентност података подразумева униформност и усклађеност података унутар једног скупа података, као и између различитих скупова података. Она укључује логичке везе и комплетност података, осигуравајући да подаци буду доследни и смислени у различитим контекстима. Кохерентност такође подразумева да структуре и формати података буду компатибилни за различите сврхе и примене, омогућавајући поуздану интеграцију и анализу података. Одржавање кохерентности података кључно је за очување квалитета података и њихову корист за доношење одлука.

2.2 Субјективне мере

Субјективне мере квалитета података нису универзално мерљиве и зависе од контекста коришћења, потреба и искустава корисника. Ове мере укључују перцепцију корисника о корисности, поузданости и разумљивости података и могу значајно утицати на њихову употребу. Иако се могу чинити мање „научним“, субјективне мере су кључне јер указују на то како подаци функционишу у стварном свету и како се доживљавају од стране крајњих корисника.

2.2.1 Уверљивост

Уверљивост података директно утиче на њихову употребу. Ако корисници немају поверења у извор података или у методе које су коришћене за њихово прикупљање и обраду, они ће се вероватно окренути алтернативним изворима. Због тога је неопходно не само обезбедити објективну тачност података, већ и комуницирати транспарентност у процесу њиховог стварања.

2.2.2 Употребљивост

Подаци који су тешко доступни или сложени за коришћење могу значајно умањити њихову вредност, чак и ако су објективно квалитетни. Корисници очекују да могу лако приступити подацима и користити их за доношење одлука или анализу. Ово подразумева лакоћу приступа, интуитивно структурирање, као и подршку у виду документације и алата који олакшавају коришћење података.

2.2.3 Објективност

Чак и када су подаци технички тачни и комплетни, њихова перципирана пристрасност може утицати на начин на који се користе. Корисници морају имати осећај да су подаци прикупљени и представљени на непристрасан начин, без скривених мотива или пристрасности. Ово може бити посебно важно у контексту медија, политике и научних истраживања.

3. Статистичке мере за квалитет података

Статистичке мере за квалитет података су квантитативни алати који се користе за процену и анализу квалитета података у датасетовима. Кроз различите статистичке анализе, могуће је идентификовати обрасце, одступања и потенцијалне проблеме у подацима, што помаже у бољем разумевању података и доношењу одлука о потребним корацима за њихово побољшање.

3.1 Расподела података

Значај расподеле података у науци о подацима и статистици:

- Расподела података помаже у разумевању основних образаца и понашања у подацима. Она пружа увид у то како су подаци распоређени по различитим вредностима, што је кључно за доношење одлука заснованих на подацима.
- Разумевање расподеле података је основа за спровођење статистичких тестова, избор модела машинског учења и визуелизацију података.

Преглед кључних концепата везаних за расподелу података:

- Дефиниције кључних појмова као што су "расподела," "функција густине вероватноће," "кумулативна функција расподеле" и "случајна променљива."
- Објашњење значаја разумевања различитих типова расподела у контексту инференцијалне статистике и предиктивног моделирања.

3.1.1 Типови расподеле података

- Нормална расподела
- Поасонова расподела
- Биномна расподела
- Експоненцијална расподела
- Униформна расподела

Нормална расподела

Нормална расподела података, такође позната као Гаусова расподела или расподела у облику звона, једна је од најчешће коришћених расподела у статистици и вероватноћи. Њена употреба је распрострањена у многим областима, укључујући науку, инжењеринг, економију, друштвене науке и машинско учење. Основни разлог за њену популарност је што многе природне и социјалне појаве, када се анализирају у великом обиму, показују тенденцију да прате нормалну расподелу.

Карактеристике нормалне расподеле:

1. **Облик звона:** График нормалне расподеле има облик звона који је симетричан око средње вредности (μ). Ова симетрија значи да су вредности једнако распоређене око средње вредности, са једнаком вероватноћом за веће и мање вредности.

2. **Средња вредност (μ), медијана и мод:** За нормалну расподелу, средња вредност, медијана и мод су једнаки и налазе се на центру расподеле.

3. **Стандардна девијација (σ):** Стандардна девијација је мера распршености података око средње вредности. Већа стандардна девијација значи да су подаци више распршени, док мања стандардна девијација значи да су подаци ближе средњој вредности.

