Klasifikacija poremećaja autističnog spektra kod odraslih

Projekat iz predmeta Sistemi odlučivanja u medicini

Vladan Bašić 2022/0395 Milica Gojak 2022/0061

Oktobar 2025.

Sadržaj

1 Uvod **3**
1.1 Opis skupa podataka . . . 3

2 Analiza skupa podataka **3**
2.1 Čišćenje i priprema podataka . . . 3
2.2 Analiza distribucije atributa . . . 5
2.3 Analiza distribucije klasa . . . 6
2.4 Normalizacija parametara . . . 6

3 Selekcija obeležja **6**
3.1 Teorijske osnove informacione dobiti . . . 6
3.2 Rezultati informacione dobiti . . . 7
3.3 Korelaciona analiza . . . 7
3.4 Analiza najkoreliranijih obeležja sa klasom . . . 9

4 Redukcija dimenzija - Linear Discriminant Analysis (LDA) **10**
4.1 Teorijske osnove LDA . . . 10
4.2 Rezultati LDA . . . 10

5 Parametarska klasifikacija - Gaussov klasifikator **10**
5.1 Teorijske osnove . . . 10
5.2 Rezultati parametarske klasifikacije . . . 11

6 Neparametarska klasifikacija - k-Nearest Neighbors (kNN) **11**
6.1 Teorijske osnove kNN . . . 11
6.2 Optimizacija parametra k . . . 12
6.3 Rezultati kNN klasifikacije . . . 12

7 Neuralne mreže **13**
7.1 Teorijske osnove neuralnih mreža . . . 13
7.2 Arhitektura neuralnih mreža . . . 13
7.2.1 Underfitted Network . . . 13
7.2.2 Overfitted Network . . . 13
7.2.3 Optimal Network . . . 14
7.3 Treniranje neuralnih mreža . . . 14
7.4 Regularizacija . . . 14
7.5 Uticaj regularizacije na trening . . . 15
7.6 Napomena . . . 15

8 Komparativna analiza **16**
8.1 Poređenje performansi . . . 16
8.2 Komentar rezultata komparativne analize . . . 16

9 Zaključak **16**

1 Uvod

Ovaj izveštaj predstavlja kompletnu analizu problema klasifikacije poremećaja autističnog spektra (ASD) kod odraslih osoba. Korišćen je skup podataka "Adult Autism Spectrum Disorder" koji sadrži informacije prikupljene kroz skrining test za detekciju ASD-a.

Cilj projekta je implementacija i poređenje različitih metoda mašinskog učenja za klasifikaciju postojanja indikatora autističnog spektra. Analizirane su tri glavne kategorije pristupa: parametarska klasifikacija (Gaussov klasifikator sa LDA), neparametarska klasifikacija (k-Nearest Neighbors) i neuralne mreže.

1.1 Opis skupa podataka

Skup podataka sadrži ukupno 704 instance i 21 atribut. Svaki red predstavlja pojedinačni slučaj (ispitanika), dok kolone predstavljaju različita obeležja vezana za demografiju, porodičnu anamnezu i odgovore na skrining pitanja.

Ključni atributi uključuju:

- **A1_Score – A10_Score**: Rezultati pojedinačnih pitanja iz ASD screening upitnika
- **age**: Starost ispitanika u godinama
- **gender**: Pol ispitanika (muški/ženski)
- **ethnicity**: Etnička pripadnost ispitanika
- **jaundice**: Da li je dete rođeno sa žuticom (da/ne)
- **autism**: Porodična istorija autizma (da/ne)
- **Class/ASD**: Ciljna promenljiva – rezultat klasifikacije

2 Analiza skupa podataka

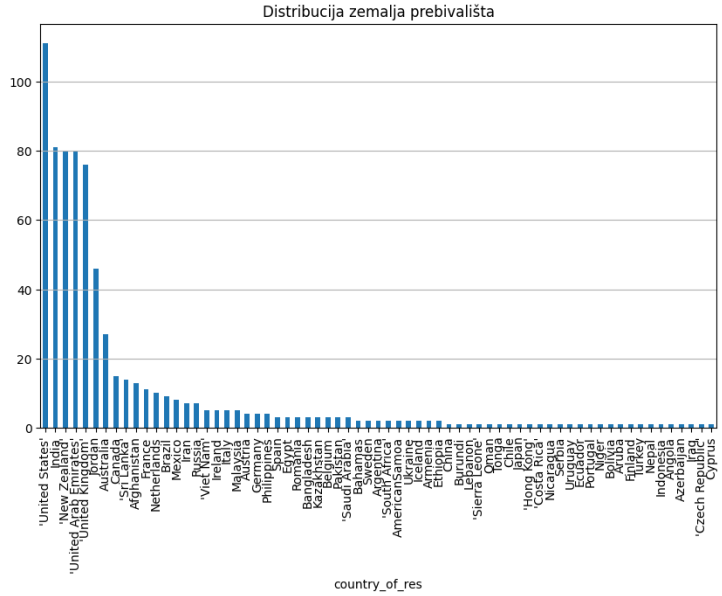
Pre primene bilo koje metode klasifikacije, izvršena je detaljna analiza skupa podataka koja je uključila identifikaciju nedostajućih vrednosti, analizu distribucije klasa i pripremu podataka za dalju obradu.

