УНИВЕРЗИТЕТ У НИШУ
ЕЛЕКТРОНСКИ ФАКУЛТЕТ

Квалитет података

Семинарски рад
Предмет: Прикупљање и предобрада података за Машинско учење

Студент:

Настасија Станковић, бр. инд. 1622

Ментор:

доц. др Александар Станимировић

Ниш,2024. година

Садржај

Ниш,2024. година

Ниш,2024. година

1. Увод

У савременом добу технолошког напретка, развој софтверских и хардверских система резултирао је у стварању огромних количина података. Ови подаци су постали темељ за функционисање модерних софтверских система, чија је корисност изражена у различитим доменима њихове примене. Централно место података у рачунарском окружењу намеће потребу за њиховим детаљним обрадама како би се унапредио њихов квалитативни ниво. Овај процес унапређења квалитета података кључан је за повећање употребне вредности података, која игра битну улогу у доношењу закључака и идентификацији узорака у скуповима података.

Процес прикупљања и предобраде података, који се одвија пре фазе анализе података или закључивања, укључује низ корака усмерених ка побољшању квалитета ових података. Током прикупљања података, посебан фокус се ставља на квалитет појединачних узорака, где је циљ да се добијене вредности узорака поклапају са општеприхваћеним опсегом вредности заснованим на доменском знању. Узорци који одступају од овог опсега могу бити укључени у коначни скуп података само након темељне анализе у процесу предобраде.

Процеси који се спроводе у оквиру прикупљања и предобраде података обухватају кораке као што су агрегација узорака, испитивања квалитативних карактеристика узорака или целокупног скупа података, провера присуства појединачних вредности унутар скупа, валидација добијених вредности узорака, анализа дистрибуције података, процена концентрације узорака око одређених мера централне тенденције, анализа међусобних односа појединачних узорака унутар целокупног скупа, као и упоређивање зависности међу различитим карактеристикама присутним у скупу података.

Ови кораци су есенцијални за целокупни процес обраде и анализе података, јер доприносе разумевању основних и суштинских својстава великих скупова података. На основу резултата добијених у овој фази, стварају се мета-подаци о анализираном скупу, који су кључни за идентификацију даљих корака у процесу анализе. Ако мета-подаци не испуњавају критеријуме квалитета података тада се тај скуп података сматра ризичним и може бити искључен из даље анализе. Критеријуми квалитета података, који би требали бити испуњени у сваком скупу, детаљно су обрађени у наставку овог рада.

2. Квалитет података

Квалитет података не представља само техничку потребу, већ стратешку компоненту која утиче на све аспекте пословања и истраживања. У процесима као што је креирање модела машинског учења, квалитет података је не само неопходан за прецизност и поузданост модела, већ и за откривање нових, дубљих увида који нису одмах уочљиви. Квалитетни подаци омогућавају моделима да ефикасније "уче" из података, идентификујући сложене обрасце и трендове који су кључни за развој прецизних предиктивних модела.

Значај квалитета података се огледа у више кључних аспеката савременог пословања и истраживачких активности:

- **Оптимизација пословних процеса:** Поуздани подаци играју кључну улогу у идентификовању оперативних неефикасности, што омогућава предузећима да побољшају своје операције и повећају продуктивност.

- **Боље доношење одлука:** Квалитетни подаци су темељ ефикасног доношења одлука. Организације које се ослањају на тачне и ажурне податке имају значајну конкурентску предност, јер су у могућности да брзо и ефикасно реагују на тржишне промене.

- **Повећање задовољства купаца:** Омогућавају боље разумевање потреба циљних купаца што је кључно за прилагођавање производа и услуга, и то даље директно доприноси повећању задовољства купаца и јачању лојалности.

Квалитет података се мери према различитим димензијама што ћемо видети у наставку рада.

2.1 Основе и принципи управљања квалитетом података

Кључне компоненте управљања квалитетом података:

- **Стандардизација података**: Имплементација стандарда попут ISO 8000 и Data Quality Management Frameworka (DQMF) помаже у постављању јасне структуре за управљање подацима. Ови стандарди дефинишу методе за процену, контролу, и побољшање квалитета података.

- **Принципи FAIR**: Ови принципи наглашавају важност да подаци буду Проналазиви (Findable), Приступачни (Accessible), Интероперабилни (Interoperable) и Поново употребљиви (Reusable).

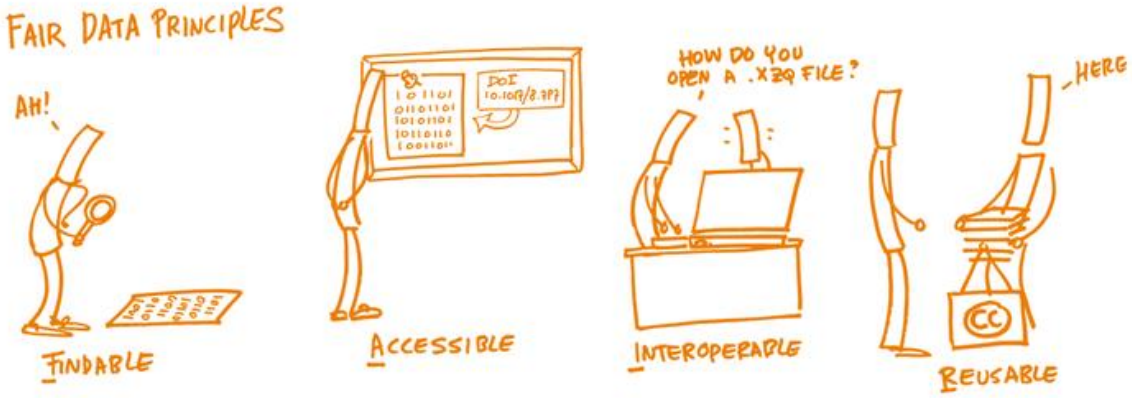
Процеси и методологије:

- **Евалуација и мониторинг квалитета података**: Редовна евалуација и мониторинг су неопходни за одржавање високог нивоа квалитета података. То укључује праћење тачности, комплетности, релеванције и доследности података.

- **Чишћење и корекција података**: Идентификација и исправљање грешака у подацима, као што су дупликати, недостајући подаци, и нетачни унос података, су кључни за одржавање интегритета података.

- **Интеграција података из различитих извора**: Управљање подацима из разних извора захтева кохерентне стратегије за осигуравање њихове конзистентности и употребљивости у различитим контекстима.

Примена ових принципа и методологија у пракси може знатно унапредити способност организација да ефикасно користе податке, што доводи до смањења ризика повезаних са лошим квалитетом података и побољшања одлучивања. Континуирана евалуација и унапређење квалитета података омогућавају организацијама да остану конкурентне, прилагођавајући се променама тржишта и технологије.



Слика 1. Илустративни приказ FAIR принципа

3. Мере квалитета података

Кључне мере квалитета података укључују:

- Тачност
- Комплетност
- Конзистентност
- Кохерентност
- Актуелност
- Релевантност
- Јединственост

Слика 2. Мере квалитета података

3.1 Тачност

Тачност података представља основни стуб квалитета података, означавајући степен у којем подаци тачно представљају реално стање или информације. Тачни подаци су темељ за ефикасно доношење одлука, поуздану аналитику и успешно управљање процесима.

У контексту машинског учења, тачност добија специфично значење. Она се користи као мера за процену перформанси класификационих модела. Тачност у машинском учењу је дефинисана као однос броја тачно предвиђених инстанци према укупном броју инстанци у тестном скупу података.

Формула за тачност је следећа:

$$A=TP+TN / TP+TN+FP+FN$$

Где су:

TP (*True Positive*) - број тачно позитивних инстанци,
TN (*True Negative*) - број тачно негативних инстанци,
FP (*False Positive*) - број лажно позитивних инстанци,
FN (*False Negative*) - број лажно негативних инстанци.

Ова мера је кључна за процену тога колико је модел добар у класификацији података. Међутим, важно је напоменути да тачност може бити обмањујућа у случајевима када је скуп података неуравнотежен, тј. када једна класа доминира над другом. У таквим случајевима, друге метрике као што су прецизност, одзив и F1 скор пружају детаљнији увид у перформансе модела.

3.2 Комплетност

Комплетност података се односи на број попуњених вредности унутар скупа података што доприноси целовитости или свеобухватности скупа података. Потпуни подаци су неопходни јер омогућавају ефикасну обраду и анализу. Када подаци нису потпуни, то отежава аналитичке процесе и може довести до закључака који нису засновани на свим релевантним информацијама.

Да би се непотпуни подаци попунили могу се применити различите технике обраде података.Ове технике укључују *интерполацију* или *импутацију*, где се недостајући подаци попуњавају процењеним вредностима на основу доступних података, или елиминацију, где се редови или колоне са недостајућим вредностима уклањају из анализе.

Избор технике за обраду непотпуних података зависи од контекста и природе података, као и од тога колико је недостајућих информација критично за анализу која се обавља. У неким случајевима, прихватљиво је користити скупове података са одређеним степеном недостајућих података, док у другим ситуацијама, присуство чак и малог броја недостајућих вредности може утицати на крајњи резултат

3.3 Конзистентност

Конзистентност података представља један од најважнијих аспеката у управљању квалитетом података. Овај појам описује степен у којем подаци остају униформни, доследни и без конфликата широм различитих система, апликација и база података у којима се користе. Јединственост и усклађеност података кроз време и разне платформе је фундаментална за обезбеђивање тачности и поузданости информација.

Да би се постигла конзистентности, користе се различити методи и алати:

- **Детекција аномалија**: Анализа ванредних вредности (*outliera*) помаже у идентификовању и исправљању података који одступају од очекиваних вредности, што може указивати на проблеме у конзистентности.

- **Програми за одржавање квалитета**: Софтверска решења која прате и верификују конзистентност података могу аутоматски да исправљају грешке и да обезбеђују континуирану конзистентност.

3.4 Кохерентност

Кохерентни подаци треба да одржавају јединствену структуру, формат и дефиницију, омогућавајући да се подаци из различитих извора могу лако комбиновати, упоредити и анализирати без конфликта или нејасноћа.

Кључни аспекти кохерентности података:

- **Логичка усклађеност**: Подаци из различитих система треба логички да се поклапају. На пример, ако се у једном систему користи термин "клијент", а у другом "купац", мора постојати јасно дефинисана веза између та два појма.

- **Усклађеност података у времену**: Кохерентност се такође односи на одржавање доследности података током времена, што значи да историјски подаци треба да остану релевантни и упоредиви са новијим подацима.

Одсуство кохерентности може довести до неспоразума, погрешног тумачења података и грешака у пословним одлукама.

3.5 Актуелност

Актуелност података представља меру квалитета података која се односи на доступност и ажурираност података у одређеном временском тренутку.Одлучивање у било којој организацији често зависи од актуелности података који су на располагању, па је значајно да се подаци који се користе у анализи редовно ажурирају и одржавају.

Уколико подаци нису благовремено ажурирани, постоји велики ризик од доношења одлука заснованих на застарелим информацијама, што може имати значајне негативне последице. Метрика правовремености се често мери у процентима и показује колики део података је доступан и ажуран у оквиру одређеног временског интервала, било да је реч о данима, недељама или месецима.

3.6 Релевантност

Релевантност података је димензија која одређује колико су информације садржане у скупу података значајне за специфичне циљеве анализе или одлучивања. Степен у којем подаци одговарају и помажу у испуњавању конкретних информационих потреба директно утиче на њихову корисност и вредност. Подаци могу бити технички тачни и комплетни, али ако нису релевантни за задатак који је пред нама, њихова укупна корисност је ограничена.

