Drugi projekat iz predmeta *Veštačka inteligencija*

Mašinsko učenje kao pomoć za predikciju malignih ćelija kod raka dojke

Đorđe Antić 17544

Emilija Ćojbašić 18026

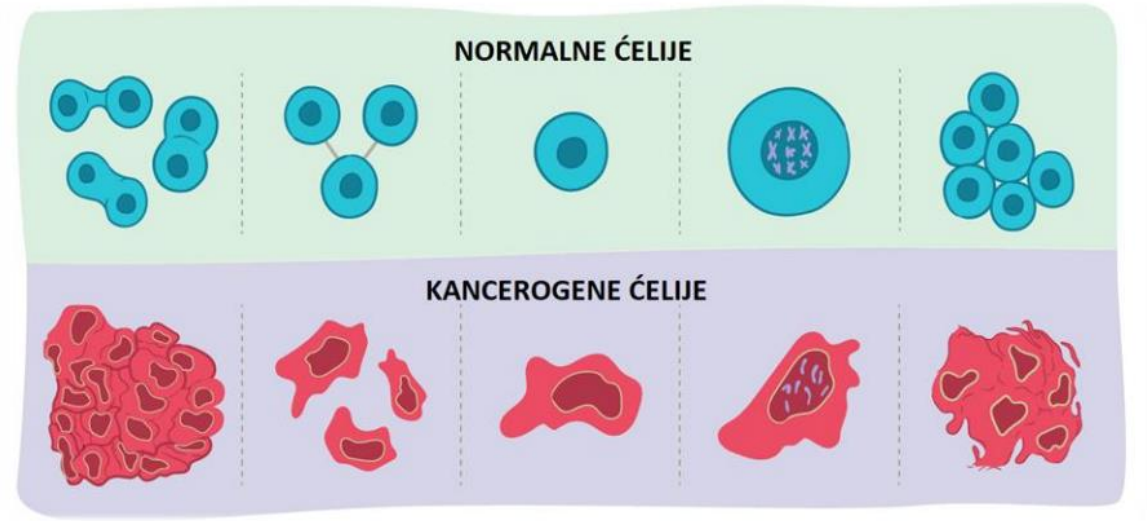
Matija Špeletić 18043

Link do repozitorijuma sa implementacijom i dataset-om:

https://github.com/djoleant/ML-for-breast-cancer-prediction

1. Kratak opis problema

Rak dojke je jedna od najčešćih malignih bolesti kod žena širom sveta. Nastaje od ćelija koje sačinjavaju tkiva dojke i u najvećem broju slučajeva zahteva hitnu intervenciju. Predikcija malignih ćelija kod raka dojke može biti od velike pomoći medicinskim stručnjacima prilikom donošenje odluke o daljem lečenju pacijenata. U ovom radu prikazano je kako se dve tehnike mašinskog učenja, stablo odlučivanja i neuronske mreže, mogu iskoristiti za klasifikaciju ćelija na benigne i maligne.



Slika 1. Razlika između normalnih i kancerogenih ćelija
Izvor: https://www.verywellhealth.com/cancer-cells-vs-normal-cells-2248794

2. Pregled korišćenih tehnika veštačke inteligencije

Mašinsko učenje je podoblast veštačke inteligencije čiji je cilj konstruisanje algoritama i računarskih sistema koji su sposobni da se adaptiraju na analogne nove situacije i uče na bazi iskustva. Razvijene su različite tehnike mašinskog učenja za različite zadatke [1]. Tehnike veštačke inteligencije (mašinskog učenja) korišćene u ovom radu su stablo odlučivanja i neuronske mreže.

2.1. Stablo odlučivanja

Stablo odlučivanja predstavlja metodu mašinskog učenja koja na osnovu kreiranog modela i ulaznih promenljivih predviđa vrednost ciljne promenljive. Stablo odlučivanja ima strukturu grafa toka. Unutrašnji čvorovi predstavljaju testiranje nekog atributa, a grane predstavljaju rezultate testiranja. Na krajevima stabla odlučivanja nalaze se listovi ili terminalni čvorovi. Listovi predstavljaju klasu koja je rezultat klasifikacije ulaznih podataka (kod klasifikacionog stabla odlučivanja) ili brojna vrednost (kod regresionog stabla odlučivanja). Stablo se konstruiše kroz deobu početnog skupa podataka na podskupove na osnovu zadatih atributa [2].

2.2. Neuronske mreže

Neuronske mreže predstavljaju model mašinskog učenja čija struktura se bazira na nervnom sistemu. Osnovna jedinica neuronske mreže je perceptron. Svakom ulaznom signalu perceptrona dodeljuje se određena težina, a zatim se vrši sumiranje svih signala i poređenje dobijene vrednosti sa pragom aktivacione funkcije. U zavisnosti od toga da li je sumirana vrednost veća ili manja od aktivacionog praga forimira se izlazni signal [3].