2.1 Čišćenje i priprema podataka

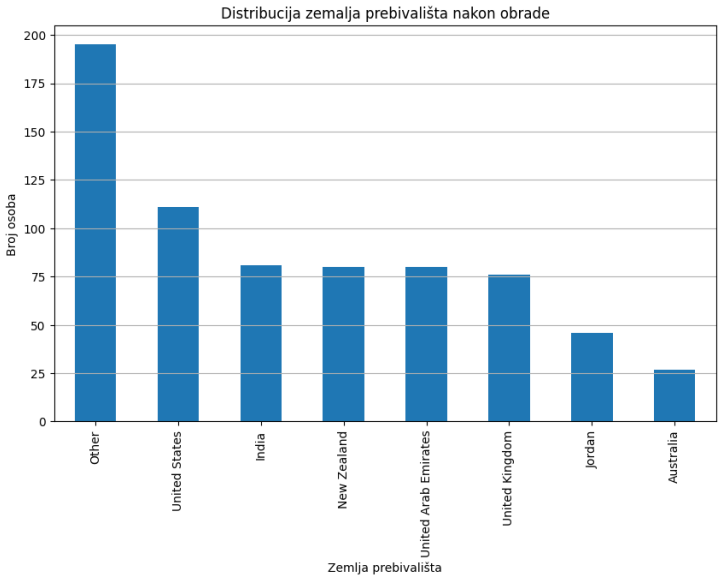
Analiza je pokazala da skup podataka sadrži mali broj duplikata i uzoraka koji imaju nedostajuće ili besmislene vrednosti nekih od atributa. Nakon uklanjanja ovih uzoraka ostalo je 696 od početnih 704 uzoraka. Neke od kolona sadrže "?" umesto potrebne prednosti za određen atribut. Ove kolone čine značajan deo baze pa umesto uklanjanja grupisane su sa vrednošću "Unknown" ili "Other" za dati atribut.

Vrednosti atributa koje se pojavljuju u manje od 3% uzoraka grupisane su u kategoriju "Other"

Na osnovu atributa *result* su formirani rezultati klase pa je taj atribut izbačen iz baze da bi analiza metoda imala smisla.



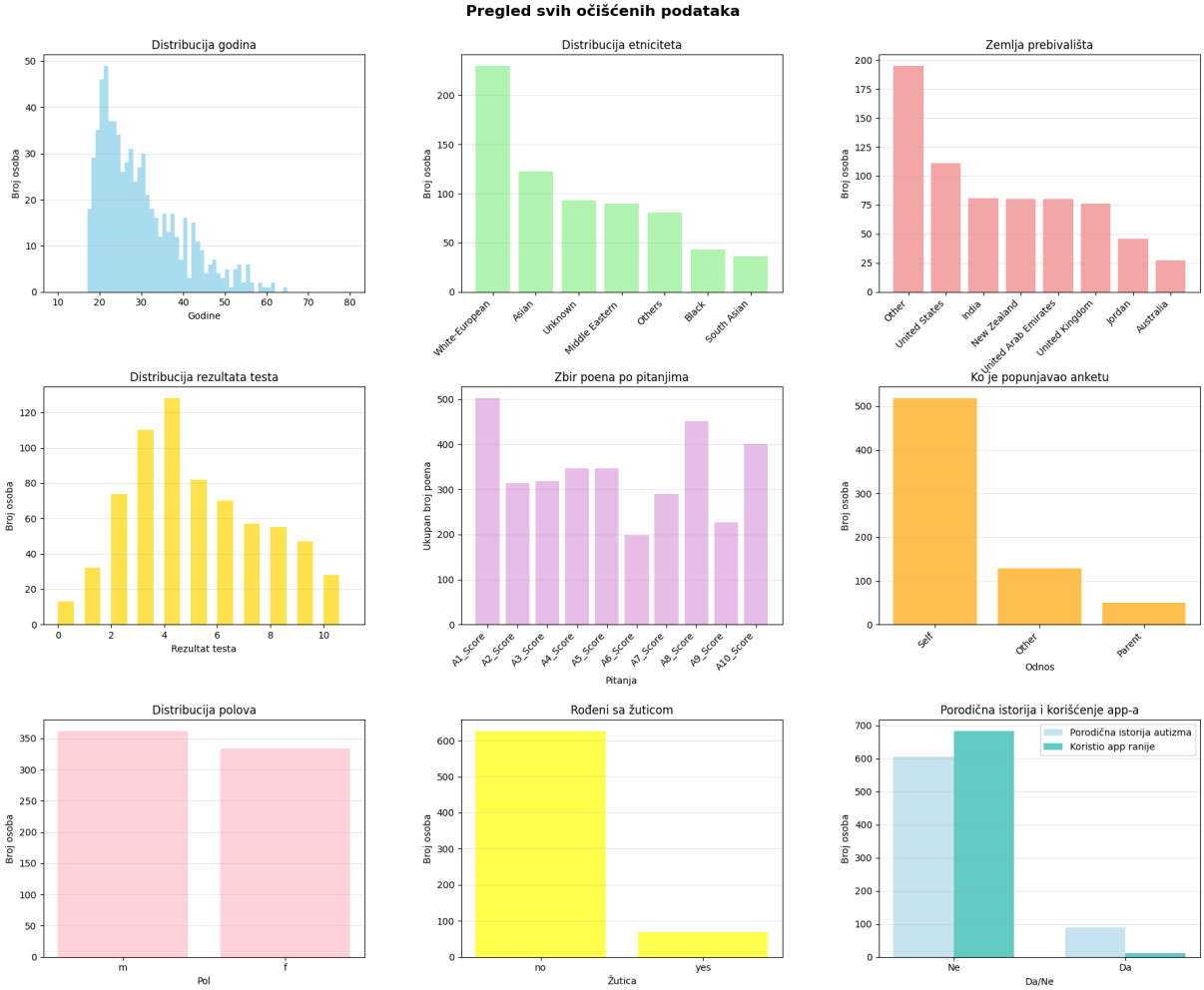
Slika 1: Distribucija atributa zemlja prebivališta u skupu podataka



Slika 2: Distribucija atributa zemlja prebivališta u skupu podataka nakon grupisanja

