Универзитет у Београду – Електротехнички факултет

Катедра за сигнале и системе

Дипломски рад

Класификација руком писаних цифара применом статистичких метода

Ментор:

Др Жељко Ђуровић, редовни професор

Кандидат:

Урош Војичић, 2018/0202

Београд, октобар 2025. године

САДРЖАЈ

1.Увод

У савременом добу подаци представљају један од највреднијих ресурса за развој интелигентних инжењерских решења која проширују могућности рачунара у готово свим областима људске делатности. Њихов значај у технолошком развоју може се упоредити са утицајем који је у историји развоја индустријске производње, па самим тим и развоја цивилизације, имао проналазак парне машине. Док су рачунари у почетку могли само да извршавају унапред дефинисане алгоритме, развој њихових меморијских капацитета омогућио им је да имитирају људску способност учења из искуства, што је довело до настанка савремених система машинског учења. Подаци, као дигитално сачуване опсервације, чине основу овог процеса, јер омогућавају рачунарима да из сирових података извлаче информације, препознају обрасце и самостално доносе одлуке. Захваљујући томе, подаци (eng. data) данас представљају кључни покретач технолошког напретка, па се с правом сматрају једним од најважнијих ресурса савременог друштва. Управо се способност препознавања ових образаца сматра суштином интелигентног понашања, а њена примена у техничким системима позната је као *препознавање облика* (eng. *pattern recognition*).

Да бисмо разумели проблем препознавања облика и класификације, потребно је најпре разумети шта представља појам *обележја* (енг. *feature*). Обележје представља мерљиву карактеристику објекта или појаве из реалног света, која се користи за њихово описивање и разликовање. Кроз обележја апстрахујемо стварне појаве тако што њихове кључне особине нумерички изражавамо и смештамо у *вектор обележја*, који се често назива и *облик*. На тај начин реалне појаве преводимо у математички домен погодан за обраду у рачунару. Алгоритми класификације и кластеризације користе управо ове векторе обележја као улазне податке, где сваки вектор представља један узорак. Квалитет екстракције обележја има пресудан утицај на успешност класификатора, јер неинформативна обележја могу ограничити његову тачност без обзира на сложеност саме методе. У традиционалним системима избор обележја врши инжењер, док савремене методе, као што је *deep learning*, омогућавају неуралним мрежама да самостално откривају и “уче” обележја из података.

Проблем препознавања облика заснива се на класификацији посматраних објеката или појава на основу скупа њихових обележја. Коришћењем вектора обележја, сложени реални подаци, било да су у облику сигнала, слика, звука или других физичких величина, представљају се у компактном облику погодном за статистичку анализу и класификацију. Квалитет и избор обележја имају пресудан утицај на успешност система, јер одређују степен у коме се класе могу раздвојити у простору обележја. Као један од честих примера проблема препознавања облика коришћењем савремених рачунара и метода машинског учења јесте управо проблем препознавања знакова (енг. *character recognition*).

Препознавања знакова један је од најстаријих и најистраженијих задатака у области рачунарског вида (енг. *computer vision*) и машинског учења (енг. *machine learning*), чији корени датирају још од средине 20. века, када су се појавили први покушаји аутоматског читања рукописаних и штампаних симбола. Први системи, развијани педесетих и шездесетих година, били су засновани на једноставним правилима и шаблонском упоређивању облика, али су брзо показали ограничења у суочавању са варијацијама у рукопису, фонтова, оријентацији и квалитету слике. Са појавом статистичких метода седамдесетих и осамдесетих година, као што су Бајесови и линеарни дискриминантни класификатори, проблем је добио формални математички оквир који је омогућио поузданије моделирање различитих класа знакова кроз њихова мерљива обележја. Увођење алгоритама машинског учења и неуронских мрежа током деведесетих година донело је још већи напредак, омогућивши системима да „уче“ из великих скупова примера и да сами откривају најрелевантније карактеристике које разликују поједине знакове. Данас, са развојем дубоког учења и конволуционих неуронских мрежа, системи за препознавање знакова достижу тачности блиске људском нивоу, чак и у условима велике варијабилности, буке и деформација.

Савремене технологије препознавања знакова имају огромну практичну вредност и налазе примену у готово свим сегментима друштва и индустрије. Аутоматско читање рукописа и штампаних докумената (OCR) омогућило је дигитализацију огромних архива и убрзало административне процесе. У банкарству се користи за читање бројева чекова, док у поштанским системима служи за аутоматско препознавање адреса. У саобраћају ове технологије омогућавају препознавање регистарских таблица, а у мобилним апликацијама омогућавају скенирање рачуна, цена или текстова у реалном времену. У индустрији се системи за препознавање знакова примењују у роботима за визуелну контролу квалитета, у интелигентним сензорима и у системима за аутоматизацију производње. Поред бројних практичних примена, развој метода за препознавање знакова имао је велики утицај и на поље вештачке интелигенције. Управо кроз овај проблем постигнут је значајан напредак у разумевању начина на који машине могу да уче, генерализују и препознају структуру у подацима, што представља суштину интелигентног понашања.

Проблем препознавања цифара може се решавати применом различитих приступа. Традиционалнији приступи биле би параметарске и непараметарске методе статистичке класификације облика, као и савремене методе дубоког учења (енг. *deep learning*), најчешће засноване на конволуционим неуронским мрежама (CNN). У овом раду биће реализована по једна метода из ове две групе, линеарни Бајесов класификатор као представник параметарских и kNN класификатор као представник непараметарских метода. Ови класификатори одабрани су због своје једноставности, ефикасности и добрих перформанси у задацима препознавања облика са ограниченим бројем обележја. Циљ рада је пројектовање ефикасних класификатора и поређење њихових перформанси и метрика.

2.Предобрада слика

Један од кључних предуслова за успешно функционисање техника машинског учења јесте употреба добро припремљених, адекватно обележених (*labeled*) и стандардизованих улазних података. Квалитет улазних података директно се одражава на квалитет резултата. Уколико су подаци непотпуни, неконзистентни или шумовити, и резултати класификације биће непоуздани и нетачни. Због тога је фаза предобраде и стандардизације базе података од изузетне важности, јер обезбеђује да се сви узорци доведу у јединствен и репрезентативан облик погодан за даљу анализу и учење модела.

У контексту проблема класификације руком писаних цифара, улазни подаци представљају слике појединачних цифара које су оригинално написане на папиру и потом дигитализоване скенирањем. Свака цифра се налази на засебној слици у `.png` формату. Међутим, овако добијене слике често нису идеалног квалитета, тачније могу имати смањену оштрину, различите нивое осветљења, додатни шум који уноси скенер. Битан фактор су такође варијације у дебљини и интензитету потеза које настају услед употребе различитих оловки, хемијских оловки итд. Сви ови фактори уносе неконзистентност у податке, што може значајно умањити тачност и поузданост класификације.

Из тог разлога, извршена је предобрада слика која обухвата примену више техника дигиталне обраде слике са циљем побољшања квалитета и стандардизације узорака. Развијен је и примењен секвенцијални процес (енг. *pipeline*) који је у потпуности аутоматизован и примењен на целокупан скуп података, тако да свака слика пролази кроз идентичне кораке обраде.

- Гаусово замућење (енг. *Gaussian blur*)
- Бинаризација
- Морфолошке операције дилатација и ерозија
- Одсецање вишка белих пиксела око цифре (енг. *Margin crop*)
- Стандардизација димензија слике
- Морфолошка операција скелетонизација

У наредним поглављима биће речи о свакој од коришћених техника дигиталне обраде слике. За потребе приказивања ових техника, коришћен је насумично одабрани пример слике цифре ‘5’ (пет).

2.1 Gausovo zamucenje slike

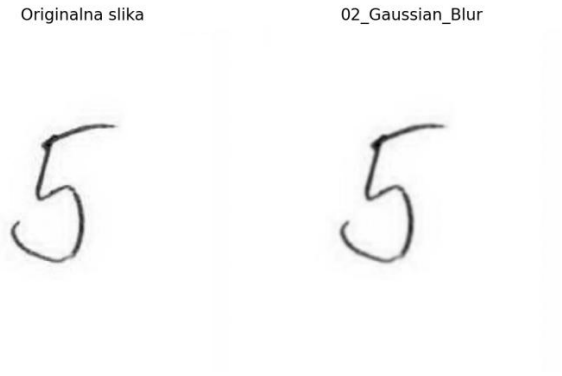
Гаусово замућење слике користи се за потребе смањења шума и детаља на слици. Оно делује као нископропусни (NF) филтар, одсецајући и ублажавајући високофреквентне компоненте у виду шума и ситних „оштрих“ промена. Поред уклањања шума, овај поступак омогућава глатко прелазење између различитих делова слике, чиме се добија визуелно мекша и природнија структура. Због својих својстава, Гаусово замућење често се користи као корак предобраде у алгоритмима рачунарског вида, на пример пре детекције ивица, сегментације или издвајања обележја, јер су те технике јако осетљиве на утицај шума.

