AKADEMIJA TEHNIČKO-VASPITAČKIH STRUKOVNIH STUDIJA
ODSEK NIŠ

Primena modela mašinskog učenja u predikciji srčanih oboljenja

Predmet: Inteligentni sistemi

Mentor:

Prof. dr Nikola Sekulović

Prof Nikola Vukotić

Student:

Anđela Mladenović 03/24

Niš, 2026.

Sadržaj

Spisak slika:

Spisak tabela:

Spisak ilustracija:

Rezime

U radu je korišćen dataset preuzet sa Kaggle platforme, namenjen klasifikaciji srčanih oboljenja, koji u originalnoj verziji sadrži približno 70.000 uzoraka. U okviru projekta sprovedena je analiza i priprema podataka, uključujući pregled karakteristika, klasifikaciju atributa i njihovo skaliranje.

Nakon faze pretprocesiranja pristupljeno je implementaciji i evaluaciji algoritama mašinskog učenja. Ispitivani su modeli Logističke regresije, K-najbližih suseda (KNN) i veštačka neuronska mreža (ANN). Kod svakog algoritma sprovedena je optimizacija relevantnih hiperparametara, a za konačnu evaluaciju odabrane su konfiguracije koje su pokazale najbolje performanse na datom skupu podataka.

S obzirom na to da je skup podataka bio uravnotežen, modeli poput logističke regresije i neuronske mreže mogli su efikasno da nauče obrasce u podacima, dok je KNN algoritam uspešno procenjivao sličnosti između instanci na osnovu distance u prostoru karakteristika.

Uvod

U savremenom društvu zabeležen je porast učestalosti hroničnih oboljenja, pri čemu kardiovaskularne bolesti predstavljaju jedan od vodećih uzroka smrtnosti na globalnom nivou. Razvoj metoda mašinskog učenja omogućava primenu naprednih analitičkih tehnika u medicini, sa ciljem unapređenja ranog otkrivanja bolesti i podrške procesu donošenja odluka.

Motivacija ovog rada zasniva se na potrebi da se ispita primenljivost i efikasnost različitih klasifikacionih algoritama na realnim medicinskim podacima. U istraživanju je korišćen dataset preuzet sa Kaggle platforme, koji obuhvata veliki broj uzoraka i relevantne medicinske atribute povezane sa srčanim oboljenjima. Rad sa realnim i obimnim skupom podataka omogućava objektivnu procenu performansi modela i analizu njihove sposobnosti generalizacije.

Cilj rada nije bio isključivo postizanje maksimalne tačnosti, već komparativna analiza ponašanja različitih algoritama u uslovima konkretnog skupa podataka. Ispitivani su modeli logističke regresije, KNN algoritma i veštačkih neuronskih mreža, uz optimizaciju njihovih hiperparametara. Na taj način omogućeno je sagledavanje razlika u pristupu učenju, osetljivosti na strukturu podataka i stabilnosti performansi.

Relevantnost istraživanja ogleda se u primeni realnih podataka dovoljne veličine, što omogućava modelima da iskažu svoje prednosti i ograničenja u praktičnom kontekstu. Dobijeni rezultati doprinose boljem razumevanju izbora optimalnog modela za probleme medicinske klasifikacije i predstavljaju osnovu za dalju primenu mašinskog učenja u oblasti zdravstvene analitike.

1. Opis podataka

Skup podataka korišćen u ovom radu odnosi se na informacije o pacijentima i faktorima rizika povezanim sa pojavom kardiovaskularnih oboljenja. Dataset obuhvata podatke o osnovnim demografskim karakteristikama ispitanika, kao što su pol i starost, kao i informacije o životnim navikama, uključujući pušenje, konzumaciju alkohola i nivo fizičke aktivnosti. Pored toga, sadržani su i klinički parametri, među kojima su vrednosti krvnog pritiska, nivo glukoze u krvi i koncentracija holesterola. Ciljna promenljiva *cardio* označava prisustvo ili odsustvo srčanog oboljenja i predstavlja osnovu za dalju analizu i razvoj prediktivnih modela.

Pre pokretanja procesa čišćenja i transformacije, napravljena je kopija originalnog skupa podataka kako bi se omogućilo poređenje između početnih i regulisanih vrednosti. Ovakav pristup omogućava jasnije praćenje svih izvršenih izmena i doprinosi većoj transparentnosti i pouzdanosti u interpretaciji rezultata.

1.1 Analiza skupa podataka

Nakon učitavanja skupa podataka, sprovedena je osnovna analiza sa ciljem boljeg razumevanja njegove strukture i kvaliteta. U ovoj fazi analizirane su dimenzije skupa podataka, odnosno broj redova i kolona, kao i tipovi podataka dodeljeni svakoj promenljivoj. Posebna pažnja posvećena je proveri postojanja nedostajućih vrednosti, kao i izračunavanju osnovnih deskriptivnih statistika za numeričke atribute.

Rezultati analize pokazali su da inicijalni skup podataka ne sadrži nedostajuće vrednosti, što predstavlja dobru osnovu za dalju obradu. Ipak, uočena je potreba za dodatnim proverama logičke konzistentnosti pojedinih atributa, kao i za određenim transformacijama podataka. Ove aktivnosti su sprovedene kako bi se unapredio kvalitet skupa podataka i obezbedila pouzdanija osnova za dalju analizu i razvoj modela mašinskog učenja.