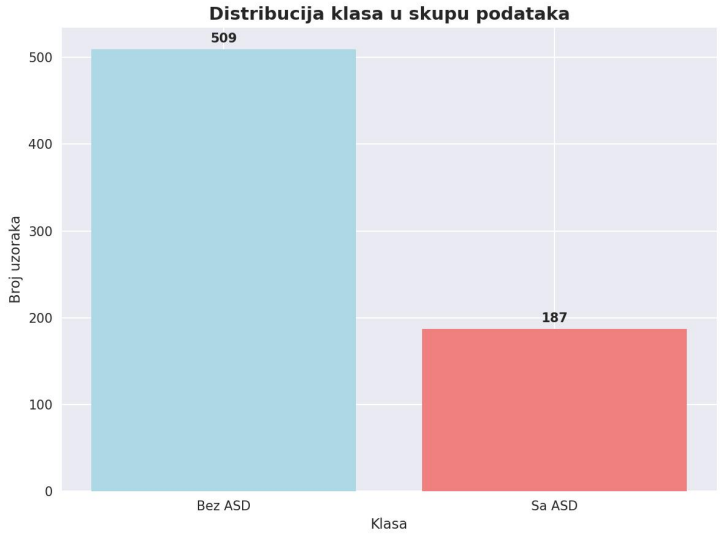
Analiza je pokazala da atribut age_desc nema varijansu u raspodeli (svaki uzorak u koloni age_desc ima vrednost ”18 or more”) pa je taj atribut uklonjen iz baze jer ne sadrži nikakvu informaciju.

2.2 Analiza distribucije atributa



Slika 3: Distribucija atributa u skupu podataka

2.3 Analiza distribucije klasa



Slika 4: Distribucija klasa u skupu podataka

Analiza distribucije klasa pokazuje da je skup podataka relativno balansiran, što je povoljno za primenu različitih metoda klasifikacije.

2.4 Normalizacija parametara

Pre nego što se vrši bilo kakva dalja analiza potrebno je normalizovati bazu. Većina baze se sastoji iz atributa koji imaj jednu od dve mogće vrednosti. Ti atributi su kodirani 0/1 kodiranjem gde je jednoj vrednosti odeljena vrednost 0 a drugoj vrednost 1. Atributi koji imaju više mogućih vrednosti su kodirani one-hot kodiranjem. Atributi koji imaju skalarnu vrednost normalizovani su na skalu [0,1].

3 Selekcija obeležja

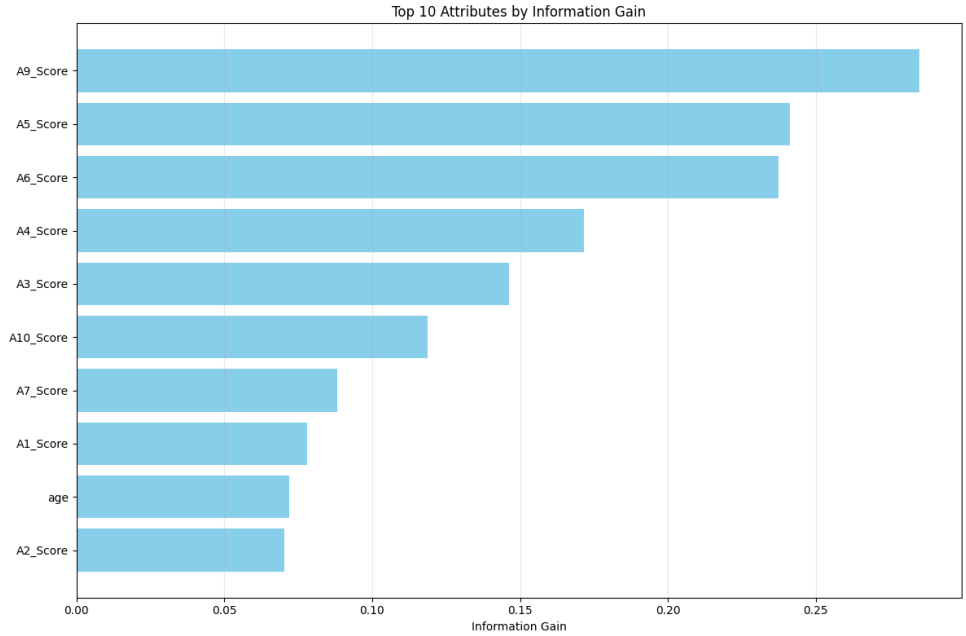
3.1 Teorijske osnove informacione dobiti

Informaciona dobit (Information Gain) je mera koja kvantifikuje koliko određeno obeležje doprinosi smanjenju entropije sistema. Matematički se definiše kao:

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \tag{1}$$

$$H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \tag{2}$$

3.2 Rezultati informacione dobiti

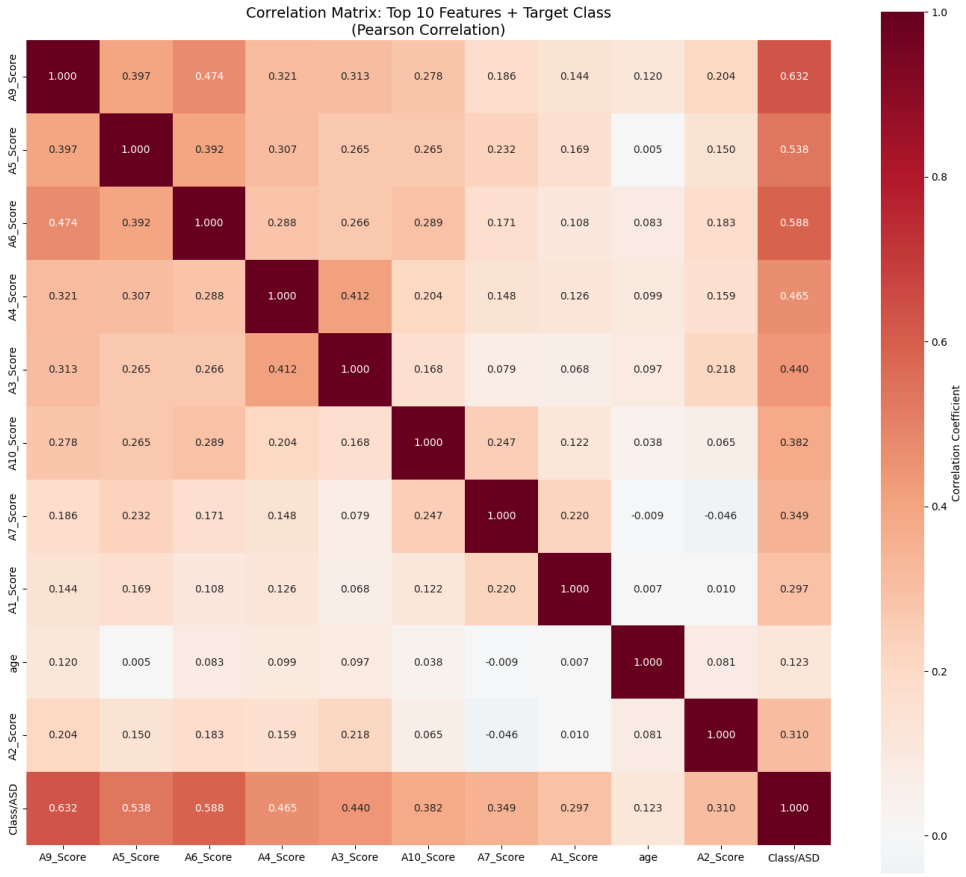


Slika 5: Grafički prikaz informacione dobiti top 10 obeležja

Primećuje se da atribut *result* ima drastično veću informacionu dobit od ostalih atributa. Odavde možemo izvući pretopstavku da se klasifikacija može raditi sa zadovoljavajućim performansama samo na osnovu tog atributa.

3.3 Korelaciona analiza

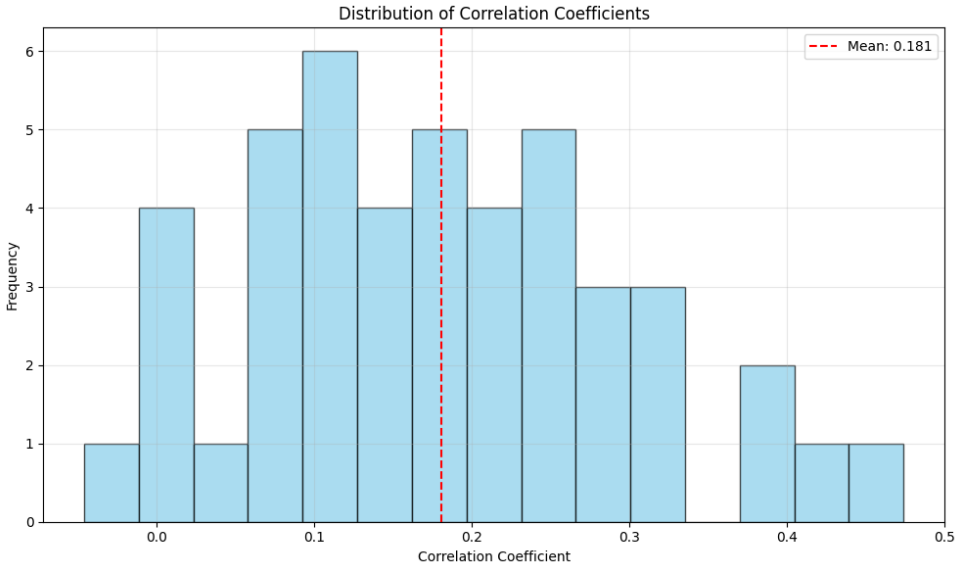
Ispitana je međusobna korelacija između odabranih 10 najinformativnijih obeležja korišćenjem Pearson-ovog koeficijenta korelacije.



Slika 6: Matrica korelacije između top 10 obeležja

Tabela 1: Top 10 obeležja prema korelaciji sa klasom

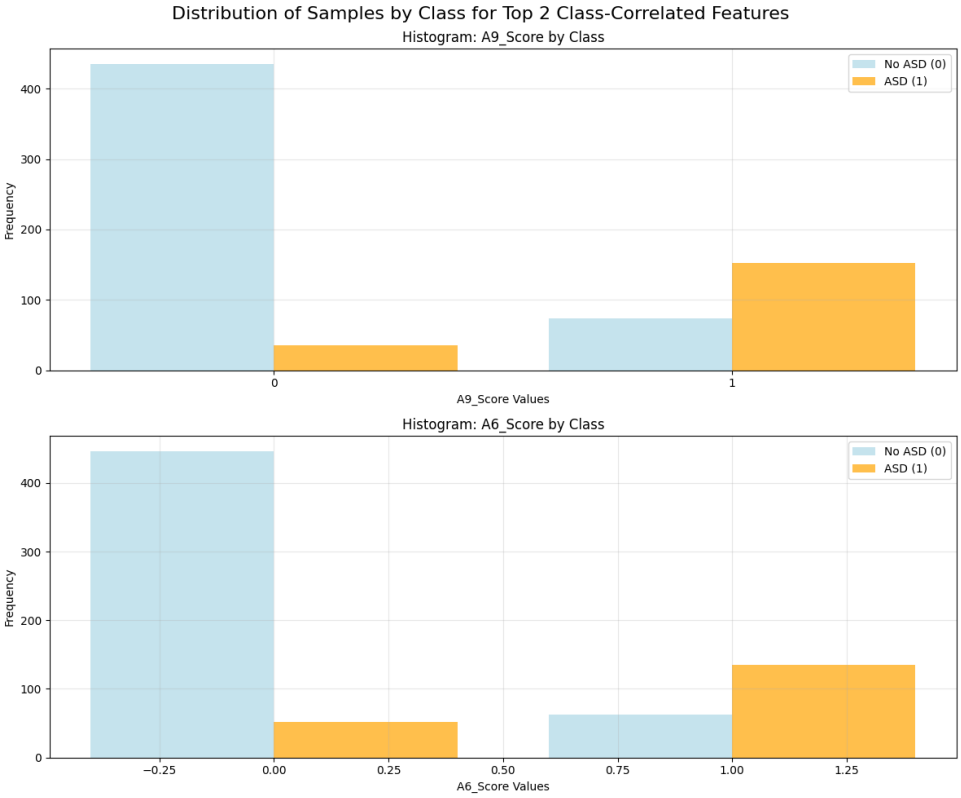
Atribut	**Korelacija sa klasom**
A9_Score	0.6318
A6_Score	0.5877
A5_Score	0.5383
A4_Score	0.4653
A3_Score	0.4396
A10_Score	0.3821
A7_Score	0.3490
A2_Score	0.3103
A1_Score	0.2973
age	0.1233