3. Formulacija problema

U svrhu rešavanja problema klasifikacije ćelija na benigne i maligne biće korišćene stabla odlučivanja i neuronske mreže. Obe tehnike mašinskog učenja kao ulaz koriste dataset koji sadrži odgovarajuće parametre ćelije na osnovu kojih se vrši predikcija da li je ćelija benigna ili maligna.

3.1. Dataset

Sa *Kaggle* sajta [4] preuzeti su podaci na osnovu kojih će biti formirani modeli mašinskog učenja koji vrše adekvatnu predikciju malignih i benignih ćelija. Prva kolona (*id*) predstavlja identifikacioni broj pacijenta, a druga (*diagnosis*) dijagnozu za svakog od pacijenata – benigno ili maligno. Preostale kolone predstavljaju parametre na osnovu kojih se zaključuje da li je promena maligna ili benigna. Parametri su proračunati na osnovu digitalizovane slike uzorka dobijenog iglenom aspiracionom dipsijom i opisuju karakteristike ćelijskih jedara na slici. Neki od parametara koji se koriste u dijagnostici raka dojke su prikazani u tabeli 1.

Tabela 1. Parametri koji se koriste u dijagnostici raka dojke.

Parametar	**Opis**
radius	srednja udaljenost tačaka na obodu od centra ćelijskog jedra
texture	standardna devijacija vrednosti sive skale
perimeter	obim
area	oblast
smoothness	lokalna varijacija u dužinama poluprečnika
compactness	perimeter^2 / area – 1.0
concavity	jačina konkavnih delova konture
concave points	broj konkavnih delova konture
symmetry	simetrija
fractal dimension	fraktalna dimenzija

4. Opis rešenja

4.1. Opis rešenja korišćenjem stabla odlučivanja

4.1.1. Korišćene biblioteke i struktura dataset-a

Kako bi se omogućilo korišćenje funkcija potrebnih za preprocesiranje podataka kao i treniranje modela mašinskog učenja, potrebno je učitati odgovarajuće biblioteke i fajl *breast-cancer.csv*, što je omogućeno kodom prikazanim u nastavku.

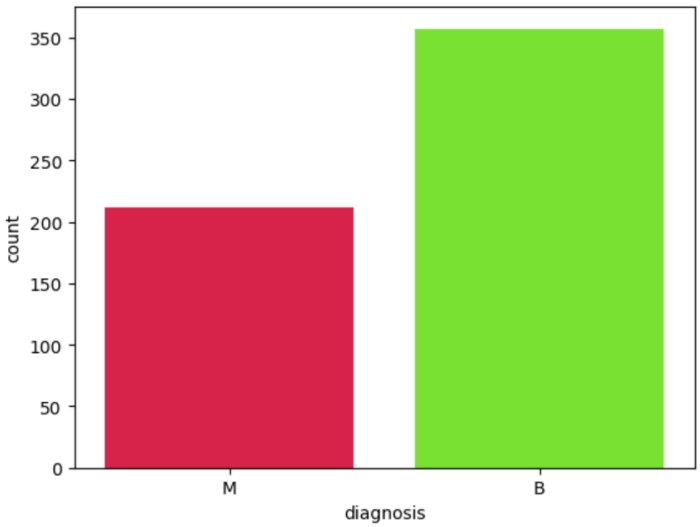
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

BCD = pd.read_csv('breast-cancer.csv')
```

Nakon učitavanja dataset-a, pažnja se može preusmeriti na ciljnu kolonu *diagnosis* i proučiti kako druge kolone utiču na vrednosti u ovoj koloni. Pomoću funkcije *countplot()* iz biblioteke *seaborn*

grafički je prikazan broj vrsta u koloni koje imaju vrednost „M“ i „B“, a funkcija *value.counts()* iskorišćena je kako bi se utvrdio tačan broj vrednosti: „M“ – 212 i „B“ – 357. Vrednost „M“ koristi se za označavanje malignih promena, dok se vrednost „B“ koristi za označavanje benignih promena. Countplot promenljive *diagnosis* prikazan je na slici 22.

```
sns.countplot(data = BCD, palette=sns.color_palette(["#f50537","#72ff13"]),
       x="diagnosis")
plt.show()
print(BCD['diagnosis'].value_counts())
```



Slika 2. Countplot promenljive diagnosis

```
B 357
M 212
Name: diagnosis, dtype: int64
```

4.1.2. Preprocesiranje ulaznog skupa podataka

Pre formiranja modela mašinskog učenja potrebno je razmotriti kako koji parametri utiču na stanje ciljne promenljive. Za početak, pomoću funkcije *corr()* i *sort_values()* na ekranu se ispisuje prvih 5 kolona sa najvećom vrednosti korelacije u odnosu na kolonu *diagnosis*, kao i 5 kolona sa najmanjom vrednosti korelacije.

