Univerzitet u Beogradu

Matematički fakultet

Klasifikacija tipa fizičke vežbe nad MEx skupom podataka

Student: **Ana Arsić**

Predmet: Istraživanje podataka 2

Godina: 2025/2026

Sadržaj

1 Uvod

Prepoznavanje fizičkih aktivnosti na osnovu senzorskih podataka predstavlja značajnu oblast primene mašinskog učenja (npr. praćenje rehabilitacije, podrška pacijentima sa muskuloskeletnim problemima, praćenje kvaliteta izvođenja vežbi). Savremeni sistemi koriste više modaliteta senzora (inercijalni senzori, pritisak, video/depth), čime se dobijaju heterogeni izvori podataka pogodni za klasifikaciju.

U ovom radu analiziran je MEx (Multi-modal Exercise) skup podataka sa ciljem rešavanja višeklasnog problema klasifikacije tipa vežbe. Fokus rada je na: (i) pretprocesiranju podataka, (ii) redukciji dimenzionalnosti (PCA i selekcija atributa), (iii) obuci najmanje pet različitih klasifikacionih algoritama i (iv) upoređivanju dobijenih rezultata.

2 Opis zadatka

Zadatak zahteva:

- obavezno pretprocesiranje podataka,
- izgradnju minimum pet modela (različiti algoritmi),
- analizu i poređenje rezultata,
- vizuelni prikaz podataka u 2D ili 3D,
- poređenje modela treniranih na svim atributima i na različitim redukovanim skupovima atributa,
- pripremu materijala za ponavljanje postupka u lokalnom okruženju (podaci pre/posle obrade i sačuvani modeli).

U ovom radu posmatra se **klasifikacija** podataka.

3 Opis skupa podataka

MEx dataset sadrži podatke za **7 fizioterapeutskih vežbi** koje izvodi **30 ispitanika**. Svaki ispitanik izvodi vežbe do 60 sekundi, bez nametnutog ritma. Podaci su prikupljeni u četiri modaliteta: **2 akcelerometra**, **pressure mat** i **depth kamera**. [1]

Prema zvaničnom opisu skupa podataka:

- **Broj instanci**: 6262 [1]
- **Senzori i frekvencije**:
  - Axivity AX3 3-osni akcelerometar, 100Hz (dva uređaja: zglob i butina) [1]

  - Sensing Tex pressure mat, 15Hz, veličina frejma 32×16 [1]
  - Orbbec Astra depth kamera, 15Hz, originalno 240×320 (u datasetu skalirano na 12×16) [1]

- **Organizacija fajlova**: posebni folderi po senzoru i ispitaniku; kod vežbe 4 postoje dve strane izvođenja pa ima više fajlova. [1]

Raspodela instanci po klasama nije potpuno uravnotežena. Klasa 4 sadrži 465 instanci, dok ostale klase imaju po 30 instanci. Ovakva neravnoteža može dovesti do pristrasnosti modela ka dominantnoj klasi, naročito kod modela osetljivih na raspodelu podataka. Zbog toga je kao primarna metrika izabrana F1-macro, koja daje jednak značaj svim klasama bez obzira na njihov broj uzoraka.

U ovom radu, iz originalnih senzorskih zapisa formirane su numeričke karakteristike (features) po instanci, pri čemu svaka instanca predstavlja jedan segment vežbe opisan skupom izračunatih atributa. Ciljna promenljiva predstavlja tip vežbe (oznake 1–7). Time je problem formulisan kao višeklasna klasifikacija nad numeričkim atributima.

4 Pretprocesiranje podataka

Pre primene modela izvršeni su sledeći koraci:

- Uklanjanje/obrada nekonzistentnih vrednosti (npr. $\pm\infty$ prevođeno u NaN), zatim imputacija nedostajućih vrednosti.Za imputaciju nedostajućih vrednosti korišćena je median strategija, jer je robustnija na ekstremne vrednosti (outlier-e) u poređenju sa aritmetičkom sredinom. S obzirom da senzorski podaci mogu sadržati nagle promene ili šum, median predstavlja stabilniju procenu centralne tendencije.

- **Podela podataka:** Podaci su podeljeni na trening i test skup u odnosu 80:20 uz stratifikaciju po klasi (`stratify=y`) i fiksiran `random_state` radi reproduktivnosti.

- **Dimenzionalnost:** Nakon formiranja ulaznih atributa dobijeno je ukupno 156 numeričkih atributa po instanci, dok ciljna promenljiva predstavlja tip vežbe (7 klasa).

- **Standardizacija** numeričkih atributa (StandardScaler) radi ujednačavanja skala, što je posebno važno za distance-based i margin-based algoritme (KNN, SVM, Logistic Regression).Iako modeli zasnovani na stablu (Decision Tree, Random Forest) nisu osetljivi na skalu, u eksperimentu je radi konzistentnosti upotrebljena ista standardizovana reprezentacija za sve modele i sve varijante atributa.

Da bi se izbeglo curenje informacija (*data leakage*), standardizacija je fitovana **samo na trening skupu**, a zatim primenjena na test skup.

4.1 Čuvanje podataka za reprodukciju

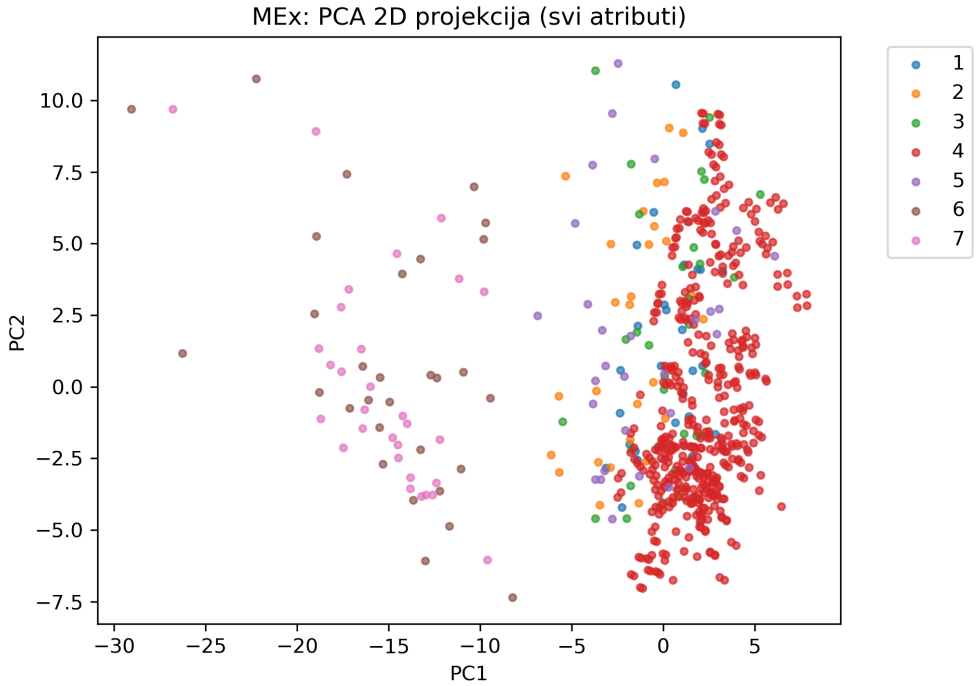
U folderu `outputs/` čuvaju se:

- podaci pre obrade (`mex_features_all_raw.csv`),
- podaci posle obrade (`mex_features_all_preprocessed.csv`),

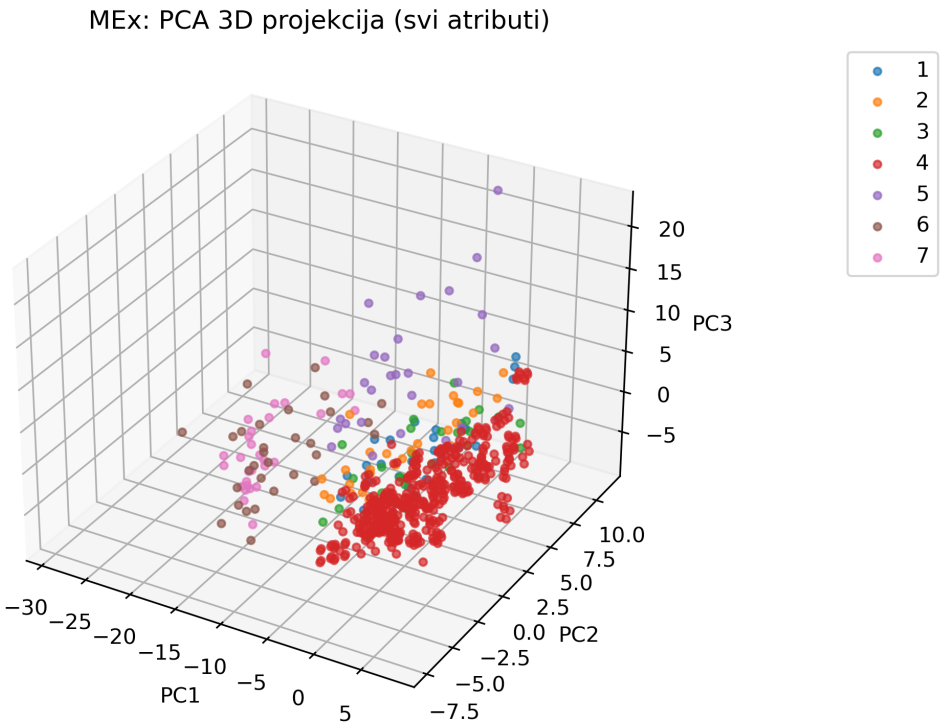
što omogućava ponavljanje eksperimenta i proveru svih koraka u lokalnom okruženju.

5 Vizuelizacija podataka (2D i 3D)

Zahtev zadatka je da se podaci prikažu u 2D ili 3D. U radu je korišćena PCA projekcija na 2 i 3 glavne komponente kako bi se vizuelno procenila separabilnost klasa. PCA projekcija pokazuje da su klase u velikoj meri separabilne u prostoru glavnih komponenti. Neke klase formiraju kompaktne i jasno izdvojene klastere, dok se pojedine delimično preklapaju, što je očekivano jer PCA predstavlja nenadgledanu metodu redukcije dimenzionalnosti koja ne koristi informacije o klasama. Ipak, vizuelna analiza potvrđuje da senzorski atributi sadrže dovoljno diskriminativnih informacija za uspešnu klasifikaciju.



Slika 1: PCA projekcija na 2 komponente.



Slika 2: PCA projekcija na 3 komponente.

6 Redukcija dimenzionalnosti

Kako bi se ispunio zahtev o treniranju modela i na redukovanim skupovima atributa, primenjena su dva pristupa:

- PCA redukcija na osnovu očuvane varijanse (90% i 95%),
- selekcija atributa metodom `SelectKBest` sa `mutual_info_classif` (KBest50).

6.1 PCA: izbor dimenzionalnosti preko očuvane varijanse

Umesto ručnog biranja broja komponenti, korišćen je kriterijum očuvanja varijanse: bira se minimalan broj komponenti takav da kumulativno objašnjava zadati procenat ukupne varijanse.

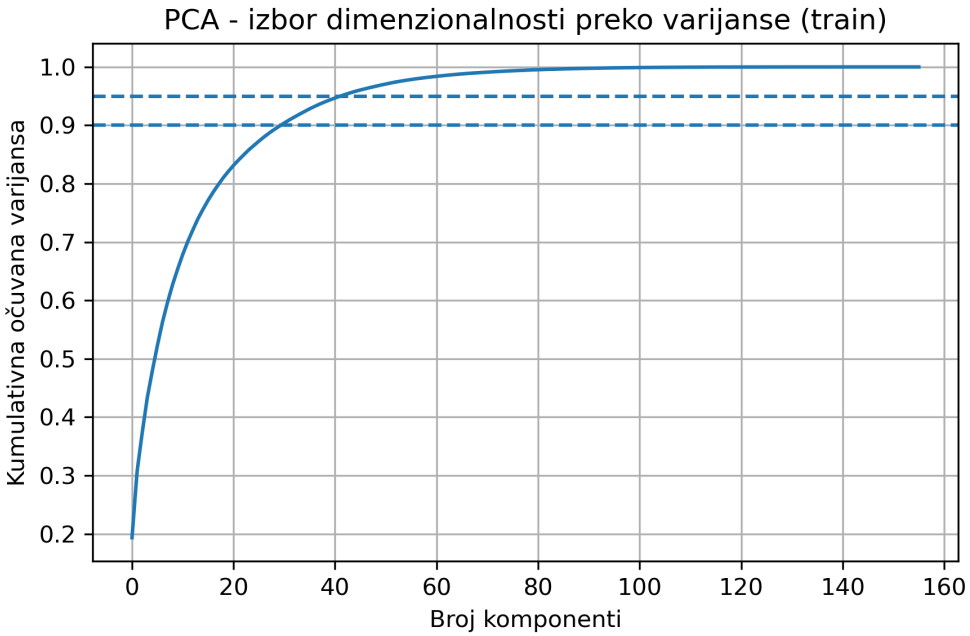
U eksperimentu su testirane dve vrednosti:

- **PCA90**: očuvanje 90% varijanse,
- **PCA95**: očuvanje 95% varijanse.

Dobijene dimenzionalnosti su:

- originalno: 156 atributa,

- PCA90: 31 komponenta,
- PCA95: 43 komponente.



Slika 3: Kumulativna očuvana varijansa u zavisnosti od broja komponenti.

6.2 KBest50: selekcija atributa

KBest50 zadržava 50 atributa sa najvećom zavisnošću od ciljne promenljive (mutual information). Ovaj pristup se razlikuje od PCA jer zadržava originalne atribute (interpretabilnost), dok PCA formira nove linearne kombinacije atributa.

7 Analiza algoritama klasifikacije

U ovom poglavlju analizirani su primenjeni klasifikacioni algoritmi. Za svaki model dat je kratak teorijski opis, način obrade podataka, performanse na punom i redukovanim skupovima atributa (PCA90, PCA95, KBest50), kao i interpretacija dobijenih rezultata.

7.1 Logistic Regression

Opšti opis modela

Logistic Regression je linearni model koji procenjuje verovatnoću pripadnosti klasi. U višeklasnom slučaju koristi se softmax:

$$P(y = k \mid x) = \frac{e^{w_k^T x}}{\sum_{j=1}^{K} e^{w_j^T x}}$$

gde x predstavlja vektor atributa, a w_k parametre za klasu k.