1.2 Čišćenje i transformacija podataka

Proces pripreme i čišćenja sproveden je u nekoliko faza.

Urađena je transformacija starosti. Atribut „age“ u originalnom csv fajlu izražen je u danima. Kako bi se povećala interperabilnost uveden je novi atribut „age_years“, koji predtsavlja starost ispitanika. Nakon toga je sprovedeno izračunavanje indeksa telesne mase (BMI). On predstavlja značajni faktor rizika za kardiovaskularna oboljenja. Ova transformacija

omogućava precizniju analizu uticaja telesne konstitucije na pojavu bolesti. Kolona id, koja ne sadrži informativnu vrednost uklonjena je iz skupa podataka. Ovim korakom sprečava se potencijalni šum, kao i nepotrebno opterećivanje modela. Izvršena je provera odnosa sistoličkog (*ap_hi*) i dijastoličkog (*ap_lo*) krvnog pritiska, pri čemu su analizirani slučajevi u kojima bi sistolički pritisak bio manji od dijastoličkog. Analiza je pokazala da takvi nelogični zapisi ne postoje u skupu podataka, što dodatno potvrđuje kvalitet podataka.

Nakon čišćenja podataka, sprovedena je i deskriptivna statistička analiza, kako bi se sagledala raspodela ključnih atributa i njihov potencijalni uticaj na pojavu srčanih oboljenja.

	Before cleaning	After cleaning
Rows	70000	68530
Columns	13	12

Slika 1. Prikaz pre i nakon čišćenja podataka

	Before cleaning	After cleaning
active	0.0	0.0
age	0.0	NaN
age_years	NaN	0.0
alco	0.0	0.0
ap_hi	0.0	0.0
ap_lo	0.0	0.0
cardio	0.0	0.0
cholesterol	0.0	0.0
gender	0.0	0.0
gluc	0.0	0.0
height	0.0	0.0
id	0.0	NaN
smoke	0.0	0.0
weight	0.0	0.0

Slika 2. Prikaz kolona pre i nakon čišćenja podataka

	mean_before	std_before	min_before	max_before	mean_after	std_after	min_after	max_after
id	49972.419900	28851.302323	0.0	99999.0	NaN	NaN	NaN	NaN
age	19468.865814	2467.251667	10798.0	23713.0	NaN	NaN	NaN	NaN
gender	1.349571	0.476838	1.0	2.0	1.348606	0.476533	1.00	2.0
height	164.359229	8.210126	55.0	250.0	164.436553	7.849141	122.00	207.0
weight	74.205690	14.395757	10.0	200.0	74.111033	14.242284	35.45	178.0
ap_hi	128.817286	154.011419	-150.0	16020.0	126.663345	16.631243	80.00	220.0
ap_lo	96.630414	188.472530	-70.0	11000.0	81.312009	9.388596	50.00	150.0
cholesterol	1.366871	0.680250	1.0	3.0	1.364643	0.678895	1.00	3.0
gluc	1.226457	0.572270	1.0	3.0	1.225741	0.571620	1.00	3.0
smoke	0.088129	0.283484	0.0	1.0	0.088005	0.283305	0.00	1.0
alco	0.053771	0.225568	0.0	1.0	0.053349	0.224730	0.00	1.0
active	0.803729	0.397179	0.0	1.0	0.803400	0.397430	0.00	1.0
cardio	0.499700	0.500003	0.0	1.0	0.494659	0.499975	0.00	1.0

Slika 3. Prikaz kolona sa podacima pre i posel čišćenja

Analiza i čišćenje podataka imali su za cilj pripremu dataseta u optimalnom obliku za dalju primenu i evaluaciju mašinskog učenja. Nakon uklanjanja netransparentnih vrednosti, transformacije relevantnih varijabli i selekcije odgovarajućih atributa, dobijen je uravnotežen i reprezentativan skup podataka za treniranje i testiranje modela.

Modeli K-najbližih suseda (KNN), logističke regresije i veštačke neuronske mreže (ANN) primenjeni su na skupu podataka koji sadrži sledeću distribuciju klasa,

	Count	Percentage (%)
cardio		
No Heart Disease	34631	50.53
Heart Disease	33899	49.47

Slika 4. Distribucija klasa posle čišćenja podataka

Prikazana distribucija predstavlja adekvatnu i proporcionalnu osnovu za objektivno ispitivanje performansi modela, jer omogućava realističnu procenu njihove sposobnosti da generalizuju i pravilno klasifikuju podatke.

Kreirane su metrike i analitički prikazi koji prikazuju u kojoj meri pojedini faktori utiču na rizik od srčanih oboljenja. Pre evaluacije modela, izvršeno je sklairanje numeričkih atributa kako bi se obezbedilo da sve promenljive imaju uporedive vrednosti. Skaliranje je posebno

važno za algoritme koji se oslanjaju na izračunavanje udaljenosti ili gradijentne optimizacije, kao što su KNN i neuronske mreže.

Standardizacija numeričkih atributa izvršena je primenom StandardScaler metode, pri čemu je skaliranje prilagođeno isključivo trening podacima, a zatim primenjeno na test skup radi očuvanja validnosti evaluacije modela.

