Primjena algoritama strojnog učenja u predviđanju srčanih bolesti

Forjan, Dario

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Dario Forjan

PRIMJENA ALGORITAMA STROJNOG UČENJA U PREDVIĐANJU SRČANIH BOLESTI

DIPLOMSKI RAD

Varaždin, 2023.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Dario Forjan

Matični broj: 0016133297

Studij: Organizacija poslovnih sustava

PRIMJENA ALGORITAMA STROJNOG UČENJA U PREDVIĐANJU SRČANIH BOLESTI

DIPLOMSKI RAD

Mentorica:

Izv. prof. dr. sc. Dijana Oreški

Varaždin, srpanj 2023.

Dario Forjan

Izjava o izvornosti

Izjavljujem da je moj diplomski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor potvrdio prihvaćanjem odredbi u sustavu FOI-radovi

_____

Sažetak

Tema ovog rada je primjena algoritama strojnog učenja u predviđanju srčanih bolesti. Za potrebe istraživanja, korišten je skup podataka s 303 zapisa, podijeljenih u dva skupa. Podaci uključuju informacije o pacijentima kao što su dob, spol, različiti biomedicinski pokazatelji (npr. krvni tlak, razina kolesterola) i prisutnost srčanih bolesti. Cilj istraživanja je primijeniti algoritme strojnog učenja na ovim podacima kako bi se izgradio model koji može predvidjeti prisutnost srčanih bolesti na temelju ostalih karakteristika pacijenta.

U teorijskom dijelu rada će biti sažeto prikazana prethodna istraživanja i rezultati u ovom području. Također, bit će detaljnije objašnjene metode koje će biti korištene u radu, uključujući stablo odlučivanja, neuronske mreže i Bayesove mreže. Nakon toga, u radu će biti izrađeni navedeni modeli nad skupom podataka, a dobiveni rezultati metoda će biti uspoređeni.

Cilj istraživanja je pružiti uvid u mogućnosti algoritama strojnog učenja u predviđanju srčanih bolesti te identificirati najučinkovitiju metodu za tu svrhu. Rezultati ovog istraživanja mogu biti korisni za daljnji razvoj sustava za dijagnosticiranje i predviđanje srčanih bolesti, pružajući podršku medicinskim stručnjacima u donošenju informiranih odluka o liječenju i prevenciji ovih bolesti.

Ključne riječi: Stablo odlučivanja; Neuronske mreže; Bayesove mreže; Strojno učenje; Srčane bolesti; Kardiovaskularne bolesti;

Sadržaj

1.Uvod 1
2.Prethodna istraživanja 2
2.1. Sustav predviđanja srčanih bolesti korištenjem rudarenja podataka . 2
2.2. Predviđanje bolesti srca kombinacijom strojnog i dubokog učenja 4
2.3. Predviđanje rizika od bolesti srca klasifikatorima strojnog učenja 7
2.4. Predviđanje bolesti srca 10
2.5. Predviđanje bolesti srca i analiza osjetljivosti klasifikatora 11
2.6. Usporedba rezultata prethodnih istraživanja 13
2.7.Zaključak na temelju prethodnih istraživanja 17
3.Opis podataka 18
4.Priprema podataka 22
4.1. Odabir, čišćenje, konstrukcija, integracija i oblikovanje podataka 22
5.Opis korištenih metoda 23
5.1. Stablo odlučivanja 23
5.2. Neuronske mreže 24
5.3. Bayseove mreže 27
6.Razvoj prediktivnih modela 29
6.1. Dizajn za testiranje 30
6.2. Stablo odlučivanja 31
6.3. Neuronske mreže 34
6.4. Bayseove mreže 36
7.Evaluacija 40
7.1. Dobiveni rezultati evaluacije podataka 41
7.2. Analiza osjetljivosti Bayesove mreže 43
7.3 Usporedba dobivenih rezultata i analiza osjetljivosti 43
8.Korištenje 45
9.Zaključak 47
10.Literatura 49
11.Popis slika 51
12.Popis tablica 52

1. Uvod

U današnjem svijetu, srčane bolesti predstavljaju značajan javnozdravstveni izazov. Prema Svjetskoj zdravstvenoj organizaciji, srčane bolesti su vodeći uzrok smrti diljem svijeta. Precizno predviđanje tih bolesti ima ključnu ulogu u ranom otkrivanju, prevenciji i pružanju pravovremenog liječenja pacijentima. U tom kontekstu, primjena algoritama strojnog učenja otvara mogućnosti za razvoj efikasnih prediktivnih modela koji mogu pomoći u identifikaciji rizičnih skupina i pravovremenom intervencijom.

Cilj ovog diplomskog rada je istražiti primjenu algoritama strojnog učenja u predviđanju srčanih bolesti. Za tu svrhu, koristit će se skup podataka koji sadrži relevantne informacije o pacijentima kao što su dob i spol, šećer, kolesterol i slični.

U ovome radu razmotriti će se relevantna istraživanja koja su se bavile sličnim temama i identificirati ključne metode i tehnike korištene u tim istraživanjima. Nakon toga, biti će opisan skup podataka koji će biti korišten u istraživanju, uključujući njegovu strukturu i karakteristike. Nadalje, biti će opisane metode strojnog učenja koje će se koristiti u ovom istraživanju: stablo odlučivanja, neuronske mreže i Bayesove mreže. Zatim proces izrade svake od metoda, te na kraju rada rezultati i biti će uspoređena učinkovitost različitih algoritama strojnog učenja u predviđanju srčanih bolesti i zaključak ovog rada.

2. Prethodna istraživanja

U ovom poglavlju biti će detaljno opisano pet sličnih istraživanja različitih autora vezanih uz ovu temu. Istraživanja imaju slične skupove podataka i slične metode predviđanja srčanih bolesti.

2.1. Sustav predviđanja srčanih bolesti korištenjem rudarenja podataka

Sažetak ovog istraživanja napisan je na temelju članka *Effective heart disease prediction system using data mining techniques* iz 2021. godine, čiji su autori Singh P., Singh S. i Pandi-Jain G.

Istraživanje se fokusira na razvoj sustava za predviđanje bolesti srca primjenom tehnika rudarenja podataka. Dijagnostika bolesti srca predstavlja izazovan zadatak, a cilj ovog istraživanja je pružiti automatizirano predviđanje stanja srca pacijenta kako bi se daljnji tretmani mogli učiniti učinkovitijima. Uobičajena dijagnostika bolesti srca temelji se na znakovima, simptomima i fizičkom pregledu pacijenta. Međutim, postoji nekoliko faktora koji povećavaju rizik od srčanih bolesti, kao što su pušenje, razina kolesterola u tijelu, obiteljska povijest srčanih bolesti, pretilost, visoki krvni tlak i nedostatak tjelesne aktivnosti. [1]

Jedan od glavnih izazova s kojima se suočavaju zdravstvene organizacije, poput bolnica i medicinskih centara, jest pružanje kvalitetnih usluga po prihvatljivim troškovima. Kvalitetna usluga uključuje pravilnu dijagnozu pacijenata i primjenu učinkovitih tretmana. U sklopu istraživanja korišten je skup podataka o bolestima srca koji sadrži brojne numeričke i kategoričke podatke. Prije daljnje obrade, provodi se čišćenje i filtriranje zapisa kako bi se uklonili irelevantni podaci iz baze. [1]

Predloženi sustav ima sposobnost otkrivanja skrivenog znanja, poput uzoraka i veza povezanih s bolestima srca, iz povijesne baze podataka o bolestima srca. Također sposoban je odgovoriti na složene upite za dijagnosticiranje bolesti srca, što može biti od pomoći zdravstvenim praktičarima u donošenju inteligentnih kliničkih odluka. Rezultati istraživanja pokazuju da predloženi sustav ima jedinstvenu sposobnost ostvarivanja ciljeva definiranih rudarenjem podataka. [1]

U ovom istraživanju koristio se javno dostupan skup podataka o bolestima srca s ukupno 303 zapisa, podijeljenih na skup za treniranje (40%) i skup za testiranje (60%). Za provedbu eksperimenta korišten je alat za rudarenje podataka nazvan Weka 3.6.11. Kao algoritam za treniranje korištena je višeslojna perceptronska neuronska mreža (MLPNN) s algoritmom propagacije unatrag (BP). [1] U tablici 1 možemo vidjeti atribute, njihove opise i vrijednosti korištene u istraživanju.

Atribut	Opis atributa	Vrijednosti
age	Dob u godinama	Kontinuirano
sex	Muško ili žensko	1 = muško, 0 = žensko
thestbps	Mirovni krvni tlak	Kontinuirana vrijednost u mmHg
cp	Vrsta bolova u prsištu	1 = tipični tip 1, 2 = tipična angina, 3 = neanginalna bol, 4 = asimptomatski
chol	Serumski kolesterol	Kontinuirana vrijednost u mm/dL
fbs	Krvni šećer natašte	1 ≥ 120 mg/dL, 0 ≤ 120 mg/dL
restecg	Rezultati elektrokardiograma u mirovanju	0 = normalno, 1 = prisutna ST-T valna abnormalnost, 2 = hipertrofija lijeve klijetke
thalach	Maksimalna postignuta brzina otkucaja srca	Kontinuirana vrijednost
old peak	ST depresija izazvana tjelesnom aktivnošću u odnosu na mirovanje	Kontinuirana vrijednost
exang	Tjelesna aktivnost izazvana angina pektorisom	0 = ne, 1 = da
ca	Broj glavnih krvnih žila obojenih fluoroskopijom	Vrijednost 0-3
slope	Nagib vrha vježbe ST segmenta	1 = ravan, 2 = ravno, 3 = nagnuto
thal	Tip defekta	3 = normalan, 6 = fiksni, 7 = reverzibilni defekt
obes	Pretilost	1 = da, 0 = ne
num	Dijagnoza bolesti srca	0% ≤ 50%, 1% > 50%

Tablica 1: Popis korištenih atributa [1]

MLPNN je jedan od najznačajnijih modela umjetnih neuronskih mreža. Sastoji se od ulaznog sloja, jednog ili više skrivenih slojeva i izlaznog sloja. Ulazni čvorovi prenose vrijednosti prvom skrivenom sloju, a zatim čvorovi prvog skrivenog sloja prenose vrijednosti drugom sloju i tako dalje sve do generiranja izlaza. [1]

BP algoritam služi kao korisna metodologija za treniranje višeslojnih perceptrona u različitim primjenama. Algoritam BP računa razliku između stvarnih i predviđenih vrijednosti, koja se širi od izlaznih čvorova prema čvorovima u prethodnom sloju. Algoritam BP učenja može se podijeliti na dvije faze: propagaciju i ažuriranje težina. Prvo, ovaj algoritam učenja pruža podatke za treniranje neuronskoj mreži i uspoređuje stvarne i željene izlaze. Zatim se računa pogreška u svakom neuronu. Na temelju toga, algoritam računa kakav bi trebao biti izlaz za svaki neuron i koliko treba povećati ili smanjiti izlaz kako bi se postigao željeni rezultat, te naposljetku prilagođava težine. Cjelokupni proces je usmjeren na poboljšanje težina tijekom obrade. [1]

Rezultati istraživanja prikazani su pomoću „matrice konfuzije“ (eng. confusion matrix). Matrica konfuzije pruža informacije o stvarnim i predviđenim klasifikacijama koje je napravio sustav klasifikacije. Iznosi u matrici ocjenjuju performanse takvog sustava. Matrica konfuzije sadrži četiri ulaza: TP (pravi pozitivni), FP (lažni pozitivni), FN (lažni negativni) i TN (pravi negativni). [1]

Zaključno, ovo istraživanje predstavlja učinkovit sustav za predviđanje bolesti srca korištenjem tehnika rudarenja podataka. Sustav se temelji na upotrebi neuronskih mreža i postiže visoku točnost u predviđanju rezultata. Ovaj sustav može biti koristan za stručnjake iz područja zdravstva u donošenju kliničkih odluka i planiranju učinkovitijih dijagnostičkih postupaka za pacijente. Rezultati istraživanja potvrđuju da se bolest srca može predvidjeti s gotovo 100% točnosti primjenom neuronskih mreža. [1]

2.2. Predviđanje bolesti srca kombinacijom strojnog i dubokog učenja

Sažetak ovog istraživanja temelji se na članku "*Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning*" objavljenom u časopisu "*Computational Intelligence and Neuroscience*" 2021. godine. Autori istraživanja su Bharti R., Khamparia A., Shabaz M., Dhiman G., Pande S. i Singh P.

U okviru ovog istraživanja, temeljita analiza različitih metoda strojnog učenja i dubokog učenja provedena je s ciljem predviđanja srčanih bolesti. Kako su kardiovaskularne bolesti vodeći uzrok smrti diljem svijeta, s godišnjom stopom

smrtnosti od 17,9 milijuna prema izvješćima Svjetske zdravstvene organizacije, važno je pronaći učinkovite načine za dijagnosticiranje i predviđanje tih bolesti. Nezdrave navike poput visokog kolesterola, pretilosti, povišenih triglicerida i hipertenzije povećavaju rizik od srčanih bolesti. Uz to, određeni simptomi kao što su poremećaji spavanja, nepravilan rad srca, oticanje nogu i brz porast tjelesne težine mogu ukazivati na prisutnost srčanih bolesti. Stoga je ključno razviti točne metode za dijagnozu i predviđanje ovih bolesti kako bi se pravovremeno interveniralo i spriječile moguće fatalne posljedice. [2]

Skup podataka koji je korišten u ovom istraživanju naziva se Javni zdravstveni skup podataka i datira iz 1988. godine. Sastoji se od četiri baze podataka: Cleveland, Mađarska, Švicarska i Long Beach V. Ukupno, skup podataka sadrži 76 atributa, od kojih se u objavljenim eksperimentima koristi podskup od 14 atributa. Ciljni atribut naziva se "target" i odnosi se na prisutnost srčane bolesti kod pacijenta. Ovaj atribut ima dvije vrijednosti: 0 za "bez bolesti" i 1 za "bolest". [2] U tablici 2 možemo vidjeti popis atributa zajedno s njihovim opisom.

Atribut	Opis atributa
age	Dob pacijenta u godinama
sex	1 = muško, 0 = žensko
Cp	Vrsta boli u prsištu
Trestbps	Mirovni krvni tlak (u mm Hg prilikom prijema u bolnicu). Normalni raspon je 120/80
Chol	Serumska razina kolesterola koja prikazuje količinu triglicerida. Trebala bi biti manja od 170 mg/dL
Fbs	Krvni šećer natašte. Veći od 120 mg/dL označava 1, manji od 100 mg/dL je normalan, a između 100 i 125 mg/dL se smatra preddijabetesom
Restecg	Rezultati elektrokardiograma u mirovanju
Thalach	Maksimalna postignuta brzina otkucaja srca. Maksimalna brzina srčanog otkucaja je 220 minus dob pacijenta
Exang	Tjelesna aktivnost izaziva anginu pektoris (1 da). Angina je vrsta boli u prsištu uzrokovana smanjenim protokom krvi do srca

Oldpeak	ST depresija izazvana tjelesnom aktivnošću u odnosu na mirovanje
Slope	Nagib vrha vježbe ST segmenta
Ca	Broj glavnih krvnih žila (0-3) obojenih fluoroskopijom
Thal	Bez objašnjenja, ali vjerojatno thalassemia (3 = normalno, 6 = fiksni defekti, 7 = reverzibilni defekti)
Target (T)	Nema bolesti = 0, bolest = 1 (stanje bolesti prema angiografiji)

Tablica 2: Popis atributa [2]

S obzirom na dostupnost velike količine istraživačkih podataka i pacijentskih zapisa, istraživači su prepoznali mogućnost korištenja različitih tehnologija strojnog učenja i umjetne inteligencije za pravilnu dijagnozu srčanih bolesti i sprječavanje fatalnih ishoda. Primjena strojnog učenja i dubokog učenja omogućuje analizu potpunih genomičkih podataka, pandemijskih prognoza i dublju analizu medicinskih zapisa radi poboljšane predikcije i dijagnostike srčanih bolesti. [2]

Međutim, jedan od izazova s kojima se suočavaju istraživači u ovom području je visoka dimenzionalnost podataka. Skupovi podataka koji se koriste za analizu srčanih bolesti često su veliki i složeni, što može rezultirati prevelikom potrošnjom memorije i mogućnošću prenaučenosti modela. U cilju rješavanja tog problema, istraživači su primijenili tehnike smanjenja dimenzionalnosti podataka, kao što su inženjering značajki i odabir značajki, kako bi smanjili redundanciju podataka i poboljšali učinkovitost modela. [2]

U literaturi su provedene studije koje su pokazale da primjena tehnika inženjeringa značajki i odabira značajki poboljšava rezultate klasifikacije i predviđanja srčanih bolesti. Primjeri takvih tehnika uključuju linearnu i nelinearnu nesuperviziranu redukciju dimenzionalnosti, kao i primjenu algoritama kao što su Princip Komponenti Analize (PCA) i Metoda Nezavisnih Komponenti (ICA). [2]

Ovaj rad također ističe da srčane bolesti nisu stanje koje treba uzimati olako, posebno jer su muškarci skloniji obolijevanju od srčanih bolesti nego žene. Istraživanja su pokazala da su muškarci dvostruko više izloženi riziku od srčanog udara tijekom života, čak i kada se uzmu u obzir tradicionalni faktori rizika kao što su visoki kolesterol,

visoki krvni tlak, dijabetes, indeks tjelesne mase i tjelesna aktivnost. Stoga je važno razviti sustav koji može pravovremeno identificirati rizik od srčanih bolesti i pružiti odgovarajuće liječenje i skrb. [2]

U zaključku istraživanja ističe se da je kombinacija strojnog učenja i dubokog učenja (*eng. Deep learning*) pokazala obećavajuće rezultate u predviđanju srčanih bolesti. Predložene metode su analizirane na skupu podataka s kliničkim parametrima pacijenata, a rezultati su evaluirani kroz matricu konfuzije, točnost, preciznost, osjetljivost i F1 mjere. Duboko učenje je postiglo najbolje rezultate, s točnošću od 94,2%. Istraživači sugeriraju daljnje proširenje skupova podataka, korištenje drugih optimizacijskih tehnika te integraciju razvijenih modela s multimedijskim sustavima kako bi se olakšala dijagnoza i praćenje pacijenata s srčanim bolestima. [2]

2.3. Predviđanje rizika od bolesti srca klasifikatorima strojnog učenja

Sažetak ovog istraživanja temelji se na članku "*Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators*" objavljenom u časopisu "*Applied Sciences*" 2021. godine. Autori ovog istraživanja su Reddy K., Elamvazuthi I., Aziz A., Paramasivam S., Chua H. i Pranavanand S.