```
#najveća korelacija
print(abs(BCD.corr()['diagnosis']).sort_values(ascending=False)[0:5])
```

```
diagnosis                 1.000000
concave points_worst      0.793566
perimeter_worst           0.782914
concave points_mean       0.776614

radius_worst              0.776454
Name: diagnosis, dtype: float64
```

```
#najmanja korelacija
print(abs(BCD.corr()['diagnosis']).sort_values(ascending=True)[0:5])
```

```
symmetry_se               0.006522
texture_se                0.008303
fractal_dimension_mean    0.012838
id                        0.039769
smoothness_se             0.067016
```

```
Name: diagnosis, dtype: float64
```

Obzirom da vrednosti određenih parametara slabo variraju između benignih i malignih ćelija, odnosno imaju suviše mali stepen korelacije u odnosu na ciljnu kolonu, iste se izbacuju iz dataset-a. Ovom koraku treba pristupiti oprezno jer ciljna kolona možda nema direktno veliku vrednost korelacije u odnosu na neku kolonu, ali postoji mogućnost da kombinacija nekoliko kolona utiče u velikoj meri na ciljnu, iako pojedinačni rezultati to ne pokazuju. Imajući to u vidu, izbačene su kolone sa korelacijom čija je apsolutna vrednost **manja od 0.3**. Ukoliko krajnji rezultati predikcije nisu zadovoljavajući, potrebno je ozbiljnije razmotriti da li neka od izbačenih kolona pripada nekoj kombinaciji koja utiče na ciljnu kolonu.

Kako bi se izbacile kolone sa suviše malom vrednošću korelacije formira se vektor korelacija koji sadrži vrednosti korelacija svih kolona u odnosu na kolonu *diagnosis*. Poređenjem vrednosti elemenata u ovom vektoru kroz *for* petlju se izbacuju sve kolone sa suviše malom korelacijom u odnosu na kolonu *diagnosis*.

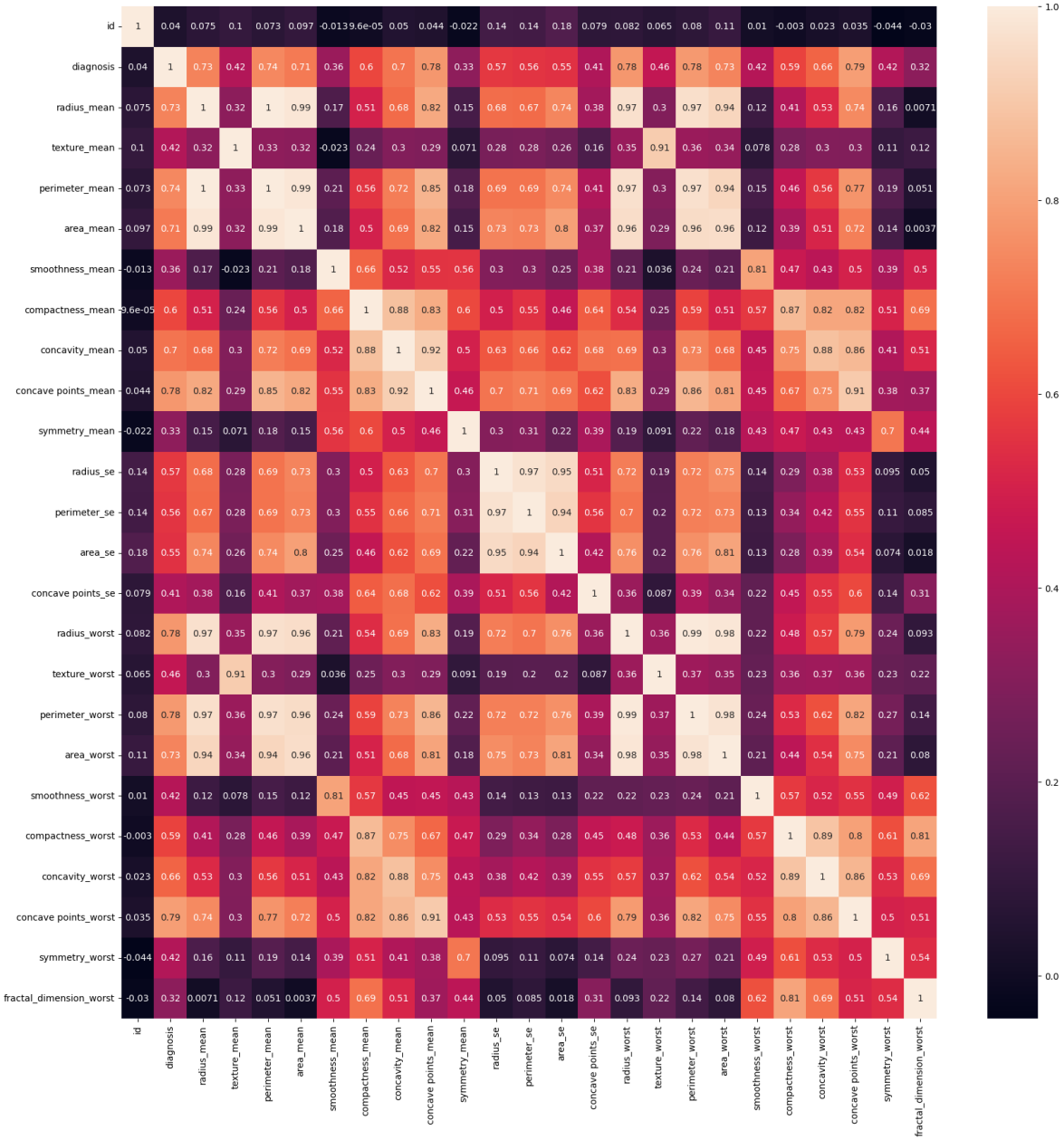
```
korelacija = BCD.corr()['diagnosis']
kolone = BCD.columns
for j in range(1, 31):
    if korelacija[j] <= 0.3 and korelacija[j] >= -0.3:
        BCD = BCD.drop(kolone[j], axis=1)
```

Pomoću funkcije *heatmap()* formirana je matrica korelacije u kojoj su prikazane vrednosti korelacija među preostalim parametrima. Proučavanjem matrice može se uočiti da postoje parovi kolona čija je korelacija jednaka ili bliska jedinici. Kada je korelacija između dve kolone jednaka jedinici podaci u tim kolonama su potpuno identični ili proporcionalni, pa je preporučljivo da se po jedna kolona iz svakog para izbaci kako bi se formirao što bolji model mašinskog učenja.