U implementaciji je korišćen model `LogisticRegression(max_iter=3000, C=0.01)`. Parametar `C` predstavlja inverznu jačinu regularizacije. Manja vrednost parametra (0.01) uvodi jaču regularizaciju, čime se smanjuje rizik od prenaučavanja u visoko-dimenzionalnom prostoru atributa. Povećana vrednost parametra `max_iter` (3000) omogućava stabilnu konvergenciju optimizacionog postupka.

Rezultati na punom skupu atributa

Na punom skupu atributa ostvareno je: Test Accuracy = 0.9845 i Test F1-macro = 0.9624. Ovi rezultati ukazuju da se klase mogu dobro razdvojiti i linearnom granicom odlučivanja.

Rezultati na redukovanim skupovima

U PCA90 i PCA95 varijantama performanse ostaju praktično iste (Test F1 $\approx$ 0.962), što znači da PCA komponente uspešno zadržavaju informacije relevantne za klasifikaciju. U KBest50 varijanti model takođe ostaje stabilan (Test F1 $\approx$ 0.9588).

Komentar

Logistic Regression se pokazao kao robustan model sa malom razlikom train/test performansi, što ukazuje na dobru generalizaciju.

7.2 K-Nearest Neighbors (KNN)

Opšti opis modela

KNN klasifikuje instancu na osnovu većine među k najbližih suseda u prostoru atributa (npr. euklidska udaljenost). Korišćen je model `KNeighborsClassifier()` sa podrazumevanim parametrima, pri čemu je broj suseda `n_neighbors=5`, a metrika rastojanja Euklidska udaljenost. Ova konfiguracija predstavlja standardnu postavku koja omogućava fer poređenje sa ostalim modelima bez dodatne optimizacije hiperparametara.

Rezultati na punom skupu atributa

Na punom skupu atributa dobijeno je: Test Accuracy = 0.9380 i Test F1-macro = 0.8297, što je slabije u odnosu na ostale modele.

Rezultati na redukovanim skupovima

U PCA i KBest varijantama performanse variraju, ali KNN i dalje zaostaje za ansambl metodama. Razlog je što u višedimenzionalnom prostoru često dolazi do problema “cur-

se of dimensionality": udaljenosti postaju manje informativne, a lokalno susedstvo može sadržati mešane klase.

Komentar

KNN je jednostavan, ali u ovakvim senzorskim podacima sa više klasa i većom dimenzionalnošću ne daje najbolje rezultate.

7.3 Decision Tree

Opšti opis modela

Decision Tree deli podatke rekurzivno na osnovu atributa sa ciljem povećanja čistoće čvorova.

Kriterijum podele (Gini). Za izbor podele korišćen je Gini indeks:

$$G = 1 - \sum_{k=1}^{K} p_k^2,$$

gde p_k predstavlja udeo klase k u čvoru. Korišćen je model `DecisionTreeClassifier(random_state=42` sa podrazumevanim kriterijumom podele `gini`. Nije postavljeno ograničenje maksimalne dubine stabla, što omogućava modelu da u potpunosti prilagodi strukturu trening podacima. Postavljanje parametra `random_state=42` obezbeđuje reproduktivnost rezultata.

Rezultati na punom skupu atributa

Decision Tree ostvaruje najbolje performanse: Test Accuracy = 0.9922 i Test F1-macro = 0.9760, uz Train Accuracy = 1.0. Perfektni trening rezultati su očekivani jer stablo može "naučiti" trening skup.

Rezultati na redukovanim skupovima

Kod PCA varijanti dolazi do značajnog pada (npr. Test F1 oko 0.71–0.76), što ukazuje da stablo najbolje funkcioniše u originalnom prostoru atributa gde može iskoristiti nelinearne odnose. Kod KBest50 varijante performanse ostaju visoke (Test F1 $\approx$ 0.9373), ali ipak ispod Full varijante.

Komentar

Stablo je interpretabilno i veoma precizno na punom skupu, ali je sklonije overfitting-u i osetljivije na promene reprezentacije (npr. PCA).

7.4 Random Forest

Opšti opis modela

Random Forest je ansambl metoda koja gradi više stabala i kombinuje njihove odluke većinskim glasanjem:

$$\hat{y} = \text{mode}(h_1(x), h_2(x), \ldots, h_T(x)).$$

Korišćen je model `RandomForestClassifier(n_estimators=200, max_depth=8, min_samples_leaf=`
`random_state=42)`.

Broj stabala (200) povećava stabilnost i smanjuje varijansu modela. Ograničenje dubine stabla (`max_depth=8`) i minimalni broj uzoraka u listu (`min_samples_leaf=5`) uvedeni su radi smanjenja prenaučavanja i bolje generalizacije na test skupu.

Rezultati na punom skupu atributa

Na Full varijanti dobijeno je: Test Accuracy = 0.9845 i Test F1-macro = 0.9725. Model ima visoke performanse i stabilnu generalizaciju.

Rezultati na redukovanim skupovima

U KBest50 varijanti Random Forest ostaje na istom nivou (Test F1-macro = 0.9725), što pokazuje da selekcija najinformativnijih atributa zadržava kvalitet klasifikacije uz nižu dimenzionalnost. U PCA varijantama performanse opadaju (npr. PCA90 Test F1 $\approx$ 0.7876), što je očekivano jer PCA transformiše atribute u linearne kombinacije, pa se gubi deo interpretabilnih nelinearnih odnosa koje stabla koriste.

Komentar

Random Forest daje odličan kompromis između tačnosti i robusnosti, posebno u kombinaciji sa selekcijom atributa (KBest50).

7.5 Support Vector Machine (SVM)

Opšti opis modela

SVM je margin-based algoritam koji traži hiperravan maksimalne margine. U linearnom slučaju rešava:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{uz uslov} \quad y_i(w^T x_i + b) \geq 1.$$

Za nelinearna razdvajanja može se koristiti kernel funkcija, npr. RBF:

$$K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2}.$$

U osnovnoj implementaciji korišćen je model `SVC(kernel="rbf")`. RBF (Radial Basis Function) kernel omogućava modelovanje nelinearnih granica odlučivanja u prostoru atributa, što je pogodno za kompleksne senzorske podatke.

Parametri `C` i `gamma` korišćeni su u podrazumevanim vrednostima (`C=1.0`, `gamma=šcale"`). Parametar `C` kontroliše kompromis između širine margine i tačnosti klasifikacije na trening podacima, dok `gamma` određuje uticaj pojedinačnih instanci na formiranje granice odlučivanja.

Rezultati na punom skupu atributa

Na Full varijanti dobijeno je: Test F1-macro = 0.8827, što je slabije od LR/DT/RF.

Rezultati na redukovanim skupovima

U PCA varijantama dolazi do poboljšanja (Test F1 $\approx$ 0.9087), što je očekivano jer PCA smanjuje dimenzionalnost i šum, a SVM često bolje radi u kompaktnijem prostoru.

Komentar

SVM je stabilan model, ali u ovom problemu LR i ansambl metode daju bolje rezultate na punom skupu atributa.