Ova istraživačka studija bavila se razvojem modela za predviđanje rizika od srčanih bolesti koristeći tehnike strojnog učenja i evaluaciju atributa. Primijenjeni su različite klasifikatore, uključujući Naive Bayes (NB), logističku regresiju (LR), sekvencijalnu minimalnu optimizaciju (SMO), instance-based classifier (IBk), AdaBoostM1, bagging, JRip i random forest (RF). Klasifikatori su trenirani koristeći cjelokupni skup atributa i optimizirane skupove atributa dobivene kroz evaluaciju atributa. [5]

Skup podataka korišten za eksperimentiranje bio je Cleveland dataset o srčanim bolestima, koji se sastoji od 303 instanci i 76 atributa i koji je korišten u drugom istraživanju. Međutim, samo 14 atributa smatrano je prikladnima za istraživačke svrhe. Skup podataka prošao je proces obrade podataka, uključujući rješavanje nedostajućih

vrijednosti i pretvaranje ciljnog atributa u binarni format (0 i 1) za predviđanje srčanih bolesti. [5]

Atributi s manje od 10 klasa smatraju se nominalnim ili kategoričkim tipovima. Atribut „sex“ ima dvije klase temeljene na spolu: 1 = muško i 0 = žensko. Atribut „cp“ ima četiri klase koje opisuju vrste boli u prsištu: 1 = tipična angina, 2 = atipična angina, 3 = neanginalna bol i 4 = asimptomatski. Atribut „fbs“ ima dvije klase koje odražavaju prisutnost ili odsutnost krvnog šećera natašte većeg od 120 mg/dL: 1 = istina i 0 = laž. Atribut „restecg“ sastoji se od tri klase koje opisuju rezultate elektrokardiografije u mirovanju: 0 = normalno, 1 = abnormalnost u ST-T valu i 2 = definitivna hipertrofija lijeve klijetke. Atribut „exang“ ima dvije klase temeljene na prisutnosti ili odsutnosti angine pektoris izazvane tjelesnom aktivnošću: 1 = da i 0 = ne. Atribut „slope“ ima tri klase koje opisuju nagib vrha vježbe ST segmenta: 1 = uzlazni, 2 = ravan i 3 = silazni. Atribut „ca“ ima četiri klase koje odražavaju broj glavnih krvnih žila (0-3) obojenih fluoroskopijom. Atribut „thal“ sastoji se od tri klase koje opisuju stanje srca: 3 = normalno, 6 = fiksni defekt i 7 = reverzibilni defekt. Atribut „target“ ima pet klasa koje predviđaju prisutnost rizika od bolesti srca: 0 = nema rizika od bolesti srca i 1 do 4 = rizik od bolesti srca u različitim fazama. Radi cilja ovog istraživanja, vrijednosti atributa „target“ u rasponu od 1 do 4 su konvertirane u 1, što rezultira samo dvije klase: 0 i 1. Atributi „age“, „trestbps“, „chol“, „thalach“ i „oldpeak“ se smatraju numeričkim/cjelobrojnim atributima. [5] Cjelokupan popis atributa možemo vidjeti u tablici 3.

Atribut	Opis atributa	Vrsta atributa	Raspon vrijednosti atributa
age	Dob u godinama	Numerički	29 do 77
sex	Spol	Nominalni	0 = žensko, 1 = muško
cp	Vrsta boli u prsištu	Nominalni	1 = tipična angina, 2 = atipična angina, 3 = neanginalna bol, 4 = asimptomatski
trestbps	Mirovni krvni tlak u mm Hg prilikom prijema u bolnicu	Numerički	94 do 200
chol	Serumski kolesterol u mg/dL	Numerički	126 do 564

fbs	Krvni šećer natašte > 120 mg/dL	Nominalni	0 = ne, 1 = da
restecg	Rezultati elektrokardiografije u mirovanju	Nominalni	0 = normalno, 1 = ST-T valna abnormalnost, 2 = hipertrofija lijeve klijetke prema Estesovim kriterijima
thalach	Maksimalna postignuta brzina otkucaja srca	Numerički	71 do 202
exang	Angina izazvana tjelesnom aktivnošću	Nominalni	0 = ne, 1 = da
oldpeak	ST depresija izazvana tjelesnom aktivnošću u odnosu na mirovanje	Numerički	0 do 6.2
slope	Nagib vrha vježbe ST segmenta	Nominalni	1 = uzlazni, 2 = ravan, 3 = nizlazni
ca	Broj glavnih krvnih žila obojenih fluoroskopijom	Nominalni	0-3
thal	Stanje srca	Nominalni	3 = normalno, 6 = fiksni defekt, 7 = reverzibilni defekt
target	Atribut predviđanja	Nominalni	0 = nema rizika od bolesti srca, 1 do 4 = rizik od bolesti srca

Tablica 3: Popis atributa. [5]

Rezultati su pokazali da je klasifikator sekvencijalne minimalne optimizacije (SMO) postigao najveću točnost od 85.148% koristeći cjelokupni skup atributa. Također, NB klasifikator postigao je točnost od 84.158% koristeći optimizirani skup atributa dobiven kroz evaluaciju atributa zasnovanu na korelaciji. Pri evaluaciji različitih tehnika evaluacije atributa, SMO klasifikator nadmašio je ostale, pokazujući najbolje rezultate u pogledu MAE (srednje apsolutne greške), osjetljivosti, greške uključivanja, preciznosti, F-mjere i specifičnosti. [5]

Osim toga, logička regresija (LR) ostvarila je visoku vrijednost površine ispod ROC krivulje (ROC area), što ukazuje na snažnu prediktivnu sposobnost. Tehnike evaluacije atributa poput evaluacije zasnovane na korelaciji, evaluacije atributa kroz χ^2-test i evaluacije atributa ReliefF korištene su za odabir najutjecajnijih atributa za predviđanje rizika od srčanih bolesti. [5]

U zaključku, ovo istraživanje koristilo je Cleveland dataset o srčanim bolestima kako bi postiglo najvišu točnost od 85,148% s SMO modelom koji se temelji na potpunom skupu atributa, te točnost od 84,158% s NB modelom koji se temelji na optimalnom skupu od sedam atributa dobivenih iz selekcije atributa temeljene na korelaciji. SMO klasifikator dalje je postigao najbolje točnosti predviđanja od 86,468% i 86,138% s optimalnim skupovima dobivenim iz tehnika chi-kvadrat (11 atributa) i ReliefF (10 atributa) redom. [5]

2.4. Predviđanje bolesti srca

Sažetak ovog istraživanja temelji se na članku "*Heart Disease Prediction*" objavljenom 2019. godine, čiji je autor Rawat S.

Primjena strojnog učenja u ovom istraživanju ima za cilj klasificirati je li osoba oboljela od srčane bolesti ili ne koristeći Cleveland Heart Disease dataset. [3]

Navedeni skup podataka sastoji se od 303 zapisa i 14 atributa koji se koriste za analizu i koji je korišten u prvom istraživanju. Uzimajući u obzir razne faktore rizika kao što su dob, spol, angina (bol u prsima), krvni tlak, razina kolesterola i drugi, cilj je primijeniti različite modele strojnog učenja poput SVM-a (eng. *support vector machine*), Naive Bayesa, logističke regresije, stabla odlučivanja, slučajnih šuma, LightGBM-a i XGboosta za klasifikaciju i usporedbu rezultata. [3]

Analizirajući podatke, uočeno je da su osobe starije od 50 godina najčešće pogođene srčanom bolešću. Također se primjećuje da su žene koje pate od bolesti obično starije od muškaraca. Kako bi se pripremili podaci za analizu, provodi se prethodna obrada, uključujući rješavanje nedostajućih vrijednosti i podjelu podataka na skupove za trening i testiranje. [3]

Nakon primjene različitih modela strojnog učenja, evaluacija se provodi pomoću matrice konfuzije, koja prikazuje točno klasificirane vrijednosti i pogrešno klasificirane

vrijednosti za svaki model. Rezultati pokazuju da su logistička regresija i SVM postigli najvišu točnost na testnom skupu, što iznosi 80,32%. Najviša točnost na trening skupu postignuta je s 100% koristeći stablo odlučivanja. [3]

Zaključno, srčane bolesti predstavljaju značajan izazov za društvo. Ručno određivanje rizika od srčane bolesti na temelju faktora rizika može biti teško, ali primjena tehnika strojnog učenja omogućuje predviđanje izlaza na temelju postojećih podataka. Ova istraživanja potvrđuju važnost strojnog učenja i rudarenja podataka u borbi protiv srčanih bolesti. [3]

2.5. Predviđanje bolesti srca i analiza osjetljivosti klasifikatora

Sažetak ovog istraživanja temelji se na članku "*Prediction of heart disease and classifiers' sensitivity analysis*" objavljenom u časopisu "*BMC Bioinformatics*" 2020. godine. Autor ovog istraživanja je Almustafa K.

Ovo istraživanje temelji se na analizi podataka o srčanim bolestima i primjeni različitih metoda strojnog učenja za klasifikaciju tih bolesti. Korišten je Cleveland skup podataka koji je prikupljen iz različitih izvora koji se sastoji od 303 zapisa i koristi se u prvom i četvrtom istraživanju, a zatim je odabran podskup od 14 atributa koji su se pokazali relevantnima za analizu. Među tim atributima su dob, spol, vrsta bolova u prsima, mirovni krvni tlak, razina kolesterola, razina šećera u krvi nakon gladi, rezultati elektrokardiografskog testiranja, maksimalna postignuta brzina otkucaja srca, angina uzrokovana vježbanjem, promjena ST segmenta tijekom vrhunca vježbanja, broj glavnih krvnih žila otkrivenih fluoroskopijom i defekt srca. [4]

Korišteni su različiti klasifikatori, uključujući K-NN, Naive Bayes, stablo odlučivanja, SVM (eng. *support vector machine*), Adaboost i druge, za klasifikaciju podataka o srčanim bolestima. Evaluacija performansi klasifikatora provedena je pomoću unakrsne validacije s 10 sklopova. Rezultati istraživanja pokazuju vrlo obećavajuće rezultate u pogledu točnosti klasifikacije za K-NN (K = 1), stablo odlučovanja J48 i JRip klasifikatore u usporedbi s drugim klasifikatorima. Na temelju analize osjetljivosti, istraživači su proučavali utjecaj određenih parametara klasifikatora na performanse. Na primjeru stabla odlučivanja J48 klasifikatora analizirana je

osjetljivost parametra "pruning confidence factor". Rezultati analize pokazuju da određene vrijednosti parametra mogu poboljšati točnost klasifikacije. [4]

Dodatno, provedena je metoda izdvajanja značajki kako bi se odabrali relevantni atributi za svaki klasifikator. Ova metoda je pridonijela poboljšanju performansi nekih klasifikatora poput K-NN (N = 1) i stablo odlučivajna J48 klasifikatora. Istraživači su identificirali kombinaciju do četiri atributa koji su dovoljni za postizanje visoke točnosti klasifikacije srčanih bolesti. [4]

U konačnici, ovo istraživanje pruža uvid u primjenu različitih metoda strojnog učenja za klasifikaciju srčanih bolesti. Korišteni su algoritmi K-Nearest Neighbor (K-NN), Naive Bayes, stablo odlučivanja J48, JRip, SVM, Adaboost, Stochastic Gradient Descent (SGD) i Decision Table (DT) klasifikatori. Pokazano je da korištenje različitih algoritama za klasifikaciju skupa podataka o bolestima srca daje vrlo obećavajuće rezultate u pogledu točnosti klasifikacije za K-NN (K = 1), stablo odlučivanja J48 i JRip klasifikatore u usporedbi s Naive Bayes, SGD, SVM, Decision Table i Adaboost klasifikatorima, s točnošću klasifikacije od 99.7073%, 98.0488% i 97.2683% Dobiveni rezultati ukazuju na važnost odabira pravih atributa i parametara klasifikatora kako bi se postigla visoka točnost klasifikacije. Ovi rezultati mogu biti korisni u području medicinske dijagnostike i prediktivne analitike za srčane bolesti. [4]

2.6. Usporedba rezultata prethodnih istraživanja

Prema prvom istraživanju [1] rezultati su dobiveni korištenjem neuronskih mreža s 100% preciznosti. Matrica konfuzije sadrži informacije o stvarnim i predviđenim klasifikacijama koje je izvršio sustav klasifikacije. Podaci u matrici se evaluiraju kako bi se utvrdila učinkovitost takvih sustava. Tablica 4 prikazuje dobivene rezultate.

	A (pacijenti s bolestima srca)	**B (pacijenti bez bolesti srca)**
A (pacijenti s bolestima srca)	109 (TP)	0 (FN)
B (pacijenti bez bolesti srca)	0 (FP)	73 (TN)

Tablica 4: Matrica konfuzije [1]

Nakon primjene neuronskih mreža na skupu podataka za treniranje, rezultati pokazuju da nema nijedan FN ili FP unos, te to možemo vidjeti u tablici 4, što ukazuje na to da sustav predviđa bolesti srca sa 100% točnosti. Ovo je kreirano naravno na eksperimentalnim podacima pa je moguće da bi na nekim novim podacima neuronska mreža možda davala neke drugačije podatke i pronašla *FP* i *FN* podatke.

Što se tiče drugog istraživanja [2], pristup je bio malo drugačiji. Naime, autori ovog istraživanja koristili su više algoritama te kombinaciju *deep learning* i *machine learning* načina za usporedbu podataka i rezultata podataka. Za to su tri načina rezultata predstavljena.

U prvom načinu, podaci rezultata nisu normalizirani pa je preciznost rezultata samo 76,7%. Važno je napomenuti da je i navedena preciznost za metodu k-susjeda, koja iznosi malo više, tj. 82,27%. Nakon deep learning-a s 128 neurona, dobivena je optimizirana preciznost 76,7%. U drugom načinu dobivena je preciznost od 86,8%. U trećem načinu podaci su normalizirani i odrađena je selekcija značajki te je u ovom načinu najveću preciznost dala metoda k-susjeda, a ona iznosi 84,86%. Nakon provedbe deep learning-a dobivena je pouzdanost od 94,2%.

U trećem istraživanju [3] korišten je sličan način kao u prvom [1], odnosno prikazan je odnos točnih i netočnih podataka, točnije, točnog predviđanja za postojanje srčane bolesti ili nepostojanje srčane bolesti te netočnog predviđanja za postojanje ili nepostojanje srčane bolesti.

Tablica 5 prikazuje način prikaza rezultata za različite načine prikazivanja rezultata koji će biti opisani i predstavljeni u nastavku.

		Stvarne vrijednosti	
		Pozitivno (1)	**Negativno (0)**
Predviđene vrijednost	**Pozitivno (1)**	TP	FP
	Negativno (0)	FN	TN

Tablica 5: Matrica konfuzije [3]

Matrica konfuzije prikazuje točno predviđene vrijednosti, kao i netočno predviđene vrijednosti od strane klasifikatora. Zbroj TP i TN, iz matrice zabune, predstavlja broj ispravno klasificiranih unosa od strane klasifikatora.

U nastavku će biti prikazan set tablica koje prikazuju samo neke od načina prikaza rezultata, npr. Naive Bayes, stablo odlučivanja, itd.

	0	**1**
0	117	20
1	12	93

Tablica 6: Matrica konfuzije za Naive Bayes (Podaci za treniranje) [3]

	0	**1**
0	30	8
1	5	18

Tablica 7: Matrica konfuzije za Naive Bayes (Testni podaci) [3]

	0	**1**
0	129	0
1	0	113

Tablica 8: Matrica konfuzije za stablo odlučivanja (Podaci za treniranje) [3]

	0	1
0	29	8
1	6	18

Tablica 9: Matrica konfuzije za stablo odlučivanja (Testni podaci) [3]

. Metode su provedene na dvije vrste podataka, testnim podacima i podacima za treniranje. Ono što je vidljivo na prvu da je stablo odlučivanja odredilo postojanje srčane bolesti nad skupom podataka za testiranje s preciznošću od 100%, dok je nad istim skupom Naive Bayes bio malo manje precizniji. Kod testnog skupa podataka obje metode daju slične rezultate, odnosno samo se jedna vrijednost razlikuje, a to je da je stablo odlučivanja napravilo jedno krivo predviđanje nad testnim skupom podataka, točnije krivo predviđanje za nepostojanje srčane bolesti. Preciznosti za ova dva načina iznose respektivnim redoslijedom: 86,77% za skup za treniranje i 78,69% za testni skup, 100% za skup za treniranje i 77,05% za testni skup. Na slici 1 možemo vidjeti rezultate za sve klasifikatore. Najpreciznji rezultati od 80.32% (na testnim podacima) postignuti su logičkom regresijom i SVM metodom, dok je preciznost od 100% (nad podacima za testiranje) dalo stablo odlučivanja.

```
Accuracy for training set for svm = 0.9256198347107438
Accuracy for test set for svm = 0.8032786885245902

Accuracy for training set for Naive Bayes = 0.8677685950413223
Accuracy for test set for Naive Bayes = 0.7868852459016393

Accuracy for training set for Logistic Regression = 0.8636363636363636
Accuracy for test set for Logistic Regression = 0.8032786885245902

Accuracy for training set for Decision Tree = 1.0
Accuracy for test set for Decision Tree = 0.7704918032786885

Accuracy for training set for Random Forest = 0.987603305785124
Accuracy for test set for Random Forest = 0.7540983606557377

Accuracy for training set for LightGBM = 0.9958677685950413
Accuracy for test set for LightGBM = 0.7704918032786885

Accuracy for training set for XGBoost = 0.987603305785124
Accuracy for test set for XGBoost = 0.7540983606557377
```

Slika 1: Prikaz točnosti za sve klasifikatore [3]

Četvrto istraživanje [4] navodi samo određene preciznosti dodijeljene različitim metodama korištenih za obavljanje predviđanja. Ono što je važno za napomenuti je da

je u 3. i 4. istraživanju korišten isti skup podataka. Unatoč tome su za obavljanje predviđanja metodom stabla odlučivanja dobivene različite preciznosti, odnosno preciznost od 100% u 3. istraživanju, a u 4. nešto manje od 100%. Napomena je da je u 4. istraživanju korišteno stablo odlučivanja J48.

U petom istraživanju [5] korišten je Cleveland skup podataka o srcu kako bi postiglo najvišu točnost od 85.148% koristeći model SMO temeljen na cjelokupnom skupu atributa, te točnost od 84.158% koristeći model NB temeljen na optimalnom skupu od sedam atributa dobivenih iz selekcije atributa temeljene na korelaciji. SMO klasifikator dalje postiže najbolje točnosti predviđanja od 86.468% i 86.138% koristeći optimalne skupove dobivene iz metoda chi-kvadrat (11 atributa) i ReliefF (10 atributa), redom. Najbolje vrijednosti drugih metrika performansi, kao što su MAE (0.135), osjetljivost (0.865), specifičnost (0.90), pogreška tipa II (0.142), preciznost (0.865) i F-mjera (0.864), postignute su s SMO klasifikatorom uz metodu chi-kvadrat.

Klasifikator	Točnost	MAE	Osjetljivost	Pogreška tipa 2	Preciznost	F-mjera	Površina ispod ROC krivulje	Specifičnost
NB	83.498	0.183	0.835	0.171	0.835	0.835	0.909	0.870
LR	84.488	0.212	0.845	0.159	0.845	0.845	0.909	0.870
SMO	86.468	0.135	0.865	0.142	0.865	0.864	0.861	0.900
IBk/KNN	77.887	0.223	0.779	0.226	0.779	0.779	0.775	0.800
AdaBoostM1 + DS	83.498	0.224	0.835	0.173	0.836	0.834	0.899	0.880
AdaBoostM1 + LR	84.488	0.214	0.845	0.159	0.845	0.845	0.854	0.870
Bagging + REPTree	82.178	0.282	0.822	0.188	0.823	0.821	0.883	0.880
Bagging + LR	85.478	0.217	0.855	0.152	0.855	0.854	0.908	0.890
JRip	76.897	0.319	0.769	0.235	0.769	0.769	0.765	0.790
RF	83.168	0.257	0.815	0.193	0.816	0.814	0.904	0.880

Tablica 10: Preciznosti za različite metode predviđanja [5]

2.7. Zaključak na temelju prethodnih istraživanja

Kako je teško je ručno odrediti izglede za dobivanje srčanih bolesti na temelju čimbenika rizika, strojno učenje ovdje pomaže te sa velikom vjerojatnošću može odrediti dobivanje srčane bolesti. Iako različita istraživanja koriste isti skup podataka, možemo vidjeti kako inteligentni sustav daje različite preciznosti. To sve ovisi o različitim algoritmima koji se koriste kako bi se odredila preciznost i pouzdanost za otkrivanje srčanih bolesti kod ispitanika.

Također korištenje strojnog učenja na manjim podacima daje bolje i kvalitetnije rezultate nego kada se koriste na velikom skupu podataka. Kako bi se smanjilo vrijeme računanja, te kako bi podaci dobili veću preciznost, potrebno ih je prvo normalizirati. Također je dobro napraviti različite analize sa različitim algoritmima kao što su K - najbližih susjeda, Naive Bayes, DT kako bi se podaci mogli usporediti. Sva istraživanja pokazuju kako strojno učenje pomažu kod otkrivanja da li će nastupiti srčana bolest ili ne, te da bi trebalo svakako unaprijediti kako bi njihova pouzdanost mogla još bolje odrediti nastupanje srčane bolesti ti time na vrijeme mogla spasiti život.

3. Opis podataka

U skladu s prethodno navedenim istraživanjima koja su analizirala rezultate različitih istraživanja u vezi s predviđanjem srčanih bolesti, važno je u ovom poglavlju detaljnije opisati odabrani skup podataka ili dataset. Ovaj skup podataka poslužiti će kao temelj za analizu i izgradnju prediktivnih modela za predviđanje srčanih bolesti.

Skup podataka koji se koristi u ovom istraživanju sastoji se od različitih varijabli koje su relevantne za predviđanje srčanih bolesti. On sadrži informacije o demografskim karakteristikama pacijenata, povijesti bolesti, biometrijskim pokazateljima te rezultatima dijagnostičkih testova. Varijable uključuju, primjerice, dob, spol, tip boli u prsima, krvni tlak, razinu kolesterola i druge relevantne čimbenike.

U ovom poglavlju biti će opisan detaljan pregled strukture i karakteristika skupa podataka, kao i detaljni opisi varijabli koje su prisutne u tom skupu poadatka. Na taj način, biti će omogućeno razumijevanje dostupnih podataka te njihova prikladna upotreba u izgradnji prediktivnih modela za predviđanje srčanih bolesti.

Skup podataka preuzet je s *Kaggle* servisa, link na dataset: https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Skup podataka je kreiran od strane više različitih izvora, odnosno kombinirani su podaci s više različitih izvora. Kombinirano je 5 različitih izvora te je odabrano 11 zajedničkih varijabli s tih 5 skupova podataka, a kreirali su ga Andras Janosi M.D., William SteinBrunn M.D., Matthias Pfisterer M.D., Robert Detrano M.D. Ph.D. Podaci dolaze od sljedećih izvora:

- Cleveland: 303 redova
- Mađarska: 294 redova
- Švicarska: 123 redova
- Long Beach: 200 redova
- Stalog (Heart) Data set: 270 redova

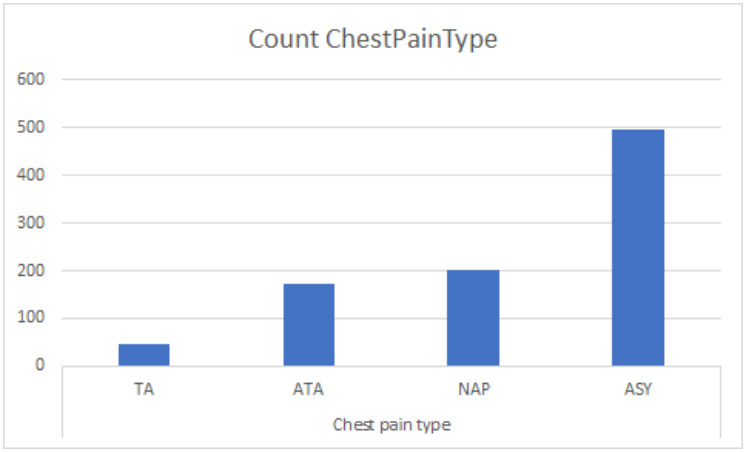
To znači da je sveukupno 918 redova podataka u konačnom skupu podataka jer je 272 redova duplikata od ukupno 1190 redova kombinacijom gornje navedenih skupova podataka.

Odabrani skup podataka sadrži 11 atributa i 1 atributa koji određuje srčanu bolest, odnosno varijabli koji su objašnjeni i prikazani u tablici 11.

Atribut	Opis	Vrijednosti
Age	Godine pacijenta	**godine**
Sex	Spol pacijenta	**M**: Male, **F**: Female
ChestPainType	Tip boli u prsima	**TA**: Typical Angina, **ATA**: Atypical Angina, **NAP**: Non-Anginal Pain, **ASY**: Asymptomatic
RestingBp	Krvni tlak u mirovanju	**mm Hg**
Cholesterol	Kolesterol	**mm/dl**
FastingBS	Šećer u krvi natašte	**1**: ako je FastingBS > 120 mg/dl, **0**: inače
RestingECG	Rezultati elektrokardiograma u mirovanju	**Normal**, **ST**, **LVH**
MaxHR	Max. brzina otkucaja srca	**60-202**
ExerciseAngina	Angina uzrokovana vježbanjem	**Y**: Yes, **N**: No
Oldpeak	Odnos vježbanja s odmorom	vrijednost mjerenja u depresiji
ST_Slope	Nagib vrha vježbanja ST segmenta	**Up**, **Flat**, **Down**
HeartDisease	Ima li pacijent srčanu bolest ili ne	**1**: srčana bolest, **0**: normalno (bez srčane bolesti)

Tablica 11: Popis atributa

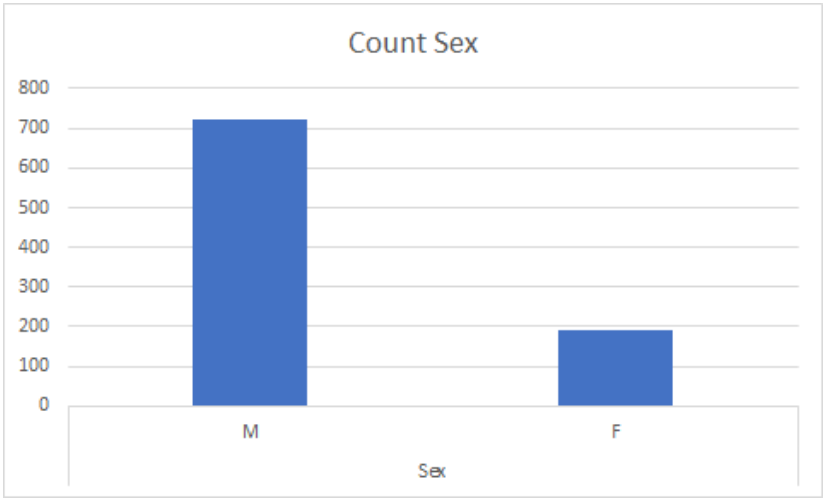
Što se kvalitete podataka tiče, podaci se čine u redu, ali takvi bi i morali biti jer su odabrani od strane više specijaliste s više izvora i prerađeni kako bi najbitniji atributi sudjelovali u predviđanju. U nastavku su opisani određeni atributi.



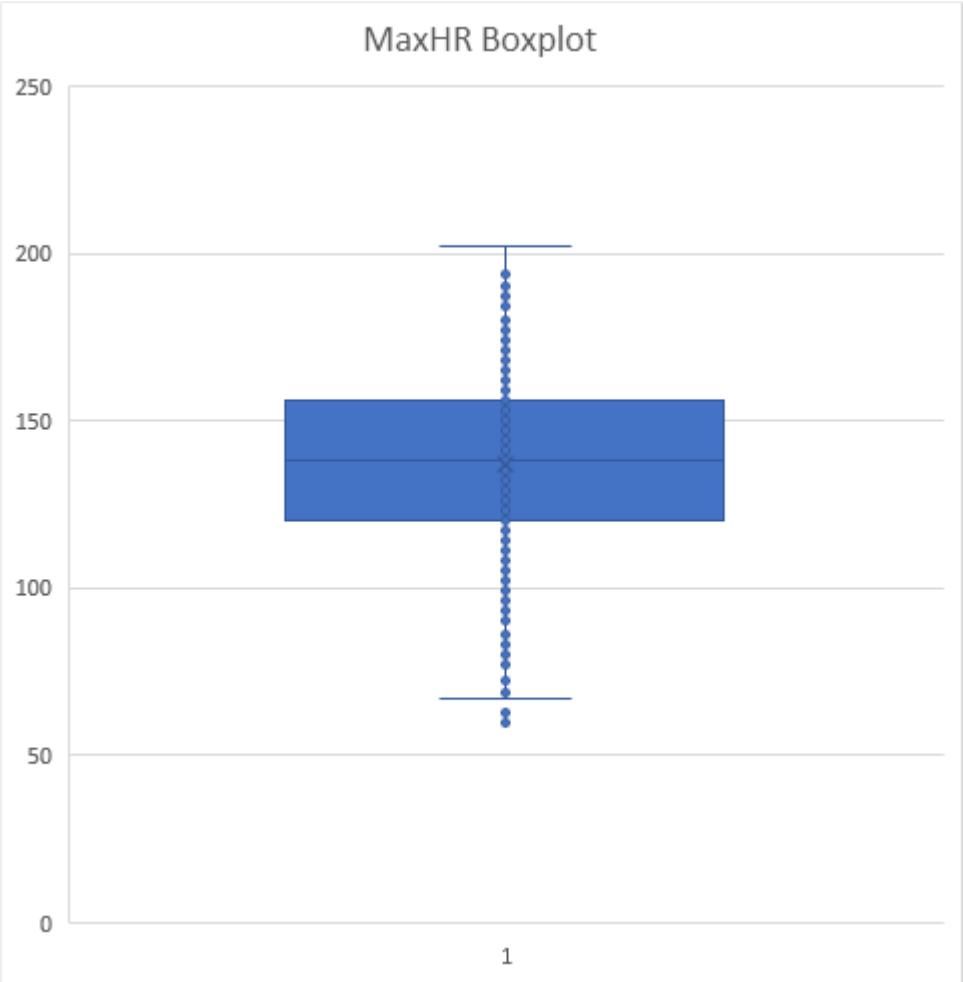
Slika 2: Prikaz prebrojenih vrijednosti za vrijednosti varijable *„ChestPainType“*

Slika 2 prikazuje broj pojavljivanja svake vrijednosti za atribut *„ChestPainType“.* Vidljivo je da je najviše zastupljena asimptomatska vrijednost *„ASY“.*

Još jedan zanimljiv graf na slici 3 koji prikazuje omjer muških i ženskih pacijenata što bi u nekim slučajevima moglo izazivati određene probleme budući da je muških pacijenata 724, a ženskih samo 193 u tom omjeru, ali naravno može biti i indikator da u većini slučajeva muškarci razvijaju simptome za srčane bolesti.



Slika 3: Prikaz prebrojenih vrijednosti za vrijednosti varijable *„Sex“*



Slika 4: Prikaz boxplot dijagrama za vrijednosti varijable „MaxHR“

Na slici 4 prikazan je boxplot dijagram koji ima svrhu pokazati medijalne, maksimalne, minimalne vrijednosti iz stupca, odnosno atributa koji određuje maksimalnu postignutu brzinu kucanja srca. Dijagram također prikazuje i inner vrijednost i outlier vrijednosti. Maksimalna vrijednost je 202, a minimalna 60, medijan iznosi 138.

4. Priprema podataka

U ovoj fazi potrebno je provesti nekoliko ključnih koraka kao što su odabir, čišćenje, konstrukcija, integracija i oblikovanje podataka, da bi u konačnici dobili potpun i uređen skup kako bi mogli dobiti što preciznije rezultate. Zbog toga, priprema podataka predstavlja vrlo važan korak pri izradi ovog projekta. Za proces pripreme podataka korišteno je svih 918 zapisa koji se nalaze u skupu podataka.

4.1. Odabir, čišćenje, konstrukcija, integracija i oblikovanje podataka

U prvom koraku su podaci sa interneta preuzeti i uvezeni u Excel. Skup podataka sastoji se od 12 atributa koji će biti iskorišteni u daljnjoj izradi projekta. Kroz primjenu funkcije Filtriranje, izvršena je provjera ispravnosti podataka za svaki atribut, uključujući provjeru praznina, nevažećih vrijednosti, neispravnih naziva i slično. Tijekom tog procesa, primijećena je jedna vrijednost (zapis) "0" za atribut "RestingBP" te je odlučeno da se ta vrijednost izbriše, budući da krvni tlak prilikom mirovanja ne može biti nula.

1	Age	Sex	ChestPain	Resting	Cholester	FastingBS	RestingEC	MaxHR	ExerciseA	Oldpeak	ST_Slope	HeartDisease
451	55	M	NAP	0	0	0	Normal	155	N	1.5	Flat	1

Slika 5: Nevažeći zapis

Ostali podaci su svi ispravni, te se skup sveo na 917 potpunih podataka koje možemo koristiti u sljedećoj fazi modeliranja.

5. Opis korištenih metoda

5.1. Stablo odlučivanja

Stablo odlučivanja je metoda koja se koristi u analizi odlučivanja i strojnom učenju kako bi se grafički predstavio proces donošenja odluka. U analizi odlučivanja, stabla se koriste za opisivanje načina donošenja odluka od strane eksperta, dok se u strojnom učenju koriste za predikciju na temelju podataka. Stabla odlučivanja mogu se koristiti za različite vrste problema, poput odabira proizvoda, lokacije tvrtke ili medicinske dijagnoze. [8]

Stabla odlučivanja temelje se na slijedu testova na deskriptivnim atributima upita. Svaki čvor u stablu predstavlja jedan atribut, a rubovi čvorova su označeni mogućim vrijednostima tog atributa. Svaki list u stablu predstavlja izlaznu vrijednost ili klasifikaciju na temelju kombinacije atributa. Stablo se konstruira "učenjem" na podacima, grananjem izvornog skupa podataka na temelju testiranja vrijednosti atributa. [8]

Stablo odlučivanja, posebno klasifikacijska i regresijska stabla (CART), je metoda rudarenja podataka koja je testirana na promatranom uzorku kako bi se izgradio uspješan model. Ova metoda pruža grafički prikaz utjecaja ulaznih varijabli na izlaznu varijablu, koja može biti izražena kao klasa ili kategorija. Svaki čvor u stablu predstavlja jednu ulaznu varijablu, a na rubovima su označeni "dječji čvorovi" za svaku moguću vrijednost te ulazne varijable. Svaki list u stablu predstavlja vrijednost izlazne varijable kada su vrijednosti ulaznih varijabli predstavljene putem od korijena stabla do tog lista. [7]

Za izgradnju stabla koristi se CART algoritam, koji koristi Gini indeks (IG) kao evaluacijsku funkciju za odabir najboljeg prijeloma. Algoritam uzima u obzir sva moguća grananja kako bi pronašao najbolje grananje za postizanje što točnijeg modela. On može uspješno raditi s kontinuiranim i kategorijalnim varijablama. [7]

Proces kreiranja stabla odlučivanja uključuje pripremu podataka, odabir algoritma i parametara stabla, generiranje grafičke i numeričke strukture stabla, te evaluaciju rezultata i tumačenje greške. Stablo odlučivanja se koristi u praksi za donošenje odluka na temelju ulaznih podataka. [8]

Stablo se konstruira "učenjem" na podacima putem procesa grananja (splitting) izvornog skupa podataka u podskupove na temelju testiranja vrijednosti varijabli. Ovaj proces se rekurzivno ponavlja na svakom podskupu dok se ne dostigne određeni uvjet zaustavljanja, kao što je jednakost vrijednosti izlazne varijable u podskupu ili nedostatak daljnjeg poboljšanja rezultata. [7]

Stablo raste sve dok se ne pronađe novo grananje koje poboljšava razdvajanje podataka u klase. Međutim, kako se brojnost populacije smanjuje s svakim novim grananjem, potrebno je smanjivati stablo (pruning) radi dobivanja točnije klasifikacije. Cilj je identificirati grane koje imaju najmanju prediktivnu sposobnost za određeni list i izbaciti ih iz stabla. Na kraju se odabire stablo odgovarajuće veličine s obzirom na točnost klasifikacije, uzimajući u obzir odnos složenosti stabla i veličine greške. [7].

Stablo odlučivanja se sastoji od sljedećih komponenti [11]:

- početnog čvora,
- unutarnjih čvorova
- čvorova listova.

Stablo odlučivanja koristi niz testova na deskriptivnim atributima upita za predikciju. Njegove prednosti uključuju interpretabilnost, rad s različitim vrstama atributa (kategorijski i kontinuirani), modeliranje interakcija među deskriptivnih atributa, robusno je na kurs dimenzionalnosti te je robusno na šum u skupu, ako se radi obrezivanje. [11]

Međutim, stablo odlučivanja ima i neke nedostatke. Kada se radi s kontinuiranim atributima, stablo može postati veliko. Također, stablo je osjetljivo na skup podataka na kojem radi te može doći do pretreniranosti ako ima puno atributa (kurs dimenzionalnosti). Također, male promjene u skupu podataka mogu uzrokovati značajne promjene u strukturi stabla (*eng. concept drift*)**.** [11]

5.2. Neuronske mreže

Neuronske mreže su računalne strukture inspirirane biološkim procesima u živčanom sustavu. One se razlikuju od drugih informacijskih tehnologija jer su adaptivne i mogu se unaprijediti s iskustvom. Neuronske mreže su također prirodno masivno paralelne. Interes za neuronske mreže proizlazi iz njihove sposobnosti obrade

informacija koja može biti vrlo snažna, čak i kod jednostavnih organizama, te iz želje za razumijevanjem bioloških sustava putem računalnog modeliranja mozga. [9]

Ulazne vrijednosti u neuronskoj mreži se šire prema naprijed, što znači da se informacija kreće od ulaza prema izlazu. Učenje neuronske mreže uključuje predstavljanje uzoraka ulaza za koje već poznajemo očekivane izlaze. Na temelju tih ulaza, neuronska mreža treba naučiti generirati pravilne izlaze. Ako neuronska mreža nije dobro naučena, može proizvesti neprikladne izlaze za dane ulaze. [9]

Umjetne neuronske mreže su moćni sustavi koji mogu obavljati složene zadatke i imaju potencijal za napredak u različitim područjima. Razlikuju se od konvencionalnih računalnih sustava po svojoj paralelnoj prirodi i suvišnosti neurona za pojedini proces. Neuronske mreže temelje se na sinapsama, koje su veze između neurona, i mogu se koristiti za rješavanje problema koji zahtijevaju veliki broj veza među pod-sustavima. Iako su umjetne neuronske mreže još u razvoju, njihova primjena je sve više prisutna u različitim područjima, a njihova sposobnost učenja i paralelne obrade informacija čini ih zanimljivim alatima za budućnost. [10]

Umjetne neuronske mreže su modeli koji imitiraju biološke neuronske mreže i koriste se za razumijevanje bioloških sustava i rješavanje problema u području umjetne inteligencije. One su građene od međusobno povezanih umjetnih neurona. Umjetne neuronske mreže se temelje na paralelnoj obradi informacija, mogu učiti i prilagođavati se te imaju široku primjenu u različitim područjima kao što su zdravstvo, obrazovanje, elektrotehnika, financije i marketing. One koriste strukturu ljudskog mozga kako bi razvile strategiju analize podataka i rješavanja složenih problema. [12]

Umjetne neuronske mreže posjeduju neka važna svojstva. Prvo, one omogućuju paralelnu obradu informacija, što znači da podaci mogu biti raspoređeni na više jedinica mreže. Drugo, umjetne neuronske mreže su otporne na kvarove, što znači da mogu nastaviti raditi čak i ako se dio mreže ošteti ili prestane funkcionirati. Treće, umjetne neuronske mreže imaju sposobnost učenja i adaptacije, što ih čini sposobnima za obradu nepreciznih podataka u nestrukturiranim okruženjima. Također, umjetne neuronske mreže su pogodne za modeliranje i upravljanje viševarijabilnim sustavima, te imaju sposobnost aproksimiranja proizvoljnih kontinuiranih nelinearnih funkcija. [12]

Model umjetnog neurona je osmišljen kako bi imitirao funkcionalnost biološkog neurona. Umjetni neuron prima signale na ulazu, koje množi težinskim faktorima, a zatim zbraja rezultate. Ako dobiveni iznos prelazi definirani prag, neuron generira izlazni signal. Umjetni neuron može imati različite funkcije, kao što su funkcija praga ili prijenosna funkcija. Ovaj model umjetnog neurona omogućuje umjetnim neuronskim mrežama obavljanje različitih računalnih zadataka. [12]

Višeslojni perceptron je jedna od najraširenijih neuronskih mreža i koristi se za učenje pomoću algoritma "širenje unatrag". Ova mreža, koju su razvili Paul Werbos, David Rumelhart, Geoffrey Hinton i Stephen Williams, ima jedan ili više skrivenih slojeva. Podaci prolaze kroz mrežu od ulaznog sloja do skrivenog sloja, a zatim do izlaznog sloja, pri čemu se greška računa unatrag i koristi za prilagodbu težina. [12]

Tijek podataka je definiran u nekoliko koraka. Prvo, podaci ulaze u ulazni sloj i šire se prema skrivenom sloju putem prijenosne funkcije. U skrivenom sloju se računaju ukupne ulazne i izlazne vrijednosti za svaku jedinicu analize. U izlaznom sloju se računa lokalna greška za svaku jedinicu, a podaci se šire unatrag od izlaznog sloja do skrivenih slojeva. U mreži se koriste ulazni signal i signal greške, koji se šire kroz mrežu prema naprijed i unatrag. Mreža "širenja unatrag" koristi se za predviđanje vrijednosti izlaznih varijabli i za probleme klasifikacije. [12]

Mreža "širenja unatrag" (*eng .backpropagation*) ima određene prednosti i nedostatke. Među prednostima se ističe mogućnost korištenja dodatnih slojeva koji omogućuju kompleksniju obradu podataka i stvaranje složenih sustava. Međutim, postoje i neki nedostaci. Treniranje mreže je dugotrajno i osjetljivo na početne vrijednosti težina. Algoritmi treniranja ne garantiraju konvergenciju, što može otežati postizanje željenih rezultata. Mreža također ima ograničenja u pogledu broja skrivenih neurona. Ako mreža nema dovoljno skrivenih neurona, može biti nedovoljno složena za rješavanje problema aproksimacije funkcije i može proizvesti veliku grešku na izlazu. S druge strane, ako mreža ima previše skrivenih neurona, postoji rizik od pretjeranog prilagođavanja (*eng. overfitting*), gdje mreža dobro radi na uzorku za treniranje, ali loše na uzorku za testiranje. Također, predviđanje greške testiranja na temelju treniranih podataka može biti teško. Određivanje optimalnog broja iteracija treniranja je također izazovno, a rješenje uključuje zaustavljanje treniranja kada greška validacije počne rasti. [9]

5.3. Bayseove mreže

Bayesove mreže su grafički modeli koji prikazuju vjerojatnosne ovisnosti između varijabli. Učenje parametara i strukture Bayesovih mreža pružaju korisne metode, uključujući učenje s nepotpunim podacima. Bayesove tehnike mogu se primijeniti na nadgledano i nenadgledano učenje te pružaju grafički prikaz ovisnosti između varijabli u sustavu. [9]

Osnovna svojstva Bayesovih mreža imaju nekoliko prednosti za analizu podataka u kombinaciji sa statističkim tehnikama. Model može riješiti situacije kada nedostaju podaci tako što kodira ovisnosti između svih varijabli. Također mogu se koristiti za učenje uzročnih ovisnosti i poboljšanje razumijevanja problemske domene te predviđanje posljedica nekih djelovanja. Budući da model ima i uzročnu i vjerojatnosnu semantiku, kombinira prethodno znanje s podacima. Bayesove statističke metode povezane s Bayesovim mrežama pružaju učinkovit i principijelan pristup za izbjegavanje pretreniranosti podataka. [9]

Bayesova mreža je probabilistički grafički model koji prikazuje zavisnosti među varijablama. Mreža se sastoji od čvorova koji predstavljaju varijable i bridova koji označavaju njihove međuzavisnosti. Ova vrsta modela omogućuje nam da razumijemo kompleksne vjerojatnosne strukture i izračunamo očekivanja (vjerojatnosti) nepoznatih varijabli na temelju već poznatih varijabli. [6]

Bayesove mreže su postale popularne osamdesetih godina prošlog stoljeća, kada su formirane njihove osnovne strukture poput belief networks, causal networks i influence diagrams. Tada su razvijeni i alati za analizu i učenje Bayesovih mreža, što je doprinijelo njihovoj širokoj primjeni u različitim područjima. Danas se Bayesove mreže koriste u računalnim sustavima za modeliranje i predviđanje u poslovnim, društvenim, ekološkim i drugim sustavima.[6]

Da bismo definirali Bayesovu mrežu, potrebno je odrediti čvorove u mreži, odnosno varijable koje se razmatraju. Svaki čvor može imati različite ishode, odnosno vrijednosti koje ta varijabla može poprimiti. Bridovi u mreži odražavaju međuzavisnosti između varijabli, dok združene distribucije vjerojatnosti definiraju vjerojatnosti ishoda svakog čvora ovisno o njegovim roditeljima u mreži. [6]

Jedan od ključnih koraka u oblikovanju Bayesove mreže je definiranje tablica uvjetnih vjerojatnosti. Te tablice sadrže vjerojatnosti ishoda za svaku varijablu, ovisno o vrijednostima njenih roditelja. Za varijable koje nemaju roditelje, dovoljno je definirati a priori očekivanja njihovih ishoda. U slučaju diskretnih binarnih varijabli, broj roditelja određuje dimenzionalnost tablice vjerojatnosti. [6]

Zaključivanje u Bayesovoj mreži omogućuje nam da izračunamo vjerojatnosti za bilo koji čvor u mreži, iako nemamo potpune informacije o ostalim varijablama. Međutim, precizno zaključivanje može biti izazovno za složene mreže s velikim brojem varijabli. U takvim slučajevima, koristimo približna rješenja koja pružaju brže rezultate s malim odstupanjima od točnih rješenja. [6]

Računanje vjerojatnosti u Bayesovoj mreži može se obavljati propagacijom informacija unaprijed i unatrag. Propagacija unaprijed koristi se za zaključivanje od roditelja prema djeci, dok se propagacija unatrag koristi za zaključivanje od djece prema roditeljima. Ovisno o postojanju dokaza, očekivanja varijabli se mijenjaju, a vjerojatnosti se računaju s obzirom na te dokaze. [13]

Bayesove mreže omogućuju uštedu broja vjerojatnosti jer se temelje na pretpostavci nezavisnosti između varijabli. Također, koriste se algoritmi poput Pearloveg algoritma koji omogućuju efikasnije računanje vjerojatnosti. Nezavisnost varijabli u mreži omogućuje smanjenje broja potrebnih vrijednosti i olakšava zaključivanje. [13]

6. Razvoj prediktivnih modela

U ovom koraku biti će opisani procesi modeliranja problema. Prva tehnika za modeliranje će biti stablo odlučivanja i neuronske mreže, a zatim Bayesova mreža. Odabir tehnika modeliranja za predviđanje mogućih srčanih bolesti je prvi korak u procesu. Za ovaj slučaj, odabrane su tri tehnike modeliranja: stablo odlučivanja, neuronske mreže i Bayesove mreže. Stablo odlučivanja i neuronske mreže biti će izrađene korištenjem programa BigML, dok će Bayesova mreža biti izrađena korištenjem programa Netica. Prvo će biti definiran dizajn za testiranje svake tehnike modeliranja, a zatim će biti izrađeni i procijenjeni modeli. Očišćeni i uređeni skup podataka biti će uvezeni u BigML kako bi se mogle primijeniti prve dvije tehnike, stablo odlučivanja i neuronske mreže. Nakon uvoza podataka u BigML, bit će generirana tablica prikazana na slici 6.

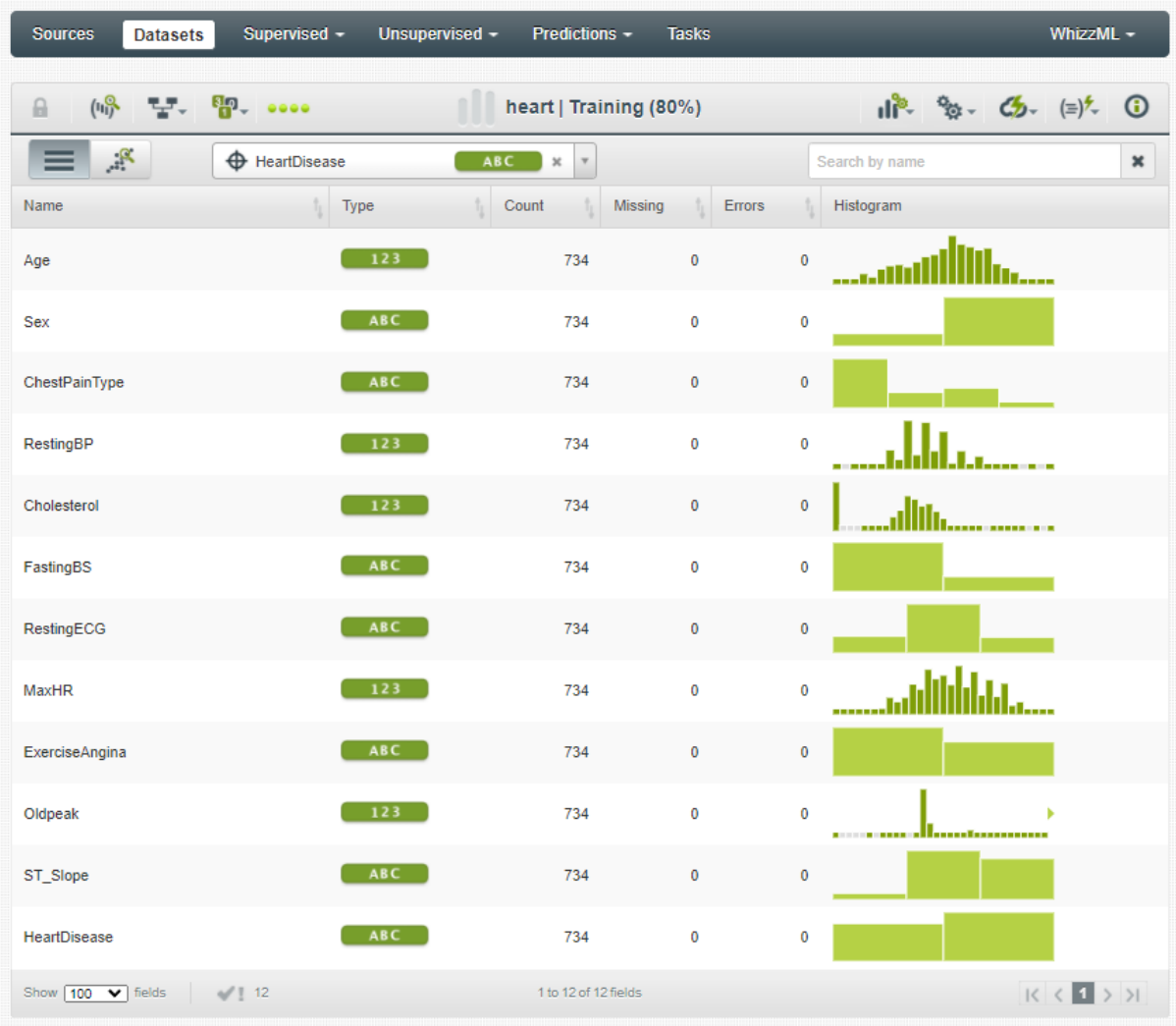
Name	Type	Count	Missing	Errors	Histogram
Age	123	917	0	0	
Sex	ABC	917	0	0	
ChestPainType	ABC	917	0	0	
RestingBP	123	917	0	0	
Cholesterol	123	917	0	0	
FastingBS	ABC	917	0	0	
RestingECG	ABC	917	0	0	
MaxHR	123	917	0	0	
ExerciseAngina	ABC	917	0	0	
Oldpeak	123	917	0	0	
ST_Slope	ABC	917	0	0	
HeartDisease	ABC	917	0	0	

Slika 6: Tablica atributa u alatu BigML

Sljedeći korak je izrada dizajna za testiranje koja nam je potrebna za vrednovanje modela stabla odlučivanja i neuronske mreže.

6.1. Dizajn za testiranje

Dizajn za testiranje biti će napravljen pomoću funkcije *Random Split* u alatu BigML. Ta funkcija će podijeliti skup podataka na skup za treniranje i skup za testiranje. U skup podataka za treniranje ulazi 80% ukupnih podataka dok u skup za testiranje ulazi 20% ukupnih podataka. Na slici 7 nalazi se skup podataka za treniranje.

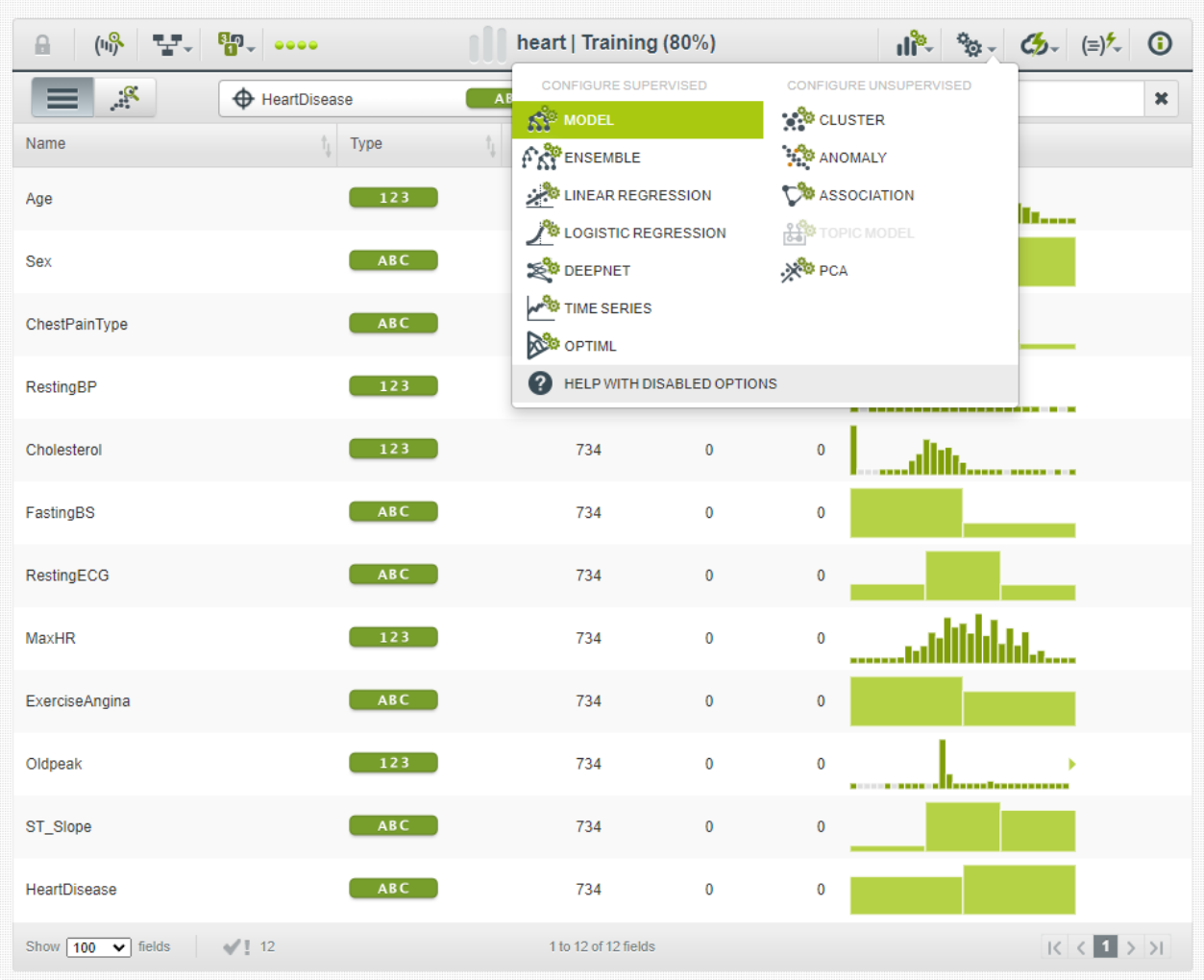


Slika 7: Skup podataka za treniranje

U sljedećem koraku biti će prikazan proces izrade modela stabla odlučivanja.

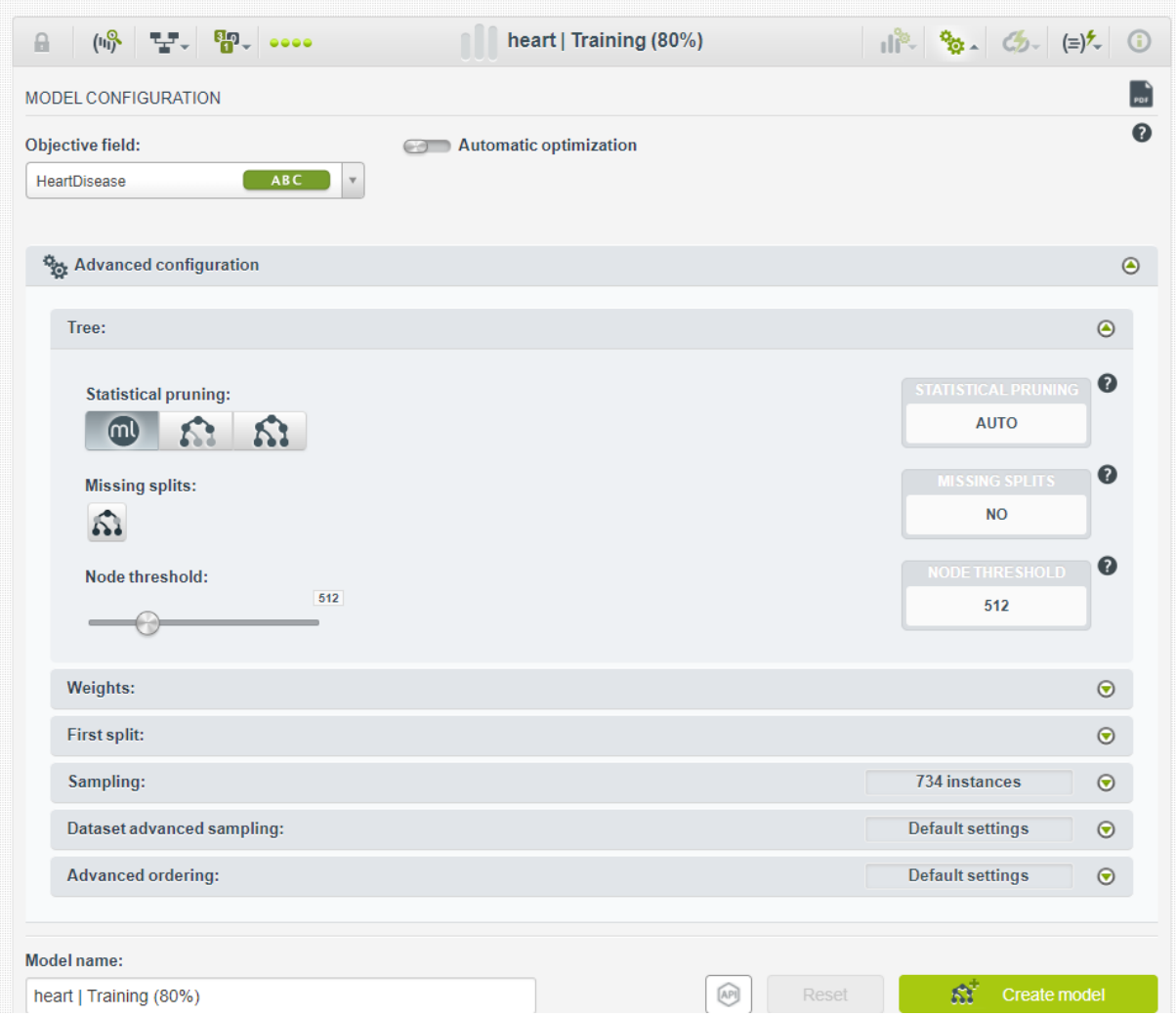
6.2. Stablo odlučivanja

Kako bi se mogao izraditi model stabla odlučivanja, prvo je potrebno ući u model za treniranje kod izbora Dataset-a, te odabrati funkcionalnost "*MODEL*". To se može vidjeti na slici 8.



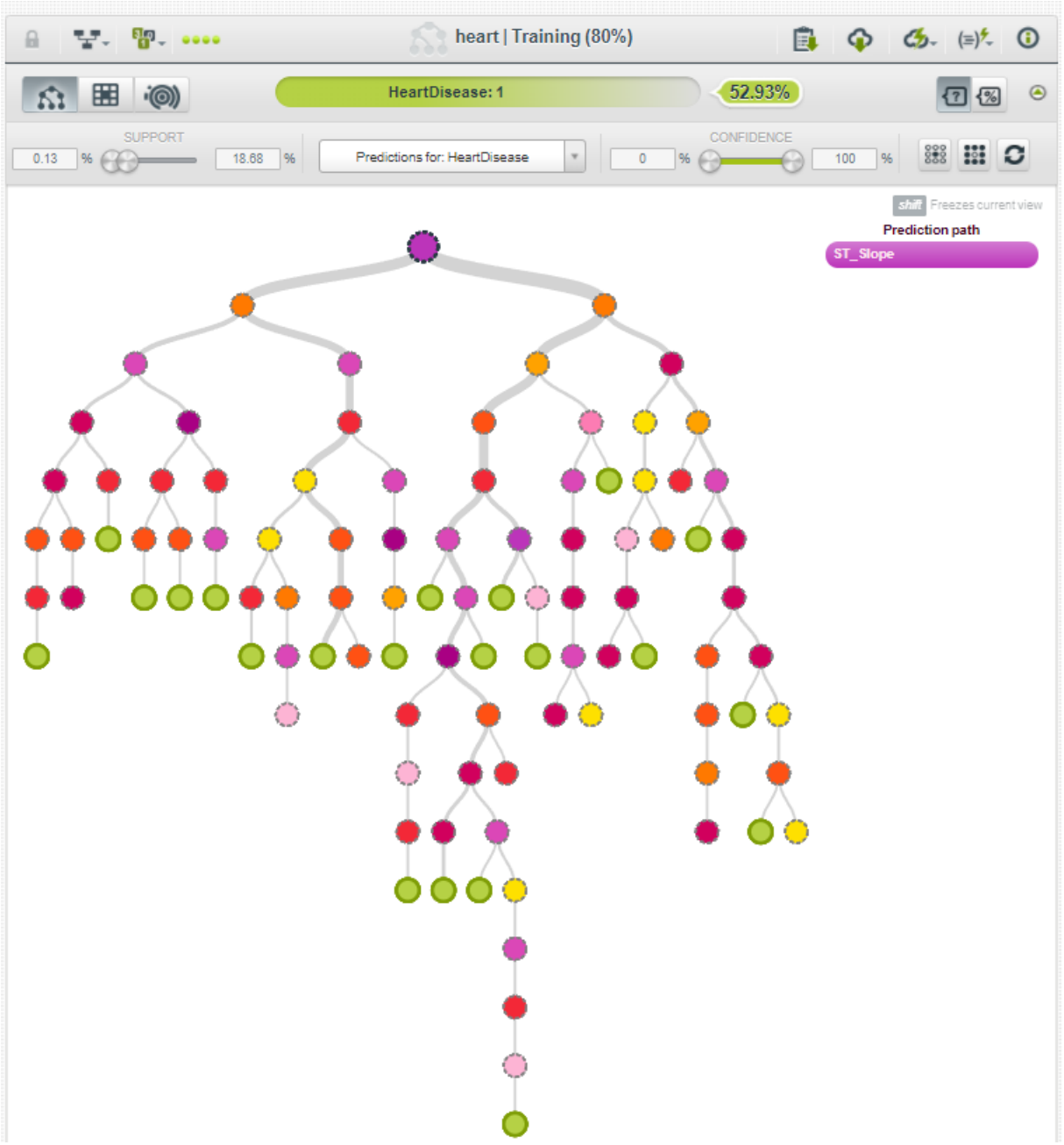
Slika 8: Izrada modela

Na slici 8 prikazano je da je odabran model za treniranje „heart | Training (80%)", te odabrana funkcionalnost "*MODEL*". Nakon toga se otvara prozor sa dodatnim mogućnostima kod kojih se može odabrati više vrsta obrezivanja. U ovom slučaju odabrano je pametno obrezivanje (smart pruning).



Slika 9: Odabiranje statističkog obrezivanja

Na slici 9 prikazana je funkcionalnost "*MODEL*", te odabiranje statističkog obrezivanja. Nakon što su svi parametri posloženi, može se napraviti model pritiskom na tipku "*Create model*". Prikaz modela može se vidjeti na slici 10. Pouzdanost ovog modela za prvu granu iznosi 52.93%, dok pouzdanost za najdublju granu iznosi 67.56%. Vjerojatnost za prvu granu iznosi 56.54%, dok za najdublju granu iznosi 95.17%.



Slika 10: Prikaz stabla odlučivanja modela za trening

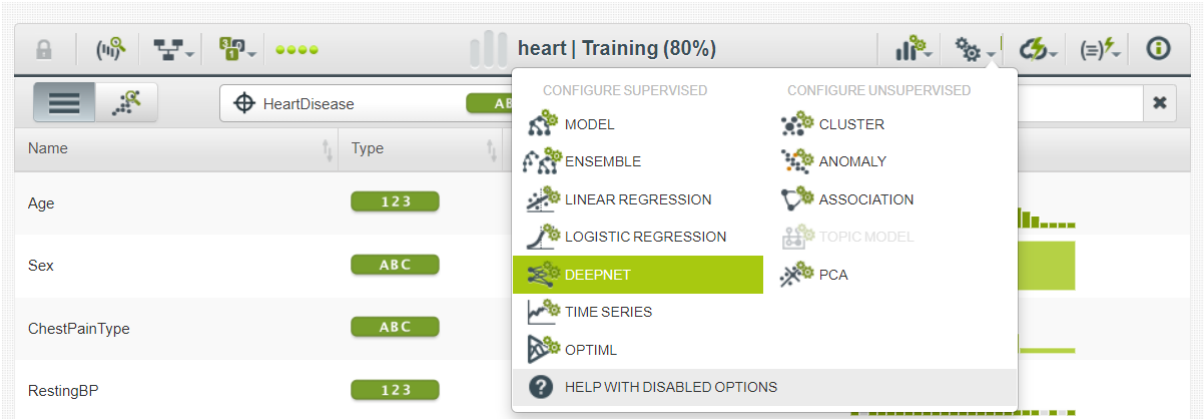
Prema izvješću kod Model Summary Report najvažniji je “*ST_Slope*” ili nagib vrha vježbanja (21.77%), te nakon toga slijede “*ChessPainType*” (16.1%) i “*OldPeek*” (15.18%) koji ukupno uzimaju više od 50% važnosti kod modela. Prikaz izvješća može se vidjeti na slici 11.

Slika 11: Prikaz izvješća Model Summary Report

Sljedeći model za izradu je model neuronskih mreža te se kao i kod izrade stabla odlučivanja treba pozicionirati na pravi skup podataka, odnosno na skup podataka za trening.

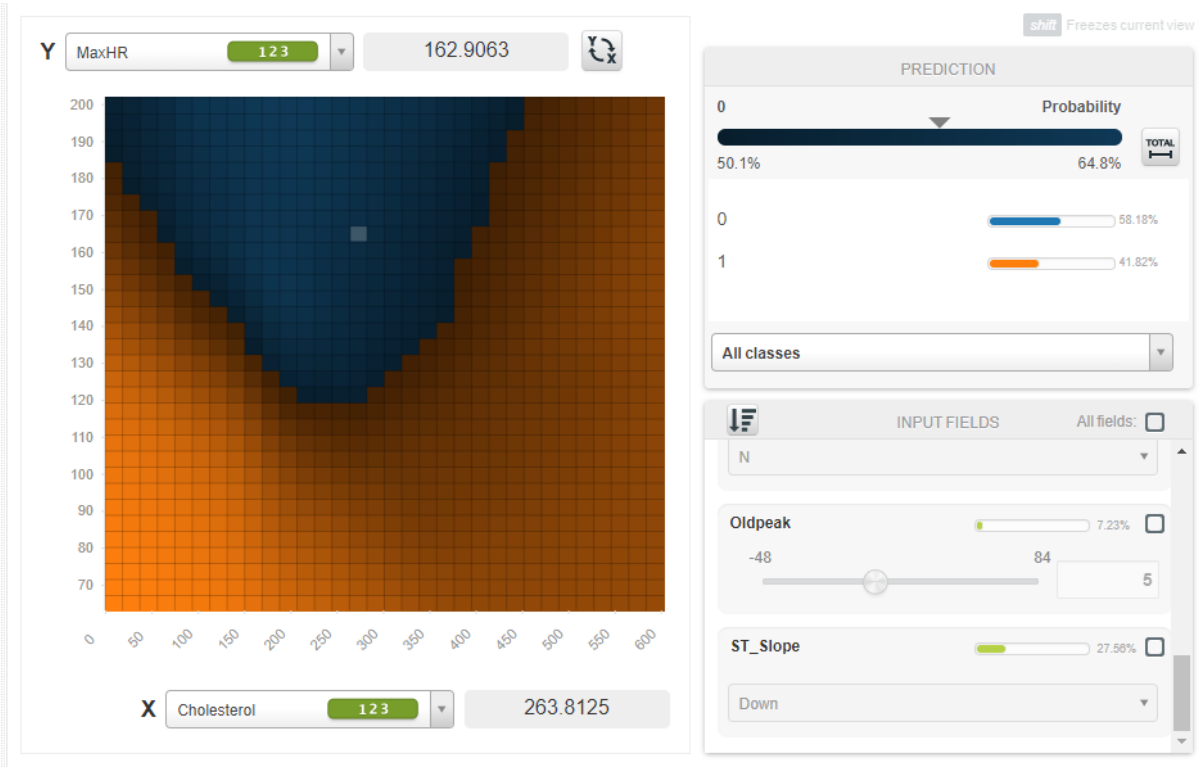
6.3. Neuronske mreže

Ovog puta odabire se funkcionalnost "*DEEPNET*", te se to može vidjeti na slici 12.



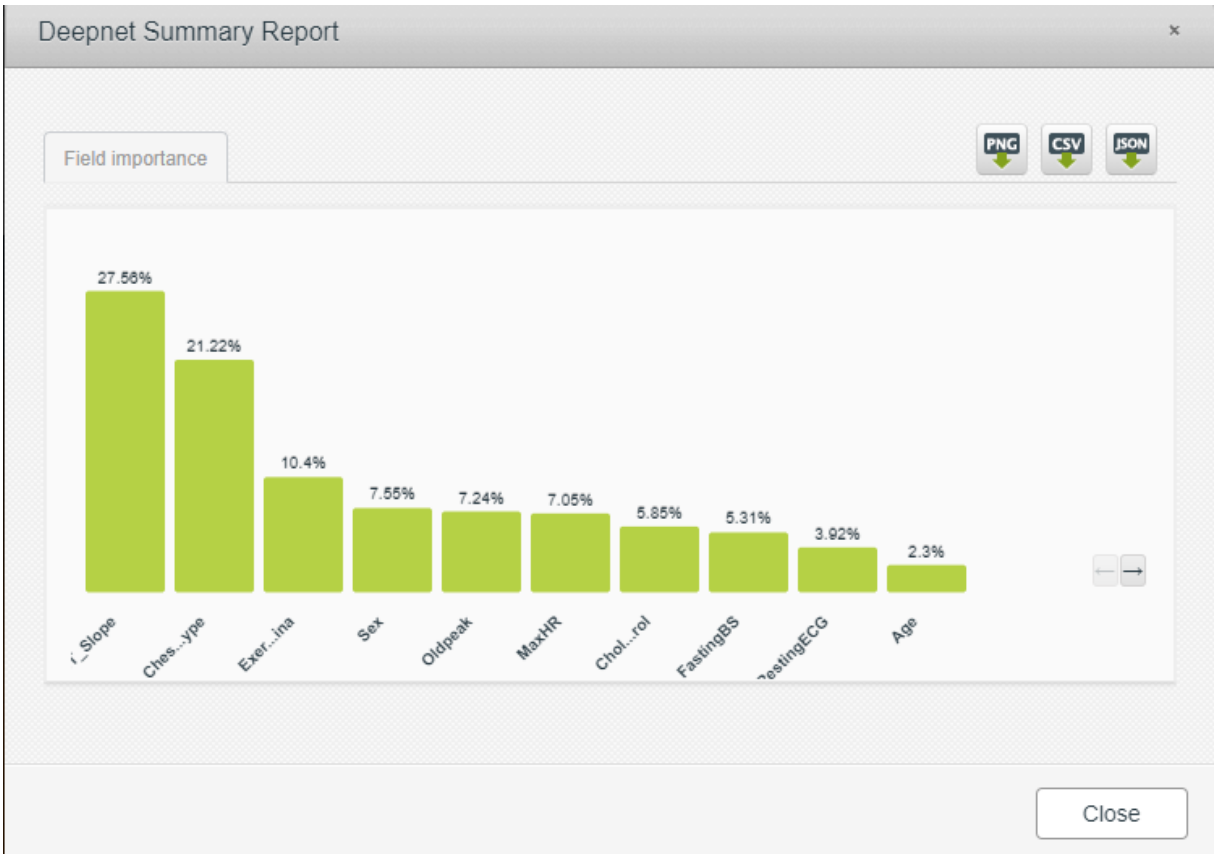
Slika 12: Odabir funkcionalnosti DEEPNET

Prilikom pritiska na “*DEEPNET”* opciju, otvara se dodatni prozor za konfiguriranje neuronske mreže. Prva opcija koja se treba odabrati je za koju varijablu se radi predikcija pomoću neuronske mreže. U ovom slučaju odabrana je varijabla “*HeartDisease*”, tj da li je ispitanik imao srčanu bolest. Također je kao i kod stabla odlučivanja moguće dodatno konfigurirati opcije, te se nakon toga pritišće gumb za izradu neuroske mreže, “*Create Deepnet”*.



Slika 13: Prikaz neuronske mreže

Na slici 13 prikazana je neuronska mreža za varijable “*Cholesterol*” i “*MaxHR*”. Narančasta boja pokazuje da je šansa za srčanu bolest veća od 50%, dok plava pokazuje da je šansa za srčanu bolest manja od 50%. Također, što je boja tamnija to je i postotak za dobivanje srčane bolesti veći ili manji, ovisno o boji. Može se vidjeti da ukoliko je kolesterol veći od 450, da je šansa za srčanu bolest veća od 50%.



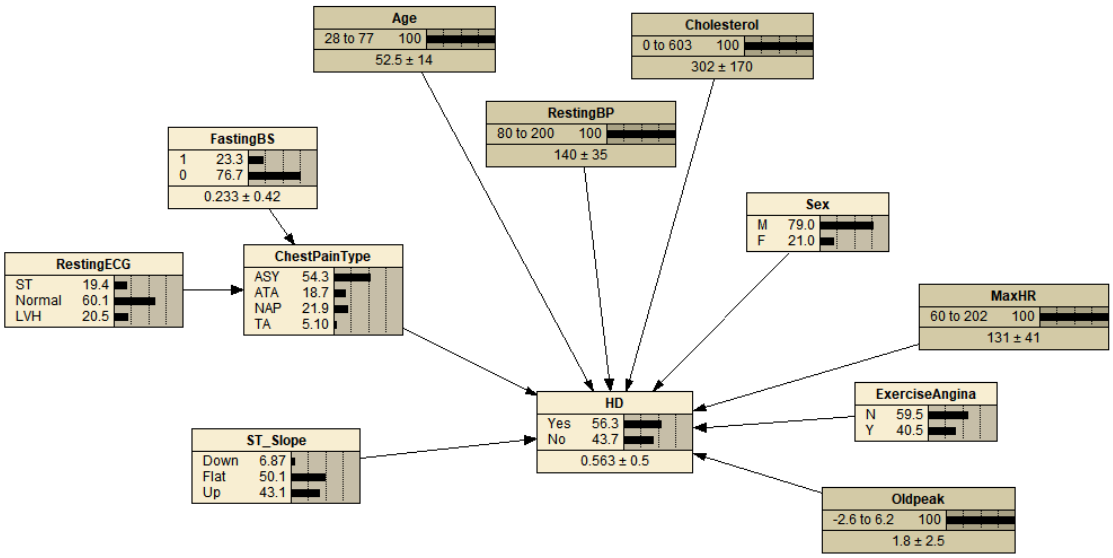
Slika 14: Izvješće Deepnet Summary Report

Na slici 14 prikazano je izvješće za Deepnet Summary Report kod kojeg se može vidjeti da je "*ST_Slope"* najvažnija varijabla (27.56%), nakon nje slijedi "*ChesPainType*" (21.22%), te "*ExcersiseAngina" (10.4%)*. Može se primijetiti na izvješću da su varijable različite po postocima od modela stabla odlučivanja, te da je kod ovog modela "*OldPeek*" ima puno manju važnost nego kod stabla odlučivanja.

Za kraj biti će još napravljen model Bayesove mreže u programu Netica.

6.4. Bayseove mreže

Prvi korak je dodavanje čvorova. Skup se sastoji od 12 atributa pa je potrebno dodati 12 čvorova i preimenovati ih u isti naziv kako se zovu Excel-u. Zatim kada smo to napravili, možemo ih povezati u jednu mrežu. To je prikazano na slici 24.

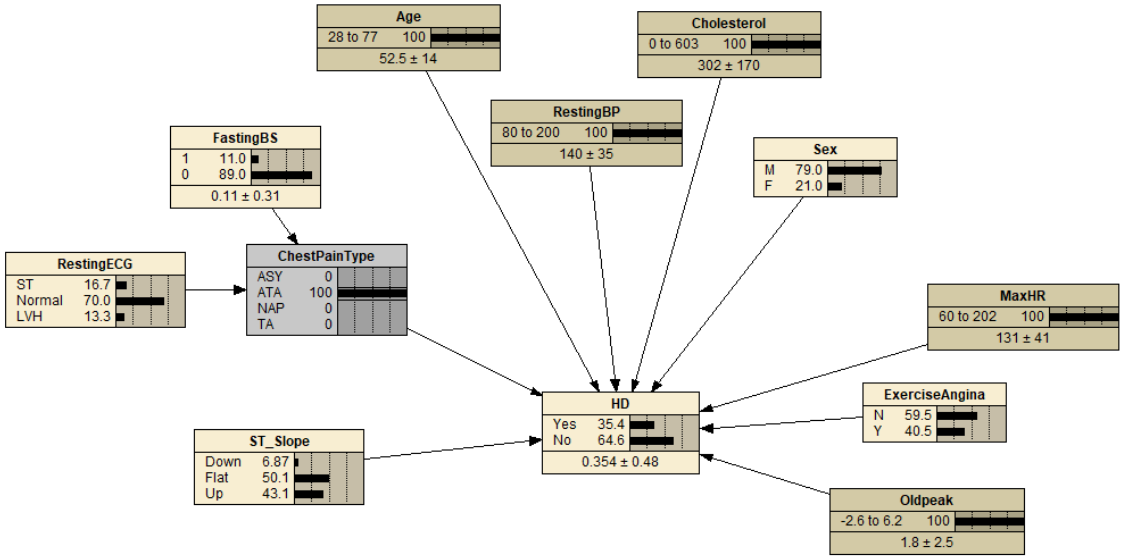


Slika 15: Bayesova mreža

Nakon toga potrebno je otići na ***CASES-> LEARN -> INCORP CASE FILE*** i odabrati *.tsv* datoteku (to je datoteka spremljena kao razmak odvojen tabulatorom) u kojoj se nalaze podaci. Kad smo to napravili, potrebno je otići na ***NETWORK -> COMPILE*** kako bi se kompajlirala mreža.

Sa slike 15 možemo zaključiti da *"FastingBS"* i *"RestingECG"* utječu na *"ChestPainType".* Za varijable koje imaju kontinuirane vrijednosti ne možemo previše uočiti koliko utječu na našu ciljnu varijablu. Neke zanimljive pojedinosti koje možemo uočiti su da je najviše ispitanika (54.3%) imalo *"ASY",* odnosno asimptotski tip boli u prsima (*"ChestPainType")* što je u konačnici povezano sa nekom vrstom srčane bolesti.

Na slici 16 prikazan je jedan slučaj ukoliko smo 100% sigurni da osoba ima "*ATA",* odnosno atipičnu anginu kao tip boli u prsima (*"ChestPainType"*), onda je 89 % vjerojatnost da ta osoba nema šećer u krvi natašte ("*FastingBS"*) , odnosno da je on manji od 120 mg/dl i da je vjerojatnost 70% da je ECG u mirovanju ("*RestingECG*") normalan, što dovodi u konačnici do rezultata da je vjerojatnost od 64.6% da osoba s tim simptomima nema srčanu bolest.



Slika 16: Bayesova mreža za specifičan slučaj

Da bi dobili točnost i preciznost Bayesovih mreža, iz skupa je odabrano 20 zapisa koji su naknadno uneseni u Neticu kako bi dobili matricu konfuzije. U alatu Netica, potrebno je kliknuti na ***CASES-> GET CASE*** i odabrati *.tsv* datoteku u kojoj se nalazi 20 zapisa. Nakon toga, dobit ćemo matricu konfuzije prikazanu u tablici 12:

		Predviđeno	
		Pozitivno (1)	**Negativno (0)**
Stvarno	**Pozitivno (1)**	15	0
	Negativno (0)	2	3

Tablica 12: Matrica konfuzije Bayesovih mreža

Točnost nam govori koliko je primjera ispravno klasificirano u odnosu na ukupan broj primjera.

Točnost = (Broj ispravno klasificiranih primjera) / (Ukupan broj primjera)

U ovom slučaju, broj ispravno klasificiranih primjera je 15 (15 pozitivnih ("1") primjera su točno klasificirana) + 3 (3 negativna ("0") primjera su točno klasificirana) = 18. Ukupan broj primjera je 15 + 0 + 2 + 3 = 20.

Točnost = 18 / 20 = 0.9 ili 90%

Preciznost nam govori koliko je pozitivnih ("1") primjera ispravno klasificirano u odnosu na ukupan broj primjera koji su predviđeni kao pozitivni ("1").

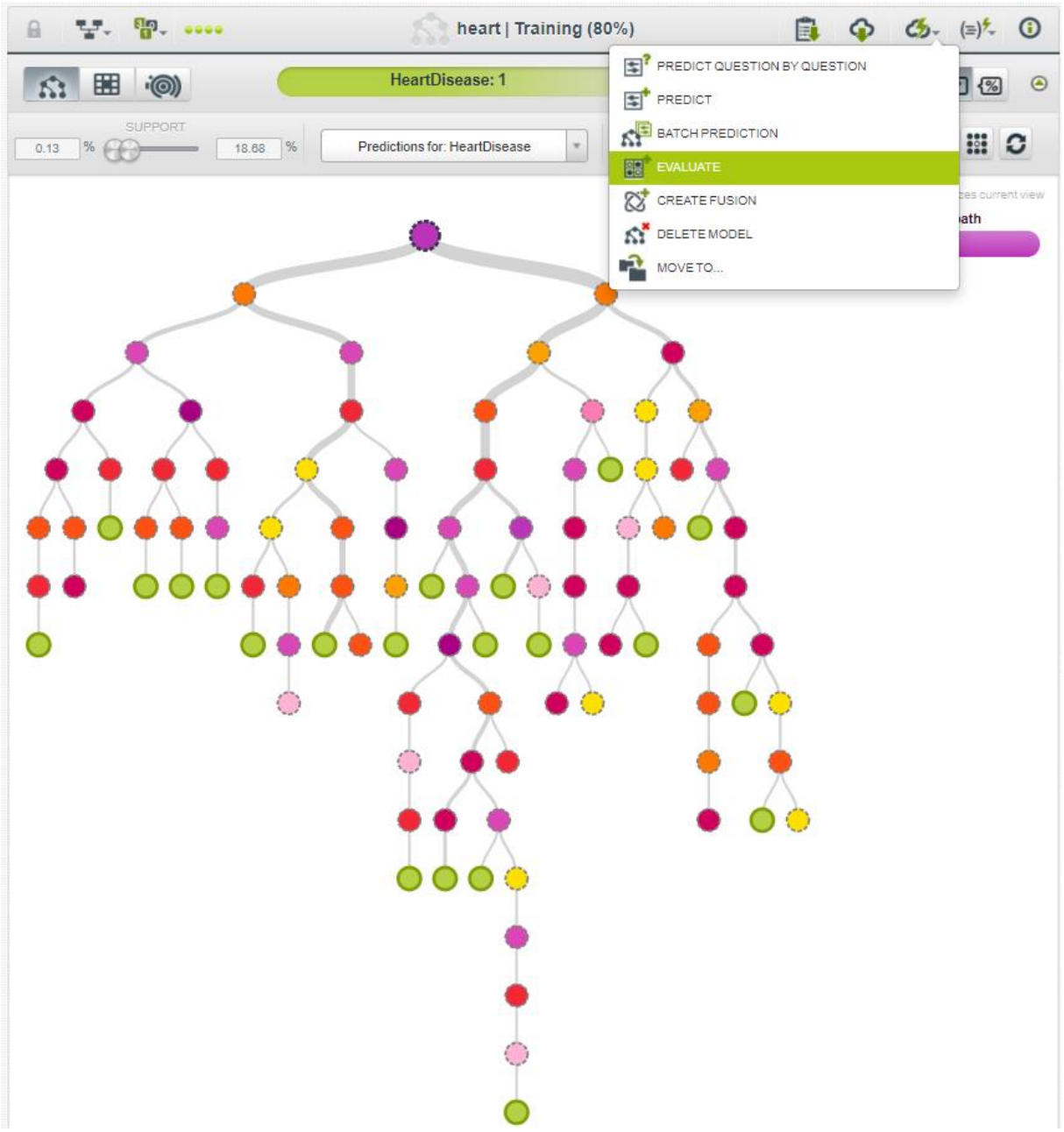
Preciznost = (Broj ispravno klasificiranih pozitivnih ("1") primjera) / (Ukupan broj primjera predviđenih kao pozitivnih ("1"))

U ovom slučaju, broj ispravno klasificiranih pozitivnih ("1") primjera je 15, a ukupan broj primjera koji su predviđeni kao pozitivni ("1") je 15 (predviđeno pozitivno "1")) + 2 (predviđeno negativno ("0")) = 17.

Preciznost = 15 / 17 ≈ 0.882 ili 88.2%

7. Evaluacija

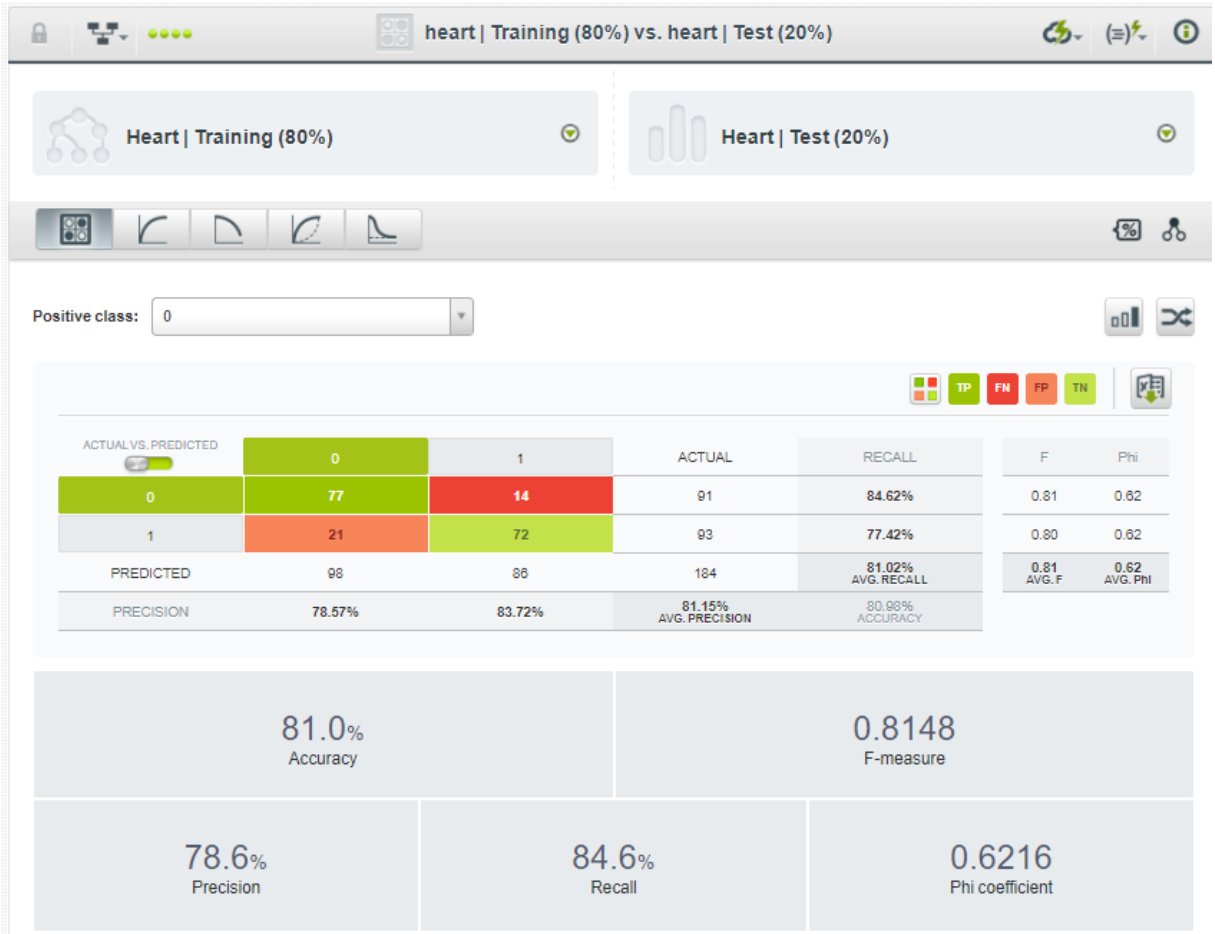
Kako bi se mogla napraviti evaluacija stabla odlučivanja i neuronske mreže prvo je potrebno pozicionirati se na onaj model koji se želi evaluirati. Nakon toga odabire se funkcionalnost "*EVALUATE"*, te se otvara novi prozor koji automatski odabire skup podataka za treniranje i testiranje. Evaluacije se pokreće pritiskom na tipku "*Evaluate"*.



Slika 17: Evaluacija stabla odlučivanja

7.1. Dobiveni rezultati evaluacije podataka

Nakon što se pritisne tipka za evaluaciju, dobiju se evaluirani podaci. Na slici 18 može se vidjeti da točnost stabla odlučivanja iznosi 81%, tj. da je stablo odlučivanja dobro klasificirao 81% instanci., dok je točnost Bayesovih mreža malo veća i iznosi 90%. Možemo vidjeti da je stablo odlučivanja predvidjelo (*eng. „predicted“)* skoro istu brojku pravih (*eng. „actual“)* broja osoba koje nemaju srčane bolest (98 vs 91). Prosječna preciznost stabla odlučivanja iznosi 78.6%, a Bayesovih mreža 88.2%. Preciznost se odnosi na mjeru koliko su rezultati mjerenja istog elementa međusobno konzistentni ili bliski.

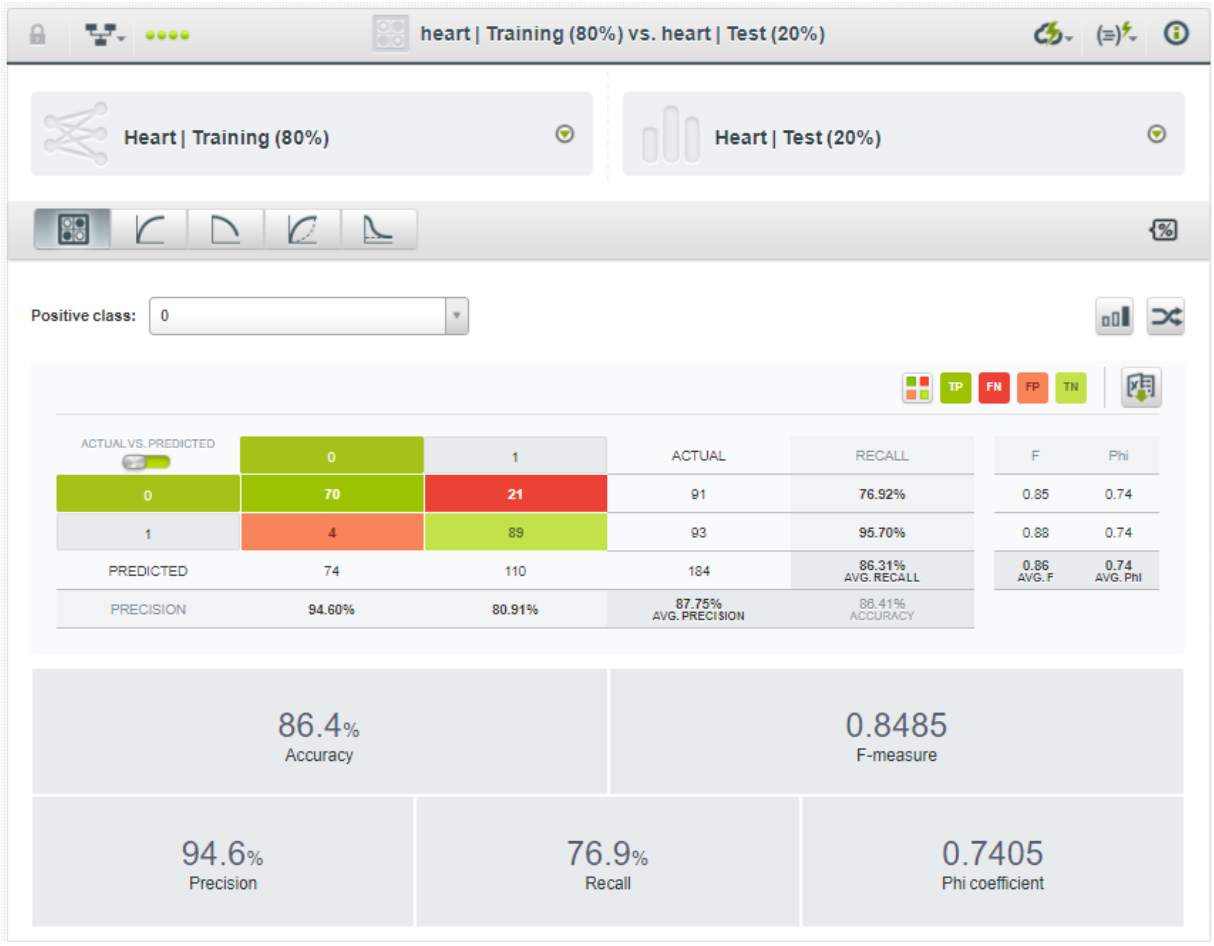


Slika 18: Matrica konfuzije stabla odlučivanja

Potrebno je napraviti i evaluaciju podataka za neuronsku mrežu, te se je prvo potrebno pozicionirati u podatke za treniranje i pritisnuti funkcionalnost “*EVALUATE*”. Kod evaluacije podataka za neuronsku mrežu dobiveno je da točnost modela neuronskih mreža iznosi 86.4%, što je malo više od točnosti stabla odlučivanja i malo

manje od Bayesovih mreža čija točnost iznosi 90%, dok preciznost iznosi 94.6% što je puno više nego kod stabla odlučivanja i više od preciznosti Bayesovih mreža.

Također svi podaci mogu se vidjeti na slici 19.



Slika 19: Matrica konfuzije neuronskih mreža

7.2. Analiza osjetljivosti Bayesove mreže

Na slici 20 prikazana je analiza osjetljivosti čvora "*ChestPainType"*. Vidljivo je da je da taj čvor ima utjecaj od 4.07% na konačni čvor "*HD"* odnosno postojanje srčane bolesti, 1.23% utjecaja na "*FastingBS*" tj. šećer u krvi natašte i 0.805% na čvor "*RestingECG*" tj. ECG u mirovanju.

Sensitivity of 'ChestPainType' to a finding at another node:

Node	Mutual Info	Percent	Variance of Beliefs
ChestPainType	1.63007	100	0.4167689
HD	0.06627	4.07	0.0130424
FastingBS	0.02004	1.23	0.0028737
RestingECG	0.01312	0.805	0.0009762
ExerciseAngina	0.00000	0	0.0000000
MaxHR	0.00000	0	0.0000000
Oldpeak	0.00000	0	0.0000000
Cholesterol	0.00000	0	0.0000000
RestingBP	0.00000	0	0.0000000
Sex	0.00000	0	0.0000000
ST_Slope	0.00000	0	0.0000000
Age	0.00000	0	0.0000000

Slika 20: Analiza osjetljivosti čvora "ChestPainType"

Iz navedenog, možemo zaključiti da čvor *„ChestPainType"* najviše utječe na čvor *„HD"*, što je i logično budući da je čvor *„HD"* dijete čvora *„ChestPainType"*.

7.3 Usporedba dobivenih rezultata i analiza osjetljivosti

U ovom radu korištene su tri različite metode strojnog učenja: stablo odlučivanja, neuronske mreže i Bayesove mreže. Primjenom navedenih metoda, dobiveni su različiti rezultati.

Dva najutjecajniji atributa na konačan ishod kod stabla odlučivanja i neuronskih mreža su *„ST_slope", „ChessPainType",* dok je na trećem mjestu *„OldPeak"* za stablo odlučivanja, a „ExerciseAngina" kod neuronskih mreža, dok su kod Bayesovih mreža to *„ST_slope", „ChessPainType"* i *„ExerciseAngina"* budući da oni ne sadrže kontinuirane vrijednosti.

U tablici 13 možemo vidjeti usporedbu rezultata točnosti i preciznosti korištenih modela:

	Točnost	Preciznost
Stablo odlučivanja	81%	78.6%
Neuronske mreže	86.4%	**94.6%**
Bayesove mreže	**90%**	88.2%

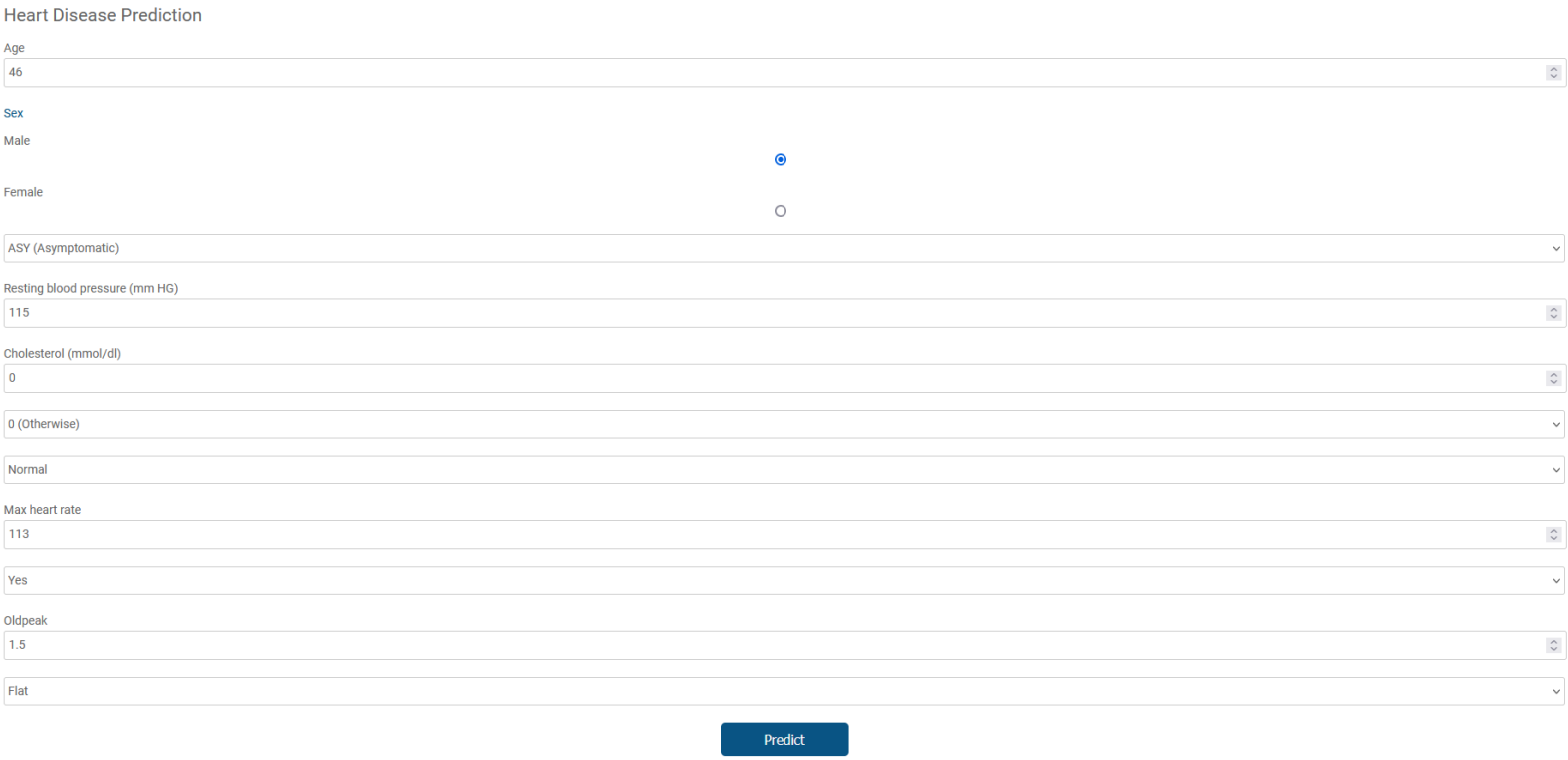
Tablica 13: Usporedba rezultata korištenih modela

Bayesove mreže pokazuju najvišu točnost od 90%, što znači da je ukupno pouzdanost njihovih predikcija relativno visoka. Neuronske mreže imaju najvišu preciznost od 94.6%, što ukazuje da je njihova sposobnost identificiranja stvarno pozitivnih primjera vrlo visoka. Stablo odlučivanja ima nižu točnost i preciznost u odnosu na Neuronske mreže i Bayesove mreže. To sugerira da su njegove predikcije manje pouzdane i precizne u usporedbi s drugim dvjema metodama, budući da se ovdje radi više o numeričkim podacima nego kategorijskim. Unatoč tome što Bayesove mreže imaju najvišu točnost, Neuronske mreže su pokazale najvišu preciznost. To ukazuje da su Neuronske mreže bolje u prepoznavanju stvarno pozitivnih primjera, dok su Bayesove mreže bolje u ukupnoj točnosti svojih predikcija.

8. Korištenje

U nastavku ovog poglavlja biti će prikazan jedan slučaj korištenja predviđanja srčane bolesti s obzirom da se zadnja faza odnosi na korištenje rudarenja podataka.

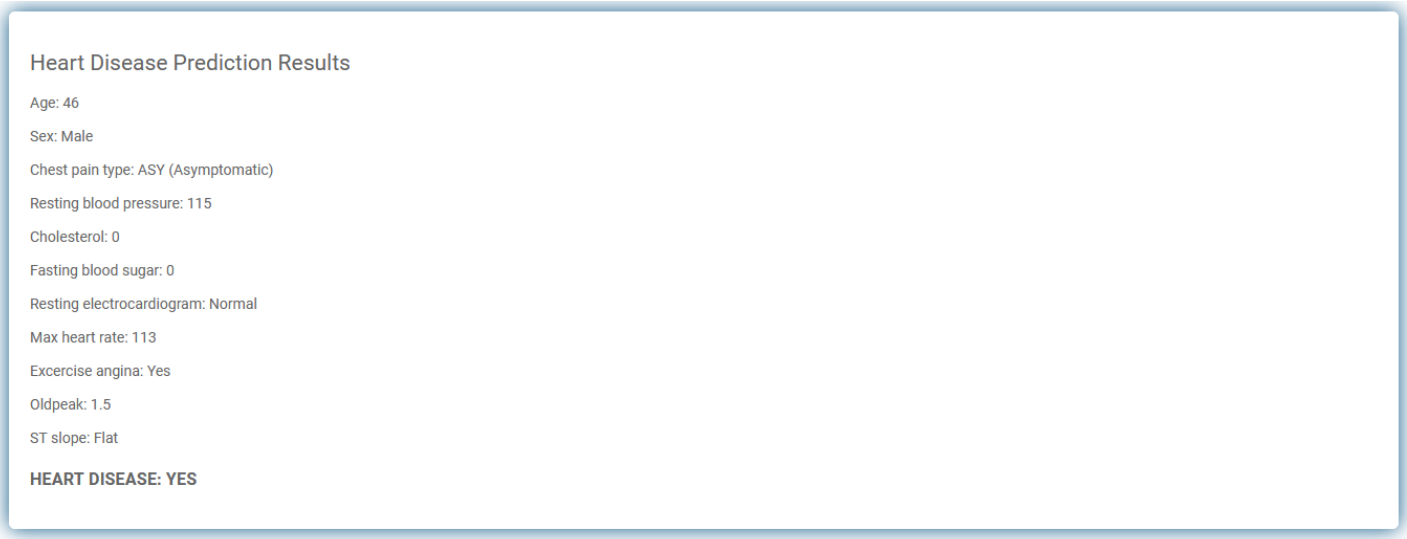
Na slici 21 prikazan je prozor koji sadrži potrebne element za određivanje predikcije srčane bolesti, a to su sve varijable navedene prethodno, osim zadnje koja određuje je li srčana bolest prisutna ili ne.



Slika 21: Primjer korištenja modela

Na slici 21 vidljivi su uneseni podaci, a oni se odnose na osobu kojoj je 46 godina, muškog spola, asimptomatske boli u prsima, krvnog tlaka u mirovanju 115, kolesterola 0, šećera u krvi natašte više od 120 mg/dl, maksimalnu brzinu otkucaja srca 113, s induciranom anginom vježbanjem, oldpeak vrijednosti 1.5 te ST nagiba *Flat*.

Nakon unesenih takvih podataka korisnika bi odvelo na stranicu rezultata koja prikazuje sve njegove unose i rezultat predviđanja. Takav prozor rezultata je prikazan na slici 22:



Slika 22: Predviđanje rezultata

Slika 22 prikazuje da muška osoba starosti 46 godina s asimptotskim tipom bolesti u prsima, maksimalnim otkucajima srca od 113 puta u minuti, anginom uzrokovanom vježbanjem i nagibom vrha vježbanja ST segmenta Flat, odnos vježbanja u mirovanju („Oldpeak“) od 1.5, ima srčanu bolest, iako osoba ima normalan elektrokardiogram u mirovanju, nema kolesterola ni šećera u krvi natašte i ima relativno dobar tlak u mirovanju od 115 mm/Hg.

9. Zaključak

Bolesti srca jedna su od najvećih briga današnjeg društva. Teško je ručno odrediti izglede za dobivanje srčanih bolesti na temelju čimbenika rizika. Međutim, tehnike strojnog učenja korisne su za predviđanje rezultata iz postojećih podataka.

Za ovaj projekt pronađeno je 5 sličnih istraživanja koji su korišteni kako bi vidjeli koje su metode klasifikacije i algoritmi korišteni u tim istraživanjima. Podaci koji su korišteni u prethodnim istraživanjima vrlo su slični podacima koji su korišteni za daljnje analiziranje. U prvom istraživanju korištene su neuronske mreže sa 100% preciznošću. Za drugo istraživanje autori su koristili više algoritama te kombinaciju *deep learning* i *machine learning* načina za usporedbu podataka te rezultata. Nakon što je bila provedena *deep learning* metoda, dobivena je pouzdanost od 94.2%. U trećem istraživanju korišten je način prikazivanja odnosa točnih i netočnih podataka. Stablo odlučivanja je za skup podataka za treniranje dalo preciznost od 100% metodom stabla odlučivanja, dok je Naive Bayes imao nekoliko pogrešaka. Kod testnog skupa podataka, obje metode su imale slične rezultate. U četvrtom istraživanju korišteno je stablo odlučivanja J48 koje je imalo najveću preciznost od 100%. Na kraju u petom istraživanju, najveću preciznost imala je metoda predviđanja sekvencijalne minimalne optimizacije sa 85.148%.

U ovom primjeru projekta, provedene su 3 tehnike modeliranja, a to su stablo odlučivanja, neuronske mreže i Bayesova mreža. Nakon što je skup podataka očišćen, uvezen je u BigML kako bi mogli provesti stablo odlučivanja i neuronske mreže. Zanimljiva činjenica iz analize je ta da su „*ST_Slope“* i „*ChestPainType“* dvije najutjecajnije varijable i kod stabla odlučivanja i kod neuronskih mreža. Nakon provedene evaluacije, rezultati pokazuju da Bayesove mreže imaju najvišu točnost od 90%, što ukazuje na visoku pouzdanost njihovih predikcija. S druge strane, neuronske mreže su pokazale najvišu preciznost od 94.6%, što znači da su vrlo sposobne u prepoznavanju stvarno pozitivnih primjera. Međutim, stablo odlučivanja ima nižu točnost i preciznost u usporedbi s neuronskim mrežama i Bayesovim mrežama, što je u jednu ruku logično, budući da skup sadrži više podataka numeričkog tipa. Kod Bayesove mreže možemo vidjeti da najviše prevladava asimptotski, odnosno „ASY“ tip boli u prsima („*ChestPainType“)* pa to znači da ako je osoba imala tu vrstu bolesti onda je velika vjerojatnost da je ta osoba imala neku srčanu bolest.

Neuronske mreže su se istaknule s najvišom preciznošću, što ukazuje na njihovu sposobnost učinkovitog prepoznavanja stvarno pozitivnih primjera. S druge strane, Bayesove mreže su postigle najvišu točnost u svojim predikcijama, što znači da su ukupno gledano pouzdanije u klasifikaciji podataka.

Na kraju možemo napomenuti da korištenjem istog skupa podataka, inteligentni sustavi daju različite rezultate, a to sve ovisi o korištenim algoritmima. Na manjim skupovima podataka dobit ćemo preciznije i kvalitetnije rezultate nego na većim.

10. Literatura

[1] Singh, P., Singh, S. and Pandi-Jain, G., 2021. *Effective heart disease prediction system using data mining techniques*. Dostupno 5.7.2023 na: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/

[2] Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S. and Singh, P., 2021. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. *Computational Intelligence and Neuroscience*, 2021, pp.1-11. Dostupno 5.7.2023 na: https://www.hindawi.com/journals/cin/2021/8387680/

[3] Rawat, S., 2019. *Heart Disease Prediction*. [Na internetu] Towards Data Science. Dostupno 6.7.2023 na: https://towardsdatascience.com/heart-disease-prediction-73468d630cfc

[4] Almustafa, K., 2020. Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinformatics* Dostupno 9.7.2023 na : https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03626-y

[5] Reddy, K., Elamvazuthi, I., Aziz, A., Paramasivam, S., Chua, H. and Pranavanand, S., 2021. Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. *Applied Sciences*, 11(18), p.8352. Dostupno 9.7.2023 na: https://www.mdpi.com/2076-3417/11/18/8352

[6] Peraić I. (2012). Primjena Bayesovih mreža u modeliranju učenika. Dostupno 13.7.2023 na: https://mapmf.pmfst.hr/~ani/radovi/diplomski/Peraic_Ivan_2012.pdf

[7] Zekić-Sušac, M., Frajman-Jakšić, A. i Drvenkar, N. (2009). Neuronske mreže i stabla odlučivanja za predviđanje uspješnosti studiranja. Ekonomski vjesnik, XXII (2), 314-327. Dostupno 10.7.2023. na: https://hrcak.srce.hr/clanak/73924

[8] Zekić-Sušac M. (bez dat.). Stabla odlučivanja. Dostupno 10.7.2023 s http://www.efos.unios.hr/poslovni-informacijski-sustavi/wp-content/uploads/sites/192/2017/10/P4_Stabla-odlucivanja-2017.pdf

[9] Kliček B. (2021). Znanje i strojno učenje. Inteligentni sustavi [Moodle]. Sveučilište u Zagrebu, Fakultet organizacije i informatike, Varaždin.

[10] Pintarić G. (2013). Umjetne neuronske mreže i mogućnosti njihove primjene u obrazovanju. Dostupno 12.7.2023 na: http://www.mathos.unios.hr/~mdjumic/uploads/diplomski/PIN12.pdf

[11] Oreški D. (2021). Algoritmi strojnog učenja. Inteligentni sustavi [Moodle]. Sveučilište u Zagrebu, Fakultet organizacije i informatike, Varaždin.

[12] Dumančić S. (2014). Neuronske mreže. Dostupno 12.7.2023 na: http://www.mathos.unios.hr/~mdjumic/uploads/diplomski/DUM05.pdf

[13] Ivančić M. (2017). Bayesove mreže. Dostupno 13.7.2023 na: https://repozitorij.foi.unizg.hr/islandora/object/foi:4605

11. Popis slika

Slika 1: Prikaz točnosti za sve klasifikatore [3] 15
Slika 2: Prikaz prebrojenih vrijednosti za vrijednosti varijable *„ChestPainType“* 20
Slika 3: Prikaz prebrojenih vrijednosti za vrijednosti varijable *„Sex“* 20
Slika 4: Prikaz boxplot dijagrama za vrijednosti varijable „MaxHR“ 21
Slika 5: Nevažeći zapis 22
Slika 6: Tablica atributa u alatu BigML 29
Slika 7: Skup podataka za treniranje 30
Slika 8: Izrada modela 31
Slika 9: Odabiranje statističkog obrezivanja 32
Slika 10: Prikaz stabla odlučivanja modela za trening 33
Slika 11: Prikaz izvješća Model Summary Report 34
Slika 12: Odabir funkcionalnosti DEEPNET 34
Slika 13: Prikaz neuronske mreže 35
Slika 14: Izvješće Deepnet Summary Report 36
Slika 15: Bayesova mreža 37
Slika 16: Bayesova mreža za specifičan slučaj 38
Slika 17: Evaluacija stabla odlučivanja 40
Slika 18: Matrica konfuzije stabla odlučivanja 41
Slika 19: Matrica konfuzije neuronskih mreža 42
Slika 20: Analiza osjetljivosti čvora “ChestPainType” 43
Slika 21: Primjer korištenja modela 45
Slika 22: Predviđanje rezultata 46

12. Popis tablica

Tablica 1: Popis korištenih atributa [1] 3
Tablica 2: Popis atributa [2]..... 6
Tablica 3: Popis atributa. [5]..... 9
Tablica 4: Matrica konfuzije [1]..... 13
Tablica 5: Matrica konfuzije [3]..... 14
Tablica 6: Matrica konfuzije za Naive Bayes (Podaci za treniranje) [3]..... 14
Tablica 7: Matrica konfuzije za Naive Bayes (Testni podaci) [3] 14
Tablica 8: Matrica konfuzije za stablo odlučivanja (Podaci za treniranje) [3].. 14
Tablica 9: Matrica konfuzije za stablo odlučivanja (Testni podaci) [3]..... 15
Tablica 10: Preciznosti za različite metode predviđanja [5]..... 16
Tablica 11: Popis atributa..... 19
Tablica 12: Matrica konfuzije Bayesovih mreža 38
Tablica 13: Usporedba rezultata korištenih modela 44