```
fig, ax = plt.subplots(figsize=(20, 20))
sns.heatmap(BCD.corr(), annot=True)
plt.show()
```

Preostale kolone su:

```
Index(['diagnosis', 'texture_mean', 'smoothness_mean', 'compactness_mean',
'concavity_mean', 'concave points_mean', 'symmetry_mean', 'radius_se',
'perimeter_se', 'area_se', 'concave points_se', 'radius_worst', 'texture_worst',
'perimeter_worst', 'smoothness_worst', 'compactness_worst',
'concavity_worst', 'concave points_worst', 'symmetry_worst', 'fractal_dimension_worst'])
```



Slika 3. Matrica korelacije

4.1.3. Formiranje modela mašinskog učenja

Nakon što je proverena zavisnost između ciljne promenljive i ostalih parametara, kao i međusobna zavisnost parametara, i nakon izbacivanja neodgovarajućih kolona moguće je započeti formiranje modela mašinskog učenja. Za početak, potrebno je formirati trening skup na osnovu koga će biti formiran model i test skup na kome će se proveriti tačnost modela. Trening i test skup formiraju se pomoću naredbe *train_test_split(X, y, test_size=0.3, random_state=41),* gde *test_size=0.3* označava da će test skup biti formiran od 30% podataka. Parametru *random_state* dodeljena je fiksna vrednost kako bi se izbeglo nasumično formiranje skupova.

```
# Formiranje trening i test skupa
X = BCD.drop('diagnosis', axis=1)
y = BCD['diagnosis']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=41)
```

Funkcija *DecisionTreeClassifier()* koristi se za formiranje stabla odlučivanja na osnovu trening skupa. Predikcija promenljive *diagnosis* vrši se na osnovu podataka u test skupu. Predviđene vrednosti *predvidjanje_stablo* upoređuju se sa stvarnim vrednostima *y_test* i formiraju se klasifikacioni izveštaj, matrica konfuzije i računa se preciznost modela. Klasifikacioni izveštaj pokazuje da je preciznost predikcije klase „0“ 100%, dok je preciznost klase „1“ 91%. Na osnovu matrice konfuzije zaključuje se da je broj lažnih jedinica (broj promena za koje je pogrešno pretpostavljeno da su maligne) 6, dok je broj lažnih nula (broj promena za koje je pogrešno pretpostavljeno da su benigne) 0. Bitno je naglasiti da ove greške nisu iste težine. Ukoliko se za neku promenu pogrešno pretpostavi da nije maligna i da lečenje nije potrebno, bolest jako brzo može da se razvije i proširi na zdrave ćelije. Iz tog razloga, broj promena za koje je pogrešno pretpostavljeno da su benigne ima veću težinu i neophodno je da bude što manji. Tačnost ove metode iznosi 96.49122807017544%.