4. **Правило 68-95-99.7:** Ово правило, познато и као емпиријско правило, указује на то да:
   - Око 68% података пада унутар једне стандардне девијације (μ ± σ) од средње вредности.
   - Око 95% података пада унутар две стандардне девијације (μ ± 2σ).
   - Око 99.7% података пада унутар три стандардне девијације (μ ± 3σ).

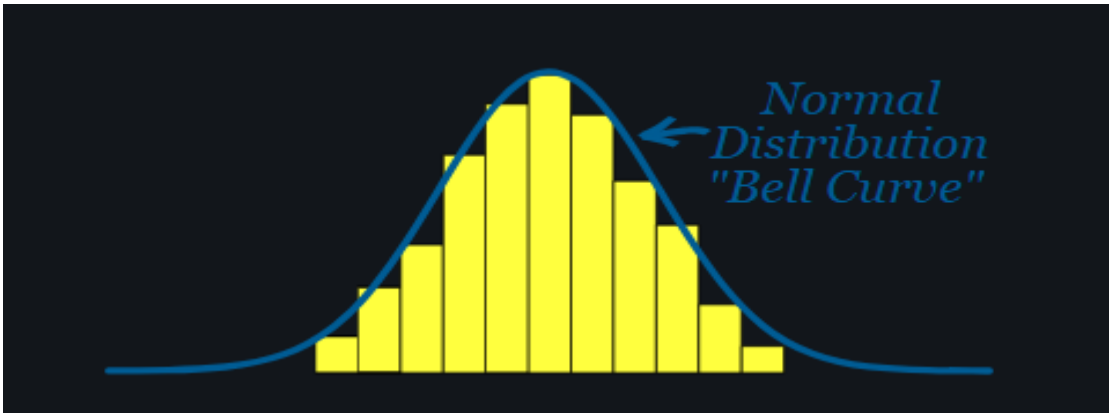
5. **Крива бесконачно иде ка нули:** Нормална расподела је асимптотска, што значи да се крива никада не додирује осу x, већ се бесконачно приближава нули како се удаљавамо од средње вредности.

Функција густине вероватноће нормалне расподеле за насумичну променљиву X је дата формулом:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

где су:

- x — насумична променљива,
- μ — средња вредност,
- σ — стандардна девијација,
- e — основа природног логаритма (приближно 2.718).



Слика 1. Пример нормалне дистрибуције података

Поасонова расподела

Поасонова расподела је једна од најважнијих дискретних расподела вероватноће у статистици и теорији вероватноће. Често се користи за моделирање броја догађаја који се дешавају у фиксном интервалу времена или простора, под условом да се догађаји дешавају независно један од другог и са константном стопом.

Карактеристике Поасонове расподеле:

1. **Дискретна природа:** За разлику од нормалне расподеле, Поасонова расподела је дискретна, што значи да моделује број појава или догађаја који су цели бројеви (0, 1, 2, 3,...).

2. **Параметар λ (ламбда):** Поасонова расподела је дефинисана једним параметром, λ, који представља просечан број догађаја у датом временском периоду или простору. Параметар λ може бити било који позитиван реалан број.

3. **Независност догађаја:** Догађаји се морају дешавати независно један од другог. На пример, појава једног догађаја не утиче на вероватноћу да ће се догодити други догађај.

4. **Реткост догађаја:** Поасонова расподела је погодна за моделе у којима се догађаји дешавају ретко или су ријетки у односу на укупно време или простор који се разматра.