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

num_cols = X_train.select_dtypes(include=['int64', 'float64']).columns

X_train_scaled = X_train.copy()
X_test_scaled = X_test.copy()

X_train_scaled[num_cols] = scaler.fit_transform(X_train[num_cols])
X_test_scaled[num_cols] = scaler.transform(X_test[num_cols])
```

Slika 5. Standardizacija numeričkih atributa

2. Arhitektura i primena modela

2.1 Logistička regresija

Logistička regresija je implementirana korišćenjem Python biblioteke scikit-learn. Model je treniran nad prethodno standardizovanim podacima, pri čemu su numeričke osobine skalirane primenom **StandardScaler** transformacije.

Za izbor optimalne vrednosti regularizacionog parametra **C**, sprovedeno je eksperimentalno testiranje više vrednosti.

```
C_values = [0.01, 0.1, 1, 10, 100]
```

Slika 6. Prikaz optimalnih vrednosti C

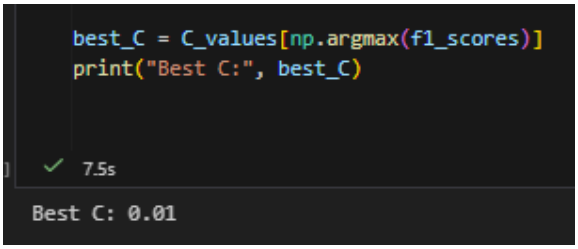
Za svaku vrednost parametra C treniran je poseban logistički regresioni model.

```
for C in C_values:
    log_reg = LogisticRegression(
        C=C,
        penalty='l2',
        solver='saga',
        class_weight='balanced',
        max_iter=5000,
        random_state=42
    )
    log_reg.fit(X_train_scaled, y_train)
    y_pred = log_reg.predict(X_test_scaled)
    f1_scores.append(f1_score(y_test, y_pred))
```

Slika 7. Prikaz logističkog regresionog modela

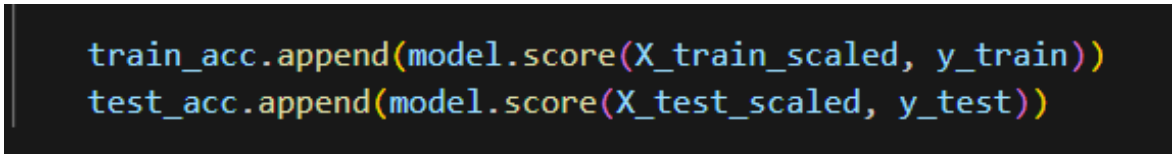
Kao kriterijum izbora optimalnog modela korišćen je F1-score, jer predstavlja balans između preciznosti (precision) i osetljivosti (recall), što je naročito važno u medicinskim klasifikacionim problemima.

Na osnovu maksimalne vrednosti F1-score metrike izabrana je optimalna vrednost parametra.



Slika 8. Prikaz optimalne vrednosti parametra C

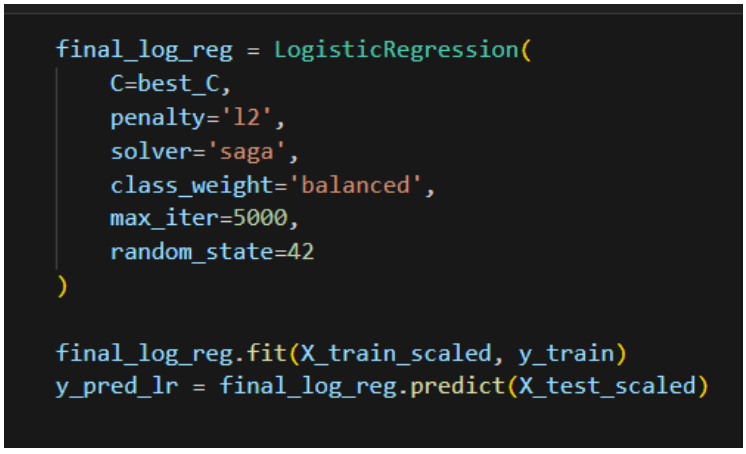
Radi provere potencijalnog overfitting-a i underfitting-a, dodatno su analizirane performanse modela na trening i test skupu za svaku vrednost parametra C.



Slika 9. Prikaz dodatne analize (overfitting/underfitting)

Upoređivanjem dobijenih vrednosti omogućeno je sagledavanje sposobnosti generalizacije modela, pri čemu je izabrana ona vrednost parametra C koja obezbeđuje najbolji kompromis između tačnosti na trening i test skupu.

Na osnovu prethodne analize, finalni model logističke regresije treniran je korišćenjem optimalne vrednosti parametra C.



Slika 10. Kod za izbor finalnog modela logističke regresije

Na ovaj način dobijen je stabilan i dobro generalizovan model, koji postiže visoke vrednosti tačnosti i F1-score metrike.

Logistička regresija je trenirana minimizacijom logaritamske funkcije gubitka (log-loss), uz primenu L2 regularizacije. Optimizacija parametara izvršena je korišćenjem SAGA optimizacionog algoritma, dok je hiperparametar C kontrolisao jačinu regularizacije. Optimalna vrednost parametra C izabrana je eksperimentalno na osnovu maksimalne vrednosti F1-score metrike.

2.1 K-nearest neighbors – KNN

U ovom radu primenjen je algoritam K-najbližih suseda (KNN) za rešavanje problema binarne klasifikacije. Princip rada algoritma zasniva se na pretpostavci da su međusobno slični uzorci bliski u prostoru obeležja, pa se klasa novog podatka određuje na osnovu klase većine među njegovim najbližim susedima. Drugim rečima, odluka o pripadnosti klasi donosi se analizom lokalnog okruženja posmatranog uzorka.

Kako KNN koristi euklidsko rastojanje kao osnovnu meru sličnosti, pre treniranja modela izvršena je standardizacija ulaznih promenljivih. Ovaj korak je realizovan primenom transformacije *StandardScaler*, čime su sve numeričke karakteristike dovedene na istu skalu. Na taj način sprečeno je da promenljive sa većim numeričkim opsegom imaju dominantan uticaj na računanje rastojanja, što doprinosi stabilnijem i pouzdanijem radu algoritma.

Početna vrednost parametra *k* postavljena je na 5, jer takav izbor predstavlja dobar kompromis između otpornosti na šum i sposobnosti generalizacije modela. Manje vrednosti *k* mogu dovesti do preosetljivosti na pojedinačne uzorke, dok veće vrednosti smanjuju sposobnost modela da prepozna fine strukture u podacima. Zbog toga je k = 5 odabran kao razumna početna tačka za dalju analizu i optimizaciju.

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report, confusion_matrix

knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train_scaled, y_train)

y_pred_knn = knn.predict(X_test_scaled)

print(classification_report(y_test, y_pred_knn))
```