```
#Stablo odlucivanja
modelstablo = DecisionTreeClassifier()
modelstablo.fit(X_train, y_train)
predvidjanje_stablo = modelstablo.predict(X_test)
print(classification_report(y_test, predvidjanje_stablo))
```

```
              precision    recall  f1-score   support

           0       1.00      0.95      0.97       110
           1       0.91      1.00      0.95        61

    accuracy                           0.96       171
   macro avg       0.96      0.97      0.96       171
weighted avg       0.97      0.96      0.97       171
```

```
accuracy_stablo = accuracy_score(y_test, predvidjanje_stablo)
print('Stablo odlucivanja: accuracy_score: ')
print(accuracy_stablo)
```

```
Stablo odlucivanja: accuracy_score:
0.9649122807017544
```

```
print(confusion_matrix(y_test, predvidjanje_stablo))
```

```
[[104   6]
 [  0  61]]
```

```
tn_stablo, fp_stablo, fn_stablo, tp_stablo = confusion_matrix(y_test,
    predvidjanje_stablo).ravel()
print('Broj promena za koje je pogresno pretpostavljeno da su maligne: ')
print(fp_stablo)
print('Broj promena za koje je pogresno pretpostavljeno da su benigne: ')
print(fn_stablo)
```

```
Broj promena za koje je pogresno pretpostavljeno da su maligne:
6
Broj promena za koje je pogresno pretpostavljeno da su benigne:
0
```

Ukoliko formiramo stablo odlučivanja bez prethodnog preprocesiranja podataka (izbacivanje kolona sa niskim stepenom korelacije), dobijena matrica konfuzije je:

```
[[105   5]
 [  1  60]]
Broj promena za koje je pogresno pretpostavljeno da su maligne:
5
Broj promena za koje je pogresno pretpostavljeno da su benigne:
1
```

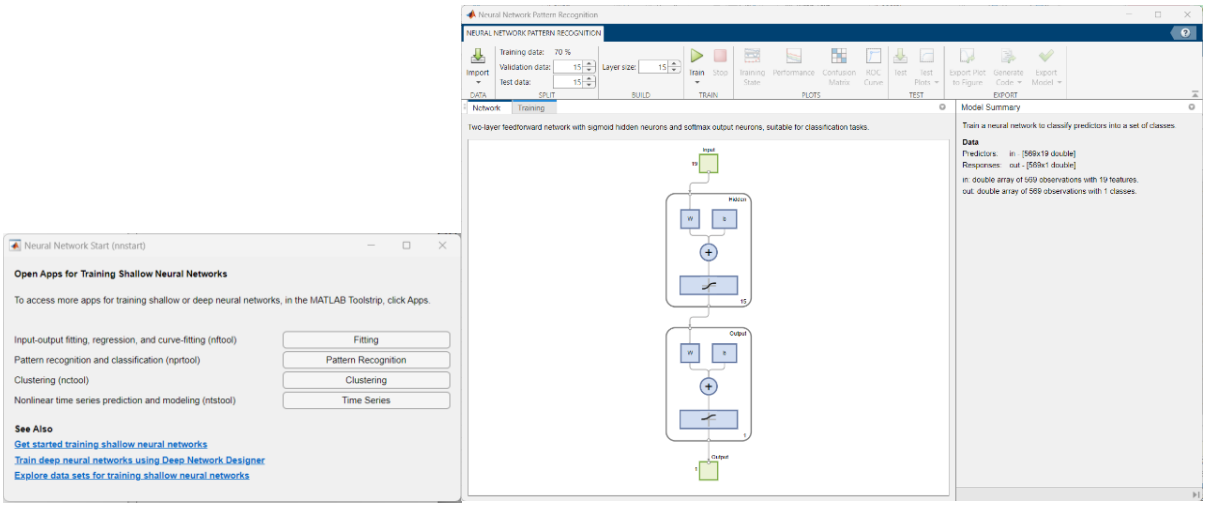
Poređenjem ova dva stabla odlučivanja, može se primetiti da se broj ispravno klasifikovanih ćelija nije promenio, međutim, uočava se da ovaj model klasifikuje jedan uzorak maligne ćelije kao benignu ćeliju, što predstavlja značajno veću klasifikacionu grešku.

Na ovom primeru se ispoljava jedna od mana stabla odlučivanja, a to je da su ona podložna prezasićenju (overfitting-u), ukoliko se na ulazu nađe velika količina podataka.

4.2. Opis rešenja korišćenjem neuronskih mreža

4.2.1. Korišćeno okruženje i struktura dataset-a

Ukupno su razvijena dva modela mašinskog učenja imlementiranih u vidu neuronskih mreža. Modeli su projektovani i obučavani Matlab softverskim paketom (The MathWorks Inc., USA), i njegovim dodatkom Deep Learning Toolbox (razvoj i obučavanje „plitkih“ neuronskih mreža u ranijim verzijama Matlaba bilo je deo sada uklonjenog Neural Network Toolbox), verzijama 2022a. Grafički interfejs za izbor rešavanja problema klasifikacije i projektovanje parametara mreže, unos ulazno/izlaznih podataka i obučavanje i testiranje mreže prikazan je na slici 4.



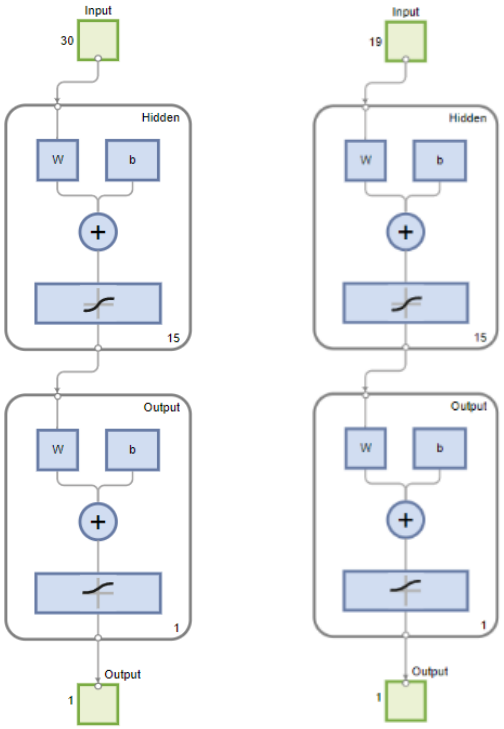
Slika 4. Grafički interfejs za izbor rešavanja problema klasifikacije i projektovanje parametara mreže, unos ulazno/izlaznih podataka i obučavanje i testiranje mreže

Prvi razvijeni neuro model koristi kompletan skup podataka iz dataset-a, dok drugi model koristi redukovani skup podataka dobijen preprocesiranjem opisanim u odeljku 4.1.2.

Skup podataka je nasumično podeljen u skupove za treniranje, validaciju i testiranje. Skup za testiranje (70%, 399 pacijenata) je prezentiran neuro-fazi mreži tokom treniranja, i mreža je algoritmom obučavanja podešena prema grešci procene. Validacioni skup (15%, 85 pacijenata) je korišćen da se u toku obučavanja procenjuje generalizacija mreže i da se prekine obučavanje kada generalizacija prestane da se poboljšava. Konačno, skup za testiranje (15%, 85 pacijenata) nije imao

uticaja na obučavanje i korišćen je da se obezbedi nezavisna mera performansi modela nakon obučavanja.

4.2.2. Struktura mreže



Slika 5a. (levo) Struktura neuronske mreže čiji je ulaz kompletan dataset
Slika 5b. (desno) Struktura neuronske mreže čiji je ulaz redukovani dataset

Na slici 5a, neuronska mreža je višeslojni perceptron koji se sastoji od jednog ulaznog sloja sa 30 neurona koji prihvata 30 ulaznih promenljivih (promenljive označavaju prethodno opisane parametre ćelije), 1 skrivenog sloja sa 15 neurona i jednog izlaznog sloja sa 1 neuronom. Izlaz modela je dihotomna promenljiva koja se može tumačiti kao predviđanje da li je ćelija benigna ili maligna (0 – benigna, 1 – maligna). Ukupna tačnost konačnog modela određena je poređenjem predviđanih vrednosti sa stvarnim ishodima. Struktura neuronske mreže na slici 5b je u potpunost analogna strukturi mreže na slici 5a, uz jedino razliku da je redukovan broj ulaznih promenljivih (redukcija je obavljena u skladu sa postupkom opisanim u odeljku 4.1.2.), te se ulaz sastoji od ukupno 19 neurona.