7.6 Naive Bayes

Opšti opis modela

Naive Bayes je probabilistički klasifikator zasnovan na Bayesovoj teoremi:

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)},$$

uz pretpostavku uslovne nezavisnosti atributa:

$$P(x_1, \ldots, x_n \mid y) = \prod_{i=1}^{n} P(x_i \mid y).$$

Korišćen je model `GaussianNB()`, koji pretpostavlja normalnu raspodelu atributa unutar svake klase. Model nema veliki broj hiperparametara, što ga čini jednostavnim i računski efikasnim za primenu.

Rezultati

Na Full varijanti dobijeno je: Test F1-macro = 0.8960. Rezultati su niži u odnosu na LR/DT/RF, što je očekivano jer su senzorski atributi često međuzavisni, pa pretpostavka

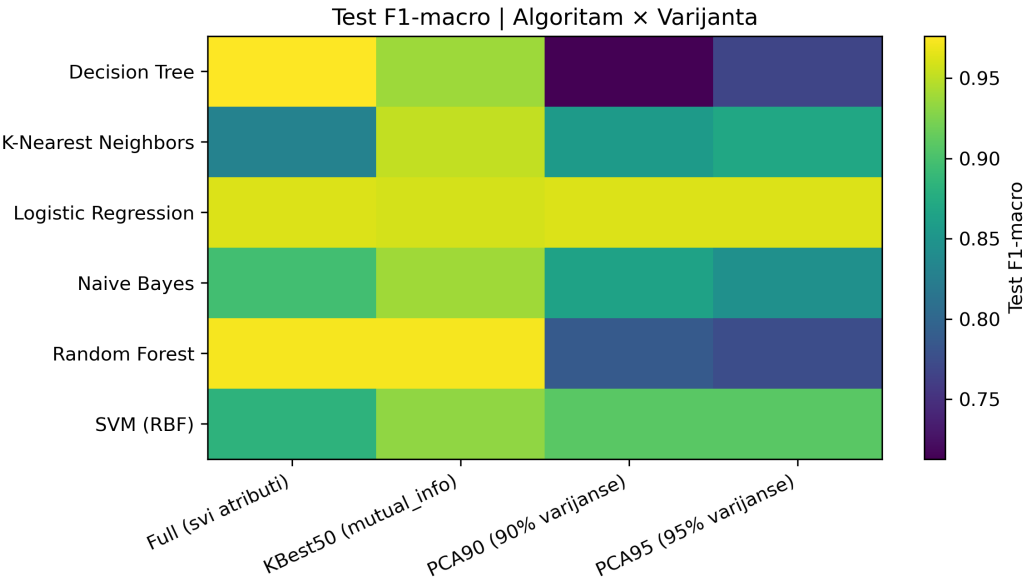
nezavisnosti nije ispunjena.

Komentar

Model je veoma brz i jednostavan, ali nije optimalan za ovakav tip podataka.

7.7 Poređenje performansi modela

Radi jasnijeg poređenja performansi algoritama nad različitim skupovima atributa, na slici 4 prikazane su vrednosti Test F1-macro metrike za sve kombinacije algoritama i varijanti atributa.



Slika 4: Test F1-macro vrednosti za sve algoritme i varijante atributa.

Sa slike se jasno vidi da Decision Tree i Random Forest postižu najviše performanse nad punim skupom atributa, dok linearni modeli poput Logistic Regression zadržavaju stabilne rezultate i nakon redukcije dimenzionalnosti. SVM pokazuje poboljšanje u PCA varijantama, dok KNN i Naive Bayes pokazuju veću osetljivost na izbor atributa.

Može se primetiti da Random Forest ostvaruje najbolje rezultate i u redukovanom prostoru, dok Logistic Regression i Decision Tree zadržavaju vrlo visoke performanse.

Na osnovu dobijenih rezultata zaključuje se:

- Najbolje performanse na punom skupu atributa ostvaruje Decision Tree (Test F1 = 0.9760, Test Acc = 0.9922).
- Random Forest je najstabilniji model i na KBest50 varijanti zadržava odlične performanse (Test F1 = 0.9725).

- PCA (90% i 95%) zadržava performanse kod linearnih modela (Logistic Regression), ali značajno narušava rezultate stabla i ansambl metoda.
- KBest50 predstavlja dobar kompromis: smanjuje dimenzionalnost, a često zadržava veoma visoke performanse.

7.8 Metrike

U višeklasnoj klasifikaciji Accuracy može prikriti slabiji rad nad manjim klasama, zato je kao primarna metrika korišćen **F1-macro**, koji računa F1 po klasi i zatim uzima aritmetičku sredinu (svaka klasa ima jednak značaj).

8 Optimizacija hiperparametara (GridSearch)

Kako performanse SVM modela značajno zavise od izbora hiperparametara, primenjena je metoda `GridSearchCV` nad trening skupom podataka.

Pretraživani su sledeći parametri:

- `C` $\in \{0.1, 1, 10, 50\}$
- `kernel` $\in \{\text{linear}, \text{rbf}\}$
- `gamma` $\in \{\text{scale}, \text{auto}\}$

Optimizacija je vršena korišćenjem unakrsne validacije, pri čemu je kao kriterijum izbora korišćena F1-macro metrika, jer je pogodna za višeklasne probleme i jednako tretira sve klase.

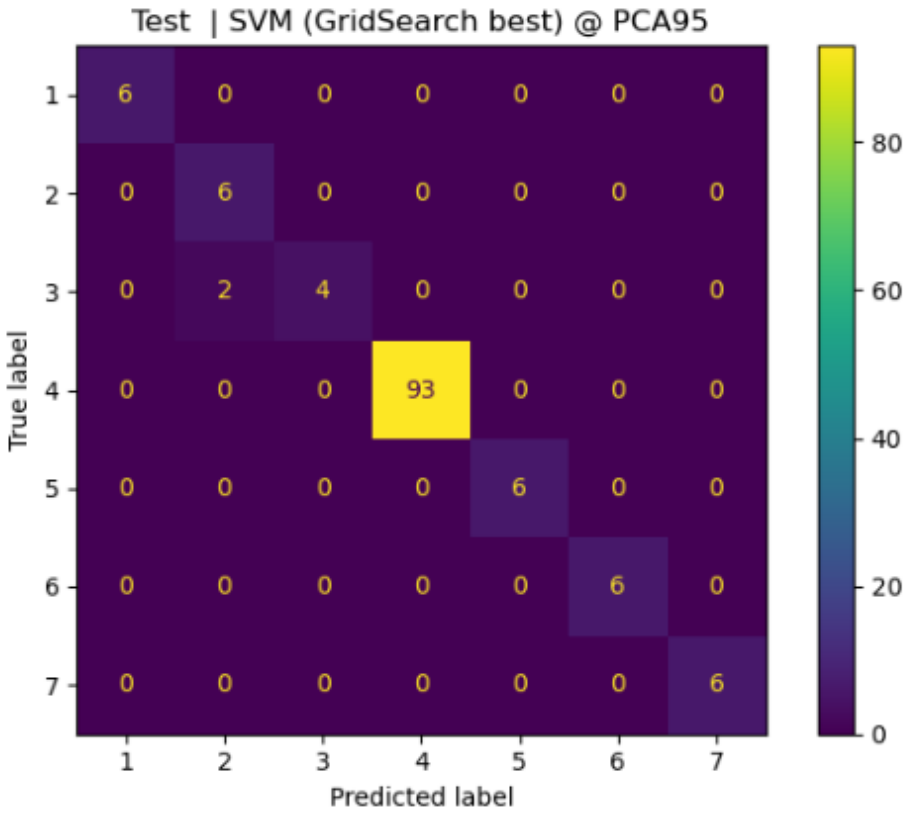
Najbolja konfiguracija dobijena pretragom bila je:

- `kernel = linear`
- `C = 0.1`
- `gamma = scale`

Dobijeni najbolji rezultat unakrsne validacije iznosio je F1-macro = 0.9292.