Slika 11.Prikaz osnovne reference KNN (k=5)

Ovaj model poslužio je kao osnovna referenca (baseline) za poređenje sa unapređenim varijantama modela.

Radi pronalaženja optimalne vrednosti broja suseda, sprovedena je sistematska analiza vrednosti parametra k u intervalu od 1 do 21.

```
k_values = range(1, 21)
f1_scores = []

for k in k_values:
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train_scaled, y_train)
    y_pred = knn.predict(X_test_scaled)
    f1_scores.append(f1_score(y_test, y_pred))
```

Slika 12. Sistematska analiza parametra (k=1,21)

Kao kriterijum za izbor optimalnog modela korišćen je F1-score, budući da predstavlja uravnoteženu meru između preciznosti (precision) i osetljivosti (recall). Na osnovu maksimalne vrednosti F1-score metrike izabrana je optimalna vrednost.

Na osnovu prethodne analize, finalni KNN model treniran je korišćenjem optimalne vrednosti parametra.

```
knn_final = KNeighborsClassifier(n_neighbors=9)
knn_final.fit(X_train_scaled, y_train)
y_pred_knn = knn_final.predict(X_test_scaled)
```

Slika 13. Kod za finalni model KNN

Radi dodatne provere generalizacije modela, analizirana je zavisnost tačnosti na trening i test skupu u odnosu na parametar k.

```
train_accuracies.append(knn.score(X_train_scaled, y_train))
test_accuracies.append(knn.score(X_test_scaled, y_test))
```

Slika 14. Prikaz dela za overfiting/ underfiting

Ovom analizom je pokazano da male vrednosti parametra k (k = 1–3) dovode do izraženog overfitting-a, velike vrednosti k rezultuju underfitting-om,optimalna vrednost k nalazi se u srednjem opsegu, što potvrđuje izabranu vrednost k = 9.

2.3 Veštačke neuronske mreže- ANN(sgd)

U ovom radu korišćena je višeslojna perceptronska neuronska mreža (MLPClassifier) iz biblioteke *scikit-learn*. Model je treniran nad standardizovanim ulaznim podacima, kako bi se obezbedila stabilna i brza konvergencija algoritma.

Tabela 1. Korišćene varijante ANN modela

Model	Optimizator	Arhitektura	Dodatne tehnike	Opis
ANN – osnovni	Adam	Standardna	–	Referentni model za poređenje
ANN – Adagrad/SGD	Adagrad / SGD	Standardna	–	Analiza uticaja optimizatora
ANN – overfitting	Adam	Proširena	–	Ispitivanje preprilagođavanja
ANN + SMOTE	Adam	Standardna	SMOTE	Balansiranje klasa

U daljoj analizi, fokus je stavljen na osnovni ANN model sa SGD optimizatorom i overfitted varijantu, dok su preostali modeli korišćeni u uporednoj analizi performansi.

Osnovni model neuronske mreže definisan je sledećim kodom.

```
ann_sgd = MLPClassifier(
    hidden_layer_sizes=(32, 16),
    activation='relu',
    solver='sgd',
    learning_rate='adaptive',
    learning_rate_init=0.01,
    max_iter=5000,
    early_stopping=True,
    random_state=42
)

ann_sgd.fit(X_train_scaled, y_train)

y_pred_sgd = ann_sgd.predict(X_test_scaled)
```

Slika 15. Prikaz osnovnog modela neuronske mreže

Tabela 2. Arhitektura neuronske mreže

Komponenta	Parametar u kodu	Vrednost	Opis / Svrha
Tip modela	**—**	**MLPClassifier (ANN)**	**Feedforward veštačka neuronska mreža**
Broj skrivenih slojeva	**hidden_layer_sizes**	**2**	**Omogućava učenje nelinearnih odnosa**
Neuroni u 1. sloju	**hidden_layer_sizes**	**32**	**Učenje osnovnih obrazaca u podacima**
Neuroni u 2. sloju	**hidden_layer_sizes**	**16**	**Smanjenje složenosti i stabilizacija**
Aktivaciona funkcija	**activation**	**ReLU**	**Brza konvergencija i stabilan tok gradijenata**
Optimizacioni algoritam	**solver**	**SGD**	**Iterativno optimizuje težine pomoću gradijenta**
Strategija učenja	**learning_rate**	**adaptive**	**Automatsko prilagođavanje koraka učenja**
Početna stopa učenja	**learning_rate_init**	**0.01**	**Kontroliše brzinu ažuriranja težina**
Rano zaustavljanje	**early_stopping**	**True**	**Sprečava prenaučenost**
Maksimalan broj iteracija	**max_iter**	**5000**	**Omogućava potpunu konvergenciju**
Kontrola slučajnosti	**random_state**	**42**	**Reproduktivnost rezultata**