Математички посматрано, поступак Гаусовог замућења слике представља конволуцију слике са Гаусовом функцијом расподеле. Сваки пиксел у новој (замућеној) слици добија вредност која је пондерисани просек вредности његових суседних пиксела, где је тежина сваког суседа одређена Гаусовом функцијом.

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

$$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

На тај начин, оштре промене и високофреквентне компоненте у слици се пригушују, док се прелази између различитих региона слике ублажавају. Због малих димензија слика, коришћен је структурни елемент (енг. kernel) величине 3×3 како би се сачувале локалне структуре и избегло прекомерно замућење или губитак детаља.

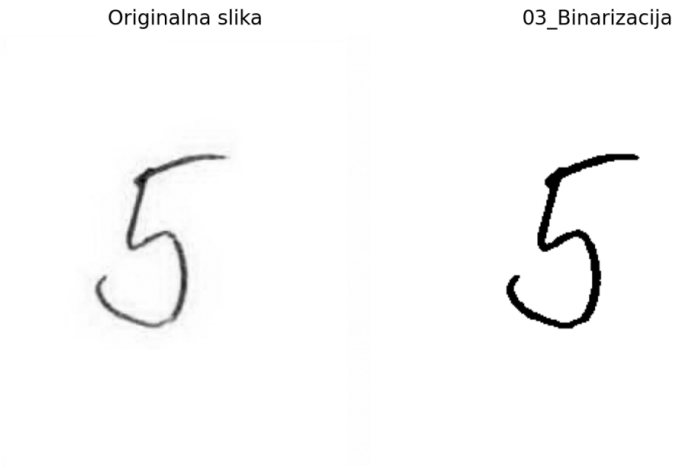


2.2 Binarizacija slike

RGB slika se moze predstaviti sivom bojom kao kombinacija crvene, zelene I plave. Jedna od glavnih karakteristika svake slike je broj nijansi sive boje. On se cesto moze predstaviti kao 2^b (ispravi me ako gresim), gde je b- broj bita za predstavljanje sive. Najcesce je to opseg celobrojnih vrednosti od 0 do 255, gde 0 predstavlja crnu boju, a 255 belu boju.

Binarizacija predstavlja smanjenje broja nijansi sive na 2, tacnije samo na crnu I belu boju. Ona se vrsi tako sto se postavlja prag (eng. *threshold*), tako da svi pikseli slike veceg intenziteta od vrednosti praga uzimaju max vrednost 255, tacnije belu boju, dok svi pikseli nizem intenziteta od praga uzimaju vrednost 0, tj. Crnu boju. Na taj nacin, slika postaje crno bela, I dolazi do gubitka informacija.

Tehnika binarizacije je veoma korisna za potrebe ovog rada, s obzirom na to da kasnija statisticka analiza postaje mnogo manje racunski zahtevna, jer slika nosi daleko manju kolicinu informacija.



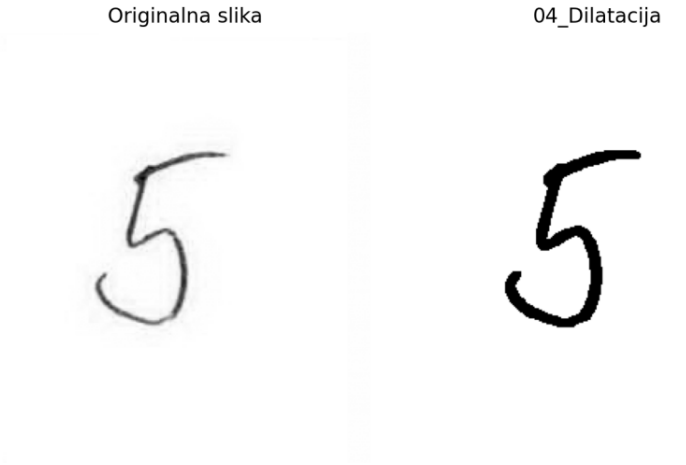
2.3 Morfoloske operacije dilatacija I erozija

Cesto se desava da nakon binarizacije slike ostanu nezeljeni fragmenti pozadine koji nisu deo glavnog objekta. Takodje, cesto se desi da zidovi samog glavnog objekta, tj. Cifre, imaju rupe nastale usled slabe olovke, ili jednosatvno loseg odabira praga binarizacije. Za popravke ovako nastalih problema najcesce se mogu koristiti morfoloske operacije poput dilatacije I erozije.

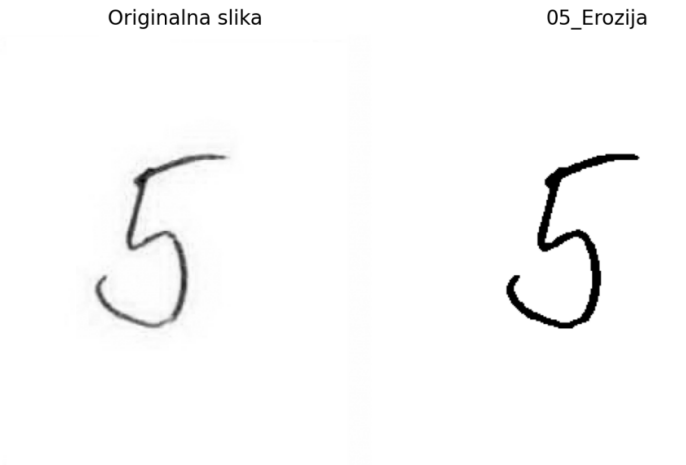
Дилатација и ерозија представљају две међусобно дуалне морфолошке операције које се користе у обради бинарних слика ради обликовања и анализе структура објеката. Њихов однос је комплементаран. Дилатација проширује објекте додавањем пиксела на њихове ивице, ерозија их сужава уклањањем пиксела са граница. У комбинацији, ове операције омогућавају извођење сложенијих

поступака као што су отварање (ерозија па дилатација) и затварање (дилатација па ерозија).

Дилатација проширује објекте у слици тако што додаје пикселе на њихове границе. Користи се за попуњавање малих празнина, спајање прекинутих делова и згушњавање танких линија. За дати пиксел, резултат дилатације је 1 (бело) ако било који пиксел под структурним елементом има вредност 1, чиме се објекат шири у свим правцима од своје границе. Ова операција омогућава обнављање непотпуних или оштећених делова објеката и побољшава њихову повезаност у слици.



Ерозија има за циљ да смањи објекте у слици тако што уклања пикселе са њихових граница. Она може елиминисати танке линије, мале артефакте и шумове, или раздвајати спојене елементе. За сваки пиксел у слици, ерозија враћа вредност 1 (бело) само ако сви пиксели под структурним елементом такође имају вредност 1; у супротном, пиксел постаје 0 (црн). На тај начин, ерозија смањује величину објеката и наглашава њихову унутрашњу структуру, док уклања ситне неправилности на површини.



2.4 Odsecanje margina

Za potrebe standardizacije podataka, neophodno je odseci bele margine sa slika koje predstavljaju belu pozadinu iznad, ispod cifre, kao I sa leve I desne strane. Ideja je da slika cifre sa svake strane pocinje crnim pikselom. Ovo se radi zbog toga sto bi razlicite velicine margina unosile nekonzistentnost u izdvojena obelezja, samim tim bi performance klasifikacije bile losije.

2.5 Smanjenje dimenzije slike

Nezeljena pojava odsecanja margina je to da dobijene slike nisu istih dimenzija, sto je glavni preduslov za dalji rad sa modelima masinskog ucenja. Iz tog razloga neophodno je standardizovati velicinu slika. Slike se ‘razvlace’ na zeljenu dimenziju. U strucnoj literaturi to su najcesce dimenzije 28x28 piksela. U ovom radu slike su standardizovane na velicinu 32x32 piksela. Ovom tehnikom se gubi odnos sirine I visine slike originalne slike, ali se poboljsava uticaj nekih obelezja u kojima ce biti vise reci u sledecem poglavlju. Ovo predstavlja neki vid kompromisa izmedju dva veoma cesto koriscena obelezja. Ovako znacajno smanjenje dimenzija slike (eng. downsampling) povlaci sa sobom pikselizaciju pri vizuelizaciji slike, sto se moze primetiti na sledecem prikazu.

2.6 Morfoloska operacija skeletonizacija

Skeletonizacija (engl. *skeletonization*) је процес који има за циљ да редукује објекат у бинарној слици на његову скелетну основу, задржавајући при томе основни облик и повезаност региона, док се већина оригиналних пиксела уклања.

Да би се то интуитивно разумело, можемо замислити да су региони предњег плана направљени од неког равномерног, споро сагоривог материјала. Ако се истовремено запали ватра дуж читаве ивице тог региона, пламен ће почети да се креће ка унутрашњости. У тачкама где се пламенови који полазе са различитих страна сретну, они ће се угасити , а линије дуж којих се то догоди називају се *линијама гашења*. Управо те линије представљају скелет слике.

Математички се може посматрати као низ морфолошких операција које конзистентно смањују бинарни објекат до његове скелетне форме, односно локације центра великих дискова уписаних унутар објекта. Формално, нека је $X \subseteq Z^2$ бинарни објекат, и нека је B структурни елемент (енг. *kernel*). Онда скуп скелетских подскупова може бити дефинисан као:

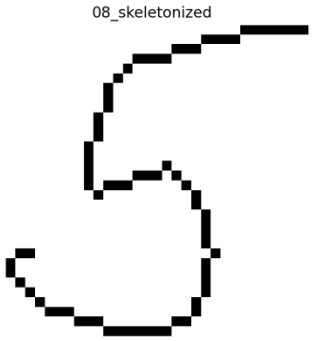
$$S\_n\,(X) = \{\, z \in X \mid\; z + nB \;\subseteq \mathrm{X} \cap \mathrm{z} + (\mathrm{n}+1)\mathrm{B} \nsubseteq \mathrm{X}\}$$

Где nB представља структурни елемент B примењен n пута, тј. облик добијен након n-тоструке дилатације или ерозије, у зависности од контекста операције. Објекат се затим може реконструисати као:

$$X = \bigcup_{n=0}^{\infty} (S_n(X) \oplus \mathrm{nB})$$

где $\oplus$ означава морфолошку дилатацију. Алгоритамски се поступак може извршавати тако што се итеративно уклањају пиксели са границе објекта све док се не задрже само тачке које имају најмање два најближа суседа на граници. Овим приступом значајно се редукује број пиксела, али се задржава већина информација од значаја.

На крају низа примена техника дигиталне обраде слике добија се финална верзија обрађене цифре.



3.Извлачење обележја

Да бисмо приступили проблему препознавања облика и изградњи ефикасних класификатора, неопходно је најпре разумети суштину појма **обележја** (енг. *feature*), јер она представљају основу сваког система за анализу и обраду података. Обележје се може дефинисати као мерљива карактеристика или својство неког објекта или појаве које омогућава њено описивање, поређење и разликовање од других. У процесу екстракције обележја, стварни објекти и појаве апстрахују се у нумерички облик тј. вектор обележја, који садржи информације релевантне за задати проблем. На овај начин сложени реални подаци, као што су слике, звуци или сигнали, преводе се у математички домен погодан за рачунарску анализу.

Алгоритми класификације и кластеризације не раде директно над сировим подацима, већ управо над овим векторима обележја, где сваки вектор представља један узорак из посматраног скупа. Због тога је поступак екстракције и одабира обележја кључан корак у сваком систему препознавања облика. Уколико изабрана обележја нису довољно информативна, чак и најнапреднији класификатори неће моћи да постигну високу тачност.

У традиционалним приступима овај избор врши инжењер који дефинише обележја на основу доменског знања и искуства. Особа која поседује дубоко разумевање конкретног проблема и може ефикасно да изабере или формулише релевантна обележја назива се експерт Улога експерта је од кључног значаја јер управо његово знање омогућава да се из великих количина података издвоје информације које најбоље описују суштину посматране појаве.

Идеалан скуп обележја треба да садржи што мањи број параметара који носе велику количину информација и који нису међусобно зависни. Степен информативности и међусобне повезаности обележја може се проценити анализом корелационе матрице, која приказује однос корелације између сваког пара обележја. Вредности корелације крећу се у интервалу од –1 до 1, где су екстремне вредности показатељ снажне линеарне везе између обележја. За потребе ефикасне класификације пожељно је одабрати обележја која имају ниску међусобну корелацију, јер се на тај начин избегава редундантност информација и постиже боља дискриминативност између класа. Уколико се међу изабраним обележјима појаве високо корелисана обележја, у пракси се најчешће једно од њих изоставља из скупа, како би се избегла редундантност информација. Алтернативно, могуће је извршити

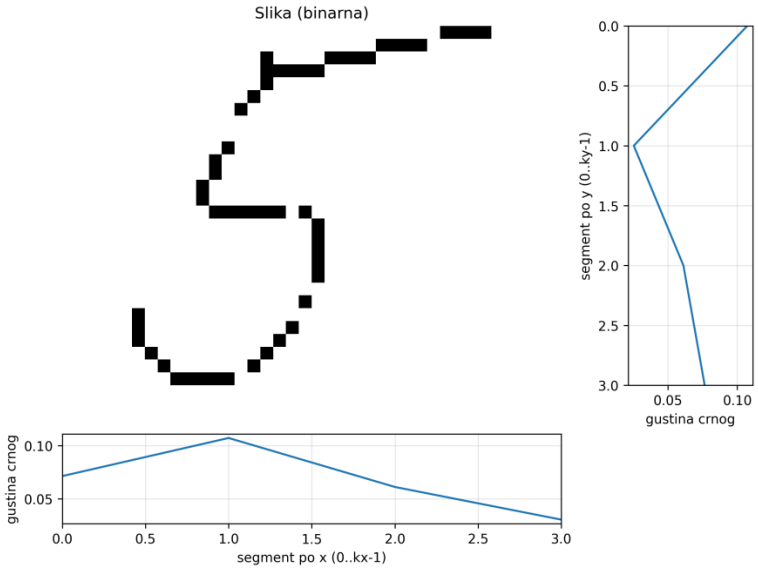
линеарну комбинацију таквих обележја и на тај начин добити ново обележје које задржава највећи део информација.

Приликом дефинисања скупа обележја за проблем препознавања цифара, полазило се од претпоставке да се различите цифре могу међусобно разликовати на основу расподеле њихових црних пиксела у различитим деловима слике, као и геометријских карактеристика облика. Цифре, иако припадају истом скупу симбола, значајно се разликују по просторној расподели пиксела. Неки облици су изражено издужени по вертикали, као што су „1“ и „7“. Друге цифре, попут „2“ и „4“, имају наглашенију хоризонталну компоненту. Постоје и симетричнији облици, као што су „0“, „8“ и „3“, чији распоред пиксела равномерније заузима површину слике.
Због тога су одабрана следећа обележја која заједно чине вектор обележја:

1. srednja vrednost gustine crnih piksela po horizontalno izdeljenim segmentima slike
2. varijansa gustine crnih piksela po horizontalno izdeljenim segmentima slike
3. varijansa gustine crnih piksela po vertikalno izdeljenim segmentima slike
4. moment slike
5. broj horizontalnih prelaza
6. broj vertikalnih prelaza
7. radijalna udaljenost centra mase od centra slike

3.1 Густине црних пиксела по сегментима слике

Prve 2 vrednosti odnose odnose se na gustinu crnih piksela. Slika se izdeli na zadati broj horizontalnih segmenata iste povrsine. Za potrebe ovog rada slika je bila izdeljena na 5 segmenata. Gustina crnih piksela jednog segmenta racuna se kao broj crnih piksela podeljen ukupnim brojem piksela u tom segmentu. Srednja vrednost I varijansa se racunaju za ovih 5 segmenata. Po istom principu se dobijaju obelezja 3 I 4, samo za vertikalno izdeljenu sliku.



3.2 Момент слике

Ово обележје се заснива на израчунавању геометријских момената слике, који описују распоред масе црних пиксела у односу на њихов центар. Прво се одређује центар масе (тежиште) бинарне слике на основу координата свих пиксела који припадају објекту. Затим се рачунају централни моменти другог реда — μ_{20}, μ_{02} и μ_{11} — који представљају мере распршености пиксела око центра по хоризонталној, вертикалној и дијагоналној оси. Комбинацијом ових вредности добија се једна бројчана мера која описује укупну распрострањеност и оријентацију облика у односу на његов центар.

Нека је B(x,y) бинарна слика, где пиксели који припадају цифри имају вредност 1 (црни пиксели), а позадина 0. Центар масе (тежиште) објекта дефинише се као:

$$c_x = \frac{\sum_{x,y} x\, B(x,y)}{\sum_{x,y} B(x,y)}\; c_y = \frac{\sum_{x,y} y\, B(x,y)}{\sum_{x,y} B(x,y)}$$

Централни моменти другог реда израчунавају се као:

$$\mu_{20} = \sum_{x,y} (x - c_x)^2 B(x,y)\,, \mu_{02}$$

$$= \sum_{x,y} \left(y - c_y\right)^2 B(x,y)\,, \mu_{11} = \sum_{x,y} (x - c_x)(y - c_y)\;\; B(x,y)\,,$$

Момент форме (комбинована мера распршености) добија се као:

$$f_{moment} = \frac{\sqrt{\mu_{20}^2 + \mu_{02}^2 + 2\mu_{11}^2}}{M^2}$$

Где је

$$\mathrm{M} = \sum_{x,y} B(x,y)$$

Укупна маса објектна, односно број пиксела који припадају цифри.

Овакво обележје је изабрано јер пружа информацију о компактности и симетрији облика, што је веома значајно код препознавања цифара. На пример, цифре као што су „0“ или „8“ имају веома уједначен и симетричан распоред пиксела око центра, док су „1“ и „7“ издужене и асиметричне. Због тога ова мера доприноси разликовању цифара које имају сличну густину пиксела, али различиту геометријску структуру. Већа вредност момента указује на издуженији и асиметричнији облик, док мања вредност означава компактнији и симетричнији распоред пиксела.

3.3 Број хоризонталних и вертикалних прелаза

Хоризонтални и вертикални прелази представљају број промена пиксела из црне у белу и обратно дуж редова и колона слике. Ова обележја описују сложеност и структурну разноликост цифре. Цифра „1“ има врло мало хоризонталних прелаза, јер већина редова садржи само једну уску вертикалну линију без значајних промена између црних и белих пиксела. С друге стране, вертикални прелази такође су ретки, јер се црни пиксели најчешће простиру континуално одозго надоле. Због тога се ова цифра одликује малим бројем промена боје по и једном и другом правцу, што је супротно од сложенијих цифара као што су „8“ или „3“, које имају више хоризонталних и вертикалних прелаза услед присуства више линија и затворених контура.

3.4 Центар масе слике

Удаљеност центра масе од центра слике представља обележје које мери колико је облик цифре „помакнут“ у односу на право средиште слике. Центар масе се одређује на основу положаја свих пиксела који припадају цифри, а затим се израчунава растојање између овог тежишта и геометријског центра слике. Добијена вредност се нормализује тако да буде између 0 и 1, чиме се омогућава поређење између различитих слика без обзира на њихове димензије.

Нека је B(x,y) бинарна слика у којој црни пиксели имају вредност 1 (припадају објекту),
а бели пиксели вредност 0 (позадина). Рачунање нормализованих координата центра масе приказано је у претходном поглављу везаном за момент слике.

Геометријски центар слике налази се у тачки (0.5, 0.5). Удаљеност центра масе од центра слике израчунава се као Еуклидско растојање између те две тачке:

$$r = \sqrt{(c_x - 0.5)^2 + \left(c_y - 0.5\right)^2}$$

Вредност *r* је нормализована у опсегу [0,1]. Што је *r* веће, то је центар масе више померен од средишта слике, што указује на асиметрију облика или његову нецентрирану позиицју.

Ово обележје је изабрано јер даје увид у положај и симетрију цифре унутар свог оквира. Центар цифара као што су „0“ или „8“ обично је близу центра слике, па оне имају малу радијалну удаљеност, док су цифре попут „6“ или „9“ асиметричне и њихов центар масе је знатно померен. На тај начин ова мера помаже у разликовању цифара које имају сличан облик, али различиту оријентацију или распоред пиксела унутар слике.

Корелациона матрица обележја може се израчунати на следећи начин. Ако су X_i и X_j вектори вредности два обележја у скупу података, тада је елемент матрице на позицији (i, j) дефинисан као:

$$\rho_{ij} = \frac{cov(X_i, X_j)}{\sigma_{Xi}\sigma_{Xj}}$$

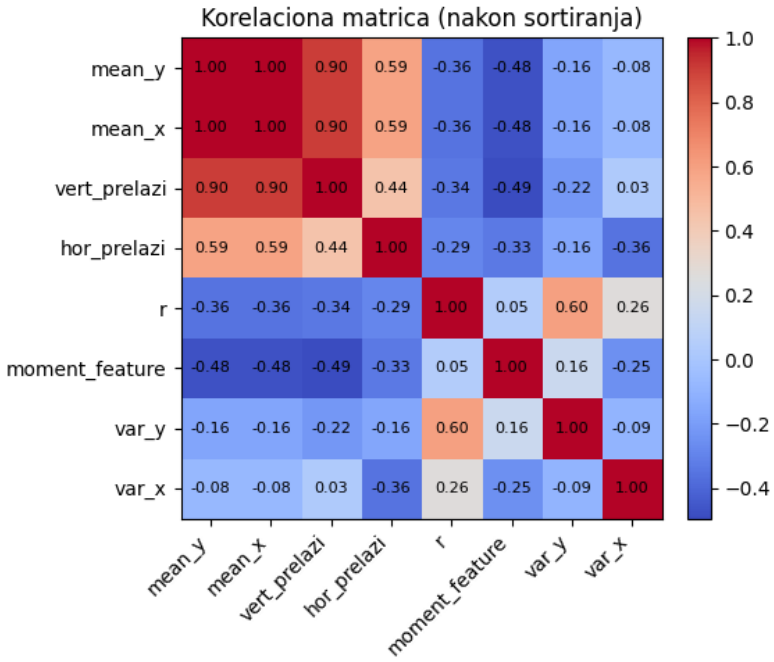
Где је

$$cov(X_i, X_j) = \frac{1}{N-1}\sum_{k=1}^{N}(x_{ki} - M_{Xi})(x_{kj} - M_{Xj})$$

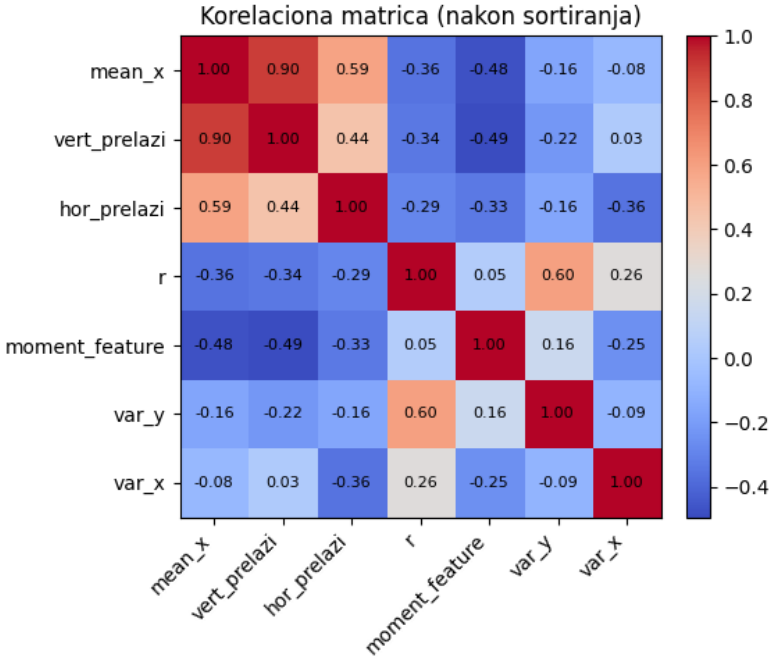
Коваријанса између обележја X_i и X_j, док су σ_{Xi} и σ_{Xj} њихове стандардне девијације, M_{X_i} и M_{X_j} средње вредности, а N број узорака. Корелациона матрица је квадратна и симетрична, димензија DxD, где D представља број обележја, а дијагонални елементи увек имају вредност 1 јер свако обележје има савршену корелацију са самим собом.

Корелациона матрица обележја представља табеларни приказ који показује степен линеарне повезаности између свих парова обележја у скупу података. Свака ћелија у матрици садржи коефицијент корелације, чија се вредност креће у интервалу од −1до 1. Вредност блиска 1 означава снажну позитивну корелацију, што значи да пораст једног обележја прати пораст другог. Вредност блиска −1 указује на негативну корелацију, односно да пораст једног обележја прати смањење другог. Вредност близу нуле означава да између обележја не постоји линеарна зависност.

У пракси, корелациона матрица служи као важан алат за анализу и избор обележја. Њена анализа омогућава откривање обележја која су међусобно снажно повезана, што указује на редундантност информација. Редундантност информација представља појаву када више обележја у скупу података носи исту или веома сличну информацију, па једно од њих не доприноси новим сазнањима већ само дуплира постојеће. Уклањањем таквих обележја постиже се једноставнији и ефикаснији модел, који боље разликује класе и избегава проблеме као што су пренаглашеност појединих карактеристика и смањење способности генерализације класификатора.

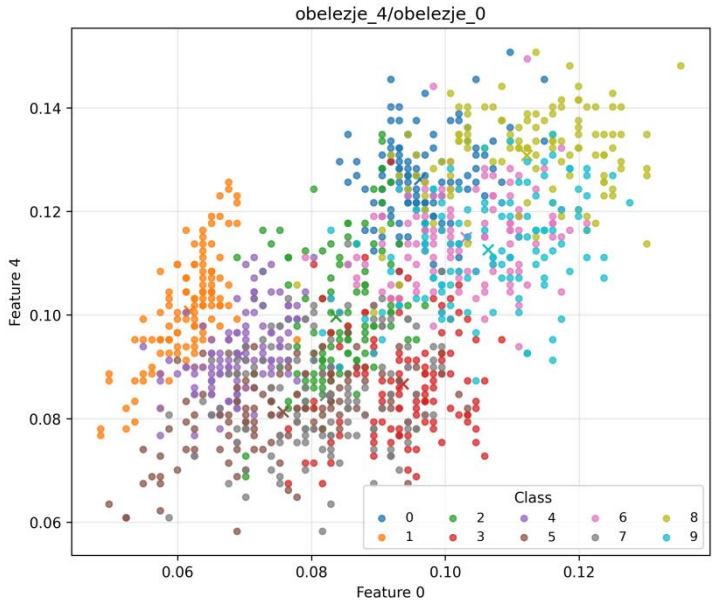


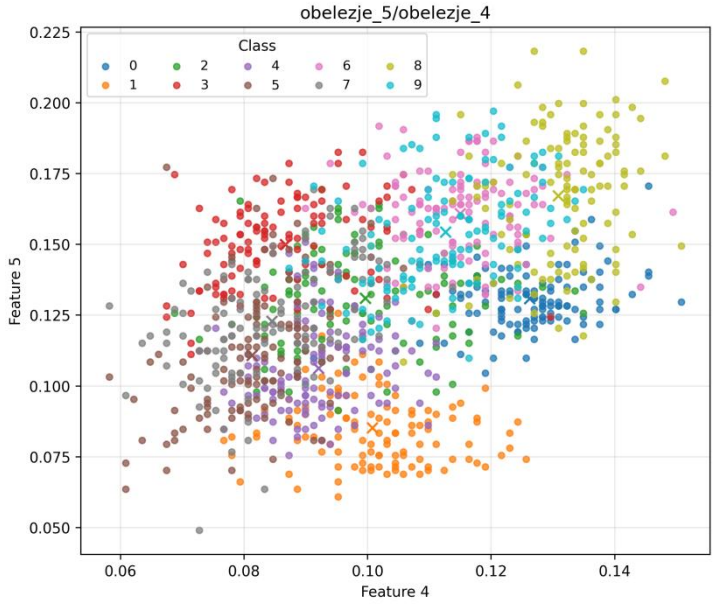
На слици је приказана корелациона матрица обележја након сортирања од највише ка најмање корелисаних. Може се уочити да средње вредности густине црних пиксела по хоризонталним и по вертикалним сегментима носе идентичну количину информација, јер се њихове корелације са свим осталим обележјима потпуно подударају. Због тога је одлучено да се из скупа обележја изостави средња вредност по вертикалним сегментима. Овиме се постиже редукција димензија са 8 на 7-димензиони вектор обележја.



Сепарабилност десет класа у простору од седам димензија, односно седам обележја, веома је тешка за интуитивну визуелизацију, јер човек може непосредно да сагледа расподелу података само у две или највише три димензије.

Због тога је за приказ одабран пример који приказује расподелу података у дводимензионалном простору за два пара обележја. На првој слици приказана је расподела свих десет класа на основу обележја која представљају густину црних пиксела у хоризонталним сегментима и број хоризонталних прелаза.





Напомена: Детаљнија расподела преосталих обележја по класама може се видети на сликама под поглављем Прилог А

4.Класификатори

Класификација представља један од кључних задатака у области анализе и обраде података, чији је циљ да се на основу познатих примера (тзв. тренинг узорака) научи правило по коме ће се нови, непознати подаци сврставати у припадајуће класе. У контексту препознавања облика, овај задатак подразумева одређивање којој класи облика или цифара припада посматрани узорак, на основу његових обележја. У зависности од приступа и нивоа претпоставки о подацима, методе класификације могу се поделити у три основне групе:

1. Методе тестирања хипотеза
2. Параметарске методе класификације
3. Непараметарске методе класификације

У овом раду, ради једноставности и јасније интерпретације резултата, одабрана су по једна метода из групе параметарских и непараметарских метода. То су линеарни Бајесов класификатор као представник параметарских и kNN класификатор као представник непараметарских метода. Ови класификатори изабрани су због своје једноставности, ефикасности и широке примене у задацима класификације слика и препознавања цифара.

4.1 Тестирање хипотеза

Технике које се заснивају на тестирању хипотеза полазе од претпоставке да су статистичке расподеле обележја унутар сваке класе потпуно познате за све могуће вредности вектора обележја. Другим речима, потребно је познавати условну функцију густине вероватноће за сваку класу $fi(X)$, као и априорну вероватноћу P_i, која одређује колико се често појављују узорци из те класе у посматраном скупу података.

У пракси, тачне функције густине вероватноће најчешће нису познате, али се уз одређене претпоставке могу естимирати на основу расположивих података. Најчешћи и уједно најједноставнији приступ јесте претпоставка да је расподела података гаусовска (нормална), што је у великом броју случајева оправдано јер се нормална расподела природно јавља у многим реалним процесима. Гаусовска расподела је дефинисана са само два параметра — средњом вредношћу и коваријационом матрицом, који се могу релативно једноставно проценити на основу података. Такође, и априорна вероватноћа појаве узорака сваке класе може се ефикасно проценити из учесталости њиховог јављања у скупу за обуку, према следећем поступку.

$$M_i = \frac{1}{N_i}\sum_{k=0}^{N_i} X_k, \;\; X_k \in \omega_i$$

$$\Sigma_i = \frac{1}{N_i - 1}\sum_{k=0}^{N\_i} (X_k - M_i)(X_k - M_i)^T, \;\; X_k \in \omega_i$$

$$f_i(X) \sim N(M_i, \Sigma_i) = (2\pi)^{-n/2}|\Sigma_i|^{-1/2} exp\left[-\frac{1}{2}(\boldsymbol{X} - M_i)^{\boldsymbol{T}}\Sigma_i^{-1}(\boldsymbol{X} - M_i)\right]$$

Где је M_i средња вредност, Σ_i kovarijaciona matrica, а f_i Gausovska raspodela klase i.

4.2 Параметарске методе

У претходном поглављу разматране су методе класификације чији је основни циљ минимизација вероватноће грешке или цене доношења одлуке. Такви приступи заснивају се на статистичком принципу тестирања хипотеза, при чему је неопходно егзактно познавање условних функција густине вероватноће за све класе које се разматрају. У пракси, међутим, ове функције ретко када су познате, већ се најчешће располаже скупом узорака који представљају емпиријске облике класа. Овај скуп, који је доступан пре или током процеса пројектовања класификатора, назива се обучавајући скуп (training set) и представља основу за процену параметара класификационог модела.

Један од начина да се препознавање облика изврши јесте да се на основу обучавајућег скупа изврши естимација условних функција густине вероватноће. Ипак, овај поступак захтева велики број података и често подразумева сложене нумеричке прорачуне, што у многим случајевима није практично. Због тога се у реалним применама често користе приближне методе класификације, које нису оптималне у строгом бајесовском смислу, али су нумерички једноставне и рачунски ефикасне.

Ове методе се називају параметарским техникама класификације, јер се заснивају на употреби ограниченог скупа статистичких параметара који описују класе — најчешће вектора математичког очекивања (средње вредности) и коваријационе матрице. На основу ових параметара, дефинишу се дискриминативне функције које омогућавају доношење одлуке о припадности узорка одређеној класи.

Међу најпознатијим представницима параметарских метода издвајају се линеарни Бајесов класификатор, квадратни класификатор, као и део-по-дело линеарни класификатор. У појединим случајевима, Бајесов класификатор се, под одређеним претпоставкама (попут једнаких коваријационих матрица), своди управо на ове једноставне линеарне или квадратне форме.

За потребе решавања проблема класификације руком писаних цифара, у овом раду је одабран линеарни Бајесов класификатор као представник параметарских метода, због своје интерпретабилности, нумеричке стабилности и доказане ефикасности у задацима препознавања облика.

4.3 Непараметарске методе

Непараметарске методе класификације представљају посебну групу техника које не захтевају претходно познавање функционалног облика густине вероватноће класа, већ настоје да је процене директно на основу доступних података. За разлику од параметарских приступа, који претпостављају одређени модел расподеле (на пример, Гаусову), непараметарске методе не намећу таква ограничења, што их чини флексибилнијим и погодним за примену у случајевима када класе немају познату или једноставну расподелу.

Основна идеја ових метода јесте да се функција густине вероватноће приближи на основу положаја и расподеле тренинг узорака у простору обележја. Одлука о припадности новог узорка одређеној класи затим се доноси применом принципа тестирања хипотеза, где се на основу процењених густина или мере сличности између узорака процењује највероватнија класа. Међу најпознатијим представницима непараметарских метода издвајају се методе засноване на кернел функцијама, које омогућавају глатко апроксимирање густине података, и метод k најближих суседа (k-Nearest Neighbors, kNN), који класификацију врши на основу близине узорка у односу на остале елементе скупа података.

4.4 Евалуација модела

За евалуацију перформанси класификатора коришћене су стандардне метрике које омогућавају кванитативну процену тачности доношења одлука. Основна мера је **тачност** (енг. *accuracy*), која представља однос броја исправно класификованих узорака и укупног броја узорака на целом тест скупу. Поред тога, израчуната је и **прецизност** (енг. *precision*) за сваку појединачну класу, која показује колико је класификатор био поуздан приликом препознавања одређене цифре, односно колики проценат узорака које је модел сврстао у ту класу заиста припада тој класи.

Прецизност по класама дефинише се као однос исправно класификованих узорака те класе и укупног броја узорака који су моделом предвиђени као та класа:

$$Preciznost = \frac{TP_i}{TP_i + FP_i}$$

Тачност модела представља удео укупно тачно класификованих узорака у односу на укупан број узорака из скупа података.

$$Tačnost = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K}(TP_i + FP_i + FN_i)}$$

Где су:

- TP_i (енг.*True Positives*) – број узорака класе *i* који су исправно класификовани
- FP_i (енг. *False Positives*) – број узорака других класа који су погрешно класификовани као класа *i*
- FN_i (енг.*False Negatives*) – број узорака класе *i* који нису препознати као та класа
- K – укупан број класа

Додатно, за визуелни приказ резултата и детаљнију анализу грешака коришћене су конфузионе матрице, које омогућавају увид у структуру погрешних класификација и показују које класе се међусобно најчешће мешају. Ова комбинација метрика обезбеђује свеобухватну процену квалитета класификатора и омогућава поређење различитих метода. Конфузиона матрица представља табеларни приказ резултата класификације, где сваки ред одговара стварним (тачним) класама, а свака колона предвиђеним класама. Елемент на *(i, j)* матрице M, означен као M_{ij} представља број узорака који припадају стварној класи i, а које је класификатор сврстао у класу j. Дијагонални елементи матрице M_{ii} представљају број тачно класификованих узорака за сваку класу, док недијагонални елементи M_{ij} представљају број погрешно класификованих узорака, односно случајеве када је класификатор заменио класу i са класом j.

4.5 Анализа главних компоненти (PCA)

Analiza glavnih komponenti (енг. PCA) je tehnika smanjenja dimenzionalnosti компоненти тј обележја koja pronalazi nove, međusobno ortogonalne ose (komponente) duž kojih podaci imaju najveću varijansu.

Историјски гледано, идеју је први formalno формулисао Karl Pirson почетком 20. века, а metod je generalizovao Harold Hoteling неколико деценија касније, i time postao standard u multivarijantnoj analizi podataka. U savremenoj praksi, PCA se koristi za kompresiju podataka, uklanjanje šuma, vizuelizaciju i pripremu podataka za dalje modele.

Напомена: U ovom radu PCA nije korišćena da poboljša klasifikaciju već isključivo kao pomoćno sredstvo: da visoko-dimenzione vektore obeležja (7 dimenzija) projektujemo u 2D prostor ради једноставније визуелизације и лакшег увида у сепарабилност класа коришћењем одређених модела.

Glavna ideja analize glavnih komponenti je da rotira koordinatni sistem tako da prva osa „uhvati“ maksimalnu varijansu podataka, druga osa maksimalnu preostalu varijansu pod uslovom ortogonalnosti na prvu, i tako redom. Ove ose su linearne kombinacije originalnih obeležja i međusobno su ortogonalne. Na taj način često mali broj prvih komponenti zadržava najveći deo informacija (varijanse), pa se ostatak može zanemariti bez velikog gubitka.

4.5 Линеарни Бајесов класификатор

Уколико се претпостави да су подаци у оквиру сваке класе распоређени према нормалној Гаусовој расподели са познатим параметрима расподеле $f_1(X) \sim N(M_1, \Sigma_1)$ и $f_2(X) \sim N(M_2, \Sigma_2)$. Дискриминациона функција Бајесовог линеарног класификатора добија следећи општи квадратни облик:

$$h(X) = -\ln\frac{h_1(X)}{h_2(X)} = \frac{1}{2}(X - M_1)^T\Sigma_1^{-1}(X - M_1) - \frac{1}{2}(X - M_2)^T\Sigma_2^{-1}(X - M_2) + \frac{1}{2}ln\frac{|\Sigma_1|}{|\Sigma_2|}$$

Бајесово правило одлучивања минималне грешке гласи:

$$h(X) \leq \ln\frac{P_2}{P_1}, \quad X \in \omega_1$$

$$h(X) \geq \ln\frac{P_2}{P_1}, \quad X \in \omega_2$$

где су P_1 и P_2 априорне вероватноће појаве класа ω_1 и ω_2.

У ситуацији када су коваријационе матрице класа једнаке, тј.

$$\Sigma_1 = \Sigma_2 = \Sigma$$

Тада израз постаје знатно једноставнији, јер се квадратни чланови поништавају и правило одлучивања поприма линеарни облик:

$$h(X) = (M_2 - M_1)^T\Sigma^{-1}\mathrm{X} + \frac{1}{2}(M_1^T\Sigma^{-1}\mathrm{M}_1 - \mathrm{M}_2^{\mathrm{T}}\Sigma^{-1}\mathrm{M}_2)$$

$$h(X) \leq \ln\frac{P_2}{P_1}, \quad X \in \omega_1$$

$$h(X) \geq \ln\frac{P_2}{P_1}, \quad X \in \omega_2$$

Ова једначина дефинише линеарни дискриминациону хипер-раван у простору обележја, која раздваја класе ω_1 I ω_2. Вектор $(M_2 - M_1)^T\Sigma^{-1}$ представља нормалу на ту граничну хиперповршину, док је константа са десне стране праг одлучивања који зависи од средњих вредности, коваријације и априорних вероватноћа.

4.6 Метод К најближих суда (KNN)

Једна од најпознатијих и најчешће примењиваних непараметарских метода класификације је метод *k* најближих суседа (енг. *k-nearest neighbours*) , скраћено *kNN*. За разлику од параметарских приступа, ова техника не захтева претходно познавање облика функције густине вероватноће класа, већ процену заснива на самим подацима. Идеја иза кНН је једноставна, класа непознатог узорка одређује се на основу већине класа његових *k* најближих суседа у простору обележја који гласају којој класи припада. Због своје интуитивности, лаке имплементације и способности да добро ради и у случајевима сложених расподела података, *kNN* метод представља један од основних и полазних приступа у непараметарској класификацији.

Приликом процене функције густине вероватноће у тачки X, постоје два приступа који се заснивају на Парзеновој процени. Први приступ подразумева да се фиксира запремина v простора око тачке X, а број узорака који у њу упада може варирати у зависности од положаја те тачке. Други, практичнији приступ подразумева да се фиксира број узорака k, док се запремина $v(X)$ мења у зависности од положаја тачке у простору.

Дакле, ако се жели проценити функција густине вероватноће у некој тачки X, потребно је ширити област око те тачке све док не обухвати тачно k узорака из скупа расположивих података. Запремина те области $L(X)$ постаје случајна променљива, једнако као и сама величина $v(X)$, а њена вредност зависи од локације тачке X унутар простора у којем се врши процена функције густине вероватноће.

Метод *kNN* може се посматрати као посебан случај Парзенове процене са једноликим униформним кернелом, чија се ширина прилагођава аутоматски у зависности од положаја тачке. На местима где је густина узорака мања, запремина $v(X)$ биће већа, а тамо где су узорци гушћи, биће мања. На тај начин, метод омогућава флексибилну локалну процену функције густине.

Функција густине вероватноће која се процењује методом *kNN* може се записати као:

$$f(X) = \frac{k-1}{Nv(X)}$$

где је N укупан број узорака у скупу, а $v(X)$запремина хиперсфере која садржи k најближих суседа тачке X.

Јасно је да параметар k има пресудан утицај на квалитет процене. Премали број суседа води ка шумовитој и нестабилној процени, док превелики број доводи до прекомерног изглађивања и губитка локалних структура у подацима. Озбиљним теоријским анализама показано је да параметар k мора задовољити следеће услове:

$$\lim_{N\to\infty} k = \infty$$

$$\lim_{N\to\infty} \frac{k}{N} = 0$$

Испуњење ових услова обезбеђује да процена функције густине буде непристрасна и конзистентна, тј. да тежи стварној функцији густине како број узорака расте. У пракси, избор параметра k за задати број расположивих узорака N врши се према емпиријском правилу које се често наводи у литератури:

$$k = [N^{\alpha}], \;\; \alpha \in (0,1)$$

где је $[\cdot]$ функција заокруживања на најближи цео број.
Најпогоднија вредност параметра k одређена је применом **k-fold крос-валидације**, технике која обезбеђује поуздану процену перформанси модела без прекомерног прилагођавања подацима (енг. *Overfitting)*. Овај поступак омогућава избор оптималног k који најбоље балансира између сложености модела и тачности класификације. Детаљнији опис принципа и примене крос-валидације биће дат у наредном поглављу.

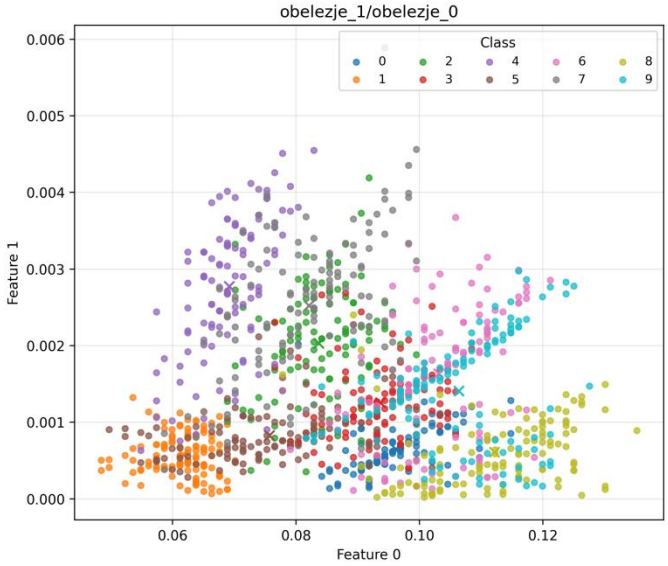
5.Резултати

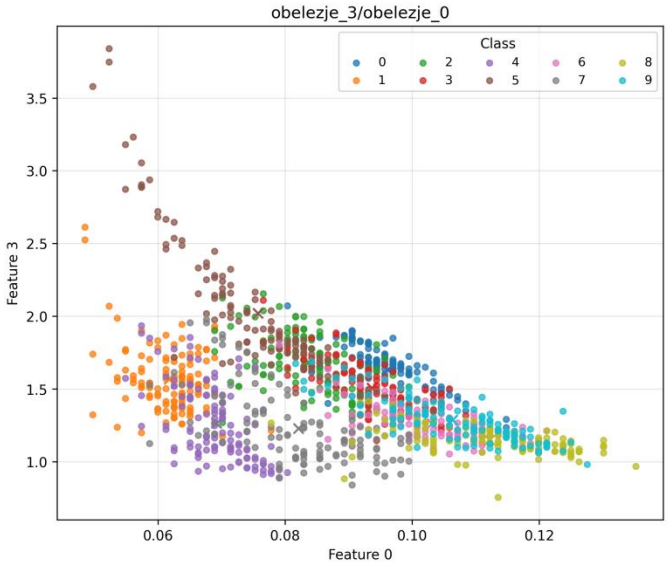
6.Закључак

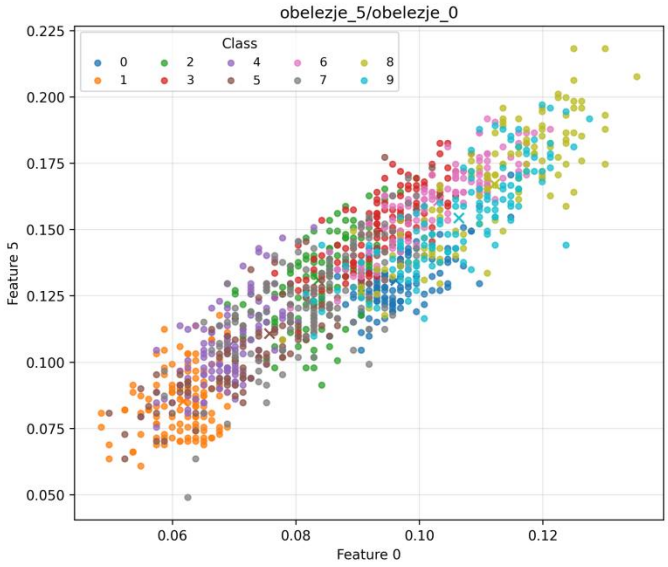
7.Литература

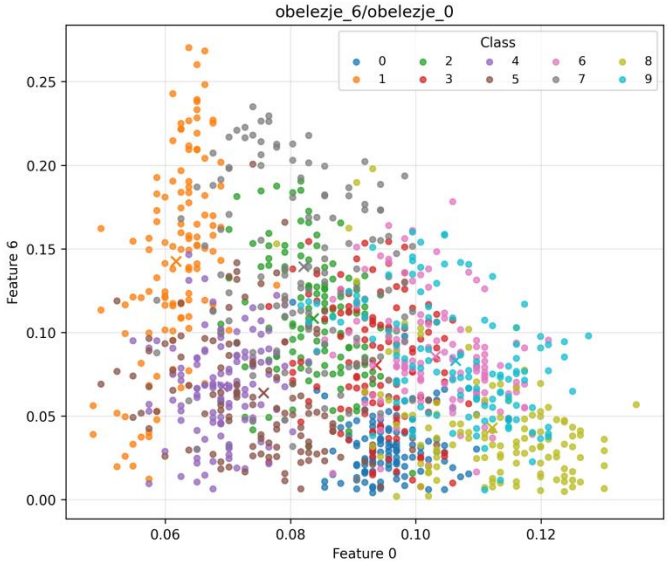
Прилог А

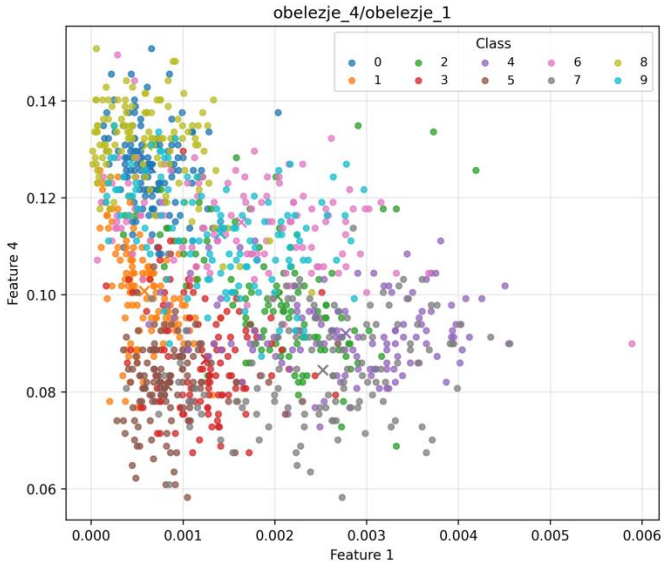
Обележје 0	Средња вредност густине црних пиксела по хоризонталним сегментима
Обележје 1	Варијанса густине црних пиксела по хоризонталним сегментима
Обележје 2	Варијанса густине црних пиксела по вертикалним сегментима
Обележје 3	Момент слике
Обележје 4	Број хоризонталних прелаза
Обележје 5	Број вертикалних прелаза
Обележје 6	Удаљеност центра масе од центра слике

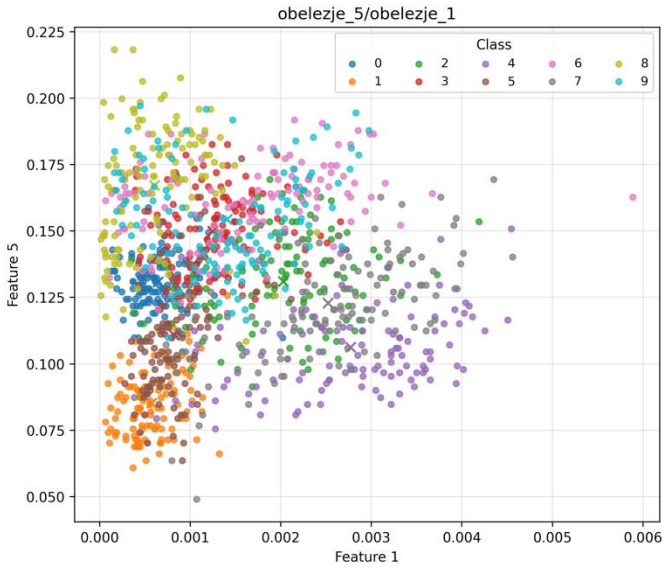


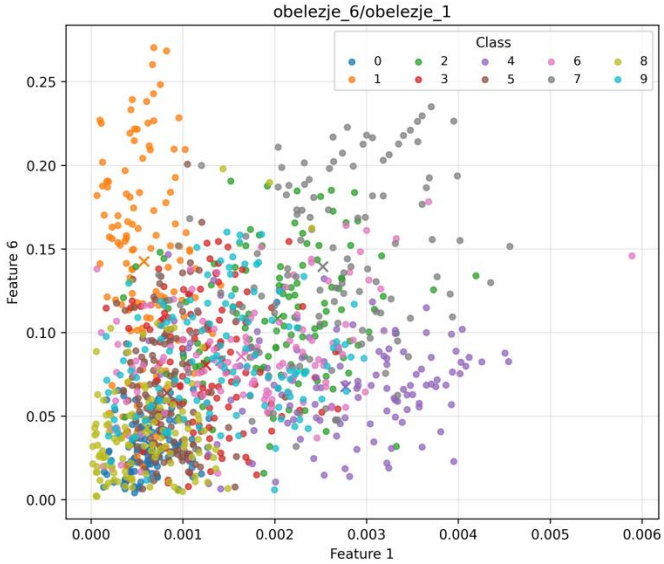

obelezje_3/obelezje_0
Class
0 1 2 3 4 5 6 7 8 9
Feature 3
Feature 0
3.5
3.0
2.5
2.0
1.5
1.0
0.06
0.08
0.10
0.12

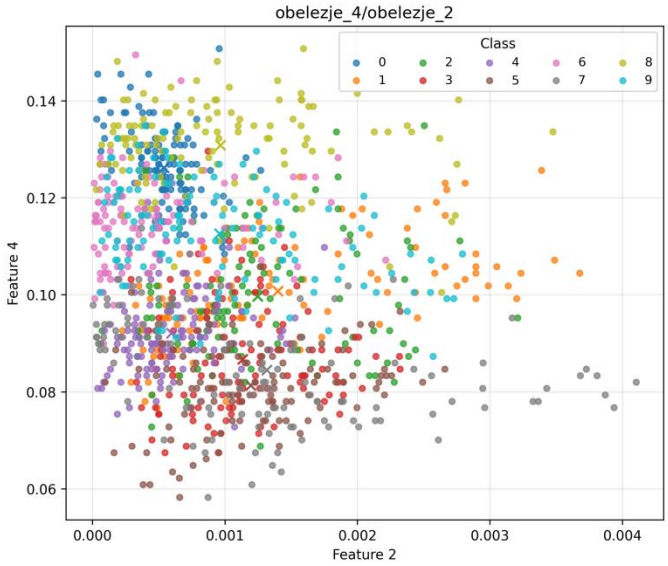

obelezje_5/obelezje_0
Class
0 1 2 3 4 5 6 7 8 9
Feature 5
Feature 0
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.06
0.08
0.10
0.12

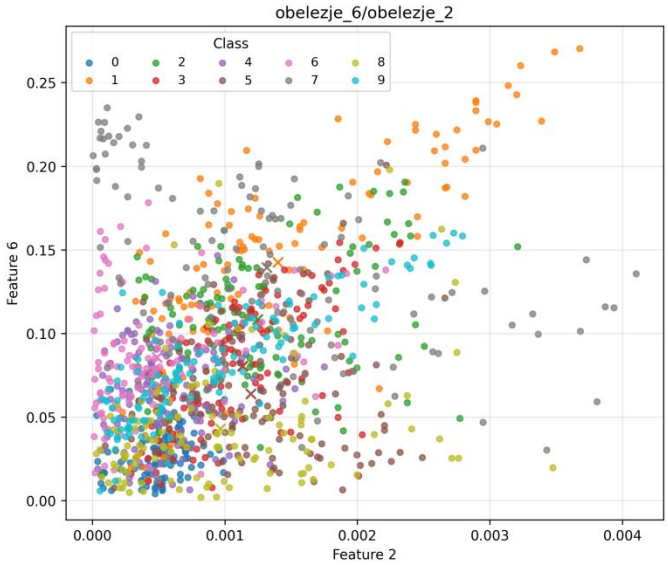

obelezje_6/obelezje_0
Class
0 1 2 3 4 5 6 7 8 9
Feature 6
Feature 0
0.25
0.20
0.15
0.10
0.05
0.00
0.06
0.08
0.10
0.12

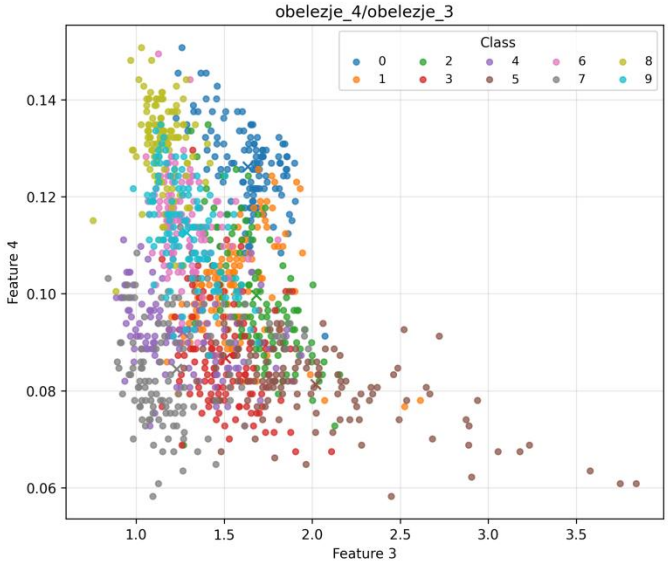

obelezje_4/obelezje_1
Class
0 1 2 3 4 5 6 7 8 9
Feature 4
Feature 1
0.14
0.12
0.10
0.08
0.06
0.000 0.001 0.002 0.003 0.004 0.005 0.006

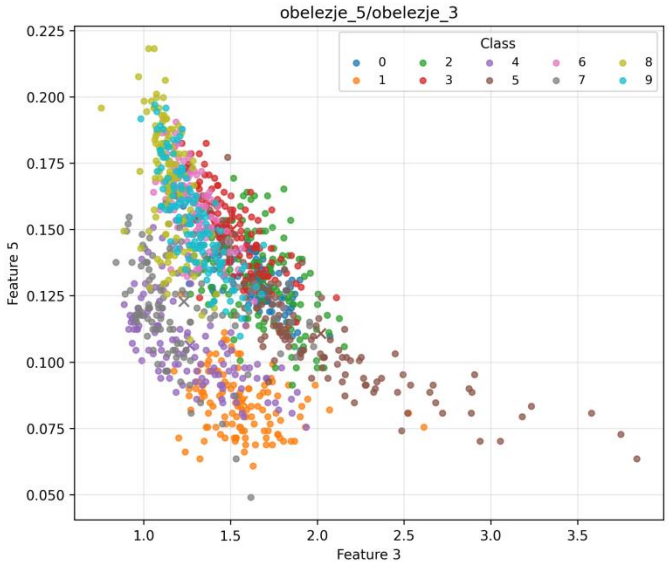

obelezje_5/obelezje_1
Class
0 1 2 3 4 5 6 7 8 9
Feature 5
Feature 1
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.000 0.001 0.002 0.003 0.004 0.005 0.006

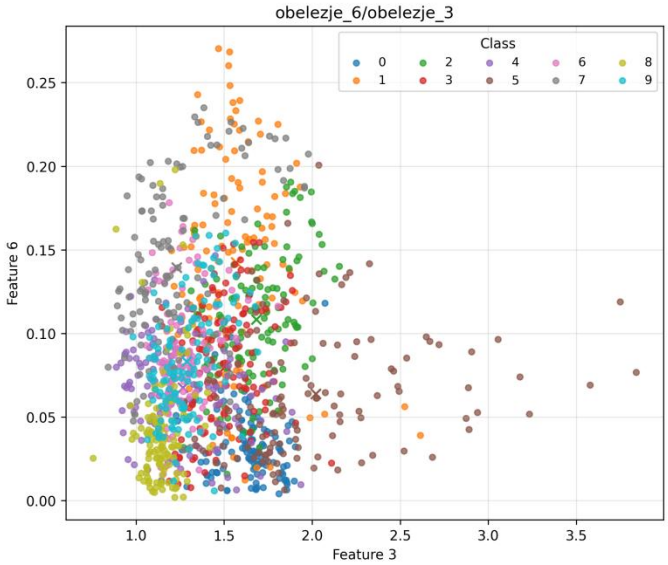

obelezje_6/obelezje_1
Class
0 1 2 3 4 5 6 7 8 9
Feature 6
Feature 1
0.25
0.20
0.15
0.10
0.05
0.00
0.000 0.001 0.002 0.003 0.004 0.005 0.006

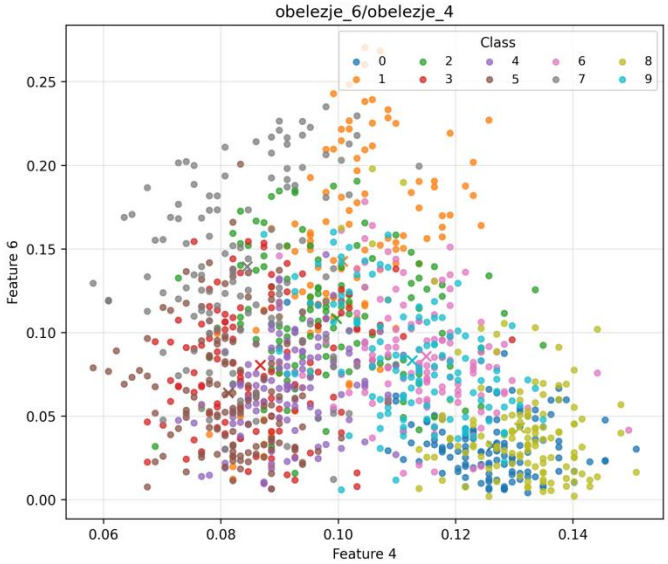

obelezje_4/obelezje_2
Class
0 1 2 3 4 5 6 7 8 9
Feature 4
Feature 2
0.06
0.08
0.10
0.12
0.14
0.000
0.001
0.002
0.003
0.004


obelezje_6/obelezje_2
Class
0 1 2 3 4 5 6 7 8 9
Feature 6
Feature 2
0.00
0.05
0.10
0.15
0.20
0.25
0.000
0.001
0.002
0.003
0.004


obelezje_4/obelezje_3
Class
0 1 2 3 4 5 6 7 8 9
Feature 4
Feature 3
0.06
0.08
0.10
0.12
0.14
1.0
1.5
2.0
2.5
3.0
3.5


obelezje_5/obelezje_3
Class
0 1 2 3 4 5 6 7 8 9
Feature 5
Feature 3
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
1.0
1.5
2.0
2.5
3.0
3.5


obelezje_6/obelezje_3
Class
0 1 2 3 4 5 6 7 8 9
Feature 6
Feature 3
0.25
0.20
0.15
0.10
0.05
0.00
1.0
1.5
2.0
2.5
3.0
3.5


obelezje_6/obelezje_4
Class
0 1 2 3 4 5 6 7 8 9
Feature 6
Feature 4
0.25
0.20
0.15
0.10
0.05
0.00
0.06
0.08
0.10
0.12
0.14