На пример, у маркетиншким истраживањима, подаци о претходним куповинама клијената могу бити веома релевантни за развој персонализованих понуда, док подаци који нису директно повезани са куповним навикама можда неће имати исту вредност.

Уколико скуп података садржи информације које не носе високу релевантну вредност, постоји могућност њиховог уклањања из анализе како би се избегли погрешни закључци и утицај на коначне резултате. Ово је нарочито значајно када се узме у обзир да процес прикупљања и складиштења података изискује ресурсе и може бити финансијски захтеван. Стога, пажљива селекција и прикупљање само оних података који су директно релевантни за предстојећу анализу може значајно смањити трошкове и повећати ефикасност целокупног процеса аналитичког рада.

3.7 Јасноћа

Јасноћа података омогућава корисницима да разумеју податке без забуне или погрешног тумачења. Она се односи на лакоћу са којом се подаци могу интерпретирати, и то не само од стране аналитичара, већ и од стране свих који се ослањају на те податке за доношење одлука.

Јасноћа података укључује:

- **Разумљивост**: Подаци би требало да буду представљени на начин који је лако разумети, без обзира на стручност корисника. То подразумева коришћење јасног и концизног језика, као и избегавање стручних назива тамо где је то могуће.
- **Доступност**: Подаци треба да буду доступни у формату који омогућава лаку анализу и обраду, као што су таблице, графикони и дијаграми који појашњавају сложене концепте на приступачан начин.
- **Документација и метаподаци**: Добро документовани подаци са јасним метаподацима омогућавају корисницима да разумеју контекст, извор, ограничења и методологију који стоје иза података.
- **Стандардизација**: Усвајање стандардизованих формата и конвенција при приказивању и објављивању података доприноси њиховој јасноћи и смањује ризик од недоследности у тумачењу.

3.8 Јединственост

Јединственост података се односи на квалитативну меру сваког појединачног податка у великом скупу различитих података. Јединственост представља особину података која се односи на сваку појединачну ставку у подацима где се већим квалитетом подразумева и већа количина јединствених података. Јединственост података јесте супротност мултипликативности података тј. дуплицирања појединачних записа (врста) у табели података.

Мултипликативност доводи до повећања обима скупа података без уношења варијабилности или различитости унутар скупа. Приликом процеса препроцесирања података, неопходно је извршити редукцију скупа података избацивањем дупликата - мултиплицираних вредности појединачних врста у табели.

4. Расподела података

Расподела података је кључни статистички концепт који илуструје како су вредности у скупу података распоређене и учестале од најнижих до највиших вредности. Ова расподела омогућава детаљан увид у структуру и карактеристике скупа података, пружајући основу за разумевање образаца и тенденција које се појављују у подацима.

Главни аспекти расподеле података:

- **Визуализација расподеле**: Графички приказ расподеле података често се користи за илустровање учесталости различитих вредности у скупу података. Ово може обухватити, на пример, хистограме где су вредности (*features*) приказане на x-оси, а њихове учесталости на y-оси.

- **Природа дистрибуције**: Разумевање како су подаци груписани може открити много о природи скупа података. Да ли су подаци концентрисани око одређених вредности, или су распоређени равномерно или неравномерно, може имати значајан утицај на интерпретацију и анализу података.

- **Типови расподела**: Постоје различити типови расподела као што су нормална, биномна, Поасонова расподела, који могу описати обрасце у подацима. Разумевање типа расподеле помаже у прецизнијем закључивању и анализи.

- **Анализа екстремних вредности**: Идентификација и анализа екстремних вредности *(„outlier"-a*) може бити важна за разумевање необичних образаца у подацима и потенцијалних грешака у скупу података.

У зависности од типова података који се обрађују, расподелу података је могуће поделити у две групе:

- **Дискретна расподела података**

- **Континуална расподела података**

4.1. Дискретне расподеле података

Дискретне расподеле података су статистичке расподеле које се користе за моделирање променљивих које узимају одређен број изолованих вредности. Оне су "дискретне" у смислу да не могу узети било коју вредност у неком континуираном опсегу уместо тога, вредности које променљива може узети су одвојене и обично се броје.

Неколико примера дискретних расподела укључује:

- **Бернулијева расподела**

- **Биномна расподела**

- **Поасонова расподела**

4.1.1 Бернулијева расподела

Бернулијева расподела је најједноставнија дискретна расподела и моделира случајеве у којима постоји само два могућа исхода неког експеримента или процеса, обично означени као "успех" и "неуспех". Ови исходи су међусобно искључиви, што значи да се мора десити један и само један од ова два исхода.

Бернулијева променљива X може узети вредност 1 са вероватноћом p која представља успех, и вредност 0 са вероватноћом 1−p која представља неуспех. Математички, функција вероватноће за Бернулијеву расподелу X је дефинисана као:

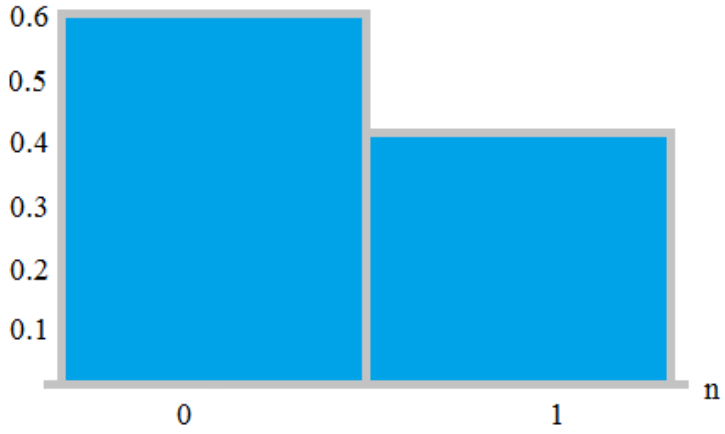
$$P(X=x)=p^{x}(1-p)^{1-x}$$

за x који узима вредност 1 или 0. Овде је:

- p вероватноћа успеха (која је уједно и средња вредност и мод ове расподеле),
- 1−p је вероватноћа неуспеха,
- x је вредност коју Бернулијева променљива може да узме.

Очекивана вредност (средња вредност) и варијанса за Бернулијеву расподелу су:

- Очекивана вредност **E[X]=p**
- Варијанса **Var(X)=p(1−p)**



Слика 3. Бернулијева расподела за случајеве успеха (p=0.4)

4.1.2 Биномиална расподела

Биномиална расподела је дискретна расподела која генерализује Бернулијеву расподелу за низ независних и идентичких испитивања. Користи се када је интересовање усмерено на бројање успеха у фиксираном броју понављања неког случајног експеримента.

Карактеристике биномиалне расподеле:

- **Број испитивања (n):** Ово је број пута који се експеримент изводи.
- **Вероватноћа успеха (p):** Вероватноћа да ће једно испитивање резултирати успехом.
- **Вероватноћа неуспеха (q или 1-p):** Вероватноћа да ће једно испитивање резултирати неуспехом.

Функција вероватноће биномиалне расподеле, која даје вероватноћу тачно k успеха у n испитивања, дата је формулом:

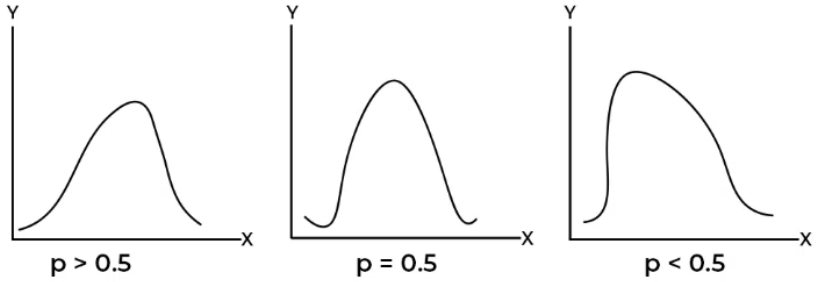
$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

где је:

- $\binom{n}{k}$ - биномиални коефицијент који представља број начина на које се може изабрати k успеха из n испитивања,
- X је случајна променљива која представља број успеха,
- k је број успеха који се разматра (креће се од 0 до n),
- p је вероватноћа успеха у једном испитивању,
- n је укупан број испитивања.

Очекивана вредност (средњи број успеха) и варијанса за биномиалну расподелу су:

- Очекивана вредност : **E[X]=np**
- Варијанса : **Var(X)=np(1−p)**



Слика 4. Облик биномиалне расподеле

Приказане су три различите биномиалне расподеле које кореспондирају са три различите вредности вероватноће успеха:

- Када је p>0.5, расподела је негативно коса, што значи да је више вредности сконцентрисано на вишим бројевима успеха.

- Када је p=0.5, расподела је симетрична, са равномерно распоређеним вредностима око средње вредности.
- Када је p<0.5, расподела је позитивно коса, што указује на то да је већина вредности сконцентрисана на нижим бројевима успеха.

Даље, опис говори о томе да је степен косине већи што је веће одступање вероватноће успеха од 0.5 и да је то такође зависно од броја испитивања n у биномиалном експерименту.

4.1.3 Поасонова расподела

Поасонова расподела се користи за моделирање броја пута који се неки догађај дешава у фиксираном временском интервалу, простору или скупу. Она је посебно корисна када су ти догађаји ретки и могу се сматрати независним један од другог унутар разматраног интервала.

Карактеристике Поасонове расподеле:

- **Ламбда (λ):** Параметар расподеле, λ, представља очекивани број догађаја у наведеном интервалу. Он је такође средња вредност и варијанса расподеле.

- **Број догађаја (k):** Количина догађаја која се дешава, који могу бити бројани цели бројеви (0, 1, 2, ...).

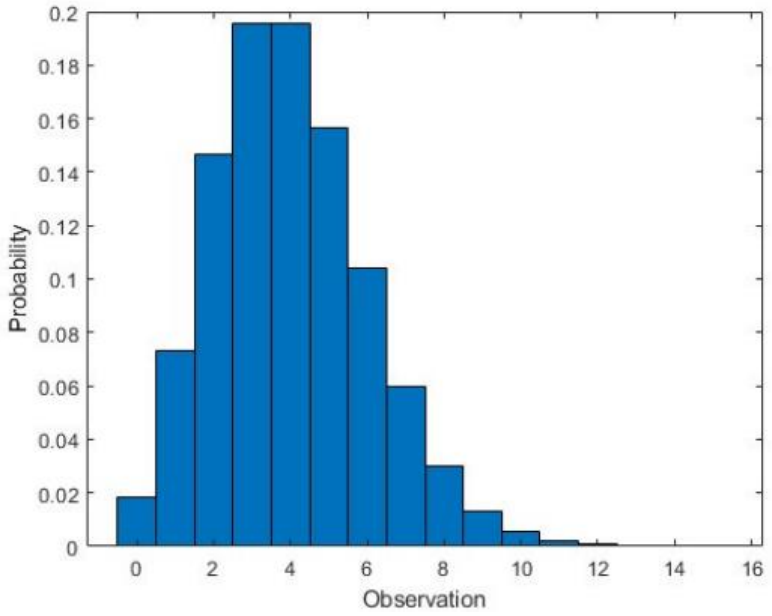
Формула за вероватноћу добијања тачно k догађаја је:

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

где је:

- e основа природног логаритма,
- k! је факторијел од k

Поасонова расподела помаже у процени вероватноће догађаја када су познате само средње вредности. Она омогућава моделирање и анализу догађаја који се могу појавити наизглед случајно, али са стабилном средњом стопом.



Слика 5. График Поасонове дистрибуције

На следећем графику је представљена вероватноћа учесталости одређеног броја догађаја у истом временском интервалу. Са графика је јасно уочљиво да се посматрани догађај у одређеном интервалу најчешће дешава три или четири пута.

4.2. Континуалне расподеле података

Континуалне расподеле података се користе за моделирање промењљивих које могу узети било коју вредност унутар одређеног опсега или интервала. Оне су супротне од дискретних расподела, које се односе на промењљиве са одвојеним, избројивим вредностима. Континуалне расподеле су кључне у статистици и машинском учењу зато што многе важне промењљиве у реалном свету, као што су време, тежина, висина и температура, могу узимати било коју вредност у опсегу.

Неке од најчешћих континуалних расподела укључују:

- **Нормална расподела**
- **Експоненцијална расподела**

4.2.1 Нормална расподела

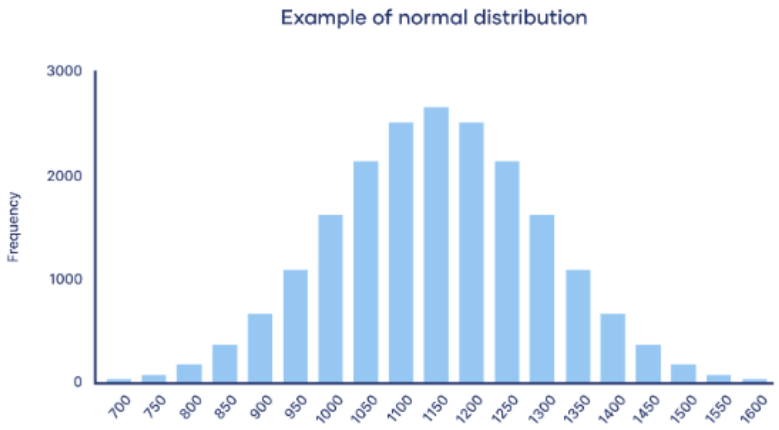
Нормална расподела, позната и као Гаусова расподела, представља један од најзначајнијих и најчешће коришћених типова расподеле у статистици и анализи података.

Карактеристике нормалне расподеле:

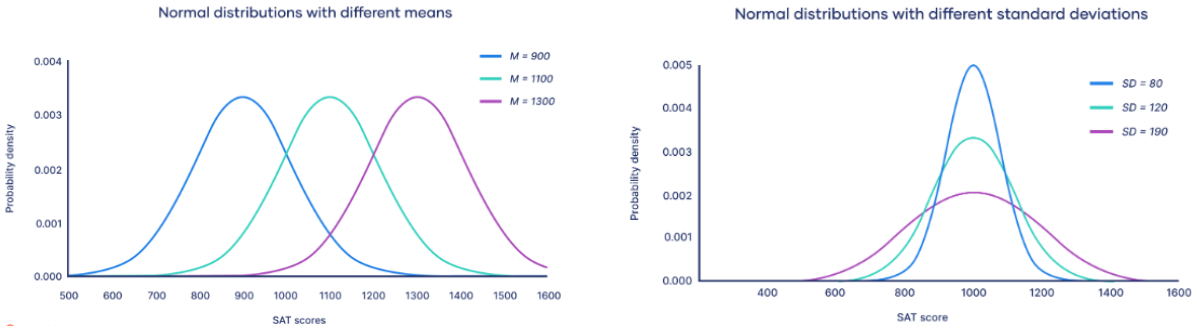
- **Облик звона**: График нормалне расподеле карактерише симетричан облик звона, где се већина вредности концентрише око централне средње вредности (средине), са вредностима које постепено опадају ка крајевима.

- **Средња вредност и Стандардна девијација**: Кључни параметри нормалне расподеле су средња вредност (μ) и стандардна девијација (σ). Средња вредност указује на централну тачку расподеле, док стандардна девијација показује колико широко се вредности распростире око средње вредности.

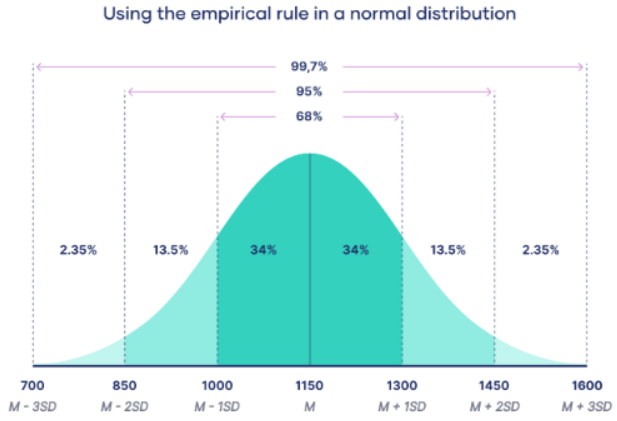
- **Правило 68-95-99.7**: Ово правило описује да у нормалној расподели око 68% вредности пада унутар једне стандардне девијације од средње вредности, око 95% унутар две стандардне девијације, и око 99.7% унутар три стандардне девијације.



Слика 6. Пример нормалне дистрибуције података



Слика 7. Средња вредност и стандардна девијација



Слика 8. Правило 68-95-99.7

4.2.2 Експоненцијална расподела

Експоненцијална расподела се користи за моделирање времена између независних догађаја који се дешавају са константном стопом. Типичан пример је време чекања за следећи долазак аутобуса, време до отказа електронског компонента, или време до следећег телефонског позива у кол-центру.

Формула која се користи за израчунавање вероватноће у експоненцијалној расподели је формула функције густине вероватноће (*probability density function* - PDF). За случајну променљиву X која описује време до следећег догађаја, функција густине вероватноће је дефинисана као:

$$f(x,\lambda)=\lambda e^{-\lambda x}$$

при чему је:

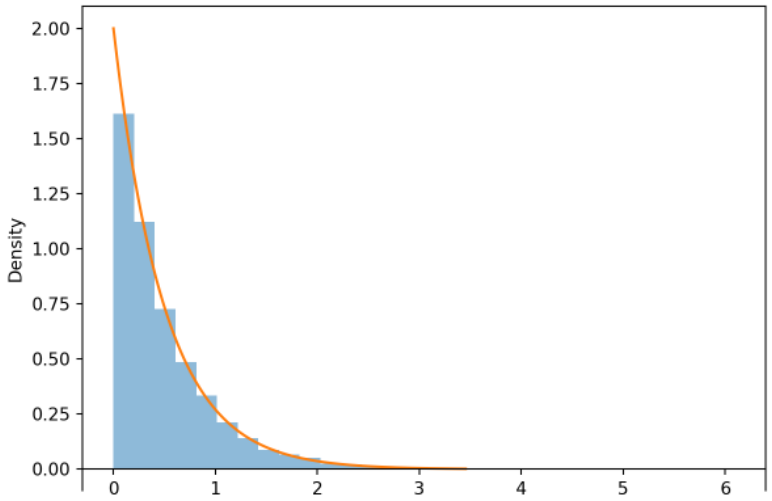
- x време до догађаја,
- λ стопа догађаја (број догађаја по јединици времена),
- e основа природног логаритма (приближно 2.71828).

Ова функција важи за $x \geq 0$, што значи да време до догађаја не може бити негативно.

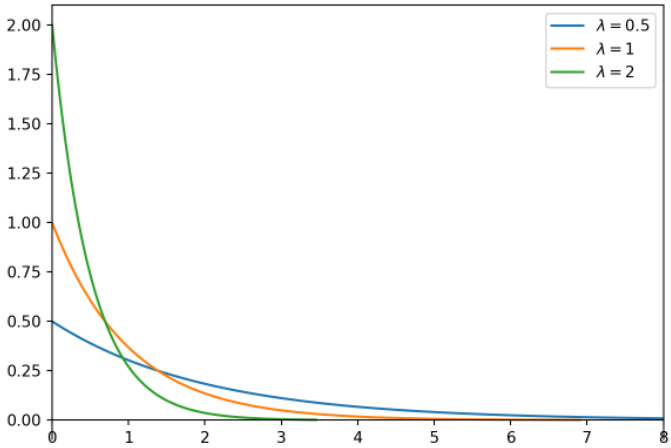
Кумулативна дистрибуциона функција (CDF), која даје вероватноћу да ће случајна променљива X узети вредност мању или једнаку x, за експоненцијалну расподелу је:

$$F(x;\lambda)=1-e^{-\lambda x}$$

С обзиром да експоненцијална расподела представља време до првог догађаја, CDF се може користити за одређивање вероватноће да ће се догађај десити у одређеном временском периоду.



Слика 9. Пример експоненцијалне расподеле



Слика 10. Пример експоненцијалне расподеле за различите вредности λ

4.3. Типови расподеле и њихова визуализација

Типови расподеле података представљају начин на који се варијабилност вредности података распоређује у оквиру целокупног скупа. Ови типови су од суштинског значаја за разумевање информација које подаци нуде и обухватају неколико кључних расподела које илуструју како се вредности у скупу података групишу и распоређују.

Када се подаци представе кроз хистограме, видљиво је како типови расподеле утичу на облик графика. Они одређују где се налази већина вредности, како су узорци груписани око одређених вредности, и како се тенденције распореда мењају. На пример, може се уочити где се налазе минималне и максималне вредности и како се распоређују најчешће вредности у односу на оне мање заступљене.

Основни типови расподеле података укључују:

- **Симетрична расподела**
- **Бимодална расподела**
- **Униформна расподела**
- **Асиметрична расподела**

Свака од ових расподела нуди различиту перспективу на податке и има своје специфичне примене у различитим областима.

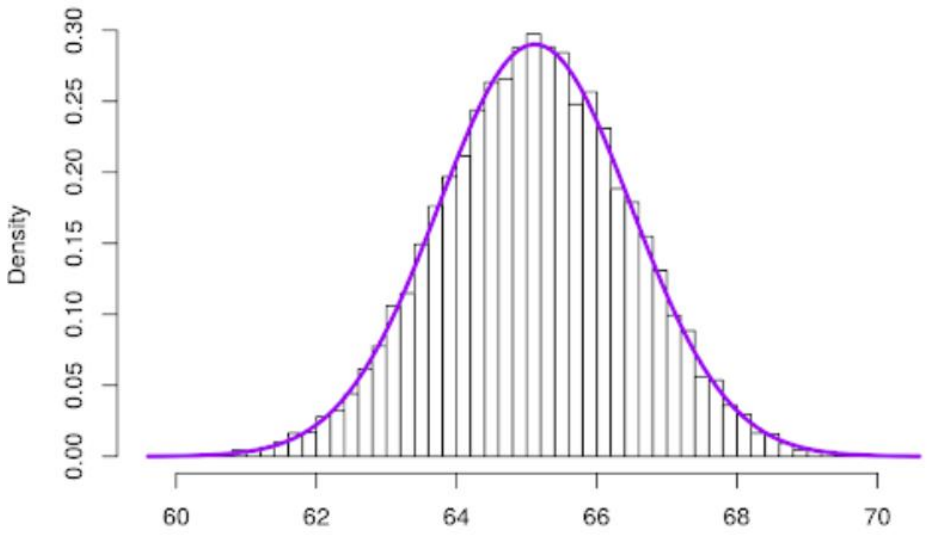
4.3.1 Симетрична расподела

Симетрична расподела података представља расподелу у којој су вредности распоређене тако да формирају симетричан образац око централне вредности, тј. средње вредности. Ова врста расподеле карактерише се тиме што су њене лева и десна половина у суштини огледало једна другој, показујући исту дистрибуцију и облик вредности на обе стране од средње вредности.

Основне карактеристике симетричне расподеле:

- **Средња вредност, медијана и мод**: У симетричној расподели, ове три мере централне тенденције се обично поклапају и лоциране су у самом центру расподеле. Таква конзистенција пружа јасну слику о централној тенденцији распоређених података.

- **Skewness је нула**: Skewness мери степен асиметрије расподеле око њене средње вредности. За симетричне расподеле, skewness је нула, што указује на то да нема одступања у асиметрији,обе стране средње вредности су равномерно распоређене.



Слика 11. Пример симетричне дистрибуције

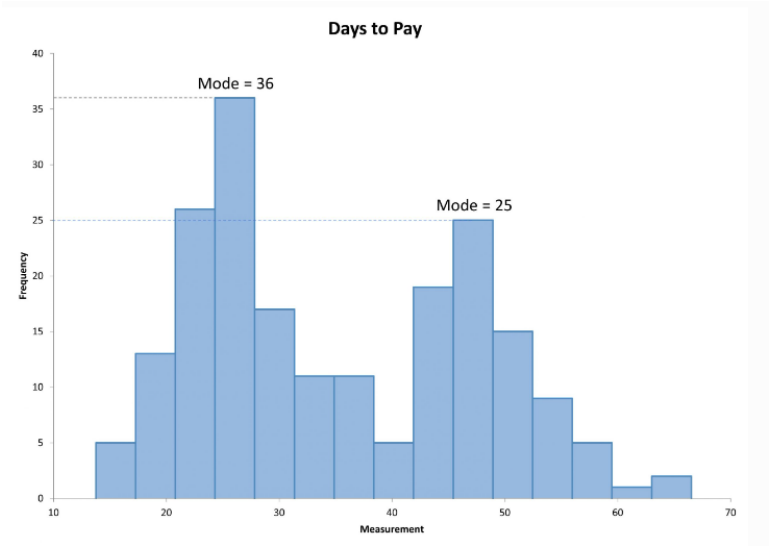
4.3.2 Бимодална расподела

Бимодална расподела података карактерише постојање два врха или модуса у расподели, указујући на присуство две главне групе вредности унутар истог скупа података. Оваква расподела може сугерисати на разноликост узорака или на постојање две различите подгрупе које се разликују по одређеним карактеристикама или величинама.

Кључна особина бимодалне расподеле је да се вредности у скупу података природно групишу око двеју доминантних области, што резултира стварањем два "врха" на графичком приказу, најчешће видљивог на хистограму. Ови врхови одражавају најучесталије вредности у скупу и служе као индикатори за разумевање структуре и динамике података.

Примери примене бимодалне расподеле се могу наћи у различитим областима, од анкетних података где бимодалност може указивати на подељене ставове испитаника, преко мерења перформанси где различите групе ученика показују различите нивое разумевања, до анализе продуктивности на радном месту где два пика могу илустровати разлику између високо и ниско продуктивних радника.

Бимодалност указује на могућност развоја специфичних стратегија за сваку од група у скупу података или чак на потребу за развојем посебних модела за анализу и предвиђање понашања унутар сваке од подгрупа.



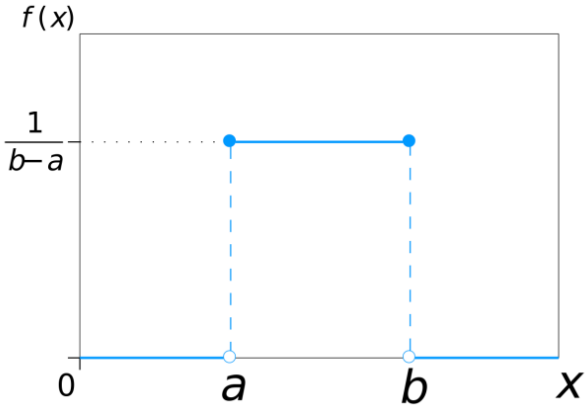
Слика 12. Пример бимодалне расподеле

4.3.3 Униформна расподела

Униформна расподела података, или равномерна расподела, јединствена је по томе што све вредности унутар одређеног интервала имају исту вероватноћу појављивања. Ово је пример континуалне расподеле, где променљива може преузети било коју вредност између две крајње тачке, означене као a и b, које представљају минималну и максималну могућу вредност у опсегу. Графички, униформна расподела представљена је хоризонталном линијом, указујући на то да је вероватноћа за било коју специфичну вредност унутар овог опсега константна.

Вероватноћа P са којом се може добити свака појединачна вредност из посматраног скупа је:

$$P = \frac{1}{b-a}$$



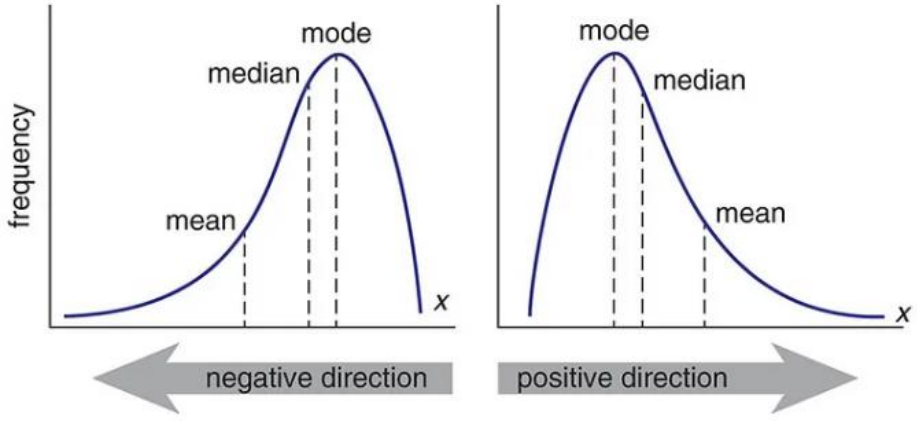
Слика 13. График униформне расподеле података на опсегу [a, b]

4.3.4 Асиметрична расподела

Асиметрична расподела представља тип расподеле података при чему график криве која описује податке није симетричан у односу мере централне тенденције скупа података. Овакав график има дугачки „реп“, односно вредност функције која умерено опада искључиво само са једне стране у односу на мере централне тенденције. Овакав „реп“ се креира на основу волумена података који се налазе изван опсега са најучесталијим вредностима у подацима.

Овакви типови података садрже велики број података чије вредности одступају од најучесталијих вредности из главног опсега. Најудаљеније вредности од главног опсега се називају „outlier“-има и такве вредности се након детаљне анализе у великом броју случајева одбацују из скупа података као невалидне.

У зависности од тога са које стране функција учесталости спорије опада, разликујемо две врсте асиметричне расподеле података. Уколико функција спорије опада са десне стране у односу на опсег најучесталијих вредности, таква асиметрична расподела се назива – *позитивно или десно кошење* расподеле података. Уколико функција спорије опада са леве стране у односу на опсег најучесталијих вредности, таква асиметрична расподела се назива – *негативно или лево кошење* расподеле података.



Слика 14. Пример позитивне и негативне расподеле

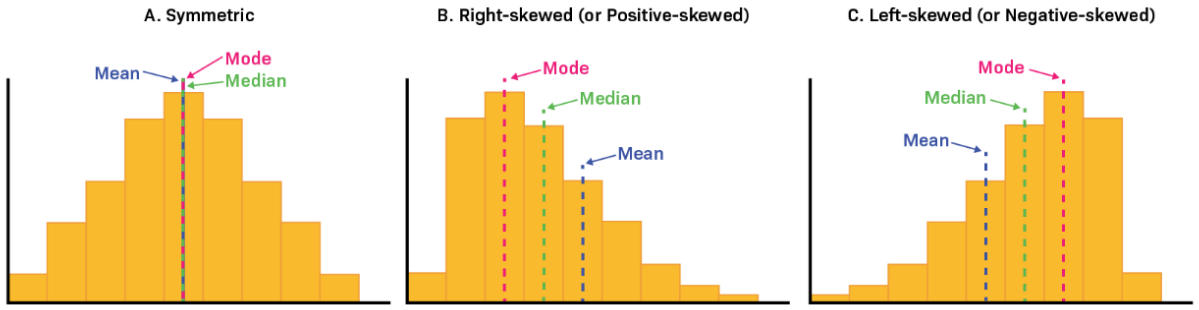
5 Мере централне тенденције

Мере централне тенденције су статистички индикатори који пружају сажет преглед сета података означавајући једну вредност која је репрезентативна за целокупан скуп. Ове мере су кључне за сумирање великих количина података, олакшавајући разумевање и интерпретацију података у једноставнијем облику. Оне омогућавају да се утврди "центар" расподеле података.

Основне мере централне тенденције су:

- Средња вредност
- Медијана
- Модуо

У даљем делу рада ће бити описане ове мере централне тенденције при чему свака носи другачију квалитативну вредност, али се свака од њих равноправно употребљава у односу на остале.



Слика 15. Мере централне тенденције

5.1 Средња вредност

Аритметичка средња вредност представља најзаступљенију и највише коришћену меру централне тенденције. Аритметичка средња вредност узима у обзир све вредности из скупа података приликом процеса рачунања коначне вредности. Управо из разлога што узима све вредности из скупа, може довести до неправилних закључивања о скупу података зато што се у процесу израчунавања користе и граничне „outlier“ вредности.Ако један податак у скупу има изузетно високу или ниску вредност, то може повући средњу вредност према тој тачки, што може довести до погрешне интерпретације централне тенденције скупа.

Ако је X скуп података са n вредности, средња вредност (x) се израчунава коришћењем формуле:

$$\overline{x}=\frac{\sum_{i=1}^{n} x_i}{n}$$

где је x_i вредност сваког појединачног податка у скупу, а n је укупан број података.

Ова мера централне тенденције има високу употребну вредност код скупова података који имају Гаусову (нормалну) расподелу података. Уколико је расподела података асиметрична тј. искошена неопходно је користити неки други приступ зато што тада средња вредност не приказује реалну вредност око које се подаци гомилају.

5.2 Медијана

Медијана је статистичка мера централне тенденције која одређује вредност која дели скуп података тако да има једнак број вредности испод и изнад себе када су подаци сортиран. Медијана је важна зато што пружа јасан увид у "средину" скупа података и ефикасна је у ситуацијама када скуп података садржи екстремне вредности или "outlier"-е,тако што их елиминише у току процеса израчунавања.

Како би се израчунала медијална вредност, најпре је неопходно сортирати добијени вектор података у растући/опадајући редослед редослед.

Затим:

- За скупове података са **непарним** бројем вредности: Медијана је једнака вредности која се налази тачно на средини сортираног низа. На пример, ако имамо скуп од 7 вредности, медијана ће бити четврта вредност у сортираном низу.

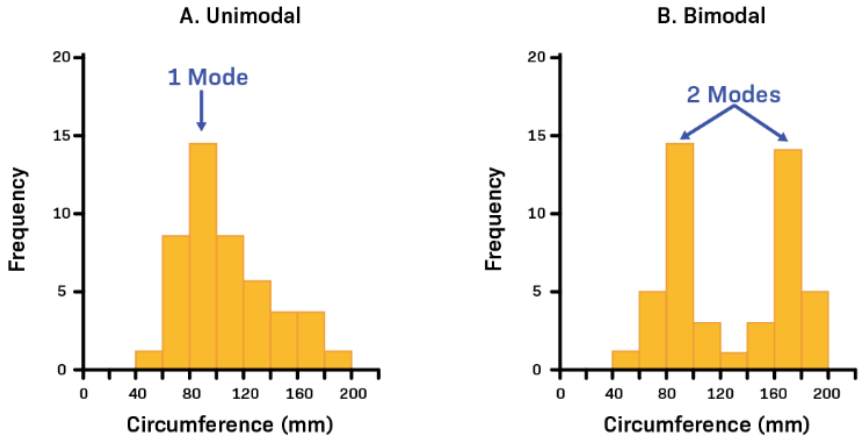
- За скупове података са **парним** бројем вредности: Медијана се израчунава као просек две средње вредности у сортираном низу. На пример, у скупу од 8 вредности, медијана ће бити просек четврте и пете вредности.

5.3 Модуо

Мод података, познат и као модус, представља још једну важну меру централне тенденције у статистици. Одликује се као вредност или вредности које се појављују најчешће у датом скупу података.

Ова мера централне тенденције се најчешће користи код фичера који могу имати мањи број могућих вредности као што су категорички подаци. Са повећањем броја могућих вредности које одређени фичер може да садржи, драстично опада квалитативна вредност коју ова мера носи са собом. Овај проблем се јавља из разлога зато што је код нумеричких података мања вероватноћа да ће вредности фичера имати потпуно идентичну вредност, јер се код мерења дешава одређен степен одступања. Због овог проблема, применљивост модуса је ниска код нумеричких типова података.

Код категоричких података код којих је број могућих вредности фичера мањи, ова мера може имати високо квалитативну вредност и може бити употребљена у различите сврхе. Такође, још један проблем који се јавља јесте да је могуће да вредност модуса буде драстично удаљена од главног опсега у којем се налазе остале вредности из посматраног скупа података, па стога вредност модуса неће носити валидну информацију о целокупном скупу података.



Слика 16. Пример унимодалног и бимодалног

6 Корелација

Корелација у представља меру која описује степен међусобне везе између две или више променљивих. Корелација може указивати на то како промена вредности једне променљиве утиче на вредност друге променљиве.

Врсте корелација:

- **Позитивна корелација**: Када вредност једне променљиве расте, вредност друге променљиве такође расте.
- **Негативна корелација**: Када вредност једне променљиве расте, вредност друге променљиве опада.
- **Нулта корелација**: Не постоји уочљива веза између променљивих.

Пирсонов коефицијент корелације између два посматрана фичера се рачуна као количник вредности коваријансе између два фичера и производа појединачних вредности стандардних девијација поменутих фичера. На овај начин се вредност коваријансе између две фичера нормализује у опсегу од [-1, 1].

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

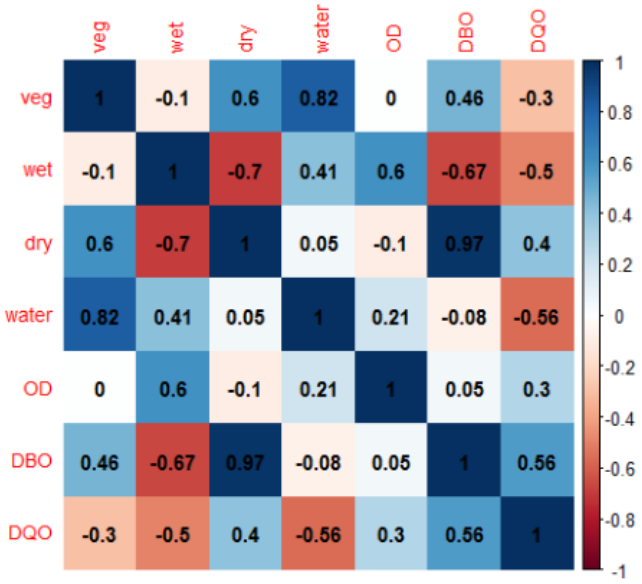
где су: x, y – улазни фичери, n – број елемената улазних фичера x и y

Вредности корелације које су у опсегу [-1, -0.5] или [0.5, 1] представљају високо условљену зависност и такву линеарну релацију између посматраних фичера треба јасно издвојити јер се на основу те релације могу извести одређени високо квалитативни закључци о скупу података.

Међутим висока корелација између атрибута може указивати на редунданцију информација или мултиколинеарност, што може довести до проблема приликом примене одређених модела машинског учења.Одлука о томе да ли треба или не треба обрисати атрибуте са високом корелацијом зависи од специфичности проблема, врсте модела који се користи и контекста података.

Приликом израчунавања вредности корелације на нивоу целокупног скупа података, применом уграђених функција, добија се матрица корелације – симетрична матрица која за врсте и колоне има улазне фичере посматраног скупа података. На главној дијагонали добијене матрице се налази вредност 1 јер је главна дијагонала пресек врста и колона истог фичера. У остала поља ове матрице се смештају вредности корелације између свих осталих фичера појединачно и то у пресеку врста и колона свих фичера понаособ.

На следећем примеру је приказана матрица корелације за скуп података,градијенталним опсегом боја је представљен степен корелације између фичера што се може видети са десне стране слике.



Слика 17. Матрица корелације скупа података

7 Варијанса

Варијанса података је мера која описује распрострањеност или дисперзију вредности у скупу података у односу на њихову средњу вредност. Другим речима, варијанса показује колико се вредности у датасету разликују једна од друге и од средње вредности. Већа варијанса указује на то да су вредности више распрострањене око средње вредности, док мања варијанса указује на то да су вредности ближе средњој вредности.

Вредност варијансе не може бити негативна, тачније налази се у скупу вредности [0, +∞).

Варијанса може имати два различита облика:

- **Варијансу популације**
- **Варијансу узорака**

Варијанса популације користи се када су доступни подаци за целу популацију која се проучава. Популација се овде односи на комплетан скуп свих елемената који заинтересују истраживача или аналитичара. Укључује сваки могући елемент који одговара одређеним критеријумима истраживања.

Формула за израчунавање варијансе за популацију (означена са σ^2) је:

$$\sigma^2 = \frac{\sum_{i-1}^{n}(x_i-\mu)^2}{N}$$

где је:

- σ^2 варијанса популације
- N укупан број вредности у популацији
- X_i вредности у скупу података
- μ средња вредност популације

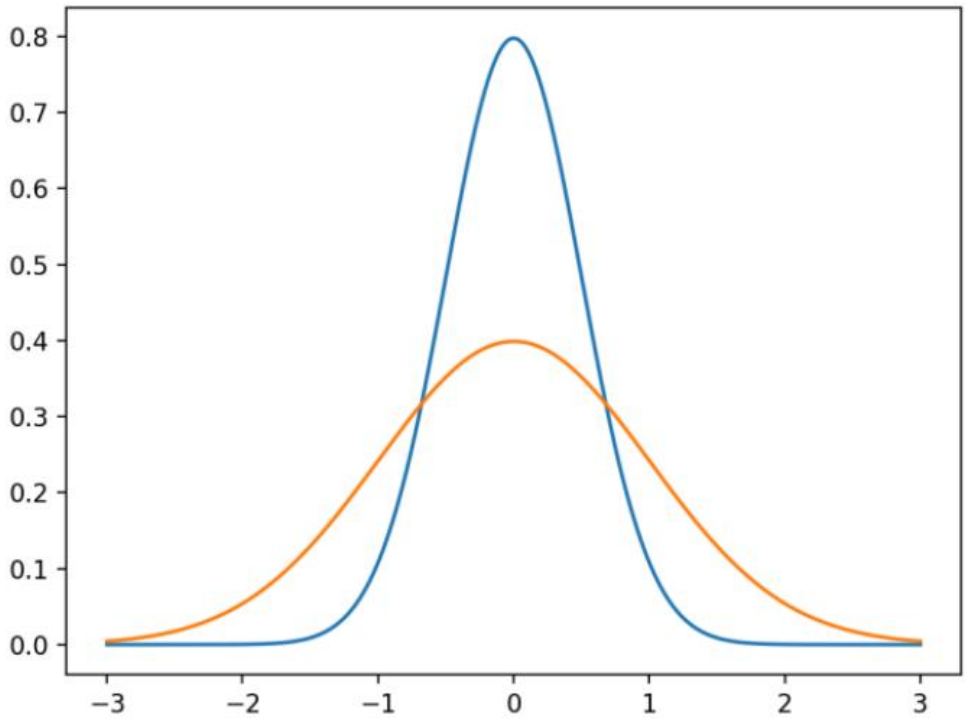
Варијанса узорка користи се када анализа подразумева само узорак одређене популације, а не целу популацију. Узорак представља мањи скуп података изабран из веће популације, и често се користи због непрактичности или немогућности анализирања сваког појединачног елемента у популацији.

За узорак из популације, формула за варијансу (означена са s^2) је мало другачија како би се узела у обзир мања величина узорка у односу на целу популацију:

$$s^2 = \frac{\sum_{i-1}^{n}(x_i-\bar{x})^2}{n-1}$$

где је:

- s^2 варијанса популације
- n укупан број вредности у узорку
- X_i вредности у скупу података узорка
- $\bar{x}$ средња вредност узорка



Слика 18. : Изглед Гаусове дистрибуције података са високом(плава крива) и ниском(наранџаста крива) варијансом

8 Примери из практичног дела рада

Практични део семинарског рада обухвата примену описаних техника провера квалитета података над два независна скупа података:

- Скуп података 1 – Овај датасет се односи на истраживање које се бавило класификацијом седам различитих врста пасуља. У истраживању су коришћене карактеристике попут облика, димензија, типа и структуре како би се разликовале седам различитих регистрованих сорти пасуља са сличним карактеристикама.

- Скуп података 2 – Подаци су повезани са директним маркетиншким кампањама португалске банкарске институције. Маркетиншке кампање су се заснивале на телефонским позивима. Често је било потребно више контаката са истим клијентом како би се утврдило да ли ће се клијент претплатити ('да') или неће ('не') на производ (депозит у банци).

Основна идеја практичног дела семинарског рада јесте испитивање квалитета података над описаним скуповима како би се добили детаљни подаци о квалитативним својствима посматраних података у циљу спровођења даљих поступака у фази предобраде података.

Процедуре обављене над свим скуповима ради утврђивања квалитативних својстава посматраних скупова података су:

- дескриптивна анализа података (провера типа података,недостајућих вредности,дупликата..)
- графички приказ расподела података различитих атрибута и уочавање вредности ван опсега као и провера балансираности датасета
- израчунавање мера централне тенденције нумеричких и/или категоричких атрибута
- израчунавање матрице корелације како би се установиле законитости/услољености које делују над подацима
- израчунавање варијансе свих нумеричких атрибута како би се проверила вредност одступања свих вредности од мера централне тенденције

8.1 Резултати над првим скупом података

Први скуп података се састоји из следећих атрибута:

- Area (A): Површина зоне пасуља и број пиксела унутар њених граница.
- Perimeter (P): Обим пасуља дефинисан као дужина њених граница.
- Major axis length (L): Растојање између крајева најдуже линије која се може повући од пасуља.
- Minor axis length (l): Најдужа линија која се може повући од пасуља док стоји вертикално у односу на главну осу.
- Aspect ratio (K): Дефинише однос између L и l.
- Eccentricity (Ec): Excentricitet елипсе.
- Convex area (C): Број пиксела у најмањем конвексном полигону који може садржати површину семена пасуља.

- Equivalent diameter (Ed): Пречник круга који има исту површину као површина семена пасуља.
- Extent (Ex): Однос пиксела у оквиру ограничавајућег оквира према површини пасуља.
- Solidity (S): Такође познато као конвексност. Однос пиксела у конвексној љусци према онима пронађеним у пасуљу.
- Roundness (R)
- Compactness (CO): Мера заобљености објекта: Ed/L
- ShapeFactor1 (SF1)
- ShapeFactor2 (SF2)
- ShapeFactor3 (SF3)
- ShapeFactor4 (SF4)

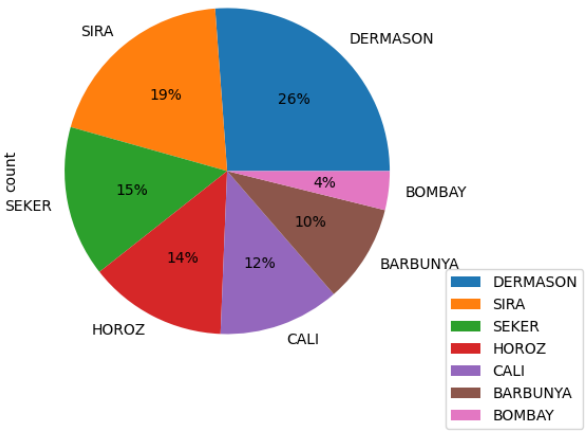
Таргет:

- Class (Seker, Barbunya, Bombay, Cali, Dermosan, Horoz i Sira) – који је тип пасуља (вишекласна класификација)

- Има 14 континуалних атрибута: MajorAxisLength,MinorAxisLength,AspectRation,Eccentricity,EquivDiameter,Extent,Solidity,roundness,Compactness,ShapeFactor1,ShapeFactor2,ShapeFactor3,ShapeFactor4
- 2 integer атрибута: Area I ConvexArea
- 1 категорички: Class

На почетку помоћу дескриптивне анализе закључили смо да имамо све нумеричке податке осим Class који је категорички,затим да нема недостајућих вредности али да имамо 68 дупликата које смо обрисали.

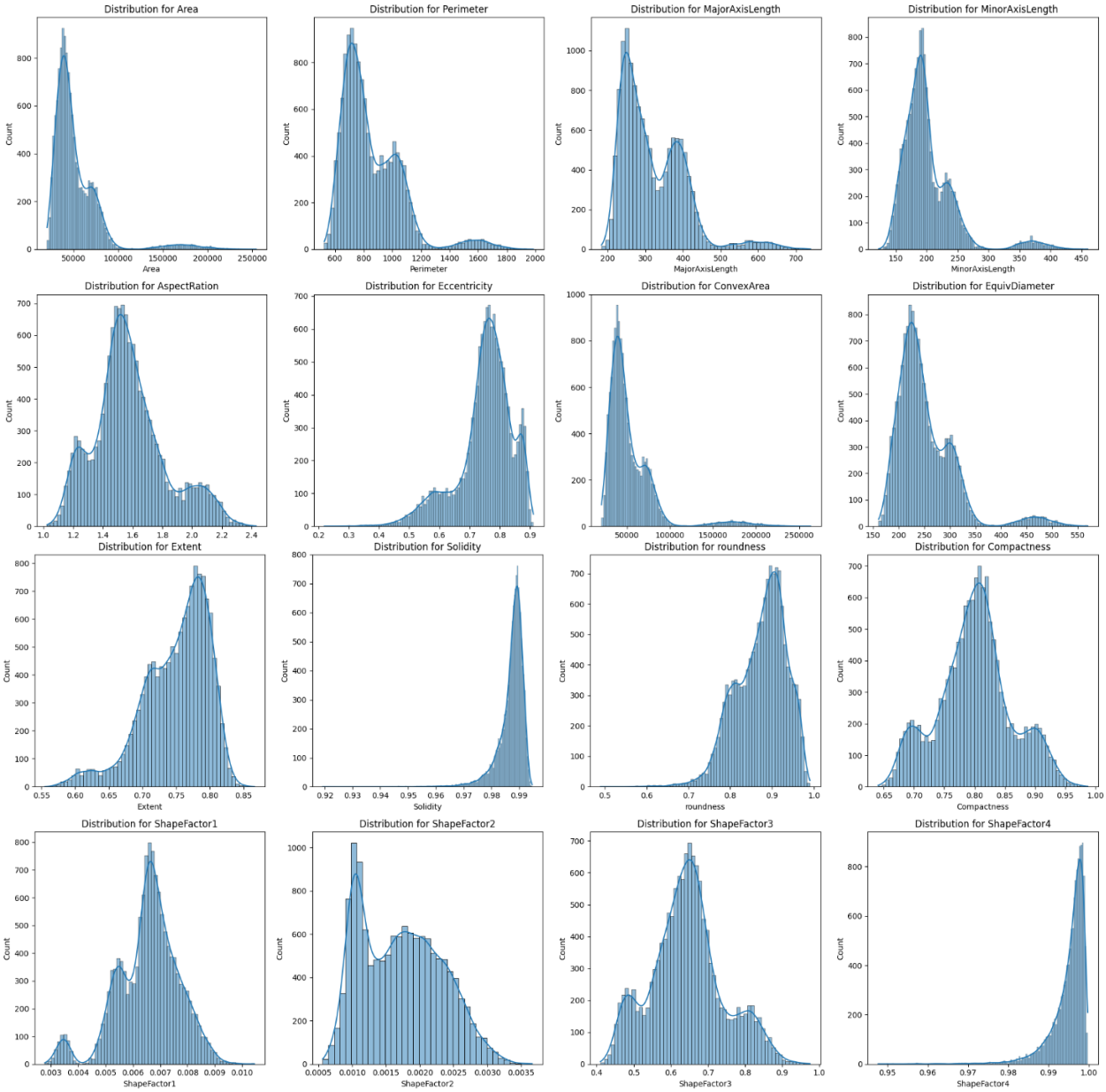
Из расподеле података видимо да је скуп података небалансиран,јер намамо једнаку расподелу таргет атрибута – Class,на слици 19. можемо да видимо каква је расподела.



Слика 19. Расподела Class атрибута

Овде већ закључујемо да би требало да употребимо неку од техника за балансирање скупа података,овде је примењена SMOTEEN техника која комбинује технике оверсамплинг-а (SMOTE) и ундерсамплинг-а (ENN).

Затим на следећој слици видимо расподелу осталих атрибута:



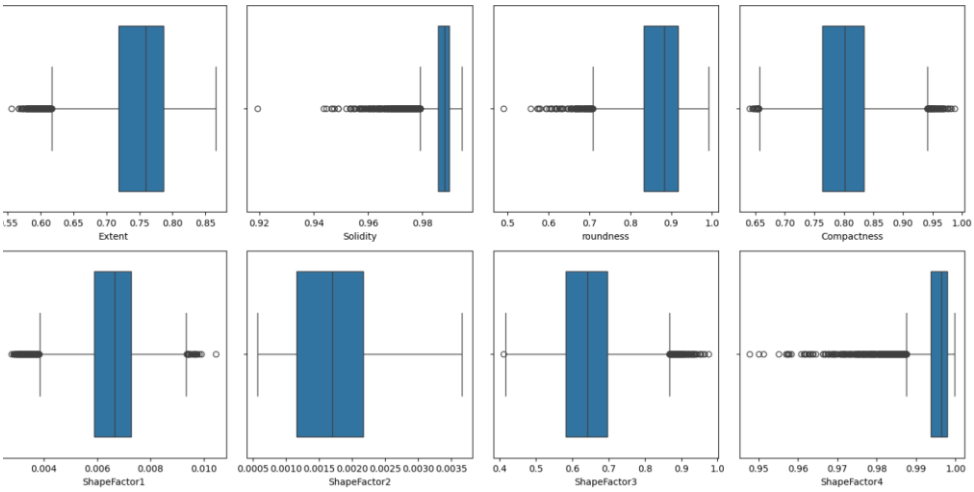
Слика 20. Расподела свих атрибута у скупу

- За Area примећује се десно накривљење асиметричне криве.
- За Perimeter примећује се десно накривљење асиметричне криве.
- За MajorAxisLength примећује се бимодална расподела са два врха.
- За MinorAxisLength примећује се десно накривљење асиметричне криве.
- За AspectRatio имамо Поасонову расподелу података.
- За Eccentricity имамо лево накривљење асиметричне криве.
- За ConvexArea и EquivDiameter имамо десно накривљење асиметричне криве.
- За Extent имамо Поасонову расподелу.
- За Solidity имамо експоненцијалну расподелу.
- За Roundness имамо лево накривљење асиметричне криве.
- За Compactness имамо нормалну односно Гаусову дистрибуцију.

- За ShapeFactor1 имамо бимодалну расподелу са три врха.
- За ShapeFactor2 имамо бимодалну расподелу са два врха.
- ShapeFactor3 имамо Гаусову расподелу.
- За ShapeFactor4 имамо експоненцијалну расподелу.

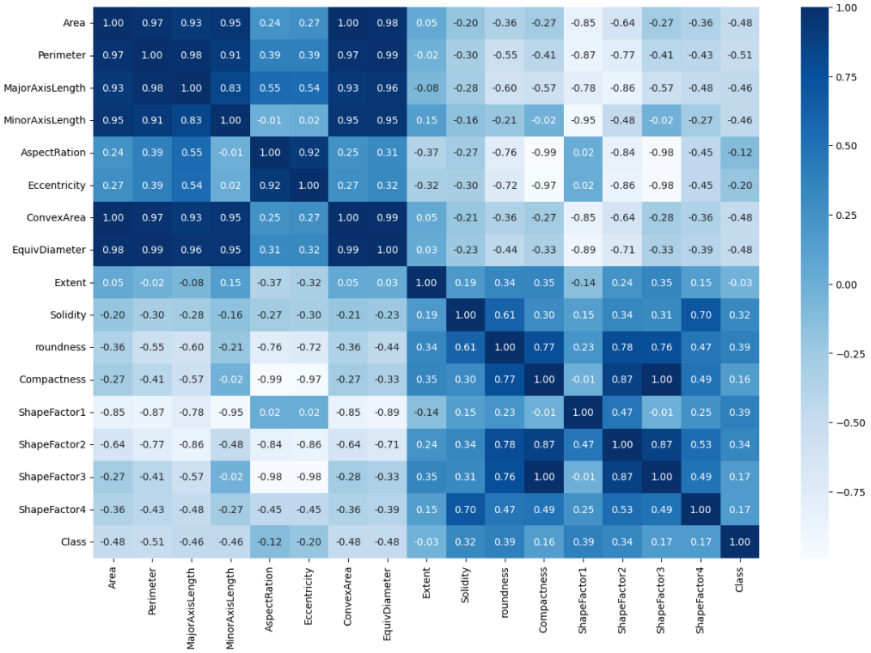
На основу расподеле података можемо извршити одређене трансформације и добити приближно нормалну дистрибуцију. Ове трансформације могу смањити варијабилност података и учинити их симетричнијим.Примењена је лог трансформација која је доста побољшала дистрибуцију појединих атрибута.

Затим на основу box plotova закључујемо да имамо outliere у скупу и они су отклоњени помоћу IQR методе.



Слика 21. Outlieri

Затим на основу матрице корелације можемо да видимо зависности између атрибута у скупу података:



Слика 22. Матрица корелације

На основу матрице закључујемо да је потребно избацити атрибуте Area, EquivDiameter,Eccentricity и ConvexArea јер постоји висока зависност међу подацима.

На крају, вредности варијансе, као и стандардне девијације:

Нулта стандардна девијација и варијанса: За колоне као што су ShapeFactor2, ShapeFactor1, ShapeFactor4, Solidity имају нулту стандардну девијацију то значи да све вредности у овим колонама имају исту вредност, односно да нема варијације у тим подацима. Ове колоне можда нису корисне за анализу или моделовање јер не доприносе разликовању инстанци.

Веће вредности стандардне девијације и варијансе: За колоне као што су Perimeter, Area, ConvexArea, MajorAxisLength, EquivDiameter, MinorAxisLength, Class и AspectRation, стандардна девијација је значајно већа то значи да су вредности у овим колонама шире распоређене око средње вредности, што указује на већу варијабилност у тим подацима. Ове колоне могу имати значајну улогу у анализи и моделовању, јер доприносе разликовању инстанци.

8.2 Закључак над првим скупом података

Да бисмо видели како све ове трансформације утичу на резултате алгоритама машинског учења,применићемо два алгоритма за класификацију Support Vector Classifier (SVC) и Random Forest алгоритам.

SVC је осетљив на неколико фактора, укључујући избор параметара, нелинеарност проблема, величину и димензионалност података, балансирање класа и присуство оутлиер-а. Осетљивост на ове факторе може утицати на тачност предикције модела и зато је погодан да видимо како све ове ствари утицу на резултат.

Док са друге стране Random Forest је енсембле алгоритам који се састоји од више стабала одлучивања отпоран је на оверфиттинг, способности руковања са великим бројем атрибута и величином података,такође може ефикасно радити са нелинеарним проблемима и не захтева пуно препроцесирања података.

Алгоритми су примењени над различитим скуповима података:

- Подела одновног скупа података
- Подела скупа на основу балансираности
- Подела скупа на основу расподеле података
- Подела скупа на основу outliera
- Подела скупа на основу матрице корелације

Резултати су следећи:

Tip podele	SVC accuracy	Random Forest accuracy
Osnovni skup	0.64	0.92
Balansirani	0.77	0.99
Raspodela podataka	0.76	0.92
Outlieri	0.64	0.92
Korelacija	0.88	0.93

Слика 23. Табела са резултатима за први скуп података

Из ових резултата можемо закључити да Random Forest алгоритам генерално постиже боље резултате од SVC алгоритма у свим сценаријима. Балансирање скупа података, примена лог трансформације и уклањање високо корелисаних атрибута имају позитиван утицај на перформансе класификације, посебно за SVC алгоритам. Међутим, уклањање outlier-а не показује значајне промене у тачности класификације.

8.3 Резултати над другим скупом података

Други скуп података се састоји из следећих атрибута:

- age
- job: тип посла
- marital : брачно стање
- education : тип школовања
- default: да ли клијент има кредит ("yes","no")
- balance: стање рачуна клијента у еврима
- housing: да ли клијент има стамбени кредит ("yes","no")
- loan: да ли има кредит ("yes","no")
- contact: тип комуникације са клијентом
- day: последњи дан контакта у месецу
- month: последњи месец контакта у години
- duration: трајање последњег контакта, у секундама
- campaign: број контаката извршених током ове кампање за овог клијента
- pdays: број дана који су прошли од последњег контакта са клијентом у претходној кампањи (-1 значи да клијент претходно није контактиран)
- previous: број контаката извршених пре ове кампање за овог клијента client
- poutcome: исход претходне маркетиншке кампање

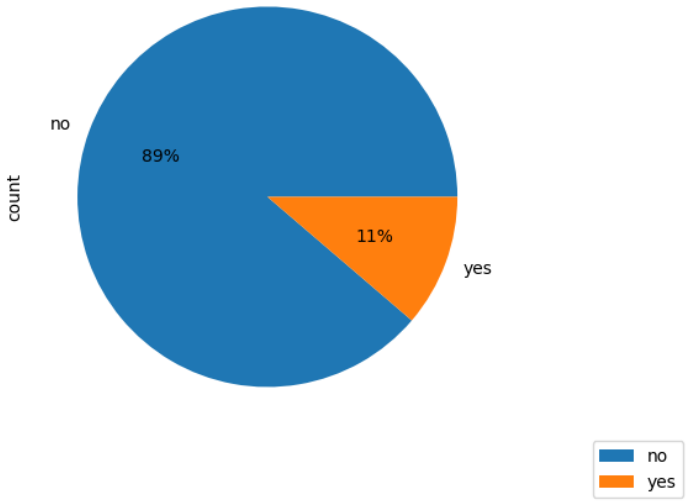
Таргет:

- y – да ли се кријент пријавио за штедни улог? ("yes","no") – бинарна класификација

- 11 integer атрибута: age,balance,duration,campaign,pdays,previous,emp.var.rate,cons.price.idx,cons.conf.idx ,euribor3m,nr.employed
- 5 категоричких атрибута: job,marital,education,contact,poutcome
- 4 бинарна атрибута: default,housing,loan,y
- 2 date атрибута: month,day_of_week

На основу дескриптивне анализе на почетку можемо да закључимо да имамо доста object атрибута и да би применили алгоритме машинског учења мораћемо да их претворимо у нумеричке. Такође што се дупликата тиче имамо их 12 и они су обрисани,недостајућих вредности нема.

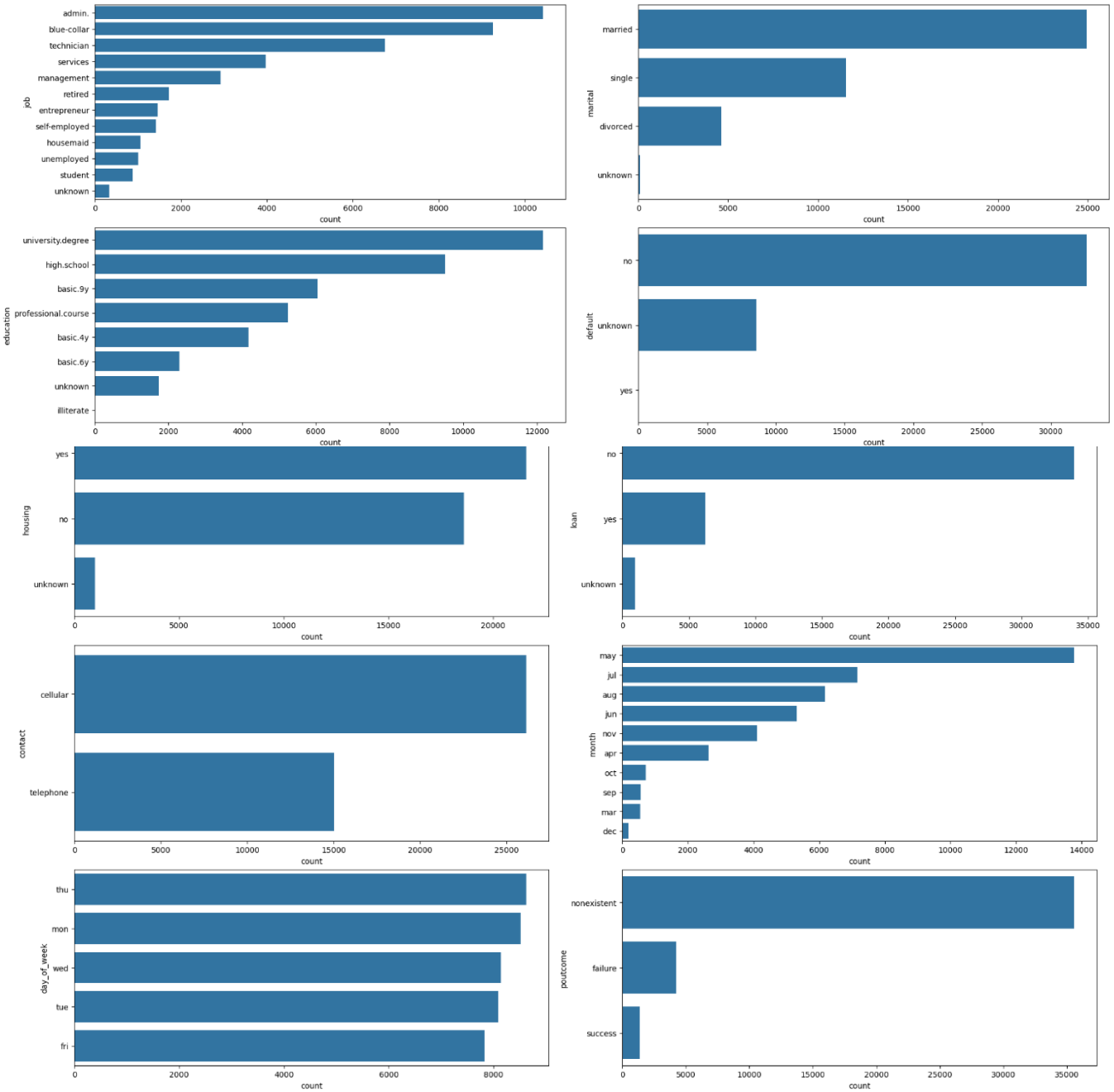
Што се балансираности скупа податак тиче видимо на слици 24. да је небалансиран.



Слика 24. Балансираност скупа података

Видимо да имамо много већи број примерака који припадају класи не. На основу овога можемо претпоставити да ће вероватно бити потребно коришћење неке од доступних техника којом ће се извршити балансирање датасета.

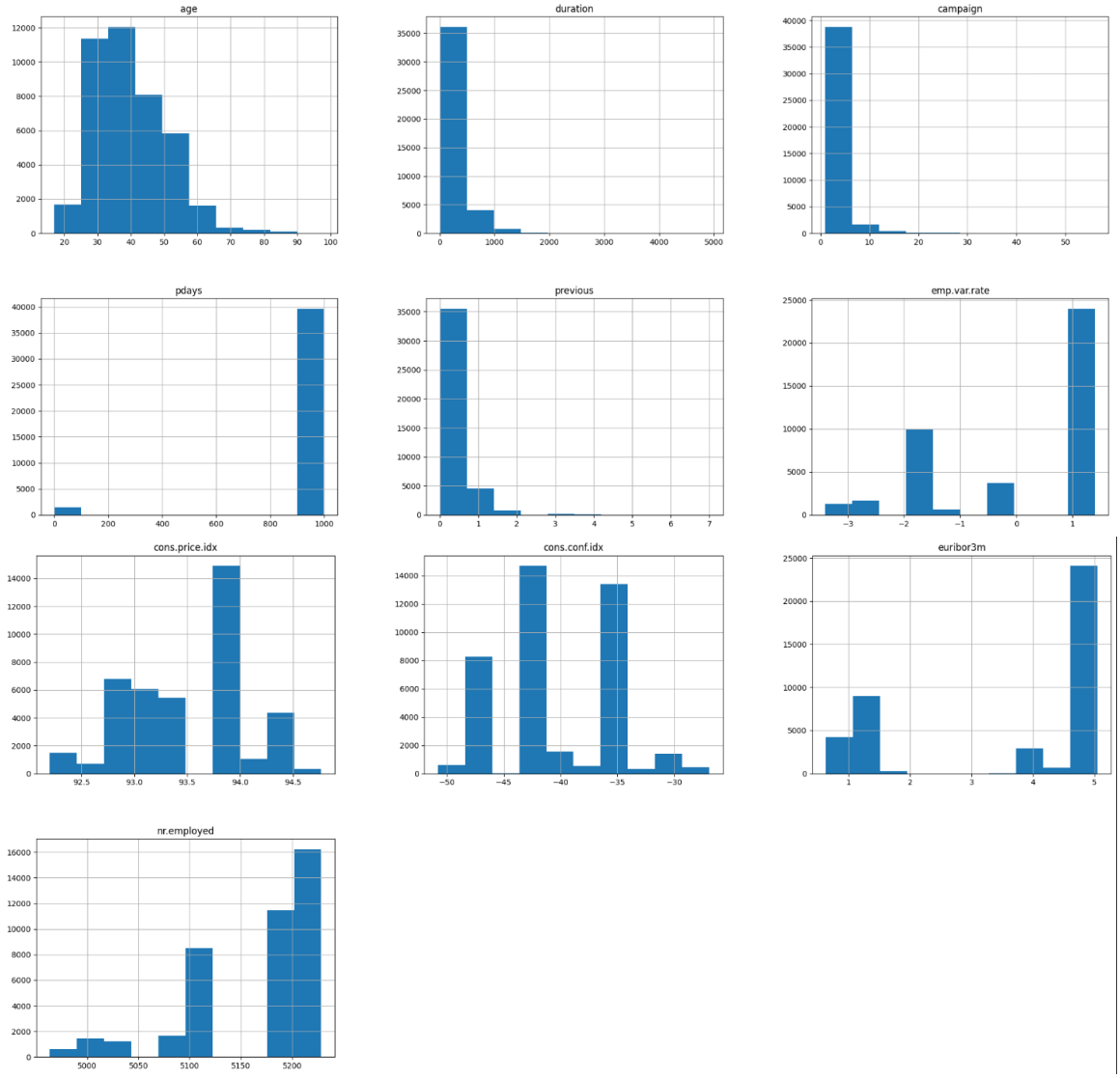
Затим на следећој слици видимо расподелу категоричких атрибута:



Слика 25. Расподела категоричких атрибута

На основу хистограма се може закључити да неке од колона садрже веома мали број примерака одређених вредности као и недостајуће вредности.

Затим погледаћемо расподелу нумеричких атрибута:



Слика 26. Расподела нумеричких атрибута

Примећујемо да неке од колона као што су duration, campaign, pdays и previous имају веома лошу расподелу вредности, а да колона age садржи доста примерака ван Гаусове расподеле.

Затим како бисмо применили алгоритме за класификацију података,мораћемо да категоричке атрибуте пребацимо у нумеричке,то је урађено на следећи начин:

- За почетак, можемо приметити да колона default скоро да нема ниједан примерак са вредношћу yes, а готово сви остали примерци (око 80%) имају непознате вредности. Закључујемо да ова колона, из наведених разлога, не носи никакве информације од значаја и због тога је можемо избацити.

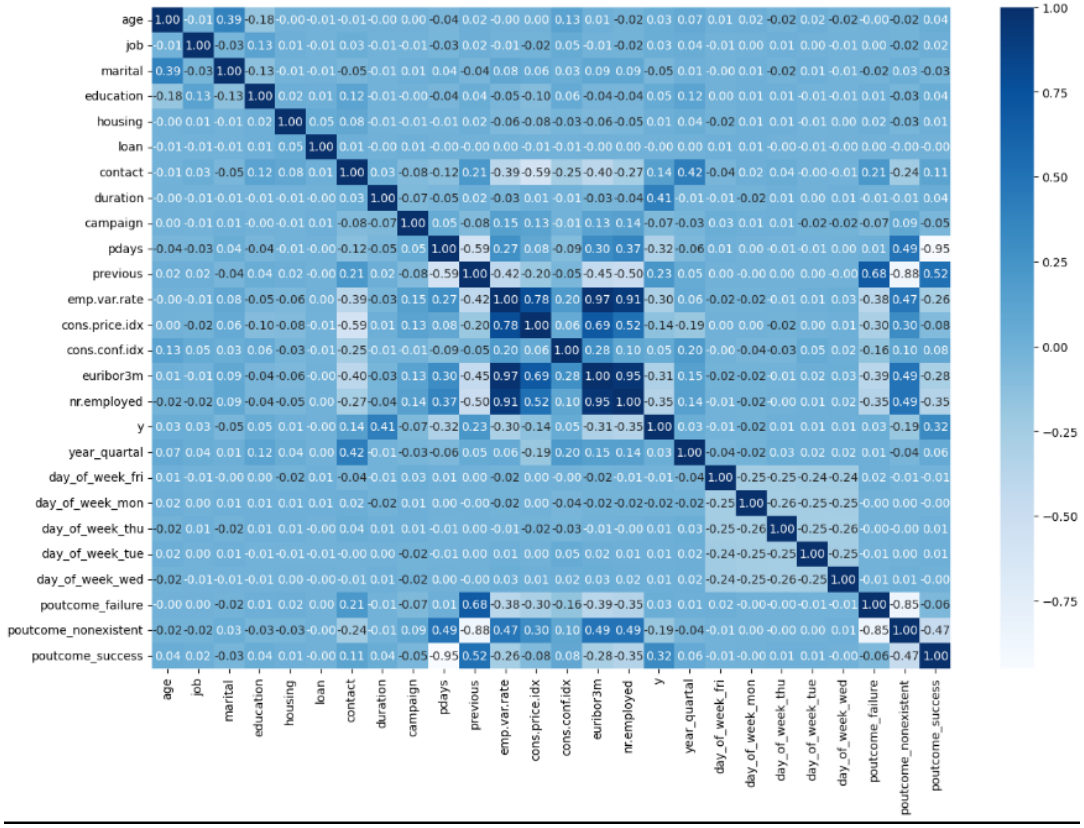
- Обзиром да колона education има веома мали број примерака са вредношћу illiterate, свим овим примерцима можемо приписати вредност basic.4y, јер је основно четворогодишње образовање следећи најближи степен образовања.

- Обзиром да неки од месеца у години имају веома мали број примерака (посматрајући колону month) у односу на остале, уместо чувања података о конкретном месецу можемо чувати податке о кварталу године. У ту сврху, уводимо нову колону year_quartal, а колону month можемо избацити.

- Што се осталих категоричких колона тиче, можемо извршити енкодирање свих вредности.

- На крају,попуњавање непознатих вредности методом KNNImputer

Затим на основу матрице корелације можемо да видимо зависности између атрибута у скупу података:



Слика 27. Матрица корелације

Закључујемо да су високо корелисани атрибути: euribor3m,emp.var.rate и nr.employed

Варијанса података је је највећа за атрибуте: age,nr.employed,pdays и duration те се стога мора водити рачуна да не дође до overfitting-а приликом креирања модела машинског учења.

8.4 Закључак над другим скупом података

Као и у прошлом скупу података,и у овом су примењена два алгоритма машинског учења за решавање проблема класификације: SVC и Random Forest.

Алгоритми су примењени над различитим скуповима података:

- Подела одновног скупа података
- Подела скупа на основу балансираности (примењена SMOTEEN метода)
- Подела скупа на основу расподеле података (над атрибутима campaign и duration је примењена логаритамска трансформација,бинарно су кодирани pdays и previous,атрибут poutcome_nonexistent је отклоњен јер су већина вредности непознате)
- Подела скупа на основу outliera (примењена IQR метода)
- Подела скупа на основу матрице корелације (уклоњени висококорелисани атрибути: emp.var.rate и nr.employed,и нискокорелисан: pdays)

Добијени су следећи резултати:

Tip podele	SVC accuracy	Random Forest accuracy
Osnovni skup	0.90	0.92
Balansirani	0.91	0.98
Raspodela podataka	0.89	0.91
Outlieri	0.90	0.91
Korelacija	0.89	0.92

Слика 28. Табела са резултатима за други скуп података

На основу мањих варијација у резултатима за оба алгоритма, можемо закључити да је овај скуп података квалитетнији. Конзистентност у резултатима указује на то да оба модела добро раде на овом датасету, што може бити резултат квалитетнијих и уједначенијих података.

9 Закључак

Разматрајући суштинске концепте квалитета података и процеса припреме података у контексту машинског учења, ово истраживање истиче кључне елементе неопходне за успешно развијање модела. Квалитетни подаци су основни темељ који омогућава изградњу поузданих и прецизних модела машинског учења, а њихова анализа пружа кључне увиде за постизање ефикасних решења.

Прва фаза у процесу припреме података подразумева детаљну анализу улазног скупа података ради утврђивања његових квалитета и репрезентативности.Одређени предуслови које подаци морају да испоштују укључују: испитивање мера квалитета података, анализа расподеле података, корелације и варијансе. Ови предуслови представљају првобитну информацију о могућностима које се могу касније спровести над посматраним скупом података.

Правилна анализа расподеле података омогућава идентификацију и третирање outlier-а, што значајно утиче на стабилност и поузданост модела. Осим тога, анализа расподеле података омогућава процену балансираности датасета, што је кључно за квалитетну класификацију.

Варијанса и корелација представљају статистичке концепте и чине битну улогу у процесу анализе података код машинског учења. Варијанса се односи на меру дисперзије података и примењује се код процене стабилности улазних података над којима се модел обучава. Корелација се користи код извођења међусобних зависности између појединачних фичера у посматраном скупу података и ближе објашњава условљености које делују између података.

За даљи развој модела, кључно је одабрати и одговарајуће алгоритме машинског учења, узимајући у обзир специфичности проблема и структуре података.

Интеграција теоријских концепта квалитета података са практичним приступом анализе и припреме података представља кључну компоненту успешне примене машинског учења у решавању различитих проблема. Схватање и примена ових принципа су од суштинског значаја за креирање поузданих модела са прецизним предикцијама, што отвара могућности за примену машинског учења у различитим областима и доменима.

10 Литература

[1] Galli, S., Python Feature Engineering Cookbook: Over 70 Recipes for Creating, Engineering, and Transforming Features to Build Machine Learning Models,2020

[2] Chris Albon,Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning,2018

[3] Julian Avila,Scikit-Learn Cookbook: Over 80 Recipes for Machine Learning in Python With Scikit-Learn,2017

[4] Salvador García , Julián Luengo and Francisco Herrera, Data Preprocessing in Data Mining, 2015

[5] Jason Brownlee, Data Preparation for Machine Learning: Data Cleaning, Feature Selection and Data Transforms in Python, 2020

[6] Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques, Third Edition, 2012, Morgan Kaufmann

[7] https://firsteigen.com/blog/6-key-data-quality-metrics-you-should-be-tracking/

[8] https://www.talend.com/resources/machine-learning-data-quality/

[9] https://www.wolfram.com/language/introduction-machine-learning/distribution-learning/

[10] https://www.scribbr.com/statistics/central-tendency/

[11] https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php

[12] https://www.wallstreetmojo.com/correlation-matrix/

[13] https://www.scribbr.com/statistics/variance/

[14] https://medium.com/@abdallahashraf90x/probability-distributions-statistics-for-machine-learning-69d0a0f21253