Radi demonstracije efekta overfitting-a, implementiran je složeniji model neuronske mreže.

```
ann_overfit = MLPClassifier(
    hidden_layer_sizes=(64, 64, 32),
    activation='relu',
    solver='adam',
    max_iter=5000,
    random_state=42
)
```

Slika 16. Prikaz koda za overfit- ANN

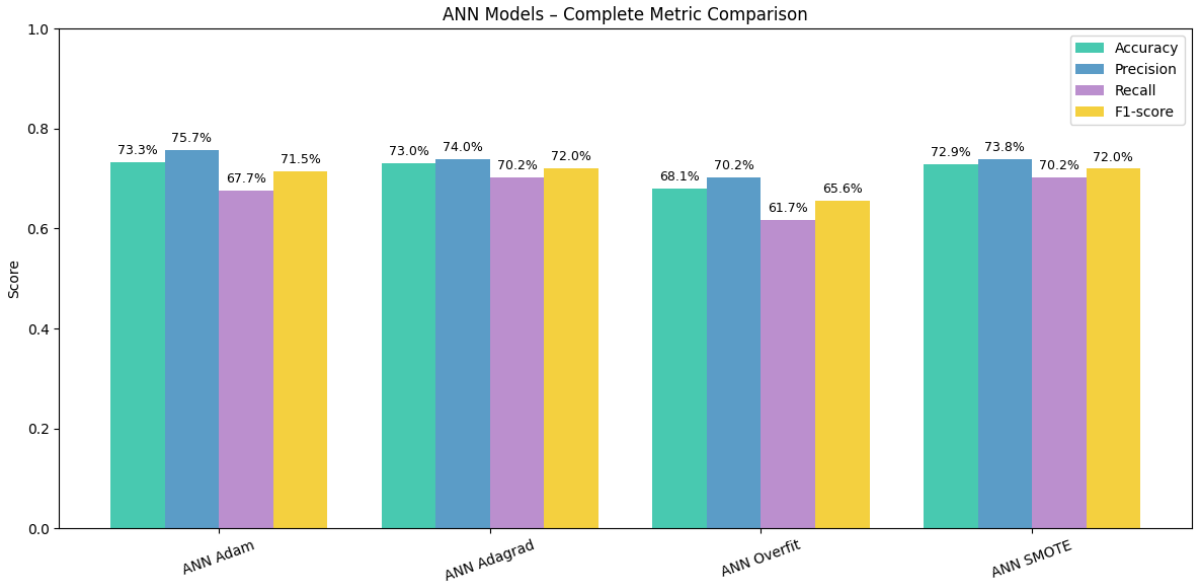
Tabela 3. Arhitektura overfiting-ANN

Komponenta	Parametar u kodu	Vrednost	Opis / Svrha
Tip modela	—	MLPClassifier (ANN)	Feedforward veštačka neuronska mreža
Broj skrivenih slojeva	`hidden_layer_sizes`	3	Veća dubina mreže povećava modelsku složenost
Neuroni u 1. skrivenom sloju	`hidden_layer_sizes`	64	Učenje kompleksnih obrazaca
Neuroni u 2. skrivenom sloju	`hidden_layer_sizes`	64	Povećana reprezentaciona moć
Neuroni u 3. skrivenom sloju	`hidden_layer_sizes`	32	Postepena redukcija dimenzionalnosti
Aktivaciona funkcija	`activation`	ReLU	Stabilan tok gradijenata i brza konvergencija
Optimizacioni algoritam	`solver`	Adam	Adaptivno podešavanje koraka učenja
Regularizacija	—	Nema eksplicitne L2 regulacije	Povećan rizik od prenaučenosti
Early stopping	—	Nije primenjen	Model može učiti predugo
Maksimalan broj iteracija	`max_iter`	5000	Omogućava dug proces treniranja
Kontrola slučajnosti	`random_state`	42	Reproduktivnost rezultata

3. Analiza rezultata i diskusija

U ovom poglavlju predstavljeni su rezultati koji su proistekli iz realizaciije projekta, sa posebnim osvrtom na to u kojoj meri su ispunjeni postavljeni ciljevi. Pored samog prikaza podataka, pažnja je usmerena na njihovo tumačenje i razumevanje u širem kontekstu projekta, kako bi se sagledali stvarni efekti sprovedenih aktivnosti i uočile eventualne prednsoti i nedostaci.

3.2 Rezultati izbora veštačke neuronske mreže – ANN



Ilustracija 1. ANN Modeli-uporedna analiza

Iako je ANN model treniran Adam optimizatorom ostvario nešto bolju ukupnu tačnost i preciznost, izbor konačnog modela nije zasnovan isključivo na ovim metrikama. U domenu medicinske dijagnostike, naročito kada je reč o srčanim oboljenjima, znatno je važnije da model pouzdano identifikuje pacijente sa prisutnim oboljenjem nego da postigne maksimalnu opštu tačnost.