Zanimljivo je da je linearni kernel pokazao bolje performanse u odnosu na RBF kernel, što ukazuje da su klase u visoko-dimenzionalnom prostoru atributa u velikoj meri linearno separabilne. Manja vrednost parametra `C` (0.1) uvodi jaču regularizaciju, čime se postiže bolja generalizacija modela i smanjuje rizik od prenaučavanja.

Evaluacijom optimizovanog modela na test skupu dobijena je tačnost od 0.9845, uz visoku vrednost F1-macro metrike, što potvrđuje stabilnost i pouzdanost modela.



Slika 5: Confusion matrica za SVM (GridSearch) na test skup.

9 Rezultati i poređenje modela

Modeli su trenirani na četiri varijante atributa: Full, PCA90, PCA95 i KBest50.

9.1 Najbolji model ukupno

Najbolji rezultat postignut je na varijanti **Full (svi atributi)** korišćenjem **Decision Tree** algoritma:

- Test Accuracy = 0.9922
- Test F1-macro = 0.9760

Decision Tree model ostvaruje perfect score na trening skupu (Train Accuracy = 1.0, Train F1-macro = 1.0), što je očekivano jer stablo bez ograničenja može u potpunosti da prilagodi pravila trening podacima. Ipak, visoka test performansa pokazuje da model generalizuje dobro i da su klase u prostoru atributa značajno razdvojive.

9.2 Najbolji model po varijanti

Najbolji modeli po varijanti (po Test F1-macro, zatim po Test Accuracy) su:

- Full: Decision Tree (Test F1-macro = 0.9760, Test Acc = 0.9922)
- KBest50: Random Forest (Test F1-macro = 0.9725, Test Acc = 0.9845)
- PCA90: Logistic Regression (Test F1-macro = 0.9624, Test Acc = 0.9845)
- PCA95: Logistic Regression (Test F1-macro = 0.9623, Test Acc = 0.9845)

Ovo pokazuje da redukcija dimenzionalnosti može zadržati veoma visoke performanse čak i uz značajno smanjenje broja atributa (npr. 156 $\rightarrow$ 31 kod PCA90). KBest50 dodatno pokazuje da selekcija informativnih atributa može biti konkurentna modelima treniranim na punom skupu atributa.

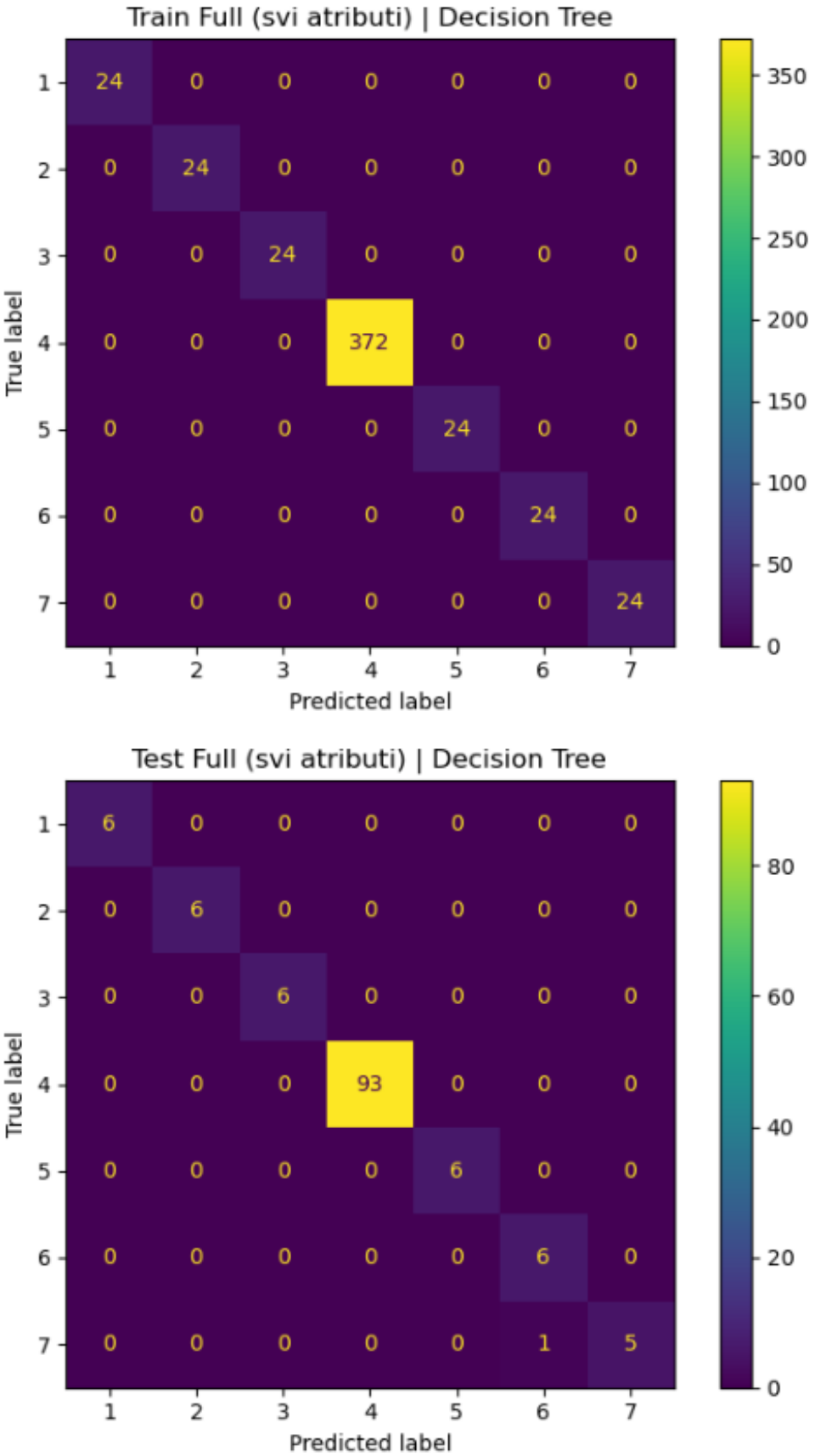
9.3 Confusion matrice

Confusion matrice kod najboljih modela pokazuju dominantne vrednosti na dijagonali, što ukazuje na vrlo visoku tačnost klasifikacije. Van-dijagonalne vrednosti su male i predstavljaju sporadične greške, uglavnom između klasa sa sličnim obrascima senzorskih merenja.

Confusion matrica za najbolji model (Decision Tree, Full varijanta) pokazuje dominantne vrednosti na glavnoj dijagonali, što ukazuje na izuzetno visoku tačnost klasifikacije. Na trening skupu model ostvaruje savršenu klasifikaciju svih instanci (Train Accuracy = 1.0), što je očekivano za stablo odlučivanja bez ograničenja dubine.