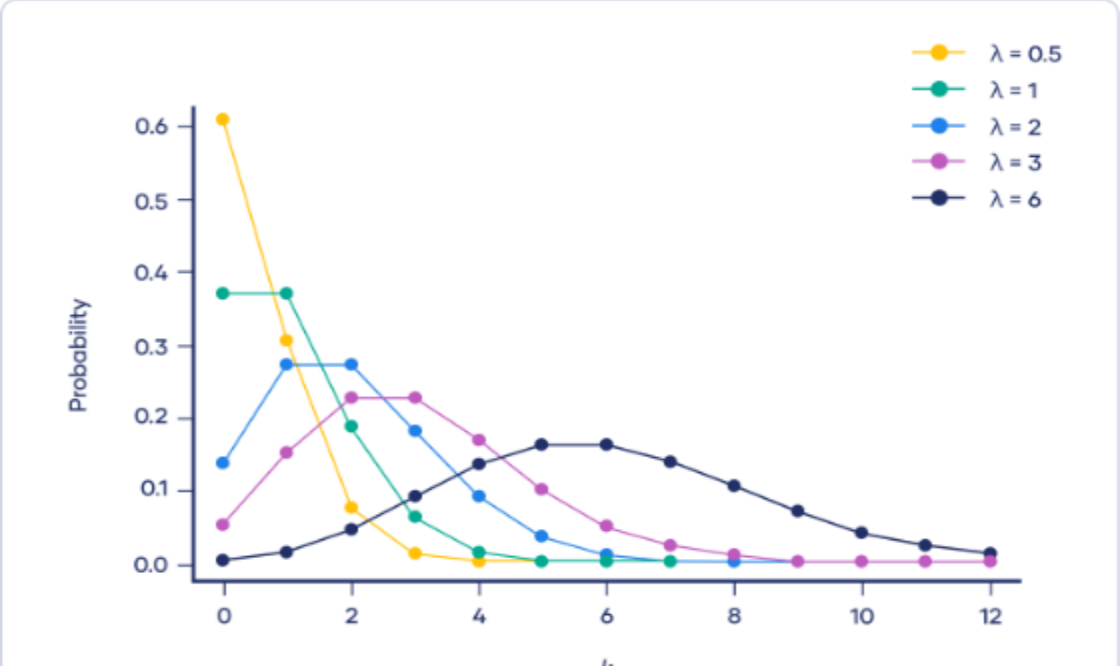
Формула Поасонове расподеле

Функција вероватноће за Поасонову расподелу, која даје вероватноћу да ће се догодити тачно k догађаја у фиксном временском периоду, је дата формулом:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

где су:

- k — број догађаја (цели број: 0, 1, 2, ...),
- X — насумична променљива која следи Поасонову расподелу,
- $P(X = k)$ — је вероватноћа да ће се догађај десити k пута,
- λ — просечан број догађаја у интервалу,
- e — основа природног логаритма (приближно 2.718).



Слика 2. Паосонова расподела са различитим вредностима λ

Биномна расподела

Биномна расподела је дискретна расподела вероватноће која описује број успешних исхода у низу независних и идентично дистрибуираних експеримената, где сваки експеримент има само два могућа исхода: успех или неуспех. Ова дистрибуција је корисна за моделирање ситуација у којима нас занима број успеха у одређеном броју покушаја, као што су број тачних одговора на тесту са питањима са два избора, број глава при бацању новчића, или број добитака у игри на срећу.

Карактеристике биномне дистрибуције

1. **Два могућа исхода:** Сваки експеримент или проба има два могућа исхода — успех (са вероватноћом p) или неуспех (са вероватноћом $1-p$).
2. **Фиксиран број покушаја:** Биномна дистрибуција се користи када постоји фиксиран број покушаја n.
3. **Независност покушаја:** Сваки покушај је независан од претходних, што значи да исход једног покушаја не утиче на исход другог.
4. **Константна вероватноћа успеха:** Вероватноћа успеха p и вероватноћа неуспеха $(1-p)$ остају константне кроз све покушаје.

Формула биномне дистрибуције

Функција вероватноће биномне дистрибуције, која даје вероватноћу да ће се појавити тачно k успеха у n независних покушаја, је дата формулом:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

где су:

- X — насумична променљива која представља број успеха,
- k — број успеха,
- n — укупан број покушаја,
- p — вероватноћа успеха у једном покушају,
- $\binom{n}{k}=\frac{n!}{k!(n-k)!}$ — биномни коефицијент.

Експоненцијална расподела

Експоненцијална расподела је континуална расподела вероватноће која се често користи за моделирање времена између догађаја у процесу који се дешава са константном стопом. Ова расподела је посебно корисна у анализи преживљавања, теорији редова, поузданости система, и другим областима где се разматра време до неког догађаја, као што су време до отказа неког уређаја, време између телефонских позива, или време чекања у реду.

Карактеристике експоненцијалне расподеле

1. **Непрекидност**: Експоненцијална расподела је непрекидна, што значи да је дефинисана за све реалне бројеве веће или једнаке нули.
2. **Параметар λ (ламбда):** Експоненцијална расподела има један параметар, λ>0, који представља стопу догађаја. Параметар λ је инверзно пропорционалан просечном времену између догађаја.
3. **Без меморије (без сећања):** Ово својство значи да вероватноћа да ће догађај да се деси у наредном временском интервалу не зависи од тога колико је времена већ прошло. На пример, ако је просечно време чекања у реду 5 минута, то што сте већ чекали 5 минута не значи да ће вам се време чекања смањити. Ово својство је јединствено за експоненцијалну расподелу.

Формула експоненцијалне расподеле

Функција густине вероватноће за експоненцијалну расподелу је дата формулом:

$$f(x;\lambda) = \lambda e^{-\lambda x}, \qquad x \geq 0$$

где су:

- x — време или интервал између догађаја,
- λ — стопа догађаја (интензитет).

Функција кумулативне дистрибуције, која даје вероватноћу да ће се догађај десити пре времена x, је дата формулом:

$$F(x;\lambda) = 1 - e^{-\lambda x}, \qquad x \geq 0$$

Униформна расподела

Униформна расподела је једна од најједноставнијих и најинтуитивнијих расподела вероватноће у статистици. Она описује ситуације у којима су сви исходи једнако вероватни. Постоје два типа униформне расподеле: дискретна и континуална.

Континуална униформна расподела

Континуална униформна расподела, често звана и "равномерна расподела", описује насумичну променљиву која има једнаку вероватноћу да поприми било коју вредност унутар одређеног интервала $[a, b]$. Ова расподела је често означена као $U(a, b)$.

Карактеристике континуалне униформне расподеле

1. **Једнака вероватноћа:** Све вредности у интервалу $[a, b]$. су једнако вероватне.
2. **Функција густине вероватноће:** Функција густине вероватноће за континуалну униформну расподелу је:

$$f(x) = \frac{1}{b - a^2}, \qquad a \leq x \leq b$$

   где су a и b доња и горња граница интервала, респективно

3. **Функција кумулативне дистрибуције:** Функција кумулативне дистрибуције која даје вероватноћу да ће насумична променљива бити мања или једнака од x је:

$$F(x) = \begin{cases} 0, & x < a \\ \dfrac{x - a}{b - a}, & a \leq x \leq b \\ 1, & x > b \end{cases}$$

Дискретна униформна расподела

Дискретна униформна расподела описује ситуацију у којој насумична променљива може да поприми једну од n дискретних вредности, а свака од тих вредности је једнако вероватна. Ова расподела је често означена као $U(\{1, 2, \dots, n\})$.

Карактеристике дискретне униформне расподеле

1. **Једнака вероватноћа:** Сваки могући исход има једнаку вероватноћу $\frac{1}{n}$, где је n укупан број могућих исхода.
2. **Функција вероватноће:** Функција вероватноће за дискретну униформну расподелу је:

$$P(X = x) = \frac{1}{n}, \qquad x \in \{x_1, x_2, \dots, x_n\}$$

3.1.2 Дескриптивна статистика за расподелу података

Дескриптивна статистика за расподелу података обухвата различите мере које омогућавају боље разумевање и описивање основних карактеристика скупа података. Ове мере се користе за сумирање и анализу важних аспеката података.

Мере централне тенденције

Мере централне тенденције су сумиране статистике које представљају средишњу тачку или типичну вредност скупа података. Примери ових мера укључују средњу вредност, медијану и модуо. Ове статистике показују где већина вредности у расподели пада и често се називају централном локацијом расподеле. Мере централне тенденције се могу схватити као склоност података да се групишу око неке средње вредности.

Средња вредност

Средња вредност (аритметичка средина) представља просечну вредност свих података у датом скупу. Израчунава се као збир свих вредности подељен са бројем вредности.

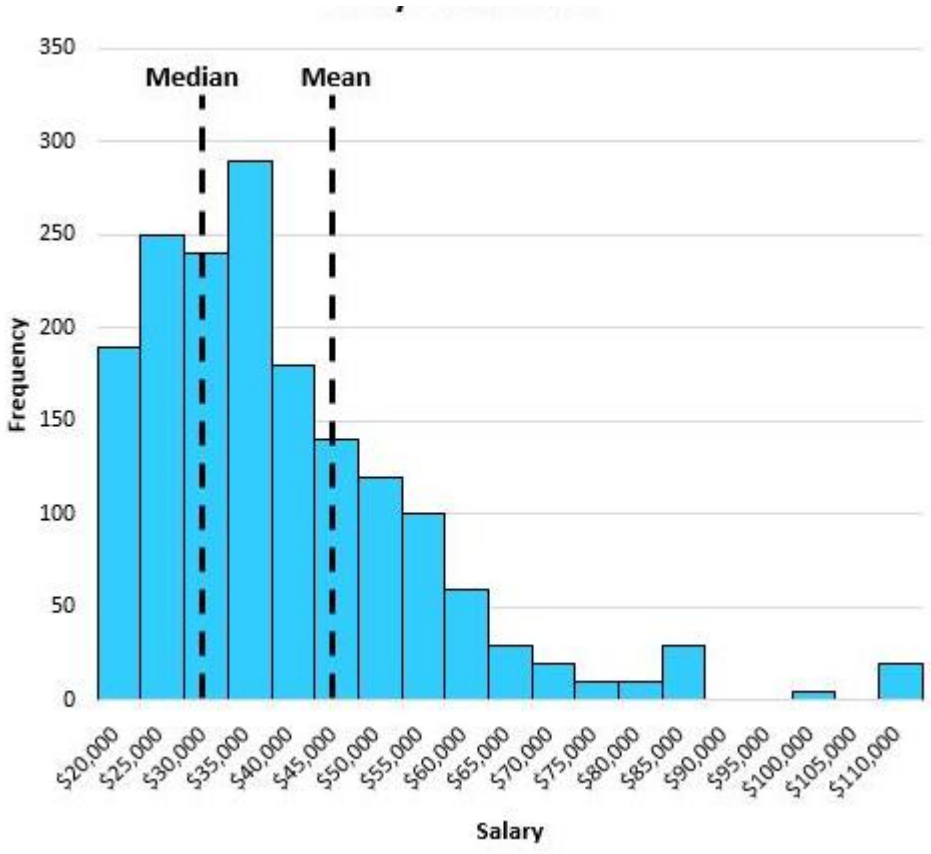
Формула је:

$$\bar{x} = \frac{x_1 + x_2 \quad \dots + x_n}{N}$$

Корисна је за скуп података који немају значајне издвојене вредности, али је осетљива на издвојене вредности (outliers).

Медијана

Медијана је средња вредност скупа података када су подаци поређани по величини. Ако је број података непаран, медијана је средња вредност; ако је паран, медијана је просек две средње вредности. Мање је осетљива на издвојене вредности и боље представља "средину" података у асиметричним расподелама.



Слика 3. Разлика између средње вредности и медијане

У примеру са слике медијана боље представља „типичну" плату појединца него средња вредност. У овом конкретном примеру, средња вредност нам показује да типичан појединац у овом граду зарађује око 47.000 долара годишње, док медијана показује да зарађује само око 32.000 долара годишње, што је много репрезентативније за типичног појединца.

Модуо

Модуо је вредност која се најчешће појављује у скупу података. Подаци могу бити унимодални, бимодални или мултимодални. Корисна је за категоријске податке и за идентификовање најчешће појављујућих вредности у скупу података.

Како идентификовати модуо: Једноставно идентификујте који податак или подаци се најчешће појављују.

Пример: Размотримо величине обуће седам особа: 7, 7, 8, 8, 8, 9, 10.

Величина обуће 8 се појављује три пута, више него било која друга величина, што чини 8 модуом.

Мере дисперзије

Мере дисперзије су статистичке мере које нам омогућавају да разумемо колико су подаци распршени или распрострањени око централне вредности, као што су средња вредност или медијана. Док мере централне тенденције (попут средње вредности, медијане и моде) показују где је "центар" података, мере дисперзије пружају информације о варијабилности или разноликости података у скупу. Ове мере су кључне за разумевање ширине расподеле и за препознавање да ли су подаци скупљени око средње вредности или широко распршени.

Опсег

Опсег је једноставна мера распршености која се израчунава као разлика између највеће и најмање вредности у скупу података:

$$\text{Опсег} = \text{Максимум} - \text{Минимум}$$

Опсег је лако разумети и израчунати, али је веома осетљив на издвојене вредности (outliers), што га чини мање поузданим за скуп података са екстремним вредностима.

Варијанса

Варијанса мери просечну квадратну разлику између сваке вредности и средње вредности. Израчунава се као збир квадрата одступања сваке вредности од средње вредности подељен са бројем података:

$$\sigma^2 = \frac{\sum(X_i - \bar{X})^2}{N}$$

Варијанса је корисна мера распршености, али је њена јединица квадрат јединице оригиналних података, што може отежати њено тумачење.

Стандардна девијација

Стандардна девијација је квадратни корен варијансе и изражава се у истим јединицама као и оригинални подаци:

$$\sigma = \sqrt{\sigma^2}$$

Стандардна девијација је најчешће коришћена мера дисперзије јер је лако интерпретирати. Она показује колико су подаци просечно удаљени од средње вредности. Мања стандардна девијација значи да су подаци скупљени ближе средњој вредности, док већа стандардна девијација указује на то да су подаци више распршени.

Интерквартилни распон

Интерквартилни распон (IQR) је мера распршености која показује разлику између 75-ог перцентила (Q3) и 25-ог перцентила (Q1) скупа података:

$$IQR = Q3 - Q1$$

IQR је мање осетљив на издвојене вредности и често се користи за мерење варијабилности података у скупу. Такође је користан за идентификацију издвојених вредности (outliers) помоћу бокс плотова.

Коефицијент варијације

Коефицијент варијације (CV) израчунава се као однос стандардне девијације и средње вредности и често се изражава као проценат:

$$CV = \frac{\sigma^2}{\bar{X}} \times 100$$

CV је користан за поређење дисперзије података који имају различите јединице мере или различите средње вредности.

Мере дистрибуције

Мере дистрибуције су статистичке мере које нам помажу да опишемо облик расподеле података у скупу. Док мере централне тенденције и дисперзије пружају информације о средишњој вредности и варијабилности података, мере дистрибуције нам дају дубљи увид у то како су подаци распоређени — да ли су симетрични, да ли имају дуге репове, да ли су концентрисани око средишње вредности, и сл.

Искошеност

Искошеност мери асиметрију расподеле података, односно показује колико је расподела „искривљена“ на једну страну.

- **Позитивна искошеност:** Када је реп расподеле дужи са десне стране, расподела је позитивно искошена. То значи да већина података има вредности мање од средње, а неколико великих вредности извлачи реп расподеле удесно.
- **Негативна искошеност:** Када је реп расподеле дужи са леве стране, расподела је негативно искошена. Већина података има вредности веће од средње, док неколико малих вредности извлачи реп улево.
- **Симетрична расподела:** Ако је искошеност 0 или близу 0, расподела је симетрична, што значи да су леви и десни реп сличне дужине.

Куртоза

Куртоза мери "дебљину" или "танкост" репова расподеле података у односу на нормалну расподелу.

- **Лептокуртична расподела:** Расподела са високом куртозом (>3). Ова расподела има "дебље" репове и оштрији врх у поређењу са нормалном расподелом. Већа куртоза указује на то да подаци имају више екстремних вредности.
- **Мезокуртична расподела:** Расподела са куртозом једнаком 3. Нормална расподела је пример мезокуртичне расподеле и има умерено дебеле репове.
- **Платикуртична расподела:** Расподела са ниском куртозом (<3). Ова расподела има "тање" репове и заравњенији врх у односу на нормалну расподелу. Мања куртоза указује на то да подаци имају мање екстремних вредности.

3.2 Корелација

Корелација представља статистички метод који се користи за мерење и анализу снаге и смера односа између две или више променљивих. У анализи података, корелација је кључна за разумевање како промене у једној променљивој могу бити повезане са променама у другој.

Корелација мери степен до којег се две променљиве међусобно односе. На пример, у контексту анализе података, може се користити корелација како би се видело да ли постоји веза између промене у продаји и оглашивању.

Типови корелације:

- **Линеарна Корелација:** Мери колико добро се подаци могу прилагодити линеарној једначини. Најчешће се мери Pearson-овим корелационим коефицијентом.
- **Монотонска Корелација:** Мери односе који нису нужно линејни, али у којима једна променљива увек расте или опада са другом. Ово се мери Spearman-овим и Kendall-овим корелационим коефицијентима.

Корелациони коефицијенти

Pearson-ов Корелациони Коефицијент:

- Дефиниција: Мери степен линечне зависности између две континуалне променљиве.
- Формула:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

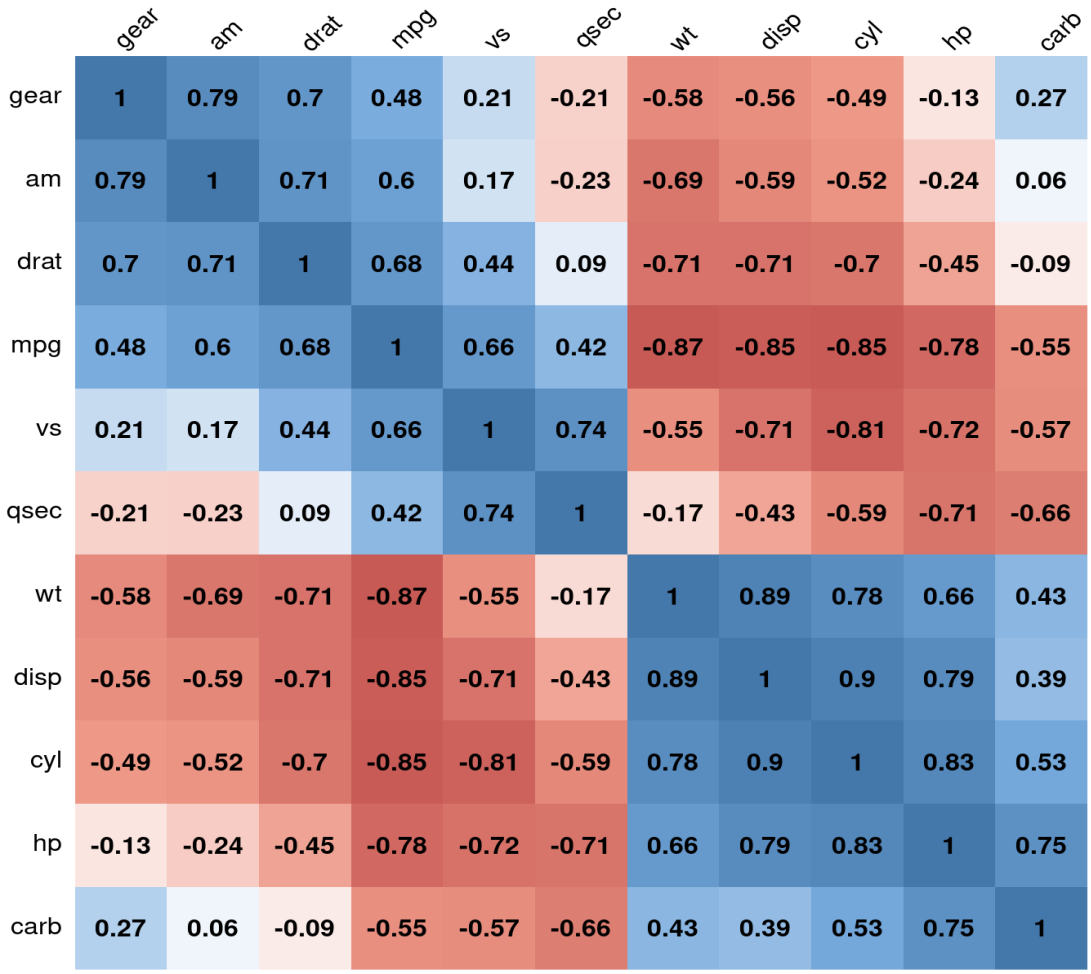
где су: x, y – улазни фичери, n – број елемената улазних фичера x и y.

Spearman-ов Ран Корелациони Коефицијент

- Дефиниција: Spearman-ов ран корелациони коефицијент мери монотонску зависност између две променљиве. Овај коефицијент користи рангове података уместо стварних вредности, што га чини корисним када подаци нису распоређени нормално или су редни.
- Интерпретација: Вредности се крећу од -1 до +1. Вредности близу +1 указују на снажну монотонску корелацију, док вредности близу -1 указују на обрнуту монотонску корелацију. Вредност 0 указује на одсуство монотонске зависности.

Kendall-ов Тау

- Дефиниција: Kendall-ов Тау мери степен слагања између рангових вредности. Овај коефицијент је посебно користан када је узорак података мали.
- Интерпретација: Вредности Kendall-овог Тау крећу се од -1 до +1. Високе вредности указују на добру сличност у рангирању, док ниске вредности указују на слабу сличност или разлике у рангирању.



Слика 4. Корелациона матрица

4. Методе за побољшање квалитета података

Квалитет података представља основу за доношење поузданих одлука и извођење прецизних анализа. Недостаци у квалитету података, као што су нетачни, непотпуни или екстремни вредности, могу довести до погрешних закључака и озбиљно утицати на резултате анализа. Због тога је кључно применити одговарајуће методе за побољшање квалитета података пре него што се они користе у било каквим аналитичким или машинским моделима.

Управљање outlier-има

Outlier-и су подаци који значајно одступају од осталих вредности у скупу података и могу утицати на резултате анализе, закључке и моделе. Они се могу јавити због грешака у уносу података, мерења, или могу представљати ретке и екстремне догађаје. Управљање outlier-има је процес идентификације, анализе и третмана ових вредности како би се минимизирао њихов утицај на аналитичке резултате.

Импутација недостајућих вредности

Недостајуће вредности су случајеви када подаци нису доступни или су непотпуни у датом скупу података. Недостаци могу бити последица различитих узрока, као што су грешке у прикупљању података, некомплетни уноси, или систематски пропусти. Импутација недостајућих вредности је процес предвиђања и попуњавања тих вредности како би се минимизирао губитак података и смањио утицај на анализе и моделе.

Одржавање квалитета подака

Редовно праћење и верификација квалитета података помаже у раном откривању проблема и обезбеђује да примењене методе остану ефикасне. Аутоматизација ових процеса кроз алате и скрипте може додатно унапредити одржавање квалитета података. Документација свих примењених метода за побољшање квалитета података, као и њихових ефеката, је кључна за транспарентност и поновљивост анализа. Добра документација такође помаже у процени ефикасности метода и побољшању будућих процеса чишћења података.

5. Закључак

Квалитет података представља један од најважнијих аспеката у области анализе података и машинског учења. Недостаци у квалитету података могу значајно утицати на тачност и поузданост аналитичких модела и одлука које се на њима заснивају. Током овог рада истражили смо различите димензије квалитета података, укључујући мере као што су тачност, комплетност, конзистентност, веродостојност и временска валидност података. Свака од ових димензија игра важну улогу у процени укупног квалитета података.

Поред тога, анализирали смо различите статистичке мере које се користе за процену квалитета података, као што су расподела, корелација и варијанса. Корелација, као једна од статистичких мера, показала се корисном за разумевање односа између различитих променљивих и откривање могућих проблема са квалитетом података.

Истраживање квалитета података такође укључује идентификацију проблема као што су недостајуће вредности, outlier-и и шум у подацима, који могу нарушити интегритет и тачност података. Адекватне методе за идентификацију и решавање ових проблема су од суштинске важности за одржавање и побољшање квалитета података. Побољшање квалитета података је континуиран процес који захтева пажњу, али пружа значајне користи у погледу тачности и поузданости аналитичких резултата и доношења одлука. За аналитичаре и истраживаче, разумевање и примена принципа квалитета података су кључни за успех било којег пројекта заснованог на подацима.

Референце

[1] https://www.fcsm.gov/assets/files/docs/FCSM.20.04_A_Framework_for_Data_Quality.pdf

[2] https://towardsdatascience.com/how-to-measure-data-quality-815076010b37

[3] https://www.dataversity.net/the-challenge-of-data-accuracy/

[4] https://www.ibm.com/topics/data-integrity

[5] https://www.ibm.com/think/topics/data-consistency-vs-data-integrity

[6] https://medium.com/analytics-mastery/understanding-data-distribution-in-data-science-and-statistics-comprehensive-guide-with-python-de5fa8735053

[7] https://brilliant.org/wiki/normal-distribution/

[8] https://www.mathsisfun.com/data/standard-normal-distribution.html

[9] https://www.statology.org/measures-central-tendency/