Model treniran uz primenu SGD optimizatora pokazao je stabilniji odnos između preciznosti i odziva, što se odražava i u vrednosti F1-mere. Ovakav rezultat ukazuje na uravnoteženije ponašanje modela prilikom klasifikacije, odnosno na njegovu veću pouzdanost u prepoznavanju pozitivnih slučajeva. U praksi, propuštanje stvarno obolelog pacijenta

predstavlja ozbiljniju posledicu od pogrešne klasifikacije zdravog, zbog čega je veći odziv od posebnog značaja.

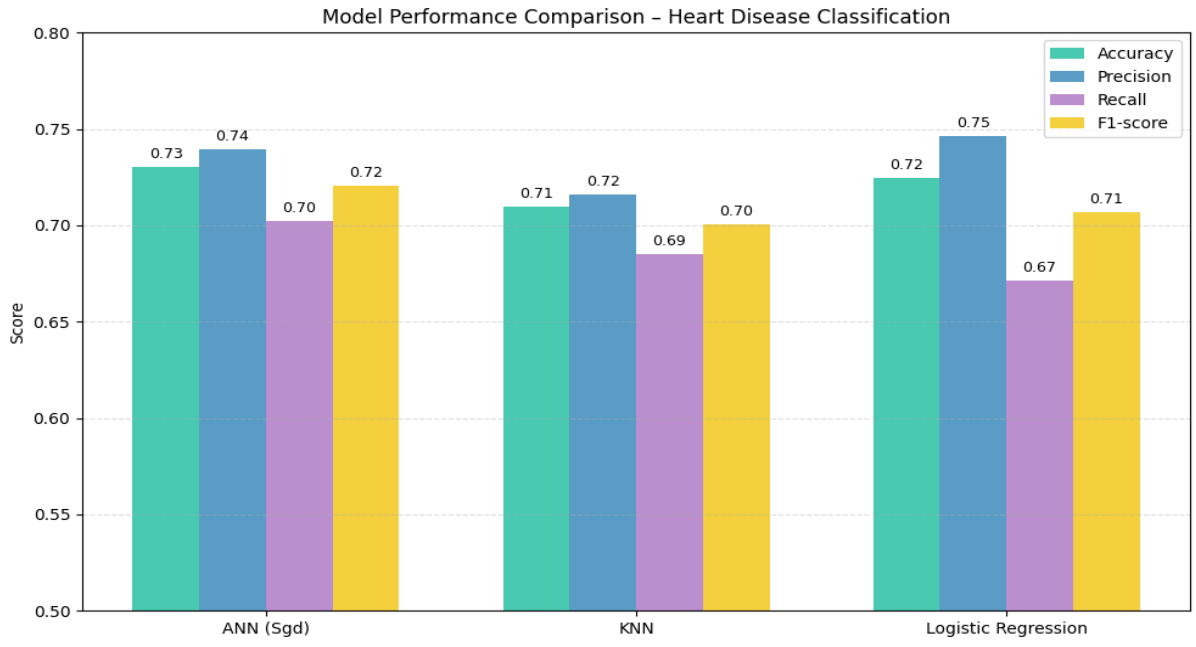
Takođe, SGD optimizacija je doprinela boljoj generalizaciji modela, omogućavajući efikasnije učenje iz podataka i smanjenje sklonosti ka favorizovanju većinske klase. Ovo je naročito relevantno u realnim medicinskim skupovima podataka, gde je neravnoteža između klasa česta i može negativno uticati na performanse modela.

Nasuprot tome, model treniran Adam optimizatorom, iako precizniji u predikciji negativnih slučajeva, pokazuje niži odziv, što ukazuje na oprezniji pristup u donošenju odluka. Takav pristup smanjuje broj lažno pozitivnih predikcija, ali istovremeno povećava rizik od neprepoznavanja pacijenata koji zaista imaju oboljenje, što predstavlja ograničenje u kontekstu zdravstvene primene.

Model sa izraženim overfitting efektom ostvario je lošije rezultate na test skupu, što potvrđuje da povećanje složenosti neuronske mreže bez odgovarajuće regularizacije ne dovodi do poboljšanja performansi, već narušava sposobnost generalizacije.

Na osnovu sveobuhvatne analize performansi, ANN model treniran uz SGD optimizator izabran je kao najprikladniji kompromis između tačnosti, stabilnosti i sposobnosti detekcije pozitivnih slučajeva, te je korišćen kao referentni model u daljem poređenju sa logističkom regresijom i KNN algoritmom.

3.2 Rezultati modela



Ilustracija 2. Uporedna analiza svih modela

Performanse razvijenih modela procenjene su primenom standardnih metrika za evaluaciju binarnih klasifikacionih problema: tačnosti (accuracy), preciznosti (precision), odziva (recall) i F1-mere. Ovakav izbor metrika omogućava sveobuhvatnu analizu ponašanja modela, posebno u kontekstu medicinskih podataka, gde je od suštinskog značaja pravilno identifikovati pozitivne slučajeve, uz istovremeno smanjenje broja lažno pozitivnih i lažno negativnih predikcija.