Na test skupu prisutna je samo jedna pogrešna klasifikacija — jedna instanca klase 7 pogrešno je klasifikovana kao klasa 6. Sve ostale klase su klasifikovane bez greške. Ovaj rezultat dovodi do Test Accuracy = 0.9922 i Test F1-macro = 0.9760. Niža vrednost F1-macro u odnosu na Accuracy posledica je disbalansa klasa i činjenice da macro prosek jednako tretira sve klase, uključujući i one sa malim brojem uzoraka.

Rezultati ukazuju na visoku separabilnost klasa u originalnom atributnom prostoru i dobru generalizaciju modela.

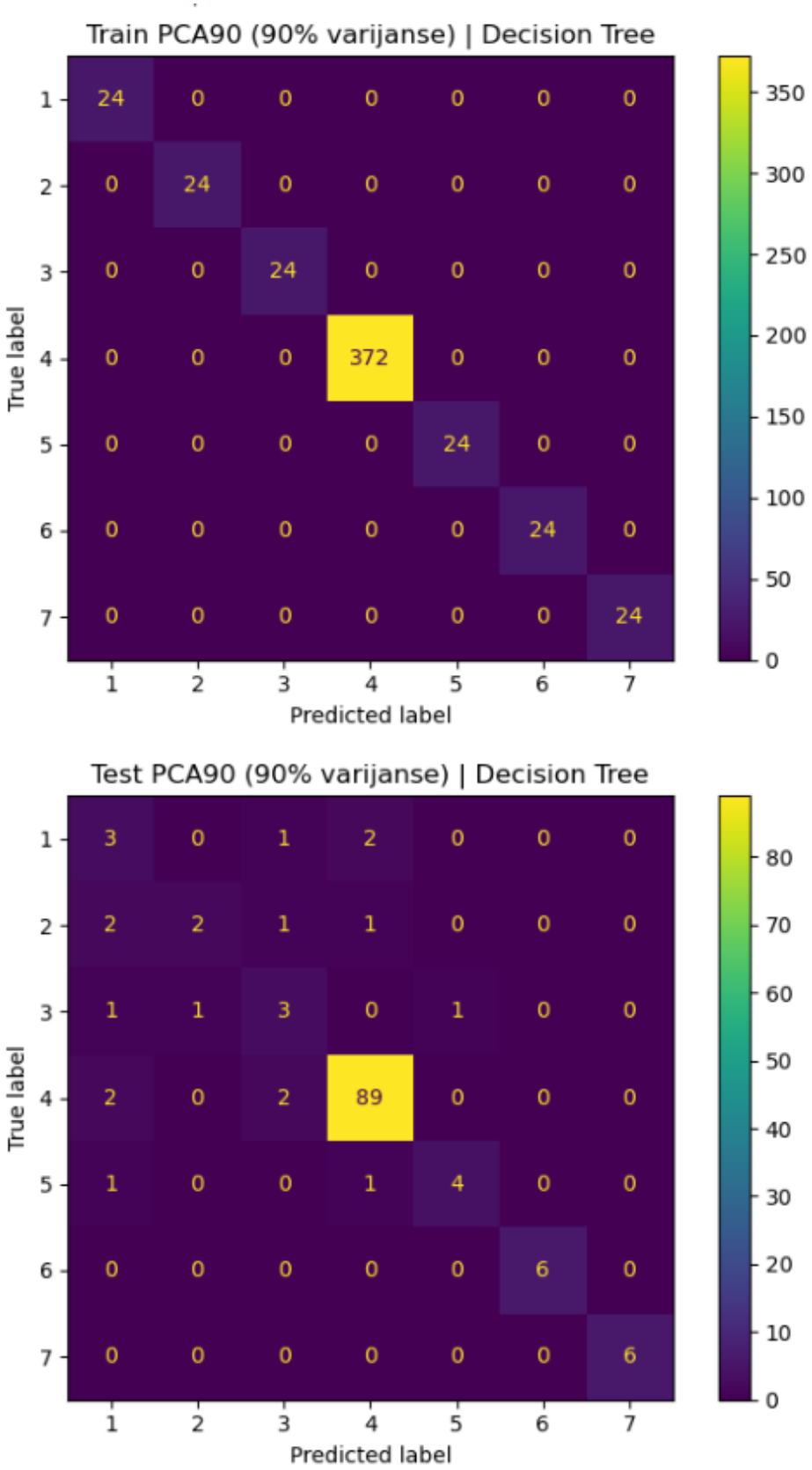


Slika 6: Confusion matrica (train i test) za najbolji model.

Za ilustraciju uticaja redukcije dimenzionalnosti na modele zasnovane na stablu, prikazana je i confusion matrica za kombinaciju PCA90 + Decision Tree, koja predstavlja jedan od slabijih rezultata u eksperimentu.

U odnosu na Full varijantu, primećuje se značajno povećanje broja pogrešnih klasifikacija, naročito kod manjih klasa (1, 2, 3 i 5). Klasa 4 ostaje dominantno pravilno klasifikovana, dok se manje zastupljene klase češće mešaju međusobno.

Ovaj rezultat potvrđuje da PCA transformacija, iako zadržava veliki procenat ukupne varijanse, može ukloniti ili izmeniti lokalne nelinearne odnose između atributa koje Decision Tree koristi za precizno razdvajanje klasa. Time se pokazuje da redukcija dimenzionalnosti ne utiče jednako na sve tipove modela.



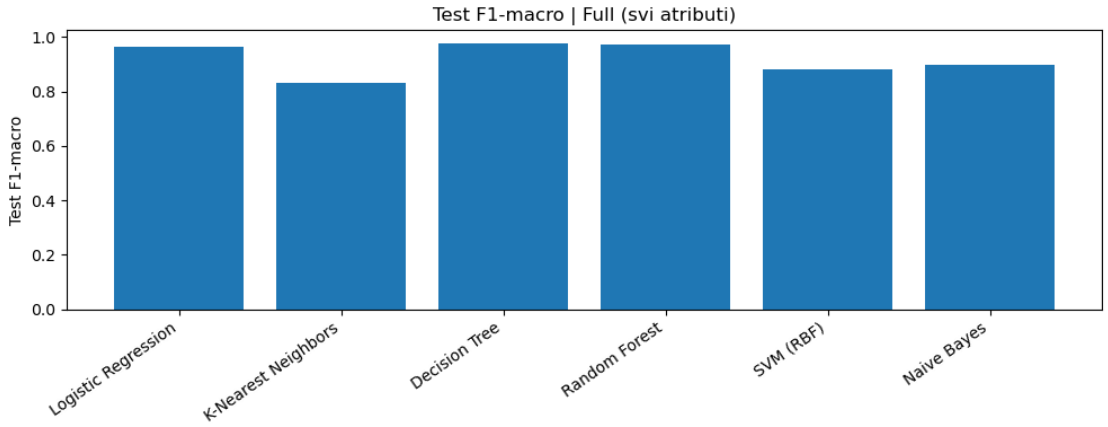
Slika 7: Confusion matrica za PCA90 varijantu sa Decision Tree modelom.

9.4 Detaljna analiza performansi po varijantama atributa

Pored tabelarnog prikaza rezultata, u nastavku je dat i vizuelni prikaz Test F1-macro metrike po algoritmima za svaku varijantu atributa. Ovakav prikaz omogućava brže uočavanje trendova i stabilnosti algoritama u zavisnosti od reprezentacije podataka (originalni atributi, PCA projekcija, selekcija atributa).

9.4.1 Full (svi atributi)

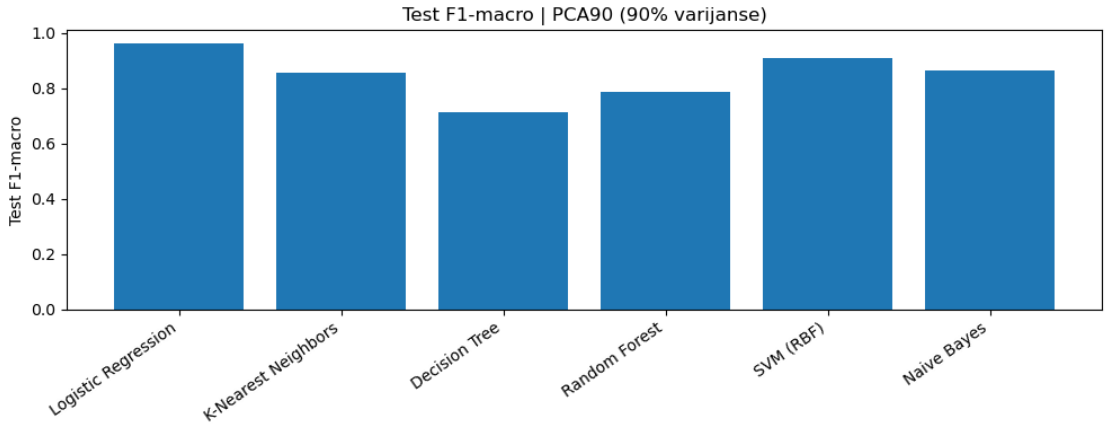
Na slici 8 prikazane su performanse modela treniranih na punom skupu atributa. Uočava se da modeli zasnovani na stablima (Decision Tree i Random Forest) ostvaruju najviše vrednosti F1-macro, što ukazuje da originalni atributni prostor sadrži nelinearne obrasce koje ovi modeli veoma uspešno iskorišćavaju. Logistic Regression takođe pokazuje stabilne i visoke performanse, što sugeriše da su klase u značajnoj meri razdvojive i linearnim granicama odlučivanja. Sa druge strane, KNN i Naive Bayes postižu niže vrednosti, što može biti posledica osetljivosti na dimenzionalnost (KNN) i narušene pretpostavke nezavisnosti atributa (Naive Bayes).



Slika 8: Test F1-macro po algoritmima za Full varijantu (svi atributi).

9.4.2 PCA90 (90% očuvane varijanse)

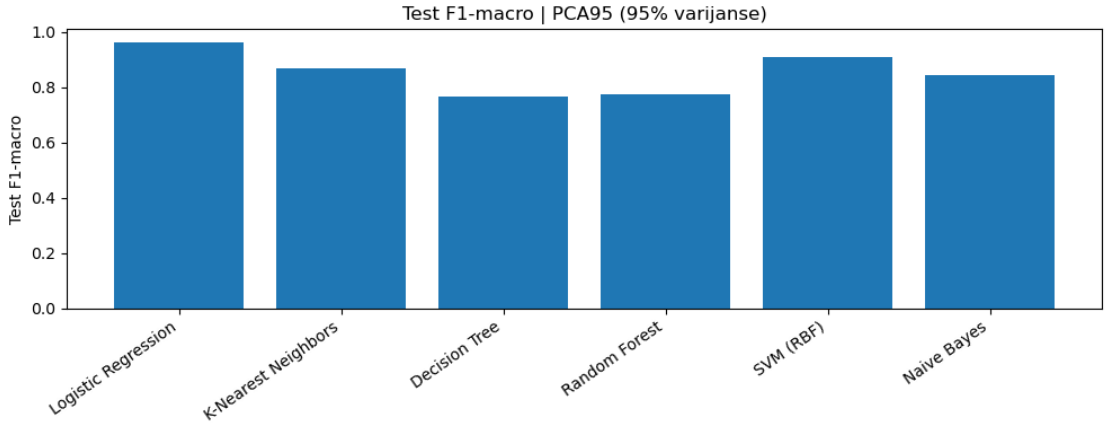
Na slici 9 prikazane su performanse modela nakon redukcije dimenzionalnosti PCA metodom uz očuvanje 90% varijanse. U ovoj varijanti se jasno vidi značajan pad performansi kod Decision Tree i Random Forest modela. Razlog je što PCA transformiše originalne atribute u linearne kombinacije, čime se menja struktura prostora na način koji nije optimalan za modele koji se oslanjaju na diskretne podele atributa i lokalne nelinearne odnose. Nasuprot tome, Logistic Regression ostaje gotovo nepromenjen, jer PCA često dobro očuva globalnu linearnu strukturu podataka, što pogoduje linearnim klasifikatorima. SVM pokazuje poboljšanje u odnosu na Full varijantu, što je očekivano jer redukcija dimenzionalnosti može smanjiti šum i olakšati formiranje stabilnije granice odlučivanja.



Slika 9: Test F1-macro po algoritmima za PCA90 varijantu (90% očuvane varijanse).

9.4.3 PCA95 (95% očuvane varijanse)

Na slici 10 prikazane su performanse modela za PCA95 varijantu. U odnosu na PCA90, očuvanjem većeg procenta varijanse performanse se delimično stabilizuju, ali se i dalje primećuje da modeli zasnovani na stablima ne dostižu nivo rezultata kao na punom skupu atributa. Ovo potvrđuje da PCA, iako zadržava veliki deo varijanse, može ukloniti ili „razmazati" lokalne strukture i nelinearne odnose koji su relevantni za stabla odlučivanja. Logistic Regression i SVM ostaju konkurentni i pokazuju stabilnost, što sugeriše da je dominantna informacija za razdvajanje klasa i dalje prisutna u relativno malom broju glavnih komponenti.

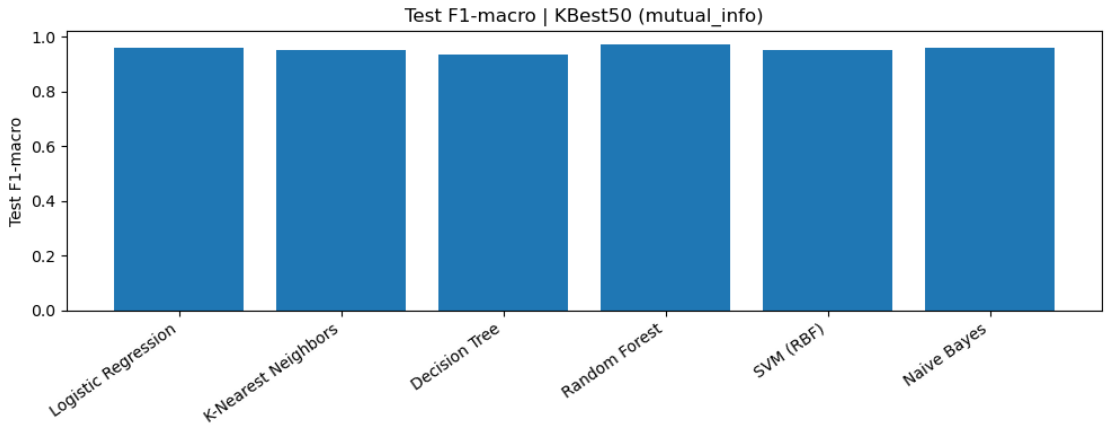


Slika 10: Test F1-macro po algoritmima za PCA95 varijantu (95% očuvane varijanse).

9.4.4 KBest50 (selekcija atributa mutual information)

Na slici 11 prikazane su performanse modela na KBest50 varijanti, gde se zadržava 50 najinformativnijih originalnih atributa (mutual information). Za razliku od PCA, selekcija atributa ne menja reprezentaciju podataka, već uklanja manje relevantne karakteristike, čime se smanjuje dimenzionalnost uz očuvanje interpretabilnosti. Uočava se da Random

Forest i u ovoj varijanti zadržava veoma visoke performanse, što ukazuje da je informacija potrebna za klasifikaciju koncentrisana u relativno malom podskupu atributa. Ovaj rezultat je značajan i iz praktičnog ugla, jer smanjenje broja atributa može ubrzati obuku i predikciju, kao i olakšati primenu u real-time scenarijima.



Slika 11: Test F1-macro po algoritmima za KBest50 varijantu (mutual information).

Sumarni zaključak

Vizuelna analiza performansi potvrđuje da izbor reprezentacije atributa značajno utiče na ponašanje algoritama. Linearni modeli (Logistic Regression) pokazuju visoku stabilnost u PCA prostoru, dok modeli zasnovani na stablima postižu najbolje rezultate u originalnom atributnom prostoru ili uz selekciju informativnih atributa (KBest50). SVM pokazuje poboljšanje nakon PCA redukcije, što je u skladu sa činjenicom da ovaj model često bolje radi u kompaktnijim prostorima sa manje šuma. KBest50 se izdvaja kao dobar kompromis, jer značajno smanjuje dimenzionalnost bez značajnog gubitka performansi, posebno kod ansambl metoda.

10 Reproduktivnost i sadržaj foldera outputs

Radi ispunjavanja zahteva da materijal sadrži sve za ponavljanje postupka, u folderu `outputs/` čuvaju se:

- podaci pre obrade i posle obrade,
- rezultati svih modela (`model_results_all.csv`),
- najbolji modeli (`best_per_variant.csv`, `best_overall.csv`),
- objekti potrebni za transformacije (scaler, PCA90, PCA95, KBest50),
- sačuvani najbolji modeli u `.joblib` formatu.

11 Zaključak

U ovom radu prikazana je kompletna procedura rešavanja višeklasnog problema klasifikacije tipa fizičke vežbe nad MEx skupom podataka, počev od pretprocesiranja, preko redukcije dimenzionalnosti, do obuke i poređenja više klasifikacionih algoritama.

Rezultati pokazuju da senzorski podaci sadržani u MEx skupu nose dovoljno informacija za vrlo precizno razlikovanje tipova vežbi. Najbolje performanse ostvario je Decision Tree model treniran na punom skupu atributa (Test Accuracy = 0.9922, Test F1-macro = 0.9760). Iako je model postigao savršene rezultate na trening skupu (Train Accuracy = 1.0), visoka tačnost na test skupu ukazuje na dobru generalizaciju i visoku separabilnost klasa u prostoru atributa. Ovakvo ponašanje je tipično za stabla odlučivanja koja su u stanju da precizno modeluju kompleksne granice odlučivanja.

Random Forest model, kao ansambl metoda zasnovana na stablu odlučivanja, pokazao je stabilne i visoke performanse na redukovanom skupu atributa (KBest50), što potvrđuje da selekcija najinformativnijih atributa može zadržati gotovo istu preciznost kao i puni skup atributa. Ova osobina je značajna u kontekstu smanjenja dimenzionalnosti i potencijalne primene modela u realnom vremenu, gde je efikasnost važan faktor.

Logistic Regression model ostvario je vrlo konkurentne rezultate u PCA90 i PCA95 varijanti. Ovi rezultati pokazuju da linearni modeli mogu postići visoku preciznost čak i nakon značajne redukcije dimenzionalnosti ($156 \rightarrow 31$ ili 43 komponente). To implicira da su glavne komponente uspešno sačuvale dominantne obrasce varijabilnosti podataka.

SVM model, uz optimizaciju hiperparametara pomoću GridSearch procedure, takođe je pokazao visoku diskriminativnu moć. Ovaj model je naročito pogodan za visokodimenzionalne prostore, što je relevantno u kontekstu senzorskih podataka sa velikim brojem atributa.

Naive Bayes model, iako jednostavan i brz, pokazao je slabije performanse u poređenju sa ostalim algoritmima. Razlog leži u pretpostavci uslovne nezavisnosti atributa, koja u realnim senzorskim podacima često nije ispunjena.

Redukcija dimenzionalnosti preko PCA (90% i 95% očuvane varijanse) pokazala je da je moguće značajno smanjiti broj atributa uz minimalan pad performansi. Time se potvrđuje da veliki deo informacija u podacima leži u relativno malom broju dominantnih komponenti. Selekcija atributa metodom KBest50 dodatno potvrđuje da informativni podskup atributa može biti gotovo jednako efikasan kao kompletan skup.

Confusion matrice pokazuju dominantne vrednosti na dijagonali, sa minimalnim brojem pogrešnih klasifikacija. Greške su sporadične i javljaju se uglavnom između klasa sa sličnim obrascima senzorskih signala, što je očekivano u višeklasnim problemima.

Na osnovu dobijenih rezultata može se zaključiti da je kombinacija adekvatnog pretprocesiranja, pažljivog izbora metrika (F1-macro), redukcije dimenzionalnosti i testiranja više algoritama omogućila izgradnju modela sa vrlo visokom pouzdanošću klasifikacije.

MEx skup podataka pokazuje visok nivo separabilnosti među klasama, što omogućava stabilne i precizne modele.

Potencijalna buduća unapređenja uključuju:

- detaljniju analizu grešaka po klasama,
- eksperimentisanje sa regularizacijom stabla odlučivanja radi smanjenja potencijalnog overfitting-a,
- ispitivanje dubokih neuronskih mreža za sekvencijalne podatke,
- evaluaciju modela u real-time scenariju.

Ukupno gledano, rezultati potvrđuju da je moguće ostvariti vrlo visoku tačnost klasifikacije fizičkih vežbi primenom klasičnih tehnika mašinskog učenja, uz pažljivo sprovedenu analizu i metodološki ispravan pristup. Iako Decision Tree ostvaruje najvišu tačnost, Random Forest predstavlja robusniji izbor u praktičnim primenama. Kombinovanjem više stabala i nasumičnim izborom podskupova atributa, Random Forest smanjuje varijansu modela i umanjuje rizik od prenaučavanja. Zbog toga se u realnim sistemima često preferira ansambl pristup, naročito kada postoji mogućnost šuma ili promena u podacima.

Literatura

[1] UCI Machine Learning Repository: MEx Dataset (id=500).
Wijekoon, A., Wiratunga, N., & Cooper, K. (2019). MEx [Dataset]. UCI Machine Learning Repository. DOI: 10.24432/C59K6T.
`https://archive.ics.uci.edu/dataset/500/mex`