Obe implementirane mreže su jednosmerne (*feedforward*) neuronske mreže, koje su pogodne za zadatke klasifikacije. Ove mreže su karakteristične po sledećem:

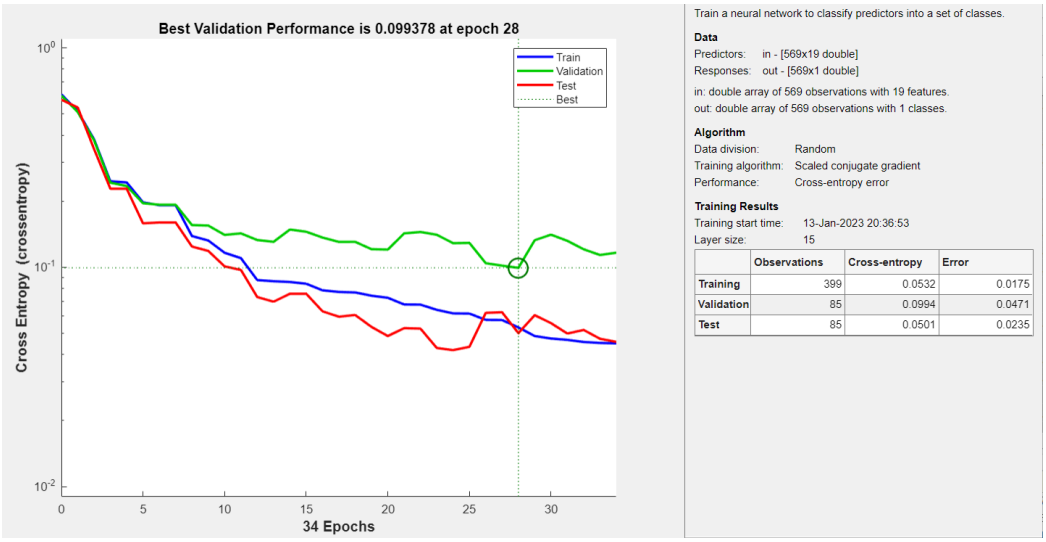
- Neuroni su organizovani u slojeve, tako da prvi sloj prihvata ulaze i poslednji sloj formira izlaz.
- Svaki neuron u jednom sloju je povezan sa svakim neuronom u narednom sloju.
- Ne postoje veze između neurona u istom sloju.

Mreža u skrivenom sloju kod oba modela koristi *Sigmoid* neurone [5], koji se razlikuju od običnih perceptrona po svoj aktivacionoj funkciji. Izlaz Sigmoid neurona je:

$$\frac{1}{1 + exp(-\sum_j w_j * x_j - b)},$$

gde je w_j – težina j-tog ulaza, x_j – j-ti ulaz, b – unutrašnja pobuda neurona (*bias*). Ova aktivaciona funkcija pogodna je za probleme klasifikacije zato što se njome ograničava izlaz na opseg [0, 1].

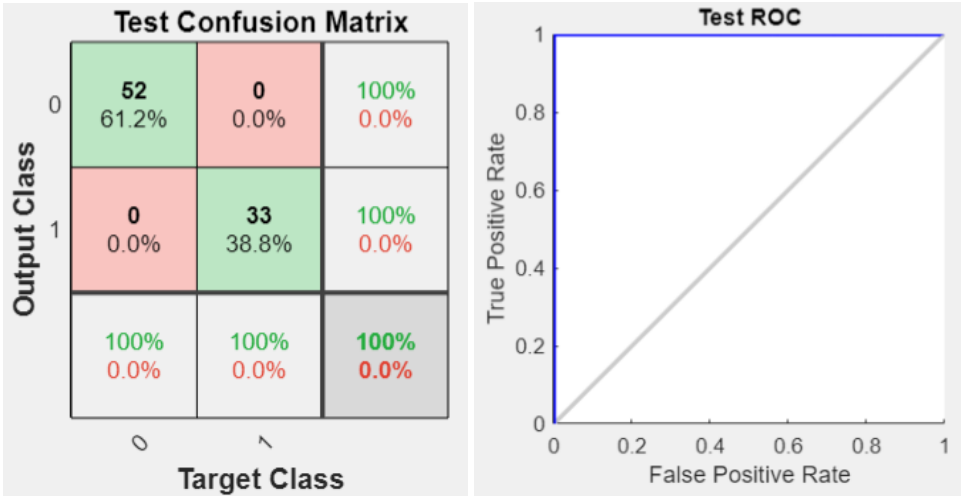
U skladu sa preporukama, od raspoloživih algoritama obučavanja neuro mreža primenjenog tipa za probleme klasifikacije, izabran je Scaled Conjugate Gradient metod zato što je on kod relativno velikih skupova podataka računski efikasniji u odnosu na alternativne algoritme. Postupak procesa obučavanja u softveru Matlab prikazan je na slici 6.



Slika 6. Rezultati procesa obučavanja neuro mreže

4.2.3. Rezultati dobijeni primenom neuronskih mreža

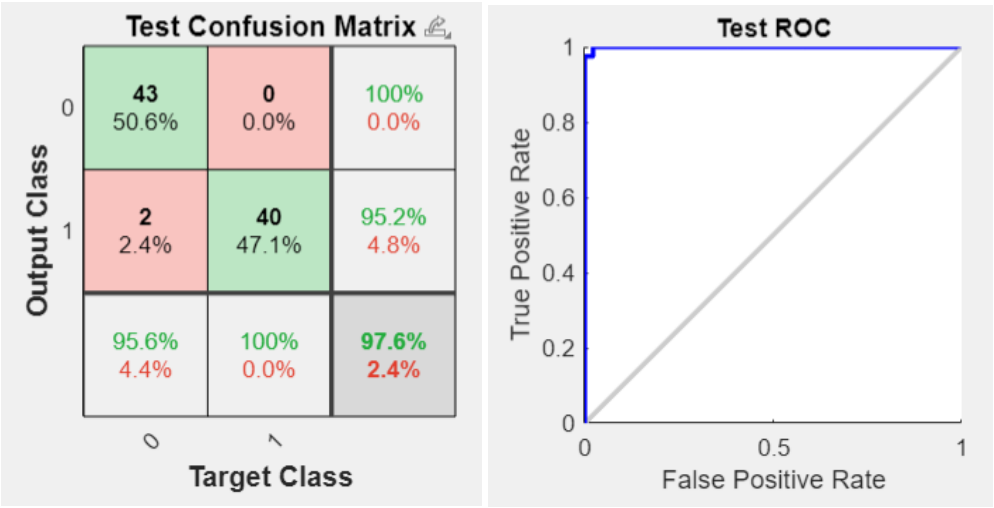
Matrica konfuzije i ROC kriva za neuronsku mrežu čiji je ulaz kompletan dataset su prikazani na slici 7. Model je na test set-u ispravno klasifikovao 100% ćelija.



Slika 7. Matrica konfuzije i ROC kriva za neuronsku mrežu čiji je ulaz kompletan dataset

Matrica konfuzije i ROC kriva za neuronsku mrežu čiji je ulaz redukovani dataset su prikazani na slici 8. Model je na test set-u ispravno klasifikovao 97.6% ćelija. Na osnovu matrice konfuzije zaključuje se da je broj lažnih jedinica (broj promena za koje je pogrešno pretpostavljeno da su maligne) 2, dok je broj lažnih nula (broj promena za koje je pogrešno pretpostavljeno da su benigne) 0. Kao što je prethodno naglašeno, ove greške nisu iste težine. Ukoliko se za neku promenu pogrešno

pretpostavi da nije maligna i da lečenje nije potrebno, bolest jako brzo može da se razvije i proširi na zdrave ćelije. Iz tog razloga, broj promena za koje je pogrešno pretpostavljeno da su benigne ima veću težinu i neophodno je da bude što manji.



Slika 8. Matrica konfuzije i ROC kriva za neuronsku mrežu čiji je ulaz redukovan dataset

Oba modela su generalno pokazala dobre performanse u predikciji malignih ćelija, sa jasnom superiornošću neuronske mreže čiji je ulaz kompletan dataset.

5. Zaključak i diskusija

Modeli zasnovani na obe tehnike mašinskog učenja (stablo odlučivanja i neuronske mreže) se dali izuzetne rezultate sa ispravnom klasifikacijom u preko 95% slučajeva. Ipak, prednost se može dati neuronskim mrežama iz sledećih razloga:

- Neuronske mreže imaju sposobnost da nauče i modeluju kompleksne veze između ulaza i izlaza, za razliku od stabla odlučivanja, koja su ograničena skupom pravila koja se ručno zadaju (tj. generišu).
- Neuronske mreže su u stanju da rade sa velikim količinama podataka efikasnije u odnosu na stabla odlučivanja, iz razloga što stabla odlučivanja mogu postati veoma velika i kompleksna, što ih čini podložnijim prezasićenju (overfitting).
- Neuronske mreže daju preciznije rezultate i bolje performanse u oblasti klasifikacije.

Treba na kraju pomenuti da u poslednje vreme puno pažnje privlači i primena dubokog učenja u dijagnostifikovanju raka dojke, što predstavlja alternativu metodama primenjenim u ovom radu. Jedan od najnovijih radova u kom je prikazan pregled literature u kojoj su primenjena takva rešenja, publikovan u januaru 2023. godine, je rad [6]. Posebno se ističe primena Konvolutivnih neuronskih mreža (CNN), koje se koriste za analizu slika tkiva dojke i identifikaciju bilo kakvih abnormalnosti koje mogu ukazivati na rak.

Konačno, možemo zaključiti nekim argumetima za primenu metoda veštačke inteligencije u dijagnostifikovanju i predikciji tumora dojke. Tradicionalno, za otkrivanje i dijagnozu raka dojke, radiolozi posmatraju slike tkiva dojke ručno (golim očima) i kroz konsultovanje drugih medicinskih stručnjaka donose svoju odluku. Ručna inspekcija slika tkiva dojke radi mogućeg otkrivanja raka dojke je široko korišćena metoda, međutim, određene neizbežne činjenice vezane za ručnu inspekciju slika mogu dovesti do netačnog otkrivanja i produžiti proces dijagnoze. Na primer:

1. Nedostupnost stručnjaka u udaljenim područjima (nerazvijene zemlje).
2. Nedostupnost stručnjaka sa dovoljnim znanjem iz domena za preciznu analizu višeklasnih slika (slike sa mogućim višestrukim karakteristikama bolesti).
3. Provera velikog broja medicinskih slika na dnevnoj bazi može biti iscrpna i glomazna praksa.
4. Suptilna priroda tumora dojke i složena struktura tkiva dojke otežavaju ručnu analizu.
5. Nivo koncentracije medicinskih stručnjaka i drugi zamor otežavaju dijagnozu i proces koji oduzima vreme.

Rezultati koje smo dobili u ovom radu potvrđuju zaključak iz brojne literature da se veštačka inteligencija može efikasno primeni kao pomoć pri dijagnozi i predikciji tumora dojke.

6. Literatura

[1] “10 Machine Learning Methods that Every Data Scientist Should Know” https://towardsdatascience.com/10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9

[2] “Decision trees”, dostupno na: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

[3] M. Nikolić, A. Zečević, "Mašinsko učenje", Matematički fakultet Univerziteta u Beogradu, Beograd 2019, dostupno na http://ml.matf.bg.ac.rs/readings/ml.pdf.

[4] “Kaggle: Your Machine Learning and Data Science Community”, dostupno na: https://www.kaggle.com/

[5] C. Bishop "Neural networks for pattern recognition", 1995.

[6] M. Nasser, U. K. Yusof, “Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction”, Diagnostics (Basel), 2023 Jan 3;13(1):161. doi: 10.3390/diagnostics13010161.