Na osnovu dobijenih rezultata, veštačka neuronska mreža trenirana pomoću SGD optimizatora ostvarila je najbolje ukupne performanse. U poređenju sa KNN algoritmom i logističkom regresijom, ANN model je pokazao nešto višu preciznost, što ukazuje na manju sklonost ka pogrešnom označavanju zdravih ispitanika kao obolelih. Iako je vrednost odziva kod KNN modela bila neznatno viša, ANN je ostvario uravnoteženiji odnos između preciznosti i odziva, što je rezultiralo najvišom F1-merom među svim testiranim modelima.

Izbor ANN(sgd) optimizatora za neuronsku mrežu pokazao se kao najpovoljnije rešenje, jer je ovaj algoritam omogućio stabilnu i efikasnu konvergenciju tokom treniranja, uz smanjenje oscilacija u procesu učenja. U poređenju sa ostalim testiranim varijantama neuronskih mreža, ANN(sgd) je obezbedio najbolju generalizaciju na test skupu, bez izraženih znakova preprilagođavanja. Zbog toga je ovaj model izabran kao reprezentativan predstavnik ANN pristupa za dalje poređenje sa logističkom regresijom i KNN algoritmom.

Kod logističke regresije posebna pažnja posvećena je izboru hiperparametra C, koji kontroliše jačinu regularizacije. Testiranjem više vrednosti ovog parametra utvrđeno je da optimalna vrednost omogućava postizanje stabilnih performansi, bez gubitka sposobnosti generalizacije. Time je izbegnuta pojava preprilagođavanja, a model je zadržao dobar balans između složenosti i tačnosti. Iako logistička regresija nije dostigla performanse neuronske mreže, ostvareni rezultati ukazuju na njenu pouzdanost i interpretabilnost, što je čini pogodnom za analizu medicinskih podataka.

Kod KNN algoritma, ključni faktor bila je optimizacija broja suseda k. Ispitivanjem različitih vrednosti parametra postignut je kompromis između lokalne osetljivosti i otpornosti na šum. Izabrana vrednost k omogućila je modelu da efikasno prepozna slične uzorke u prostoru obeležja, pri čemu su ostvarene stabilne performanse na test skupu. Iako KNN nije postigao najbolje ukupne rezultate, njegova jednostavnost i intuitivnost čine ga pogodnim za uporednu analizu i validaciju složenijih modela.

Prilikom formiranja ulaznog skupa podataka, iz modela su isključena određena obeležja koja nisu direktno doprinosila predikciji ili su mogla izazvati redundanciju i smanjenje performansi. Takođe, izvršena je standardizacija numeričkih promenljivih, kako bi se

obezbedila uporedivost njihovih vrednosti i stabilnost algoritama zasnovanih na udaljenosti i gradijentnim metodama. Prag verovatnoće za klasifikaciju zadržan je na standardnoj vrednosti od 0.5, jer se pokazalo da omogućava najuravnoteženiji odnos između preciznosti i odziva, bez značajnog favorizovanja bilo koje klase.

Na osnovu ukupnih rezultata može se zaključiti da ANN model sa sgd optimizacijom predstavlja najpouzdaniji pristup za posmatrani problem, dok logistička regresija i KNN algoritam služe kao relevantni referentni modeli koji potvrđuju stabilnost i opravdanost dobijenih nalaza.

4. Rezultati u kontekstu istraživačkog problema

Na osnovu konačne evaluacije modela, može se uočiti da su sva tri primenjena algoritma ostvarila relativno slične, ali ne identične performanse, što ukazuje na složenost samog problema klasifikacije srčanih oboljenja. Iako su vrednosti tačnosti za sve modele u rasponu od približno 71% do 73%, detaljnija analiza ostalih metrika otkriva značajne razlike u njihovom ponašanju, naročito u pogledu odnosa između preciznosti i odziva.

Veštačka neuronska mreža trenirana na uravnoteženom skupu podataka primenom ANN(sgd) tehnike ostvarila je najbolje ukupne performanse, sa najvišom tačnošću i F1-merom. Ovakav rezultat ukazuje na dobar balans između preciznosti i odziva, odnosno na sposobnost modela da istovremeno smanji broj lažno pozitivnih i lažno negativnih predikcija. Posebno je značajna relativno visoka vrednost odziva, koja sugeriše da model uspešno identifikuje veći broj stvarno obolelih pacijenata, što je od izuzetne važnosti u medicinskim sistemima za podršku odlučivanju. Primena ANN(sgd) tehnike doprinela je boljoj reprezentaciji pozitivne klase, čime je smanjena pristrasnost modela prema većinskoj klasi i poboljšana njegova diskriminativna sposobnost.

Kod logističke regresije zabeležena je nešto niža vrednost F1-mere u poređenju sa neuronskom mrežom, ali istovremeno i najviša preciznost među svim testiranim modelima. Ovakav rezultat ukazuje na to da je model naročito oprezan pri donošenju pozitivnih odluka, što dovodi do manjeg broja lažno pozitivnih klasifikacija. Međutim, ovakav pristup ima za posledicu i niži odziv, odnosno veći broj propuštenih pozitivnih slučajeva. U kontekstu dijagnostike, ovakva osobina može predstavljati ograničenje, jer je često poželjnije identifikovati što veći broj potencijalno obolelih pacijenata, čak i po cenu blagog povećanja broja lažnih alarma. Ipak, stabilnost rezultata i dobra interpretabilnost čine logističku regresiju pogodnom za situacije u kojima je transparentnost odlučivanja od posebnog značaja.