Slika 7: Raspodela korelacionih koeficijenata

3.4 Analiza najkoreliranijih obeležja sa klasom

Na osnovu analize korelacije sa ciljnom promenljivom, identifikovana su dva obeležja sa najvećom korelacijom: *A6_Score* i *A9_Score*.



Slika 8: Raspodela uzoraka prema obeležjima sa najvećom korelacijom sa klasom

4 Redukcija dimenzija - Linear Discriminant Analysis (LDA)

4.1 Teorijske osnove LDA

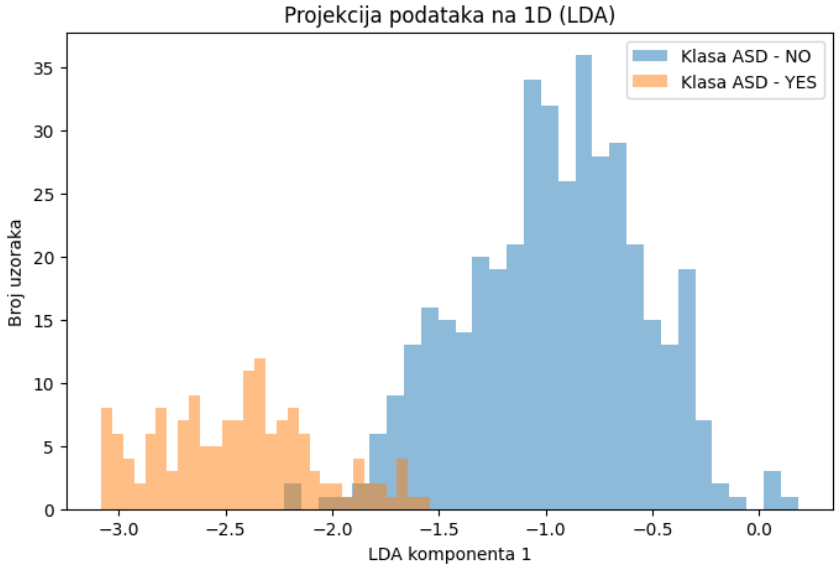
Linear Discriminant Analysis (LDA) je statistička metoda koja se koristi za redukciju dimenzionalnosti i klasifikaciju. LDA pronalazi linearnu kombinaciju obeležja koja najbolje razdvaja klase u podacima.

Cilj LDA je da maksimizuje odnos između inter-klase i intra-klase varijanse:

$$J = \frac{|W^T S_B W|}{|W^T S_W W|} \tag{3}$$

S obzirom da data baza ima samo dve klase (ima ili nema ASD) LDA redukcija dimenzinalnosti će vratiti 1D fičer space.

4.2 Rezultati LDA



Slika 9: Vizualizacija klasa u 1D LDA prostoru

Analiza je pokazala da je LDA komponenta dovoljna za objašnjavanje preko 80% varijanse u podacima, što ukazuje na dobru separabilnost klasa.

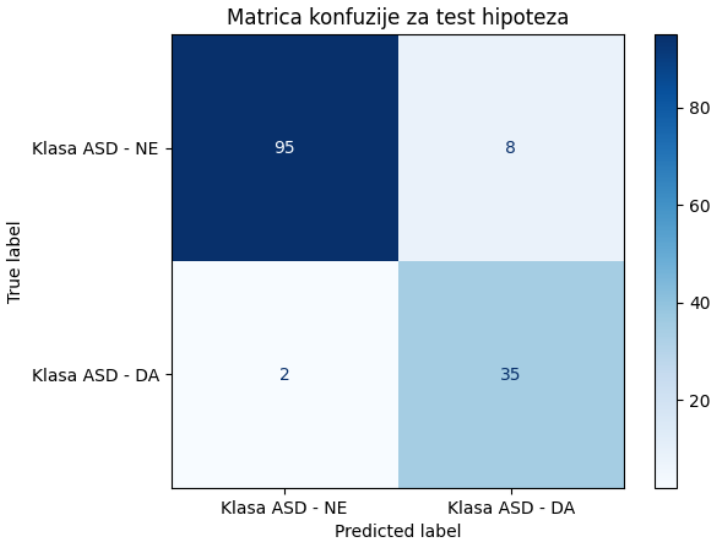
5 Parametarska klasifikacija - Gaussov klasifikator

5.1 Teorijske osnove

Gaussov (Bayes-ov) klasifikator je parametarska metoda koja pretpostavlja da se podaci svake klase mogu modelovati Gaussovom (normalnom) distribucijom. Klasifikacija se vrši na osnovu Bayes-ove teoreme:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} \tag{4}$$

5.2 Rezultati parametarske klasifikacije



Slika 10: Confusion matrix za Gaussov klasifikator

Tabela 2: Performanse Gaussovog klasifikatora

Metrika	**Vrednost**
Preciznost	81.40%
Senzitivnost (Recall)	94.59%
Specifičnost	92.23%
Ukupna tačnost	92.86%
F1 Score	87.50%

6 Neparametarska klasifikacija - k-Nearest Neighbors (kNN)

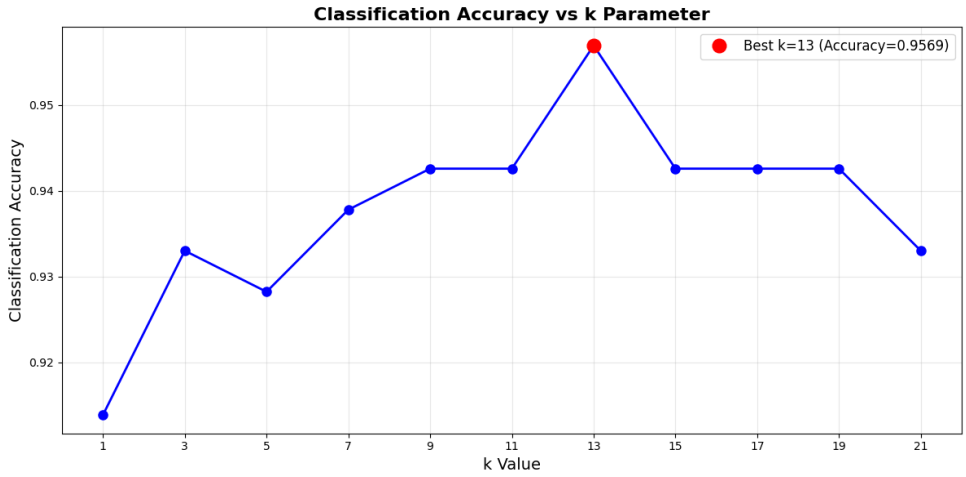
6.1 Teorijske osnove kNN

k-Nearest Neighbors (kNN) je jedna od najjednostavnijih i najintuitivnijih metoda mašinskog učenja. Za svaki uzorak koji klasifikuje posmatra k najbljižih suseda u euklidskom prostoru atributa i na onsovu toga koja klasa je najzastupljenija u tih k suseda klasifikuje dati uzorak.

Euklidsko rastojanje se računa kao:

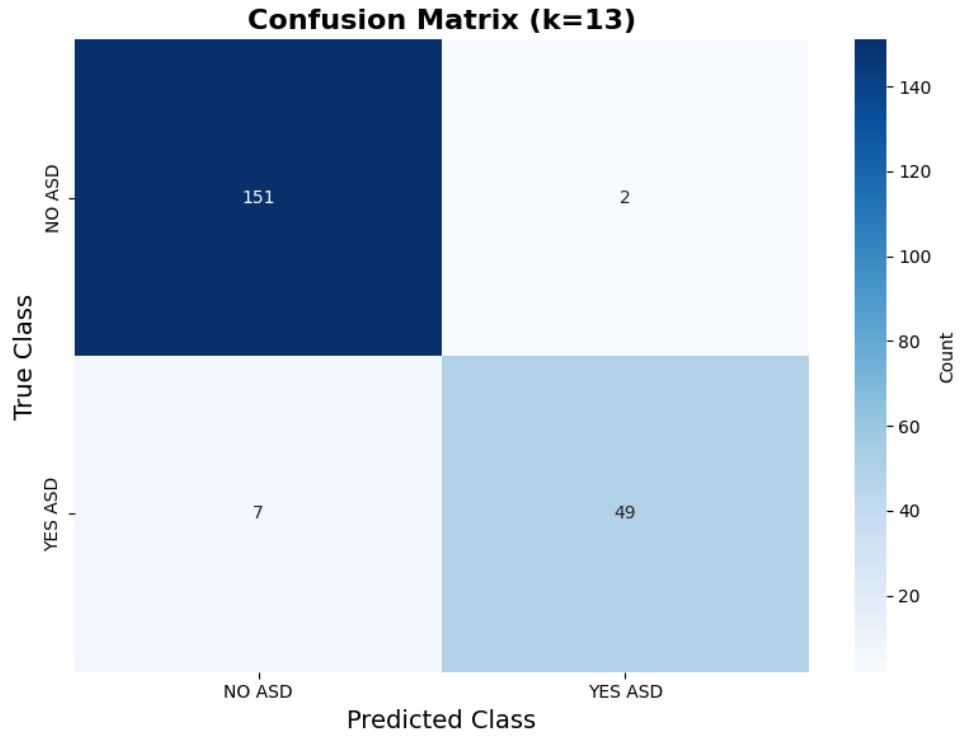
$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{d} (x_{il} - x_{jl})^2} \tag{5}$$

6.2 Optimizacija parametra k



Slika 11: Zavisnost tačnosti klasifikacije od vrednosti parametra k

6.3 Rezultati kNN klasifikacije



Slika 12: Confusion matrice za optimalnu vrednost k

Tabela 3: Performanse kNN klasifikatora (optimalna k)

Metrika	Vrednost
Optimalna k vrednost	13
Preciznost	96.08%
Senzitivnost (Recall)	87.50%
Specifičnost	98.69%
Ukupna tačnost	95.69%

Slika 13: Performanse kNN klasifikatora po klasi za optimalnu vrednost k

7 Neuralne mreže

7.1 Teorijske osnove neuralnih mreža

Neuralne mreže su računski modeli inspirisani biološkim neuralnim mrežama. Osnovna jedinica je veštački neuron (perceptron) koji prima ulaze, primenjuje težinske koeficijente i aktivacionu funkciju. Neuralna mreža se sastoji iz više slojeva ovakvih neurona koji su simultano trenirani da nađu lokalni optimum funkcije koju zajendo čine.

Fully-Connected neuralna mreža je moreža koja se sastoji od više slojeva neurona. Svaki neuron iz jednog sloja prosleđuje svoj izlaz na svaki neuron u sledećem sloju. Prvi sloj prima ulazne fičere dok poslednji sloj predstavlja izlaz klasifikacije. U ovomslučaju finalni sloj sadržaće jedan neuron koji će klasifikovati ulaz kao jednu klasu ako mu je vrednost bliža nuli ili kao drugu klasu ako mu je vrednost bliža jedinici.

7.2 Arhitektura neuralnih mreža

7.2.1 Underfitted Network

Kreirana je mreža premale kompleksnosti koja nema ni jedan skriveni sloj veličine. Ova mreža ukupno sadrži 34 parametara.

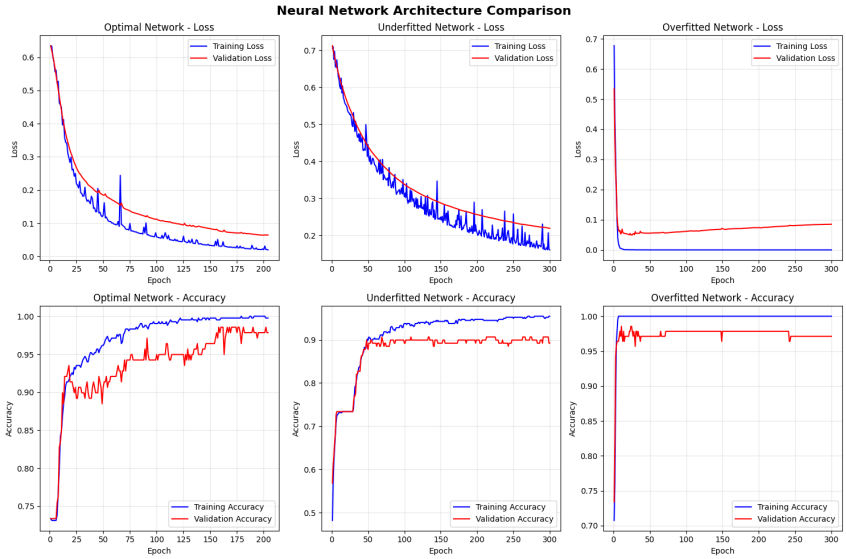
7.2.2 Overfitted Network

Kreirana je mreža prevelike kompleksnosti koja sadrži 3 skrivena sloja veličina 256, 256 i 64. Ova mreža ukupno sadrži 91009 parametara.

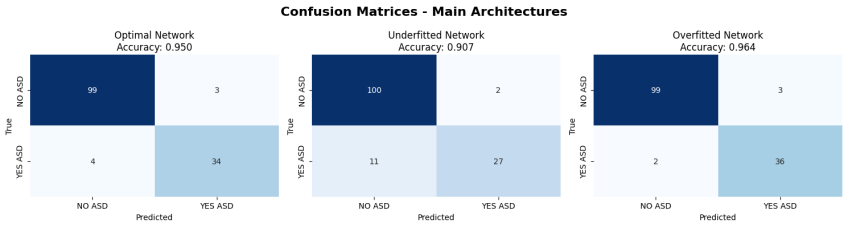
7.2.3 Optimal Network

Kreirana je mreža procenjene optimalne kompleksnosti koja sadrži 1 skriveni sloj od 16 neurona. Ova mreža ukupno sadrži 561 parametara.

7.3 Treniranje neuralnih mreža



Slika 14: Trening mreža bez regularizacije



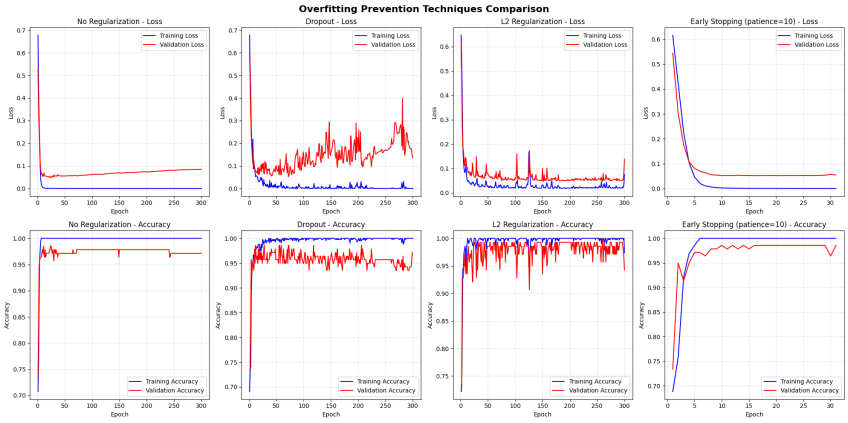
Slika 15: Matrice konfuzije mreža bez regularizacije

Može se primetiti da mreža nedovoljne kompleksnosti ne uspeva da dostigne tačnost koju dostiže mreža optimalne kompleksnosti. Sa druge strane, primećuje se da pretrenirana mreža ima najbolju tačnost iako zaista jeste overfitovana i loss validacije raste kroz vreme.

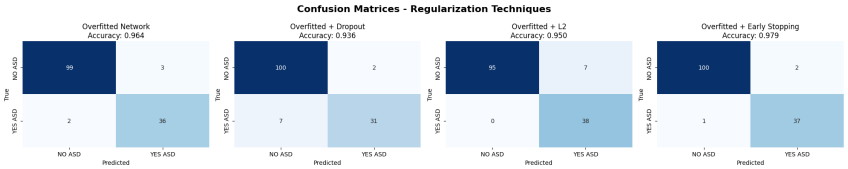
7.4 Regularizacija

Da bi se smanjio uticaj pretreniranja mreže korišćene su 4 metode regularizacije pri treniranju mreže.

7.5 Uticaj regularizacije na trening



Slika 16: Trening prevelikih mreža sa regularizacijom



Slika 17: Matrice konfuzije prevelikih mreža sa regularizacijom

Tabela 4: Rezime performansi neuralnih mreža

Model	**Tačnost**	**Preciznost**	**Senzitivnost**	**Specifičnost**
Optimalna mreža	95.00%	91.89%	89.47%	97.06%
Nedovoljno obučena mreža	90.71%	93.10%	71.05%	98.04%
Overfitted mreža	96.43%	92.31%	94.74%	97.06%
Overfitted + Dropout	93.57%	93.94%	81.58%	98.04%
Overfitted + L2 regularizacija	95.00%	84.44%	100.00%	93.14%
Overfitted + Rano zaustavljanje	97.86%	94.87%	97.37%	98.04%

Iz datih rezultata se primećuje da pretrenirani model postiže najbolje rezultate. Ovaj rezultat je nestandardan ali očekivan kada se uzme u obzir struktura baze. Baza je forirana tako da se klasa određuje kao zbir ocena na pitanja upoređena sa 0.6. Pošto je veza između ulaza i izlaza direktna i ista za svaki član baze pretrenirani model će na trening setu pretrenirati na datu vezu ali pošto je data veza tačna za celu bazu pretrenirani model će imati najtačniji rezultat.

7.6 Napomena

Tokom treniranja neuralnih mreža bilo je izuzetno teško ne *overfit*-ovati bilo koju mrežu koja ima skriveni sloj.

8 Komparativna analiza

8.1 Poređenje performansi

Tabela 5: Komparativni pregled performansi svih metoda

Metoda	Preciznost	Senzitivnost	Specifičnost	Ukupna tačnost
Gaussov klasifikator (LDA)	81.4%	94.6%	92.2%	92.9%
kNN (k=13)	96.1%	87.5%	98.7%	95.7%
Neuralna mreža (optimalna)	0.904	0.940	0.900	0.920
Optimalna mreža	91.89%	89.47%	97.06%	95.00%
Overfitted mreža	92.31%	94.74%	97.06%	96.43%

8.2 Komentar rezultata komparativne analize

Na osnovu prikazanih rezultata može se uočiti da sve analizirane metode postižu visoku tačnost u klasifikaciji prisustva autizma kod odraslih, ali sa određenim razlikama u balansu između **senzitivnosti** (tačnost u detekciji pozitivnih slučajeva) i **specifičnosti** (tačnost u detekciji negativnih slučajeva).

Gaussov klasifikator (LDA) pokazuje izuzetno visoku senzitivnost (94.6%), što znači da je vrlo efikasan u prepoznavanju osoba sa autizmom. Međutim, njegova preciznost (81.4%) i ukupna tačnost (92.9%) su nešto niže u poređenju sa ostalim metodama, što ukazuje da ima veći broj pogrešnih pozitivnih klasifikacija. Gaussov klasifikator, iako lošijih performansa od ostalih, ima najbolju računarsku efikasnost i bio bi idelan izbor u slučajevima gde nam je optimizacija izvršavanja ili treniranja vrlo bitna.
kNN klasifikator (k=13) ostvaruje najvišu ukupnu tačnost (95.7%) i izuzetno visoku specifičnost (98.7%), što znači da vrlo retko pogrešno klasifikuje osobe bez autizma kao autistične. Ipak, senzitivnost (87.5%) je nešto niža. To može biti posledica korišćenja većeg broja suseda, što „izravnava“ granice između klasa i smanjuje osetljivost na manje grupe pozitivnih uzoraka.
Neuralne mreže postižu balans između ova dva pristupa. Optimalna mreža pokazuje visoku preciznost (91.89%), dobru senzitivnost (89.47%) i vrlo visoku specifičnost (97.06%), što ukazuje na uspešnu generalizaciju. *Overfitted* mreža pokazuje najvišu tačnost (96.43%) i vrlo visoku senzitivnost (94.74%). Razlozi za ovo su već komentarisani.

Sveukupno, može se zaključiti da neuronske mreže i kNN daju najbolje performanse na ovom skupu podataka, dok LDA ostaje solidna i interpretabilna alternativa za slučajeve gde je važna jednostavnost modela i brzina izvršavanja. U praktičnim primenama, izbor metode zavisi od prioriteta — ako je cilj **maksimalna detekcija pozitivnih slučajeva**, LDA ili overfitted mreža su pogodniji; ako je cilj **minimizacija lažnih alarma**, kNN ili optimalna mreža daju bolji kompromis.

9 Zaključak

U projektu je izvršena implementacija i poređenje tri različita pristupa klasifikaciji poremećaja autističnog spektra kod odraslih kroz detaljnu analizu podataka, selekciju obeležja, redukciju dimenzionalnosti i primenu različitih algoritama mašinskog učenja.

Glavna opservacija projekta bila je neadekvatnost baze za realane medicinske aplkacije. Baza je formirana kroz upitnik unutar mobilne aplikacije, a *ground truth* je formiran iz zbira odgovora na pitanja. Da je *ground truth* formiran na osnovu propratnih izveštaja medicinskog osoblja koja

je individualno ispitala učesnike i dala profesionalnu procenu prisustva ASD kod ispitanika baza bi mogla da se iskoristi kao procena validnosti upitnika kao metoda predikcije autizma. Kako trenutno stoji baza nema mnogo validnosti sama po sebi.