KNN algoritam ostvario je najniže vrednosti svih posmatranih metrika, što se može pripisati njegovoj zavisnosti od lokalne strukture podataka i osetljivosti na šum. Iako KNN pokazuje zadovoljavajuću sposobnost prepoznavanja sličnih obrazaca, njegovi rezultati ukazuju na ograničenu generalizaciju u prostoru sa većim brojem obeležja. Niža vrednost odziva sugeriše da model teže identifikuje pozitivne slučajeve, dok relativno skromna preciznost ukazuje na prisustvo većeg broja pogrešnih klasifikacija. Zbog toga se KNN može smatrati pogodnim za brzu i jednostavnu analizu, ali ne i kao primarni izbor u sistemima gde je neophodna visoka pouzdanost.

Sa aspekta poređenja metrika, može se zaključiti da sama tačnost ne predstavlja dovoljan pokazatelj kvaliteta modela, naročito u medicinskim problemima, gde su posledice različitih tipova grešaka nejednako značajne. F1-mera se pokazala kao najrelevantniji kriterijum za izbor optimalnog modela, jer objedinuje informacije o preciznosti i odzivu, pružajući realniju sliku o ukupnoj uspešnosti. Na osnovu toga, ANN model treniran uz (sgd) tehniku može se smatrati najpouzdanijim rešenjem, budući da ostvaruje najbolji kompromis između prepoznavanja obolelih i izbegavanja lažnih dijagnoza.

Analiza grešaka ukazuje na to da se najveći broj pogrešnih klasifikacija javlja kod pacijenata sa graničnim vrednostima kliničkih parametara, kod kojih razlika između zdravih i obolelih nije jasno izražena. Ovakvi slučajevi dodatno otežavaju proces učenja i predstavljaju glavni izvor nesigurnosti kod svih modela. Takođe, moguće je da određena obeležja ne nose dovoljnu količinu diskriminativnih informacija, ili da među njima postoji visok stepen međusobne korelacije, što dodatno komplikuje proces klasifikacije.

U cilju unapređenja performansi, buduća istraživanja mogla bi obuhvatiti primenu naprednijih tehnika selekcije i ekstrakcije obeležja, kao i korišćenje ansambl metoda koje kombinuju predikcije više modela. Takođe, dodatna optimizacija praga odlučivanja mogla bi doprineti povećanju odziva, čime bi se smanjio rizik od neotkrivanja obolelih pacijenata. Na taj način moguće je dodatno poboljšati pouzdanost sistema i prilagoditi ga zahtevima realnih medicinskih primena.

Zaključak

U okviru ovog rada analizirana je uspešnost različitih pristupa mašinskog učenja u predviđanju pojave srčanih oboljenja, sa posebnim akcentom na veštačke neuronske mreže kao modele sposobne da prepoznaju kompleksne obrasce u podacima. Nakon detaljne obrade skupa podataka – koja je obuhvatila čišćenje, proveru konzistentnosti i pripremu za modelovanje – sprovedeno je poređenje više algoritama: K-najbližih suseda, logističke regresije i nekoliko arhitektura neuronskih mreža. Evaluacija je izvršena primenom standardnih klasifikacionih metrika kako bi se dobila što potpunija slika o njihovim performansama.

Rezultati su pokazali da neuronska mreža optimizovana primenom SGD algoritma postiže najstabilniji odnos između tačnosti, preciznosti i odziva. U kontekstu medicinske dijagnostike, ovakav balans ima poseban značaj, jer je prioritet smanjiti broj lažno negativnih rezultata odnosno slučajeva u kojima bolest ostane neprepoznata. Čak i uz blago odstupanje u ukupnoj tačnosti, model koji pouzdano identifikuje rizične pacijente ima veću praktičnu vrednost.

Analiza je takođe ukazala na to da povećanje složenosti modela ne vodi nužno ka boljim rezultatima. Bez adekvatnih mehanizama regularizacije, kompleksnije strukture pokazuju sklonost ka preprilagođavanju trening podacima, što negativno utiče na njihovu sposobnost generalizacije. Ovaj nalaz dodatno potvrđuje važnost pažljivog podešavanja hiperparametara i kontrolisanja kapaciteta modela.

Na osnovu dobijenih rezultata može se zaključiti da pravilno konstruisani i optimizovani modeli mašinskog učenja imaju potencijal da pruže značajnu podršku u ranom otkrivanju srčanih oboljenja. Dalji rad u ovoj oblasti mogao bi obuhvatiti proširenje baze podataka, integraciju dodatnih kliničkih pokazatelja, kao i razvoj interpretabilnijih modela, čime bi se povećalo poverenje stručnjaka i olakšala njihova primena u svakodnevnoj medicinskoj praksi.

Literatura

1. GeeksforGeeks. (n.d.). *Artificial Neural Networks and Its Applications*. Preuzeto sa GeeksforGeeks: https://www.geeksforgeeks.org/deep-learning/artificial-neural-networks-and-its-applications/
2. IBM. (n.d.). *K-Nearest Neighbors (KNN)*. Preuzeto sa IBM Think: https://www.ibm.com/think/topics/knn
3. *Logistic Regression*. (n.d.). Preuzeto sa IBM Think: https://www.ibm.com/think/topics/logistic-regression
4. Sulianova, A. (2018). *Cardiovascular Disease Dataset*. Preuzeto sa Kaggle: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset