



UNIVERZITET U NOVOM SADU
FAKULTET TEHNIČKIH NAUKA



DOKTORSKA DISERTACIJA

INTELIGENTNI SOFTVERSKI SISTEM ZA DIJAGNOSTIKU METABOLIČKOG SINDROMA

Darko Ivanović

Mentori:

Prof. dr Aleksandar Kupusinac

Prof. dr Rade Doroslovački

2018



УНИВЕРЗИТЕТ У НОВОМ САДУ • ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА
21000 НОВИ САД, Трг Доситеја Обрадовића 6

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број, РБР:			
Идентификациони број, ИБР:			
Тип документације, ТД:	Монографска документација		
Тип записа, ТЗ:	Текстуални штампани материјал		
Врста рада, ВР:	Докторска дисертација		
Аутор, АУ:	Дарко Ивановић		
Ментор, МН:	Др Александар Купусинац, Др Раде Дорословачки		
Наслов рада, НР:	ИНТЕЛИГЕНТНИ СОФТВЕРСКИ СИСТЕМ ЗА ДИЈАГНОСТИКУ МЕТАБОЛИЧКОГ СИНДРОМА		
Језик публикације, ЈП:	Српски		
Језик извода, ЈИ:	Српски/Енглески		
Земља публиковања, ЗП:	Република Србија		
Уже географско подручје, УГП:	АП Војводина, Нови Сад		
Година, ГО:	2018.		
Издавач, ИЗ:	ауторски репринт		
Место и адреса, МА:	Факултет техничких наука, 21000 Нови Сад, Трг Доситеја Обрадовића 6		
Физички опис рада, ФО: (поглавља/страна/ цитата/табела/слика/графика/прилога)	8/108/85/12/35/0/4		
Научна област, НО:	Електротехничко и рачунарско инжењерство		
Научна дисциплина, НД:	Примењене рачунарске науке и информатика		
Предметна одредница/Кључне речи, ПО:	Машинско учење, Метаболички синдром, Дијагностика, Вештачке неуронске мреже, Стабла одлучивања, Линеарна регресија		
УДК			
Чува се, ЧУ:	Библиотека Факултета техничких наука у Новом Саду		
Важна напомена, ВН:	---		
Извод, ИЗ:	Докторска дисертација разматра проблем алгоритамске дијагностике метаболичког синдрома на основу лако мерљивих параметара: пол, старосна доб, индекс телесне масе, однос обима струка и висине, систолни и дијастолни крвни притисак. У истраживању су примењене и експериментално испитане три различите методе машинског учења: стабла одлуке, линеарна регресија и вештачке неуронске мреже. Показано је да вештачке неуронске мреже дају висок ниво предиктивних вредности довољан за примену у пракси. Коришћењем добијеног резултата дефинисан је и имплементиран интелигентни софтверски систем за дијагностику метаболичког синдрома.		
Датум прихватања теме, ДП:			
Датум одбране, ДО:			
Чланови комисије, КО:	Председник:	Др Драган Иветић	Потпис ментора
	Члан:	Др Душан Малбашки	
	Члан:	Др Едита Стокић	
	Члан:	Др Дубравко Ђулибрк	
	Члан, ментор:	Др Александар Купусинац	
	Члан, ментор:	Др Раде Дорословачки	



UNIVERSITY OF NOVI SAD • FACULTY OF TECHNICAL SCIENCES
21000 NOVI SAD, Trg Dositeja Obradovića 6

KEY WORDS DOCUMENTATION

Accession number, ANO :	
Identification number, INO :	
Document type, DT :	Monographic publication
Type of record, TR :	Textual printed document
Contents code, CC :	Ph.D. thesis
Author, AU :	Darko Ivanović, M.Sc.
Mentor, MN :	Aleksandar Kupusinac, Ph.D. and Rade Doroslovački, Ph.D.
Title, TI :	INTELLIGENT SOFTWARE SYSTEM FOR METABOLIC SYNDROME DIAGNOSTICS
Language of text, LT :	Serbian
Language of abstract, LA :	Serbian/English
Country of publication, CP :	Republic of Serbia
Locality of publication, LP :	AP Vojvodina, Novi Sad
Publication year, PY :	2018.
Publisher, PB :	Author's reprint
Publication place, PP :	Faculty of Technical Sciences, 21000 Novi Sad, Trg Dositeja Obradovića 6
Physical description, PD : (chapters/pages/ref./tables/pictures/graphs/appendixes)	8/108/85/12/35/0/4
Scientific field, SF :	Electrical and Computer Engineering
Scientific discipline, SD :	Applied computer science and informatics
Subject/Key words, S/KW :	Machine Learning, Metabolic Syndrome, Diagnostics, Artificial Neural Networks, Decision Trees, Linear Regression
UC	
Holding data, HD :	Library of the Faculty of Technical Sciences in Novi Sad
Note, N :	---
Abstract, AB :	<p>The doctoral dissertation examines the problem of algorithmic diagnostics of the metabolic syndrome based on easily measurable parameters: sex, age, body mass index, waist and height ratio, systolic and diastolic blood pressure. In the study, three different methods of machine learning were applied and experimentally examined: decision trees, linear regression and artificial neural networks. It has been shown that artificial neural networks give a high level of predictive value sufficient to be applied in practice. Using the obtained result, an intelligent software system for the diagnosis of metabolic syndrome has been defined and implemented.</p>
Accepted by the Scientific Board on, ASB :	
Defended on, DE :	
Defended Board, DB :	President: Dragan Ivetić, Ph.D.
	Member: Dušan Malbaški, Ph.D.
	Member: Edita Stokić, Ph.D.
	Member: Dubravko Čulibrk, Ph.D.
	Member, Mentor: Aleksandar Kupusinac, Ph.D.
	Member, Mentor: Rade Doroslovački, Ph.D.
	Mentor's sign

Sažetak

Ova doktorska disertacija razmatra problem algoritamske dijagnostike metaboličkog sindroma na osnovu lako merljivih parametara: pol (*GEN*), starosna dob (*AGE*), indeks telesne mase (*BMI*), odnos obima struka i visine (*WhtR*), sistolni krvni pritisak (*SBP*) i dijastolni krvni pritisak (*DBP*). Rano otkrivanje metaboličkog sindroma ima važnu ulogu u pravovremenoj prevenciji brojnih hroničnih bolesti, gde na prvom mestu se nalaze kardiometaboličke bolesti – aterosklerotske kardiovaskularne bolesti i dijabetes tipa 2.

Veza između posmatranih parametara i prisustva metaboličkog sindroma kod pacijenta je složena i analitičko rešenje nije poznato. Zbog toga su u istraživanju primenjene i eksperimentalno ispitane tri različite metode mašinskog učenja: stabla odluke (*DT*), linearna regresija (*LR*) i veštačke neuronske mreže (*ANN*). Na osnovu dobijenih rezultata i ostvarene maksimalne tačnosti predikcije metaboličkog sindroma, konstrukcija algoritma je urađena primenom veštačke neuronske mreže.

Postignut je visok nivo prediktivnih vrednosti *PPV* i *NPV* dovoljnih za primenu u praksi. Korišćenjem dobijenog rezultata definisan je i implementiran prototip *inteligentnog softverskog sistema za dijagnostiku metaboličkog sindroma*, koji omogućava jednostavnu, neinvazivnu i ekonomičnu dijagnostiku i samodijagnostiku sindroma.

Abstract

This PhD thesis examines the problem of algorithmic diagnostics of the metabolic syndrome based on easily measurable parameters: gender (*GEN*), age (*AGE*), body mass index (*BMI*), waist and height ratio (*WHtR*), systolic blood pressure (*SBP*) and diastolic blood pressure (*DBP*). Early detection of metabolic syndrome has an important role in the timely prevention of numerous chronic diseases, where cardiometabolic diseases are the first – atherosclerotic cardiovascular diseases and type 2 diabetes.

The relationship between the observed parameters and the presence of metabolic syndrome in the patient is complex and the analytical solution is not known. Therefore, three different machine learning methods have been applied and experimentally tested in the research: decision trees (*DT*), linear regression (*LR*) and artificial neural networks (*ANN*). Based on the results obtained and the maximum achieved accuracy of the prediction of metabolic syndrome, construction of the algorithm was done using an artificial neural network.

A high level of predictive values *PPV* and *NPV* have been achieved which can be applied in practice. Using this result, a prototype of *intelligent software system for the diagnosis of metabolic syndrome* is defined and implemented, which allows simple, non-invasive and low-cost diagnostics and self-diagnostics of the syndrome.

Slike

1	Metabolički sindrom	15
2	Poređenje srednjih vrednosti za <i>BMI</i> i <i>WHtR</i> kada je prisutan metabolički sindrom i kada nije	18
3	Poređenje srednjih vrednosti za <i>SBP</i> i <i>DBP</i> kada je prisutan metabolički sindrom i kada nije	18
4	Komponente učenja	24
5	Verovatnoća dodata u osnovni model učenja	29
6	Uopšteni model nadgledanog učenja	31
7	Linearna klasifikacija podataka - perceptron	36
8	Linearna separabilnost podataka	37
9	Linearna klasifikacija podataka - perceptron	38
10	Linearna regresija	41
11	Kriva učenja za linearnu regresija (E_{in} i E_{out} u odnosu na broj uzoraka N)	42
12	Stablo odluke	42
13	Šematski prikaz neurona	45
14	Sigmoidalna (a) i odskočna (b) funkcija	46
15	Linearna funkcija	46
16	Veštačka neuronska mreža	47
17	Grafik vrednosti <i>PPV</i> i <i>NPV</i> u odnosu na broj skrivenih neurona	64
18	Poređenje <i>ANN90</i> , <i>DT</i> i <i>LR</i> na osnovu <i>PPV</i> i <i>NPV</i>	65
19	<i>PPV</i> i <i>NPV</i> za validacioni skup u odnosu na broj epohe	66
20	<i>PPV</i> i <i>NPV</i> za trening, validacioni i test skup	67
21	<i>PPV</i> i <i>NPV</i> za obučenu mrežu nad test skupom	68
22	Konceptualni dijagram sistema za dijagnostiku metaboličkog sindroma	71
23	Koncept korisničkog interfejsa	73
24	Hijerarhija pogleda za <i>web</i> portal	74
25	<i>Web</i> stranica <i>LOGIN</i>	75
26	<i>Web</i> stranica <i>HOME</i>	75
27	<i>Web</i> stranica <i>PREDICTION</i>	76
28	<i>Web</i> stranica <i>TEST DATA</i>	77
29	<i>Web</i> stranica <i>ALL TESTS</i>	78
30	<i>Web</i> stranica <i>ANN</i>	79
31	Pogledi za mobilnu aplikaciju sa tranzicijama	81
32	Pogled <i>WELCOME</i> (a) i Pogled <i>ABOUT</i> (b)	82

33	Pogled <i>DATA ENTRY</i>	83
34	Pogledi <i>POSITIVE TEST</i> (a) i <i>NEGATIVE TEST</i> (b)	84
35	Softverska arhitektura sistema	86

Tabele

1	Prisustvo metaboličkog sindroma u formi DA/NE	17
2	Deskriptivna statistika u formi MEAN (MIN.–MAX.)	17
3	Primer grananja stabla odluke	44
4	Grafik vrednosti <i>PPV</i> i <i>NPV</i> za ANN90, DT i LR	65
5	Rezultati Wilcoxon-Mann-Whitney testa	65
6	<i>PPV</i> i <i>NPV</i> za trening, validacioni i test skup	67
7	Poređenje <i>PPV</i> i <i>NPV</i> po algoritmima za test skup	68
8	Format za razmenu podataka	87
9	<i>PPV</i> i <i>NPV</i> na skupu za obuku (<i>TR</i>), validaciju (<i>VL</i>) i testiranje (<i>TS</i>) za broj skrivenih neurona $HN = 1, 2, \dots, 100$	100
10	<i>PPV</i> i <i>NPV</i> na skupu za obuku (<i>TR</i>), validaciju (<i>VL</i>) i testiranje (<i>TS</i>) za ANN90, DT i LR.	102
11	Primer obuke veštačke neuronske mreže ANN90.	105
12	Primeri primene ANN90, DT i LR u predikciji metaboličkog sindroma.	108

Algoritmi

1	Algoritam za određivanje <i>MetS</i> koeficijenta	19
2	<i>Pocket</i> algoritam za obuku <i>perceptrona</i>	39
3	Algoritam najmanjih kvadrata	41

Programski kod

1	Izračunavanje prediktivnih vredosti	52
2	Normalizacija podataka	53
3	Implementacija IDF definicije metaboličkog sindroma	54
4	Implementacija stabla odluke	55
5	Implementacija linearne regresije	57

Oznake i skraćenice

ANN	Veštačka neuronska mreža (<i>Artificial Neural Network</i>)
DT	Stablo odluke (<i>Decision Tree</i>)
LR	Linearna regresija (<i>Linear regression</i>)
TR	Trening skup (<i>Training set</i>)
VL	Validacioni skup (<i>Validation set</i>)
TS	Test skup (<i>Test set</i>)
PPV	Pozitivna prediktivna vrednost
NPV	Negativna prediktivna vrednost
AGE	Starosna dob (<i>Age</i>)
GEN	Pol (<i>Gender</i>)
WC	Obim struka (<i>Waist circumference</i>)
WT	Telesna težina (<i>Body weight</i>)
HT	Telesna visina (<i>Body height</i>)
BMI	Indeks telesne mase (<i>Body Mass Index</i>)
WHtR	Odnos obima struka i visine (<i>Waist circumference to Height Ratio</i>)
SBP	Sistolni krvni pritisak (<i>Systolic Blood Pressure</i>)
DBP	Dijastolni krvni pritisak (<i>Diastolic Blood Pressure</i>)
MetS	Prisustvo metaboličkog sindroma (<i>Metabolic Syndrome</i>)

Sadržaj

1	Uvodna razmatranja	8
1.1	Opis problema	8
1.2	Cilj i potreba za istraživanjem	9
1.3	Metodologija i ostvareni rezultati	9
1.4	Struktura rada	10
2	Metabolički sindrom	12
2.1	Definicija metaboličkog sindroma	12
2.2	Faktori rizika	15
2.3	Skup podataka	16
2.4	Reprezentacija <i>IDF</i> definicije u vidu algoritma	18
3	Mašinsko učenje	20
3.1	Mašinsko učenje u medicinskoj dijagnostici	20
3.2	Problem mašinskog učenja	22
3.2.1	Model mašinskog učenja	22
3.2.2	Vrste mašinskog učenja	24
3.2.3	Izvodljivost mašinskog učenja	26
3.2.4	Greška	29
3.2.5	Šum	30
3.3	Generalizacija	31
3.3.1	Kompleksnost uzorka	32
3.3.2	Kazna za kompleksnost modela	33
3.3.3	Test skup	34
3.4	Overfitting	34
3.5	Linearni model	35
3.5.1	Linearna klasifikacija	35
3.5.2	Linearna regresija	39
3.6	Stabla odluke	42
3.7	Veštačke neuronske mreže	44
4	Metodologija - implementacija algoritama	50
4.1	Kriterijum performanse algoritma	50
4.2	Normalizacija podataka	52
4.3	Izračunavanje koeficijenta <i>MetS</i> na osnovu <i>IDF</i> definicije	53

4.4	Stablo odluke	54
4.5	Linearna regresija	55
4.6	Veštačka neuronska mreža	57
5	Ekperimenti i analiza rezultata	62
5.1	Određivanje optimalne arhitekture neuronske mreže	62
5.2	Poređenje performansi algoritama	64
5.3	Obuka mreže optimalne arhitekture	66
5.4	Testiranje obučene mreže	67
6	Funkcionalni opis i arhitektura softverskog sistema	70
6.1	Konceptualni opis sistema	70
6.2	Web portal za lekare	72
6.3	Mobilna aplikacija za pacijente	80
6.4	Softverska arhitektura sistema	85
6.5	Format za razmenu podataka	86
7	Zaključak	88
	Literatura	90
A	Rezultati eksperimenta br. 1	98
B	Rezultati eksperimenta br. 2	101
C	Rezultati eksperimenta br. 3	103
D	Rezultati eksperimenta br. 4	106

1 Uvodna razmatranja

U okviru ovog poglavlja u prva dva odeljka predstavljen je problem koji rešavamo i definisan cilj i potreba za istraživanjem. Nakon toga navedena je metodologija koja će se koristiti i ostvareni rezultati. Na kraju, ukratko se upoznajemo sa osnovnom strukturom rada i sadržajem pojedinih poglavlja.

1.1 Opis problema

Metabolički sindrom čine međusobno povezani metabolički poremećaji koji povećavaju rizik za razvoj brojnih hroničnih bolesti, a to su pre svega aterosklerotske kardiovaskularne bolesti i dijabetes tipa 2. Rana dijagnoza ovog stanja omogućava da se na vreme vrši uticaj na faktore rizika, kako ne bi došlo do razvoja ozbiljnijih oboljenja. Shodno tome, dijagnostika metaboličkog sindroma predstavlja problem od velikog značaja, jer su kardiovaskularne bolesti jedan od glavnih uzroka smrtnosti u svetu i vodeći uzrok smrtnosti u Republici Srbiji.

U rutinskoj kliničkoj praksi dijagnoza metaboličkog sindroma počinje uzimanjem anamneze (pol, starost). Zatim se merenjem osnovnih antropometrijskih parametara procenjuje stepen uhranjenosti i distribucija masnog tkiva, a nakon toga sledi merenje krvnog pritiska. Dijagnoza se postavlja nakon urađenih laboratorijskih analiza nivoa lipida, lipoproteina i glukoze u krvi naše.

Doktorska disertacija razmatra da li je moguće opisani dijagnostički postupak učiniti jednostavnijim, bržim i ekonomski povoljnijim i kako ga kao takvog učiniti dostupnim što je to moguće široj populaciji.

Centralna hipoteza istraživanja je da postoji matematička relacija između dijagnoze metaboličkog sindroma i lako dostupnih parametara: pol, starost, antropometrijske mere i krvni pritisak. Pronalaženje ove veze je prvi i osnovni problem koji rešavamo. Drugi problem koji se razmatra je primena te veze u praksi u vidu alata za algoritamsku dijagnostiku.

1.2 Cilj i potreba za istraživanjem

Prvi i osnovni cilj doktorske disertacije je konstruisanje algoritma za predikciju metaboličkog sindroma na osnovu sledećih parametara:

- pol (*GEN*)
- starosna dob (*AGE*)
- indeks telesne mase (*BMI*)
- odnos obima struka i visine (*WHtR*)
- sistolni krvni pritisak (*SBP*)
- dijastolni krvni pritisak (*DBP*)

Drugi cilj je definisanje i implementacija prototipa *inteligentnog softverskog sistema za dijagnostiku metaboličkog sindroma*, koji se oslanja na gore pomenuti algoritam. Sistem treba da je takav da omogući stručnim korisnicima jednostavno, brzo i ekonomično postavljanje dijagnoze. Sa druge strane, uzevši u obzir laku dostupnost parametara potrebnih za test, nameće nam se zahtev da se omogući *samodijagnostika* od strane pacijenta.

Napominjemo da pod pojmom dijagnostika ovde podrazumevamo *algoritamsku dijagnostiku* čiji rezultati moraju biti verifikovani od strane stručnjaka.

Povezanost metaboličkog sindroma sa razvojem hroničnih bolesti koje su ujedno i među glavnim uzrocima smrtnosti, nameće potrebu za efikasnom i ranom detekcijom stanja. U slučaju prisustva sindroma kod pacijenta, pravovremenom intervencijom u vidu izmene životnih navika moguće je korigovati faktore rizika koji mogu dovesti do razvoja ozbiljnijih oboljenja. Stoga, unapređenjem dijagnostičkog postupka i njegovom širokom primenom, možemo da napravimo pozitivan uticaj na sveukupno zdravlje populacije. Omogućavanjem samodijagnoze mogli bismo dalje da unapredimo ranu detekciju sindroma i da povoljno utičemo na visinu medicinskih troškova.

1.3 Metodologija i ostvareni rezultati

Veza između posmatranih parametara i prisustva metaboličkog sindroma kod pacijenta je složena i analitičko rešenje nije poznato. Osnovni preduslov za pronalaženje rešenja za osnovni problem je upotreba metodologije koja je efikasna u uslovima složenih

visoko nelinearnih problema.

Metode mašinskog učenja ispunjavaju dati uslov, pa se zbog toga u ovom istraživanju razmatraju stabla odluke, linearna regresija i veštačke neuronske mreže. Performanse algoritama su ispitane eksperimentalnim putem, poređenjem tačnosti predikcije na skupovima podataka za testiranje, a dobijeni rezultati su analizirani korišćenjem metoda statističke obrade podataka.

Doktorska disertacija prikazuje softverski sistem za algoritamsku dijagnozu metaboličkog sindroma pomoću metoda mašinskog učenja. Prototip sistema koji je prikazan baziran je na mobilnim i *cloud* tehnologijama i implementiran je u vidu mobilne aplikacije za pacijente i *web* portala za lekare. Obe aplikacije se oslanjaju na zajednički *backend* sa kojim komuniciraju putem definisanog *REST API*-ja. Kao razvojna platforma za *backend* koristi se *.NET Core* platforma i *ASP.NET Core* biblioteka. Implementacija *web* portala je urađena pomoću jezika *HTML* i *TypeScript* upotrebom biblioteke *Angular*. Za implementaciju algoritama korišćena je biblioteka za mašinsko učenje *Accord.NET*.

U okviru teze koristi se reprezentativan skup podataka od 2928 osoba (1494 žena i 1434 muškaraca). Podaci su prikupljeni u skladu sa Helsinškom deklaracijom, na Klinici za endokrinologiju, dijabetes i metaboličke poremećaje Kliničkog centra Vojvodine.

Primenom gore navedene metodologije prikazano je da je moguće konstruisati algoritam čije performanse su zadovoljavajuće za praktičnu primenu. Implementacijom prototipa sistema pokazano je da je ovakav sistem primenjiv svakodnevno, kako u kliničkoj praksi, tako i od strane pacijenata.

1.4 Struktura rada

Prvi deo disertacije posvećen je teorijskim razmatranjima i u njemu je predstavljen metabolički sindrom (poglavlje 2) i izložene teorijske osnove mašinskog učenja. (poglavlje 3). Opisi algoritama koje koristimo su dati u odeljcima 3.5.2, 3.6 i 3.7.

Praktični aspekti rada su predstavljeni u drugom delu, gde prikazujemo implementaciju algoritama (poglavlje 4), navodimo eksperimente koji su izvršeni i analiziramo dobijene rezultate (poglavlje 5). Opis inteligentnog softverskog sistema za dijagnostiku metaboličkog sindroma i njegova arhitektura dat je u poglavlju 6.

Zaključak i osvrt na doprinos doktorske disertacije nalazi se u okviru poslednjeg poglavlja 7.

2 Metabolički sindrom

Metabolički sindrom je skup međusobno povezanih metaboličkih poremećaja koji povećavaju rizik za razvoj brojnih hroničnih bolesti, gde se na prvom mestu nalaze kardiometaboličke bolesti – aterosklerotske kardiovaskularne bolesti i dijabetes tipa 2 [1, 2]. Pojam *metabolički sindrom* je nastao u drugoj polovini 20. veka, ali je danas veoma aktuelan, jer kardiometaboličke bolesti predstavljaju jedan od glavnih uzroka morbiditeta i mortaliteta u svetu [3]. Prema podacima za 2006. godinu, kardiovaskularne bolesti predstavljaju vodeći uzrok smrtnosti u Republici Srbiji (52,0% kod muškaraca i 62,8% kod žena) [4]. Poznato je da za osobe sa metaboličkim sindromom postoji veća verovatnoća dobijanja kardiovaskularne bolesti i/ili dijabetesa, na osnovu čega proizilazi da je metabolički sindrom značajan za ranu identifikaciju osoba sa povećanim rizikom [5]. Takođe je poznato da metabolički sindrom ima značajan uticaj i na razvoj drugih hroničnih bolesti [6, 7], što je od velike važnosti za zdravlje cele populacije [8].

2.1 Definicija metaboličkog sindroma

Razne grupe eksperata su definisale različite kriterijume za dijagnozu metaboličkog sindroma (*World Health Organization (WHO)*, *European Group for the Study of Insulin Resistance (EGIR)*, *National Cholesterol Education Program Adult Treatment Panel III (NCEP ATP III)*, *Chinese Diabetes Society (CDS)*, *International Diabetes Federation (IDF)* [9, 10, 11, 12, 13]), ali se svi slažu da u osnovi definicije metaboličkog sindroma su četiri komponente: gojaznost (naročito abdominalna (centralna) gojaznost), arterijska hipertenzija, dislipidemija i insulinska rezistencija [14].

WHO je 1998. godine prva utvrdila sledeće kriterijume za postavljanje dijagnoze metaboličkog sindroma [15]:

- Povišeni nivo glukoze našte, smanjena glukozna tolerancija, dijabetes tipa 2 ili insulinska rezistencija potvrđena *HOMA* indeksom (*homeostasis model insulin resistance assessment index*),
- i barem dva od ponuđenih kriterijuma:
 - Odnos obima struka i kukova (eng. *waist-to-hip ratio (WHR)*) iznad 0,85 *cm* kod žena, odnosno iznad 0,90 *cm* kod muškaraca i/ili indeks telesne mase (*body mass index (BMI)*) iznad 30 kg/m^2 ,

- Nivo triglicerida iznad $1,7 \text{ mmol/l}$ i/ili nivo *HDL* holesterola ispod $1,0 \text{ mmol/l}$ kod žena, odnosno ispod $0,9 \text{ mmol/l}$ kod muškaraca,
- Mikroalbuminurija (stopa urinarne ekskrecije albumina iznad $20 \mu\text{g/min}$) ili odnos albumina i kreatinina iznad 20 mg/g ,
- Krvni pritisak iznad $140/90 \text{ mmHg}$.

WHO kriterijumi podrazumevaju primenu *HOMA* indeksa za određivanje insulinske rezistencije, što u celini smanjuje njihovu upotrebnu vrednost u kliničkoj praksi, jer se radi o relativno skupoj metodi, koja nije dostupna u svakodnevnom radu i nepraktična je za velike epidemiološke studije. Međutim, treba istaći da *WHO* definicija predstavlja hronološki prvi pokušaj da se uvedu jasni kriterijumi za dijagnozu metaboličkog sindroma i predstavljaju polaznu osnovu za sva dalja istraživanja [14]. Iste godine, *EGIR* (*European Group for the Study of Insulin Resistance Guidelines*) je objavila definiciju metaboličkog sindroma koja se bazira na modifikovanim *WHO* kriterijumima i orijentisana je ka osobama koje nisu obolele od dijabetesa. Kao meru insulinske rezistencije, *EGIR* definicija predlaže određivanje insulina našte. *EGIR* kriterijumi za dijagnozu metaboličkog sindroma su sledeći [10]:

- Insulinska rezistencija ili hiperinsulinemija našte (nivo insulina našte treba da bude u gornjoj četvrtini nivoa insulina našte kod osoba bez dijabetesa),
- i barem dva od ponuđenih kriterijuma:
 - Obim struka iznad 80 cm kod žena i iznad 94 cm kod muškaraca,
 - Nivo triglicerida iznad $2,0 \text{ mmol/l}$ ili nivo *HDL* holesterola ispod $1,0 \text{ mmol/l}$ ili od ranije lečena dislipidemija,
 - Nivo glukoze našte iznad $6,1 \text{ mmol/l}$, ali kod osoba koje nisu obolele od dijabetesa,
 - Krvni pritisak iznad $140/90 \text{ mmHg}$ ili od ranije lečena hipertenzija.

Najvažnije komponente metaboličkog sindroma su centralna, odnosno abdominalna gojaznost, hipertrigliceridemija, snižena vrednost *HDL* holesterola, povišen krvni pritisak i hiperglikemija. Detaljnija istraživanja metaboličkog sindroma su bila otežana do 2004. godine zbog nedostatka koncenzusa o definiciji i referentnim vrednostima pojedinih komponenti [14]. Kriterijumi za dijagnozu metaboličkog sindroma, koji ne uključuju direktno merenje insulinske rezistencije, formulisani su 2001. godine u okviru naučnog programa *NCEP-ATPIII* (*National Cholesterol Education Program - Adult Treatment Panel III*) [11]. U celini, *NCEP-ATPIII* definicija je jednostavnija za primenu u kliničkoj praksi, a 2004. godine je modifikovana od strane *American Heart Association* (*AHA*), tako da je za dijagnozu metaboličkog sindroma neophodno prisustvo barem tri

od ponuđenih kriterijuma:

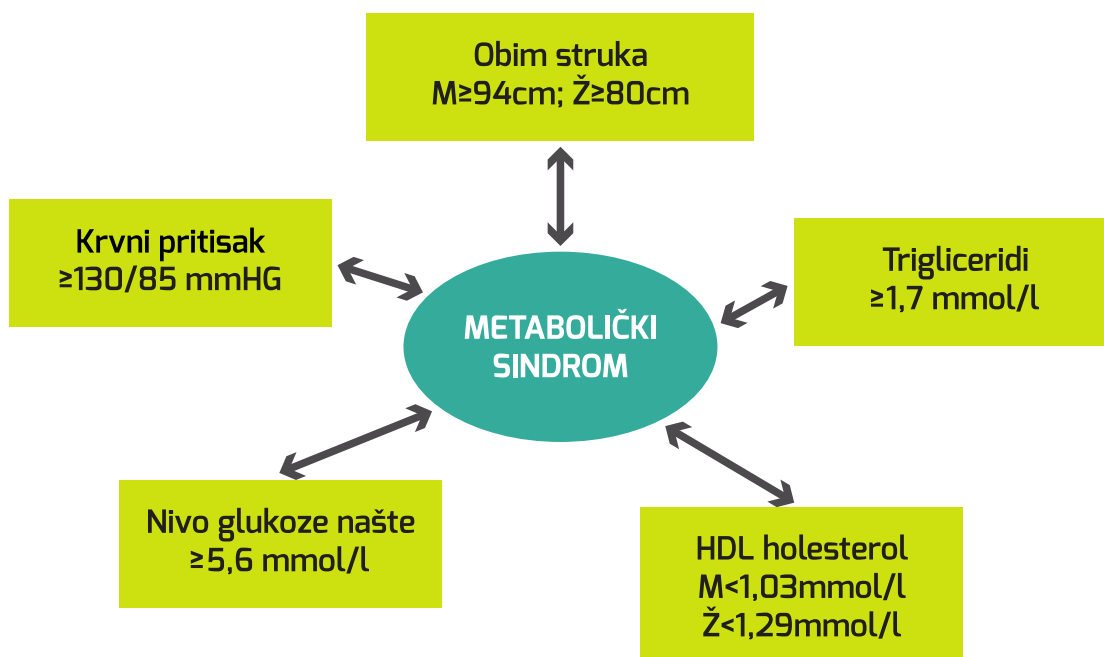
- Obim struka iznad 88 *cm* kod žena i iznad 102 *cm* kod muškaraca,
- Nivo triglicerida iznad 1,7 *mmol/l* ili od ranije lečen poremećaj,
- Nivo HDL holesterola ispod 1,29 *mmol/l* kod žena i ispod 1,03 *mmol/l* kod muškaraca ili od ranije lečen poremećaj,
- Krvni pritisak iznad 130/85 *mmHg* ili od ranije lečena hipertenzija,
- Nivo glukoze našte iznad 5,6 *mmol/l* ili prisustvo diabetesa tipa 2.

Definicija metaboličkog sindroma za populaciju Narodne Republike Kine je objavljena 2004. godine od strane *Chinese Diabetes Society (CDS)*, u kojoj je neophodno prisustvo barem tri od sledećih ponuđenih kriterijuma [13]:

- Indeks telesne mase (*BMI*) iznad 30 *kg/m²*,
- Nivo triglicerida iznad 1,7 *mmol/l*,
- Nivo HDL holesterola ispod 1,0 *mmol/l* kod žena i ispod 0,9 *mmol/l* kod muškaraca,
- Krvni pritisak iznad 140/90 *mmHg* ili od ranije lečena hipertenzija,
- Nivo glukoze našte iznad 6,1 *mmol/l* ili prisustvo diabetesa tipa 2.

International Diabetes Federation (IDF) je 2005. godine objavila definiciju metaboličkog sindroma u kojoj se u prvom planu nalazi abdominalna (centralna) gojaznost, izražena preko obima struka koji zavisi od etničke pripadnosti sa izuzetkom osoba kod kojih je indeks telesne mase (*BMI*) preko 30 *kg/m²* [12]. *IDF* kriterijumi za postavljanje dijagnoze metaboličkog sindroma su sledeći (slika 1):

- Obim struka iznad 80 *cm* za žene, odnosno 94 *cm* za muškarce poreklom iz Evrope, iznad 80 *cm* za žene, odnosno 90 *cm* za muškarce poreklom iz južne Azije, Kine i Japana itd.,
- i barem dva od ponuđenih kriterijuma:
 - Nivo triglicerida iznad 1,7 *mmol/l* ili od ranije lečen poremećaj,
 - Nivo HDL holesterola ispod 1,29 *mmol/l* kod žena i ispod 1,03 *mmol/l* kod muškaraca ili od ranije lečen poremećaj,
 - Nivo glukoze našte iznad 5,6 *mmol/l* ili prisustvo diabetesa tipa 2,
 - Krvni pritisak iznad 130/85 *mmHg* ili od ranije lečena hipertenzija.



Slika 1 – Metabolički sindrom

2.2 Faktori rizika

Među najznačajnije faktore rizika za razvoj kardiovaskularnih bolesti ubrajaju se gojaznost (naročito centralnog tipa), potom lipidni i lipoproteinski poremećaji, poremećaji glikoregulacije i insulinska rezistencija, kao i hipertenzija [16, 17]. Među pomenutim faktorima rizika gojaznost zauzima posebno mesto, jer uvećanje masne mase pokreće adipokinima posredovanu kaskadu metaboličkih, inflamatornih i hemostaznih poremećaja, koji dovode do ubrzanja procesa ateroskleroze [18]. Pri tome, najintenzivnije promene dešavaju se unutar visceralnog masnog depoa abdomena zbog čega je abdominalna ili centralna gojaznost označena kao posebno rizična [19] [20] [21].

Pol, starost i genetske predispozicije predstavljaju faktore rizika na koje je nemoguće uticati. S obzirom na to da masno tkivo kod muškaraca pokazuje tendenciju ka centralnoj ili abdominalnoj akumulaciji, muški pol već nosi potencijalno veći rizik od kardiovaskularnih bolesti, dok se kod žena tokom starenja rizik povećava usled centralizacije masnog tkiva. Poznato je da starenje dovodi do ubrzanja ateroskleroze i da se kardiometabolički rizik sa godinama povećava [22]. Indeks telesne mase (*BMI*) se računa kao odnos telesne mase i kvadrata telesne visine. *BMI* predstavlja pokazatelj stepena uhranjenosti i koristi se u definisanju gojaznosti: $25 \leq BMI < 30 \text{ kg/m}^2$

označava predgojaznost, a $BMI \geq 30 \text{ kg/m}^2$ označava gojaznost [23]. Abdominalna (centralna) gojaznost se definiše povećanjem obima struka ($WC \geq 80 \text{ cm}$ za žene i $WC \geq 94 \text{ cm}$ za muškarce). Obim struka dobro korelira sa veličinom visceralnog masnog tkiva abdomena, ali i sa drugim metaboličkim parametrima: nivoom lipida, lipoproteina i insulina, i značajan je prediktor komorbiditeta gojaznosti. Indeks dobijen kao proporcija obima struka i telesne visine (*waist-to-height ratio* ($WHtR$)) se pokazao kao precizniji indikator rizika, a vrednost $WHtR \geq 0,5$ se smatra rizičnom [24] [25]. Arterijska hipertenzija dovodi do poremećaja u mikro i makrocirkulaciji, povećane krutosti i smanjene elastičnosti arterija, povećanja periferne rezistencije, i drugih hemodinamskih poremećaja koje doprinose ubrzanju i komplikacijama ateroskleroze, te manifestaciji kardiovaskularnih bolesti.

Prema podacima iz 2006. godine, više od polovine odraslog stanovništva Republike Srbije ima prekomernu telesnu masu (54,5%), tačnije 36,2% stanovnika je predgojazno, a 18,3% gojazno [4]. Najviša prevalencija prekomerne telesne mase i gojaznosti beleži se na teritoriji Autonomne Pokrajine Vojvodine (58,5%), tačnije 35,5% odraslih je u kategoriji predgojaznih, dok je 23% gojaznih [26].

2.3 Skup podataka

Skup podataka koji se koristi u ovoj doktorskoj disertaciji je nastao u okviru istraživanja u kojem je učestvovalo 2928 volontera (1494 žena i 1434 muškaraca) na Klinici za endokrinologiju, dijabetes i metaboličke poremećaje Kliničkog centra Vojvodine. Podaci su prikupljeni u skladu sa Helsinškom deklaracijom. Svi parametri su izmereni u toku jutarnjih časova, nakon poludnevnog gladovanja.

Telesna masa je merena medicinskom decimalnom vagom, dok je telesna visina merena *Harpenden* antropometrom (*Holtain Ltd, Crowell, UK*) sa preciznošću merenja od 0,1 *cm*. Obim struka je meren mernom trakom sa preciznošću merenja od 0,1 *cm*, na sredini rastojanja najniže tačke rebarnog luka i najviše tačke na bedrenom grebenu karlične kosti. Vrednosti krvnog pritiska su merene sfingomanometrom (*Riva-Rocci*) nakon 10–15 minuta odmora. Za određivanje ukupnog holesterola i ukupnih triglicerida korišćen je enzimski postupak, vrednosti *HDL* holesterola određene su postupkom precipitacije sa natrijum-fosfo-volframatom, a vrednosti *LDL* holesterola izračunate su pomoću *Friedewald*-ove formule [27]. Vrednosti glikemije naše određene su *Dialab glucosa GOD-PAP* metodom.

Da bi dobijeni rezultati mogli da se uporede sa [28], u ovom istraživanju koristimo sličan skup podataka. Tabela 2 prikazuje karakteristike skupa podataka: starost (*AGE*), indeks telesne mase (*BMI*), obim struka (*WC*), obim struka kroz visina (*WHtR*), sistolni krvni pritisak (*SBP*), dijastolni krvni pritisak (*DBP*), trigliceridi (*TG*), HDL holesterol (*HDL*) i glikemija (*FPG*). Podaci su prikazani u odvojenim kolonama za ženski i muški pol (*GEN*), a zatim za sve zajedno. Na slikama 2 i 3 je prikazana zavisnost dijagnoze metaboličkog sindroma od parametara koje koristimo za predikciju. Tabela 1 prikazuje prisustvo metaboličkog sindroma u skupu podataka.

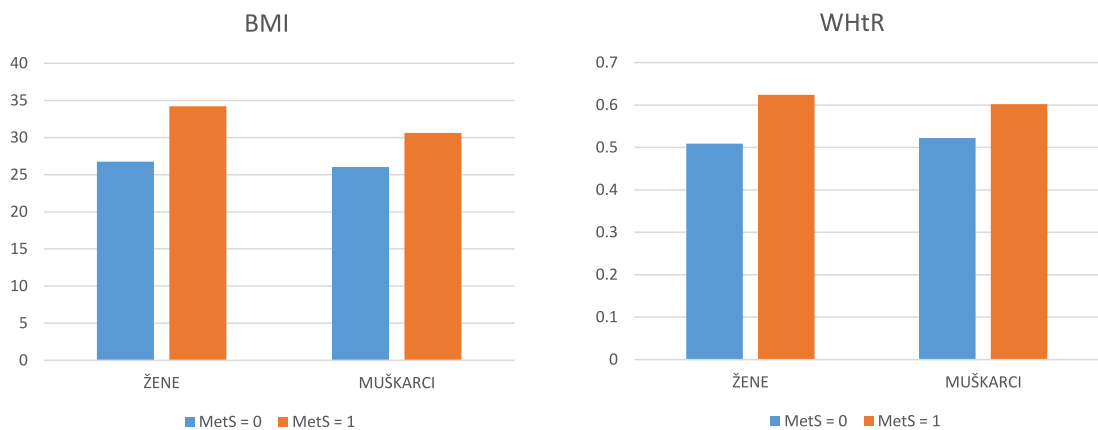
	Žene N=1494 (51,02%)	Muškarci N=1434 (48,98 %)	Svi N=2928 (100,0 %)
DA	756 (50,60%)	642 (44,77%)	1398 (47,75%)
NE	738 (49,40%)	792 (55,23%)	1530 (52,25%)

Tabela 1 – Prisustvo metaboličkog sindroma u formi DA/NE

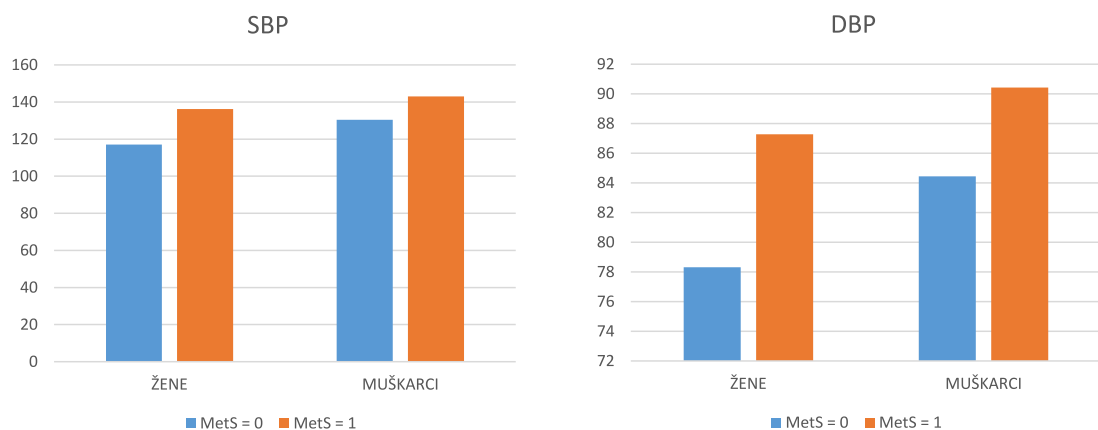
	Žene N=1494 (51.02%)	Muškarci N=1434 (48.98 %)	Svi N=2928 (100.0 %)
<i>AGE</i>	41,6 (18–76)	45,3 (18–67)	43,4 (18–76)
<i>BMI</i>	30,53 (16,64–48,10)	28,09 (19,42–45,95)	29,33 (16,64–48,10)
<i>WC</i>	92,55 (59,5–145,0)	97,4 (61,2–146,0)	94,93 (59,5–146,0)
<i>WHtR</i>	0,57 (0,36–0,88)	0,56 (0,36–0,84)	0,56 (0,36–0,88)
<i>SBP</i>	126,8 (90–185)	136,1 (100–200)	131,3 (90–200)
<i>DBP</i>	82,9 (50–130)	87,1 (60–120)	84,9 (50–130)
<i>TG</i>	1,79 (0,38–26,0)	2,15 (0,39–24,0)	1,97 (0,38–26,0)
<i>HDL</i>	1,20 (0,46–2,43)	1,13 (0,49–1,99)	1,16 (0,46–2,43)
<i>FPG</i>	4,98 (3,0–13,1)	5,25 (3,0–13,1)	5,11 (3,0–13,1)

Tabela 2 – Deskriptivna statistika u formi MEAN (MIN.–MAX.)

Slike 2 i 3 prikazuju srednje vrednosti parametara *BMI*, *WHtR*, *SBP* i *DBP* odvojeno za muškarce i za žene, i odvojeno za grupe *MetS* = 0 i *MetS* = 1. Sa ovih slika je očigledno da razmatrane srednje vrednosti su više u grupi u kojoj je prisutan metabolički sindrom (*MetS* = 1) nego u grupi u kojoj nema prisustva metaboličkog sindroma (*MetS* = 0). Ova zakonitost postoji kod svih razmatranih parametara i to kako kod žena, tako i kod muškaraca.



Slika 2 – Poređenje srednjih vrednosti za *BMI* i *WHtR* kada je prisutan metabolički sindrom i kada nije



Slika 3 – Poređenje srednjih vrednosti za *SBP* i *DBP* kada je prisutan metabolički sindrom i kada nije

2.4 Reprezentacija *IDF* definicije u vidu algoritma

U okviru ovog rada kao i u [28] *MetS* koeficijenti su određeni na osnovu *IDF* definicije metaboličkog sindroma koja je objašnjena u odeljku 2.1. *MetS* koeficijent može imati vrednost 1 ili 0, u zavisnosti od toga da li razmatrana osoba ima, ili nema metabolički sindrom. Na osnovu toga možemo formulisati algoritam za određivanje *MetS*

koeficijenta za svih 2928 osoba:

```
Rezultat: MetS koeficijenti
for  $i = 1$  to 2928 do
   $MetS(i) := 0$ 
  if ( $BMI(i) > 30$ ) or ( $GEN(i) = FEMALE$  and  $WC(i) \geq 80$  or
     $GEN(i) = MALE$  and  $WC(i) \geq 94$ ) then
     $p := 0$ 
    if  $SBP(i) \geq 130$  or  $DBP(i) \geq 85$  then
       $p := p + 1$ 
    end
    if ( $GEN(i) = FEMALE$  and  $HDL(i) < 1.29$ ) or
      ( $GEN(i) = MALE$  and  $HDL(i) < 1.03$ ) then
         $p := p + 1$ 
      end
    if  $TG(i) \geq 1.7$  then
       $p := p + 1$ 
    end
    if  $FPG(i) \geq 5.6$  then
       $p := p + 1$ 
    end
    if  $p \geq 2$  then
       $MetS(i) := 1$ 
    end
  end
end
```

Algoritam 1: Algoritam za određivanje *MetS* koeficijenta

3 Mašinsko učenje

Mašinsko učenje je sistematska studija algoritama i sistema koji unapređuju svoje znanje ili performanse sa iskustvom [29, 30]. Danas, mašinsko učenje ima široku primenu u različitim oblastima: prepoznavanje teksta i govora, finansije, autonomna vozila, dijagnostika u medicini, detekcija prevara, filtriranje informacija i sl. Multidisciplinarno po prirodi, mašinsko učenje se oslanja na koncepte i rezultate iz različitih oblasti, od kojih su neke: statistika, veštačka inteligencija, filozofija, teorija informacija, biologija, kognitivna nauka i teorija optimizacije [30].

Ovo poglavlje je posvećeno teorijskim osnovama mašinskog učenja. Prvi deo poglavlja govori o opštem problemu učenja i njegovim komponentama, dok se u drugom delu razmatraju pojedini modeli i algoritmi koji se koriste u ovoj doktorskoj disertaciji.

3.1 Mašinsko učenje u medicinskoj dijagnostici

Tipičan dijagnostički postupak teče na sledeći način: lekar uzima anamnezu i vrši pregled pacijenta; na osnovu utvrđenog statusa, lekar može pacijenta da uputi na dalje laboratorijske analize; kada su rezultati analiza gotovi, postavlja se dijagnoza i prepisuje terapija; ukoliko je potrebno, postupak može da se ponovi na kontrolnim pregledima, a dijagnoza može da se promeni.

U nekim slučajevima prva dijagnoza je i finalna, ali dešava se da konačnu dijagnozu nije moguće odrediti.

Dijagnoza u mnogome zavisi od lekara, njegovog iskustva, podataka kojima raspolaže, njegove intuicije i sklonosti. Anamneze, izveštaji o pregledima, rezultati laboratorijskih analiza i postavljene dijagnoze dobijeni u svakodnevnoj kliničkoj praksi sadrže skriveno dragoceno znanje koje nam stoji na raspolaganju. Metode mašinskog učenja omogućavaju da se to znanje otkrije i upotrebi za razvoj dijagnostičkih i drugih alata koji dijagnostički proces mogu da učine efikasnijim, objektivnijim i pouzdanijim [31].

Možemo reći da se poslednjih godina nalazimo u dobu upotrebe veštačke inteligencije u medicini, kao što je to najavljeno u [32]. Objavljuje se sve veći broj radova na temu

upotrebe mašinskog učenja za detekciju i dijagnozu bolesti [33]. Značaj primene ovih metoda u medicini je toliki da se postavlja pitanje da li će računari jednog dana moći da zamene lekare [34]. Metode mašinskog učenja su poznate već duže vreme, ali stalni porast procesorske snage učinio ih je široko primenljivim u svakodnevnoj praksi. Ovaj trend prate i softverski alati za razvoj takvih rešenja. Postoji veliki broj *open-source* i komercijalnih biblioteka za mašinsko učenje kao što su *Scikit-learn* (Python), *Accord.Net* (C#), *Apache Mahout* (Scala), *Spark MLlib* (Python, Scala, R), *TensorFlow* i *MATLAB*. Svi isporučiooci *cloud* usluga (*Microsoft Azure*, *Google Cloud*, *Amazon AWS*) danas u svojoj ponudi imaju instance za mašinsko učenje i okruženja za razvoj rešenja baziranih na ovoj metodologiji [35, 36, 37, 38].

Upotreba mašinskog učenja u dijagnostičke i druge svrhe razmatrana je za različite oblasti medicine. U onkologiji, metode mašinskog učenja su upotrebljavane za probleme kao što su: detekcija oboljenja [39], postavljanje diferencijalne dijagnoze [40], predviđanje metastaza [41], predikcija ishoda bolesti [42], analiza *DNA* tkiva [43] i analiza patohistoloških uzoraka [44]. Istraživanja u kardiologiji su usmerena na analizu ehokardiograma [45, 46] i elektrokardiograma [47]. Proučavaju se načini dijagnostike hitnih fatalnih stanja [48], kao i predikcija komplikacija [49]. Kada je reč o dijabetesu i stanjima koja mu prethode naponi su usmereni ka identifikaciji faktora rizika [50], pronalaženju skrining metoda za predijabetes [51], predikciji oboljenja kod pacijenta [52, 53] i sl.

Predikcija metaboličkog sindroma pomoću metoda mašinskog učenja razmatrana je u [54, 55, 56], gde su ostvareni rezultati slični ili nešto bolji nego u ovoj doktorskoj disertaciji, ali su dobijeni korišćenjem značajno šireg skupa ulaznih parametara koji uključuju i rezultate laboratorijskih ispitivanja, odnosno zahtevaju primenu invazivnih metoda. Slično je urađeno i u [57] gde su postignuti rezultati koji su na nivou nižem od onih u ovom istraživanju.

Na osnovu gore izloženog zaključujemo da je mašinsko učenje moćno sredstvo koje ima značajnu ulogu u medicini i medicinskoj dijagnostici. Iako rešenja i dalje nisu na dovoljno visokom nivou da bi mogla da zamene stručnu procenu, njihova primena u praksi doprinosi kvalitetnijoj, objektivnijoj i bržoj dijagnostičkoj proceduri.

3.2 Problem mašinskog učenja

Ukoliko sliku na kojoj je prikazano drvo pokažemo trogodišnjem detetu, i pitamo ga da li je na toj slici drvo, najverovatnije ćemo dobiti potvrđan odgovor. S druge strane, ukoliko pitamo odraslog čoveka da nam definiše šta je to drvo, vrlo verovatno ćemo dobiti nepotpun odgovor. Ova pojava je posledica toga što smo naučili šta je to drvo prostim posmatranjem, a ne učenjem formalnih definicija. Drugim rečima, možemo da kažemo da smo učili iz podataka [58].

Učenje iz podataka koristimo u slučajevima kada ne postoji analitičko rešenje problema, ali imamo na raspolaganju dovoljno podataka da možemo da konstruišemo empirijsko rešenje. U ovom odeljku se bavimo formalizacijom problema učenja i razmatramo glavne koncepte vezane za učenje.

3.2.1 Model mašinskog učenja

Pretpostavke koje neki problem mora da ispunjava da bi bio rešiv metodama mašinskog učenja su:

- obrazac (*pattern*) postoji u okvirima problema
- analitičko rešenje problema ne postoji
- postoji adekvatan skup podataka relevantnih za dati problem

Komponente učenja definišemo posmatrajući osnovni problem iz hipoteze ovog rada:

Pronaći funkciju koja na osnovu datih ulaznih parametara kao rezultat daje odgovor da li je kod pacijenta prisutan metabolički sindrom, ili nije.

Parametri funkcije su:

- *GEN* - pol, vrednost 1 označava muški pol, vrednost 2 označava ženski pol
- *AGE* - starosna dob
- *BMI* - indeks telesne mase
- *WtR* - odnos obima struka i visine
- *SBP* - sistolni krvni pritisak
- *DBP* - dijastolni krvni pritisak

Vrednosti tražene funkcije za zadate parametre su ili 0 (sindrom nije prisutan), ili 1 (sindrom je prisutan) i obeležavamo ga sa $MetS$.

Traženu funkciju koja predstavlja idealnu formulu za dijagnostiku metaboličkog sindroma obeležićemo sa f i nazivamo je *ciljna funkcija*. Za konkretan slučaj predikcije metaboličkog sindroma ova funkcija preslikava skup svih šestorki ulaznih parametara ($GEN, AGE, BMI, WHtR, SBP, DBP$) u skup $\{0, 1\}$. U opštem slučaju koji u nastavku razmatramo možemo da pišemo da je $f : \mathcal{X} \rightarrow \mathcal{Y}$, gde je \mathcal{X} *prostor ulaznih parametara* i predstavlja skup svih mogućih *ulaza* $\mathbf{x} = (x_1, x_2, \dots, x_N)$ gde $x_i \in \mathbb{R}$, a \mathcal{Y} je *prostor izlaznih parametara* i predstavlja skup svih mogućih *izlaza* $\mathbf{y} = (y_1, y_2, \dots, y_M)$ gde $y_i \in \mathbb{R}$. Jasno je da je za predikciju metaboličkog sindroma $N = 6$ i $x_1 = GEN$, $x_2 = AGE$, $x_3 = BMI$, $x_4 = WHtR$, $x_5 = SBP$, $x_6 = DBP$ gde $x_1 \in \{1, 2\}$, a $x_2, x_5, x_6 \in \mathbb{Z}^+$ i $x_3, x_4 \in \mathbb{R}^+$, dok je $M = 1$ i $y_1 = MetS$ gde $y_1 \in \{0, 1\}$.

Analitički oblik ciljne funkcije nije poznat. Mašinsko učenje podrazumeva da postoji značajan skup podataka \mathcal{D} , koji je restrikcija f i sastoji se od uređenih parova $(\mathbf{x}_i, \mathbf{y}_i), i \in \{1, 2, \dots, n\}$, gde je $\mathbf{y}_i = f(\mathbf{x}_i)$. Ovaj skup sadrži primere ulaza i izlaza ciljne funkcije, na osnovu kojih je moguće empirijski odrediti približan oblik rešenja koji ćemo obeležiti sa $g : \mathcal{X} \rightarrow \mathcal{Y}$ i nazivamo ga *finalna hipoteza*. Skup \mathcal{D} nazivamo *trening skup*.

U okviru ove teze koristimo trening skup dobijen u okviru istraživanja metaboličkog sindroma populacije u Srbiji, koje je izvršeno na Departmanu za endokrinologiju, dijabetes i metaboličke poremećaje, Kliničkog centra Vojvodine.

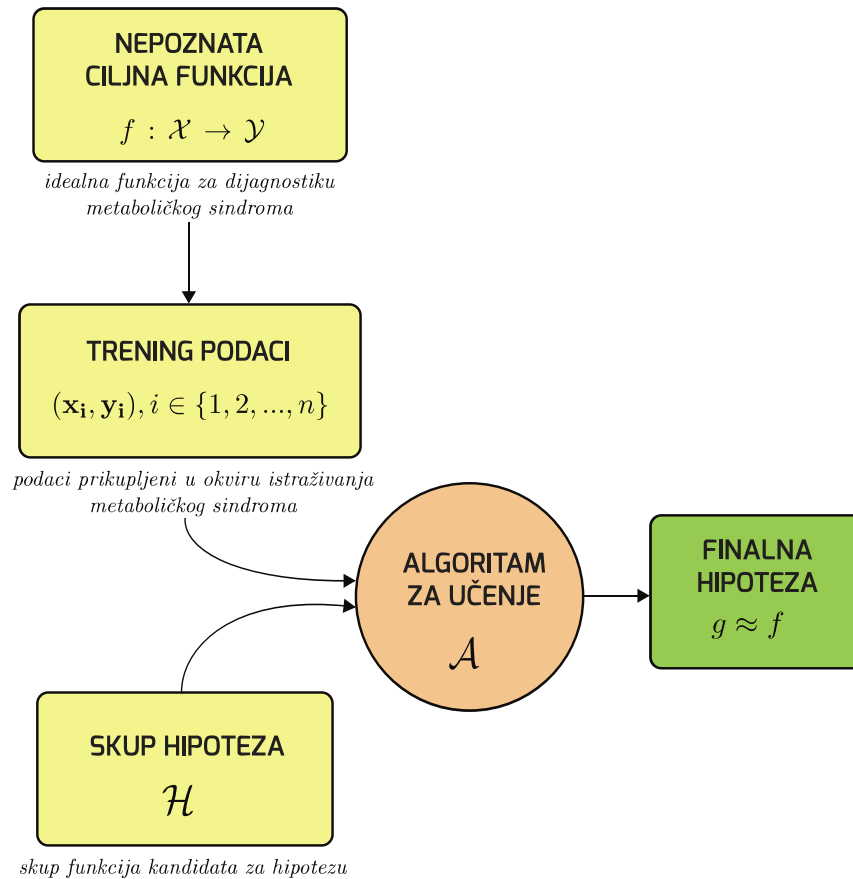
Postupak pronalaženja finalne hipoteze je definisan algoritmom \mathcal{A} koji na osnovu ulaznog skupa podataka \mathcal{D} odabira funkciju h koja aproksimira f . Algoritam odabira g iz skupa funkcija kandidata za rešenje \mathcal{H} koji nazivamo *skup hipoteza*.

Određivanje da li pacijent ima metabolički sindrom ili ne, vrši se na osnovu dobijene finalne hipoteze g , a ne na osnovu ciljne funkcije f koja ostaje nepoznata. Tačnost ovako dobijenih rezultata zavisi od toga koliko g verno aproksimira f , odnosno $g \approx f$.

Neformalno, skup hipoteza \mathcal{H} i algoritam \mathcal{A} se nazivaju *modelom učenja*.

Na slici 4 u nastavku, dat je grafički prikaz komponenti učenja i njihovih međusobnih odnosa:

- Potrebno je naučiti/aproksimirati ciljnu funkciju koja je nepoznata
- Poznati su trening podaci generisani pomoću ciljne funkcije
- Algoritam učenja koristi trening podatke da odabere finalnu hipotezu iz skupa hipoteza



Slika 4 – Komponente učenja

3.2.2 Vrste mašinskog učenja

Suština mašinskog učenja je otkrivanje osnovnog procesa na osnovu skupa podataka dobijenih posmatranjem. Zavisno od prirode procesa i dostupnih podataka, postoje različite vrste mašinskog učenja koje se koriste u različitim situacijama i imaju različite pretpostavke:

- Nadgledano učenje (*supervised learning*)

- Podržano učenje (*reinforcement learning*)
- Nenadgledano učenje (*unsupervised learning*)

Priroda podataka o metaboličkom sindromu koji su nam na raspolaganju u okviru ove teze omogućava nam da koristimo nadgledano učenje, koje je najviše istraživano i u širokoj upotrebi.

Nadgledano učenje podrazumeva da se trening skup sastoji od parova ulaza i ispravnih izlaza. Učenje je *nadgledano* u smislu da postoji *nadzornik* (*supervisor*) koji je pregledao sve ulaze i odredio tačne izlaze. U našem slučaju, podaci o pacijentima koje koristimo za dijagnostiku metaboličkog sindroma su upotpunjeni dodatnim podacima koji definitivno potvrđuju/negiraju dijagnozu.

Tipično, trening skup je unapred poznat, ali u slučajevima kada nije, možemo da koristimo dve varijacije nadgledanog učenja: *aktivno učenje* i *online učenje*.

Aktivno učenje podrazumeva da biramo skup ulaza za koje nam nadzornik daje ispravne izlaze i na taj način formiramo trening skup. Ovo nam otvara prostor da napravimo izbor ulaza takvih da maksimizujemo njihovu informacionu vrednost.

Online učenje nam omogućava da sistem unapređujemo sa pristizanjem novih podataka. Na primer, u slučaju dijagnostike metaboličkog sindroma, sistem bi mogao da dobije nove podatke zajedno sa rezultujućom dijagnozom odmah nakon sprovedene dijagnostičke procedure za novog pacijenta. Ovakvo učenje se najčešće koristi kada postoje velike količine podataka i ograničeni kapaciteti za njihovo čuvanje i obradu.

Podržano učenje se upotrebljava kada trening skup eksplicitno ne sadrži ispravne izlaze, nego sadrži nekakve izlaze i ocenu koliko je dati izlaz dobar. Za primer uzmimo dete koje treba da nauči da ne treba dirati vruću šolju čaja. Iskustvo takvog deteta se sastoji od izvesnog broja situacija u kojima je dete bilo ispred vruće šolje čaja, i donosilo je odluku da li da je dodirne ili ne. Svaki put kada bi je dodirnulo, doživelo bi visok nivo bola, svaki put kada ne bi dodirnulo šolju, bol bi izostao. Trening primeri nisu rekli detetu šta treba da uradi u pojedinim slučajevima, ali dete je dobilo ocenu postupka (bol ili odsustvo boli). Na taj način dete koristeći ocene pojedinih postupaka, na kraju nauči koji postupak je ispravan [58].

Poredeći ga sa nadgledanim učenjem čiji trening skup se sastoji od parova (ulaz, ispravan izlaz), trening skup u slučaju podržanog učenja se sastoji od trojki (ulaz, nekakav izlaz, ocena izlaza). Važno je reći da, pojedini trening primer ne govori koliko bi neki

drugi izlaz bio dobar za dati ulaz.

Podržano učenje je veoma dobro za učenje igara kao što je *backgammon*. Zamislimo neku poziciju na tabli, i da želimo da odredimo koji sledeći potez je najbolji. Kako oceniti koji je potez najbolji nije jednostavno, dobijanje dobrog trening skupa za nadgledano učenje je složeno. Zato ukoliko upotrebimo nadgledano učenje dovoljno je da odigramo neki potez i da naknadno izvestimo da li nas je taj potez doveo do pobeđe ili ne. Na taj način dobijamo jedan trening primer. Algoritam podržanog učenja treba da na osnovu ovako dobijenih trening primera dođe do najboljih poteza koje treba povlačiti [58].

Nenadgledano učenje podrazumeva da trening skup ne sadrži nikakve podatke o izlazima. Dati su samo ulazi x_1, x_2, \dots, x_N . Ova vrsta učenja najčešće se koristi za kategorizaciju. Njen cilj je da pronađe kategorije sličnih ulaza i na taj način nam pruži uvid u podatke na višem nivou. Broj kategorija i način kategorizacije nisu definisani. Možemo da kažemo da proces nenadgledanog učenja podrazumeva spontano pronalaženje obrazaca i strukture u podacima.

Uzmimo osobu koja ne govori španski. Pretpostavimo da ta osoba svakodnevno, duži period vremena ima priliku da sluša španski jezik bez znanja šta pojedine reči znače. Nakon određenog vremena ta osoba će početi da prepoznaje tipične zvuke i strukture jezika. Ovo je primer nenadgledanog učenja [58]. Ukoliko ta osoba odluči da pohađa časove španskog jezika, imaće mnogo bolju početnu poziciju nego da nije imala priliku da sluša španski jezik. Upravo tako i nenadgledano učenje može da prethodi nadgledanom učenju, ali može i da se koristi kao samostalna metoda.

3.2.3 Izvodljivost mašinskog učenja

Osnovna osobina ciljne funkcije f jeste da je nepoznata. Postavlja se pitanje da li je moguće aproksimirati je na osnovu ograničenog trening skupa \mathcal{D} koji nam je na raspolaganju.

Jasno je da u opštem slučaju, za dati trening skup \mathcal{D} postoji više od jedne hipoteza g koja verno replikuje f u svim tačkama iz \mathcal{D} . Da bismo mogli da kažemo da smo naučili, moramo da pokažemo da je moguće na osnovu trening skupa \mathcal{D} zaključiti nešto što je van tog skupa. U suprotnom možemo da kažemo samo da smo zapamtili (memorizovali).

Uzmimo proizvoljnu hipotezu $h \in \mathcal{H}$. Pretpostavimo da biramo slučajne uzorke $\mathbf{x} \in \mathcal{X}$ u skladu sa nekom raspodelom verovatnoće $P(\mathbf{x})$. Kako je f nepoznata, ne možemo da znamo da li je $h(\mathbf{x}) = f(\mathbf{x})$ ili da li je $h(\mathbf{x}) \neq f(\mathbf{x})$, ali možemo da tvrdimo da postoji neka verovatnoća μ da je $h(\mathbf{x}) \neq f(\mathbf{x})$. Onda je jasno, da je verovatnoća da je $h(\mathbf{x}) = f(\mathbf{x})$ jednaka $1 - \mu$.

Pretpostavimo da su uzorci $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ iz trening skupa \mathcal{D} odabrani nezavisno, a u skladu sa raspodelom P . Kako nam je za svako $\mathbf{x}_i \in \mathcal{D}$ poznato da li je $h(\mathbf{x}_i) \neq f(\mathbf{x}_i)$, možemo da odredimo verovatnoću ν kada je to slučaj.

Verovatnoća ν predstavlja meru koliko hipoteza h dobro aproksimira f u okviru trening skupa \mathcal{D} (ukoliko je ν blisko nuli onda h dobro aproksimira f). Sa druge strane, verovatnoća μ predstavlja meru koliko hipoteza h dobro aproksimira f van okvira trening skupa \mathcal{D} .

Hoeffding-ova nejednakost [59] uspostavlja vezu između ν i μ , i omogućava nam da na osnovu ν ocenimo μ :

$$\mathbb{P}[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \text{ za svako } \epsilon > 0$$

Ova nejednakost nam govori da kako veličina uzorka N raste, eksponencijalno se smanjuje verovatnoća da će ν odstupati od μ više nego odabrana tolerancija ϵ .

Koristeći ν da procenimo μ saznaćemo nešto o f , ali ne i to kakva je f . Ono što nam μ i ν govore jeste koliko grešku h pravi u aproksimaciji f . Tako da, ukoliko je ν blizu nule, možemo da očekujemo da će h dobro aproksimirati f na celom prostoru ulaznih parametara.

Na gore opisan način dobili smo mogućnost da ocenimo koliko je neka hipoteza dobra ili loša. Kao što je ranije navedeno učenje podrazumeva pronalaženje najbolje hipoteze iz skupa hipoteza. Da bi rezultat koji smo dobili iz *Hoeffding-ove nejednakosti* primenili na proces učenja uvodimo pojmove *in-sample* i *out-of-sample* greške.

Grešku u okviru *trening skupa* \mathcal{D} nazivamo *in-sample* greška E_{in} i definišemo je sledećim izrazom:

$$E_{in}(h) = \frac{1}{N} \sum_{i=1}^N \llbracket h(\mathbf{x}_i) \neq f(\mathbf{x}_i) \rrbracket$$

gde je $\llbracket \cdot \rrbracket$ jednak 1 ukoliko je izraz \cdot tačan, odnosno jednak 0 ako je izraz \cdot netačan. E_{in}

odgovara ν iz prethodne formule.

Slično, definišemo grešku za ulazne vrednosti izvan trening skupa \mathcal{D} koju nazivamo *out-of-sample greška* E_{out} i definišemo kao:

$$E_{out}(h) = \mathbb{P}[h(\mathbf{x}) \neq f(\mathbf{x})]$$

Gornja verovatnoća je u odnosu na prethodno definisanu raspodelu P nad skupom \mathcal{X} , pomoću koje su dobijene tačke uzoraka \mathbf{x}_i . Ova greška odgovara μ iz gore navedene *Hoeffding*-ove nejednakosti.

Sada, *Hoeffding*-ovu nejednakost možemo da zapišemo kao:

$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \text{ za svako } \epsilon > 0$$

Proces učenja za cilj ima da pronađe finalnu hipotezu g u odnosu na zadati trening skup \mathcal{D} , takvu da je verovatnoća $\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon]$ što manja. Da bismo dobili granicu za prethodnu verovatnoću koja ne zavisi od konačno odabranog g , pretpostavimo da smo u toku učenja ispitali hipoteze h_1, h_2, \dots, h_M . Kako je na kraju g jedna od hipoteza h_m , možemo da kažemo da događaj $|E_{in}(g) - E_{out}(g)| > \epsilon$ implicira događaj $|E_{in}(h_1) - E_{out}(h_1)| > \epsilon \vee |E_{in}(h_2) - E_{out}(h_2)| > \epsilon \vee \dots \vee |E_{in}(h_M) - E_{out}(h_M)| > \epsilon$. Na osnovu elementarnih pravila verovatnoće, možemo da napišemo sledeću nejednakost:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq \sum_{m=1}^M \mathbb{P}[|E_{in}(h_m) - E_{out}(h_m)| > \epsilon]$$

Ukoliko sada primenimo *Hoeffding*-ovu nejednakost na elemente sume sa desne strane, dobijamo:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

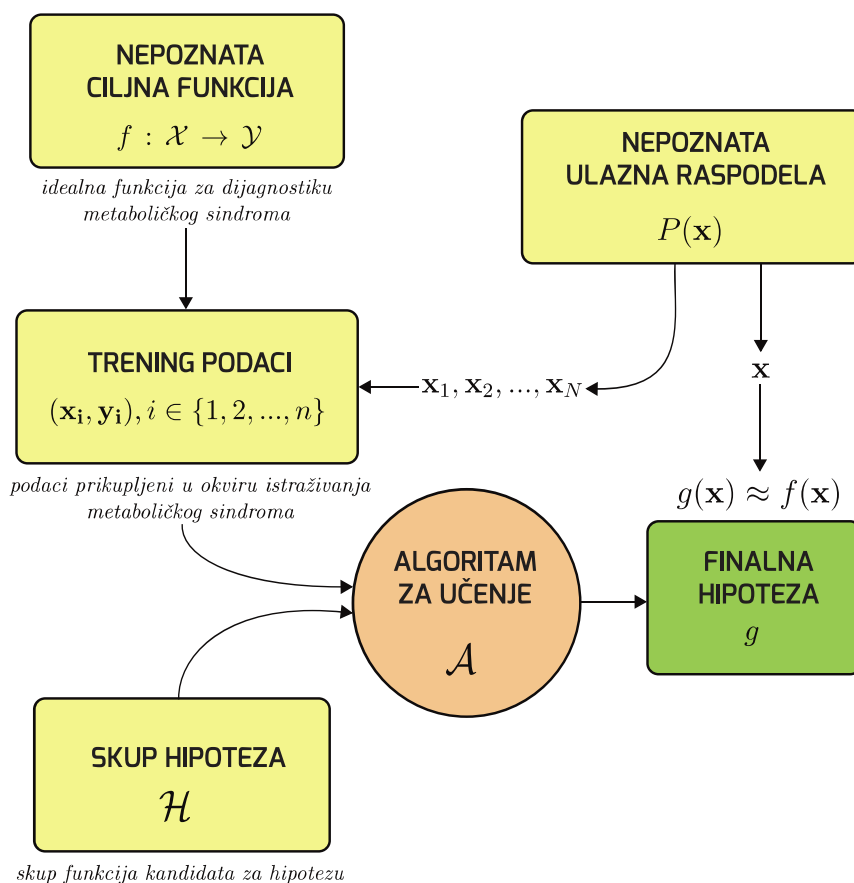
Ako kažemo da je učenje izvodljivo ukoliko finalna hipoteza g dobro aproksimira f , to znači da je $E_{out}(g) \approx 0$. Kako je $E_{out}(g)$ nepoznato, iz gornjeg izraza sledi da je $E_{out}(g) \approx E_{in}(g)$. Tako da nepoznat uslov $E_{out}(g) \approx 0$ možemo da zamenimo sa poznatim uslovom da je $E_{in}(g) \approx 0$.

Dakle, da bi učenje bilo izvodljivo imamo dva uslova:

1. Moguće je približiti E_{out} dovoljno blizu E_{in} , odnosno $E_{out}(g) \approx E_{in}(g)$.
2. Moguće je postići dovoljno malo $E_{in}(g)$.

Iz poslednje nejednakosti vidimo da E_{in} zavisi od $2Me^{-2\epsilon^2N}$. Bitan faktor je broj hipoteza koje ispitujemo M , zato što njegovim povećanjem povećavamo rizik da će g loše aproksimirati f . M se može smatrati merom kompleksnosti skupa hipoteza \mathcal{H} .

Poboljšani model učenja koji uključuje verovatnoću je dat na slici 5 u nastavku.



Slika 5 – Verovatnoća dodata u osnovni model učenja

3.2.4 Greška

Učenjem nismo u mogućnosti da idealno replikujemo ciljnu funkciju. Kao što je rečeno, finalna hipoteza je aproksimacija f . Kako bismo kvantifikovali koliko dobro g aproksimira f , potrebna nam je mera za grešku koja nam govori koliko smo daleko od ciljne funkcije.

Očigledno je da izbor mere za grešku direktno utiče na krajnji rezultat učenja. Idealno, mera za grešku se bira u skladu sa problemom koji se rešava.

Razmotrimo primer identifikacije osobe pomoću otiska prsta. Ukoliko se identifikacija vrši za potrebe kontrole pristupa, greška se dešava ukoliko se osobi kojoj je pristup dozvoljen taj pristup onemogućava (*false reject*) ili ukoliko se osobi kojoj pristup nije dozvoljen taj pristup omogućava (*false accept*). Po prirodi stvari, u slučaju kontrole pristupa, dozvoliti pristup osobi koja ga nema je značajno veća greška nego ne dozvoliti pristup osobi koja ga ima.

Sa druge strane, pogledajmo primer gde se identifikacija pomoću otiska prsta koristi za odobravanje popusta kupcima u supermarketu. Ukoliko u ovom slučaju kupcu koji ima odobren popust taj popust ne odobrimo, suočavamo se sa mogućnošću da tog kupca izgubimo u budućnosti. Dok, odobriti popust kupcu koji ga nema predstavlja značajno manju grešku, gledano iz ugla budućeg profita.

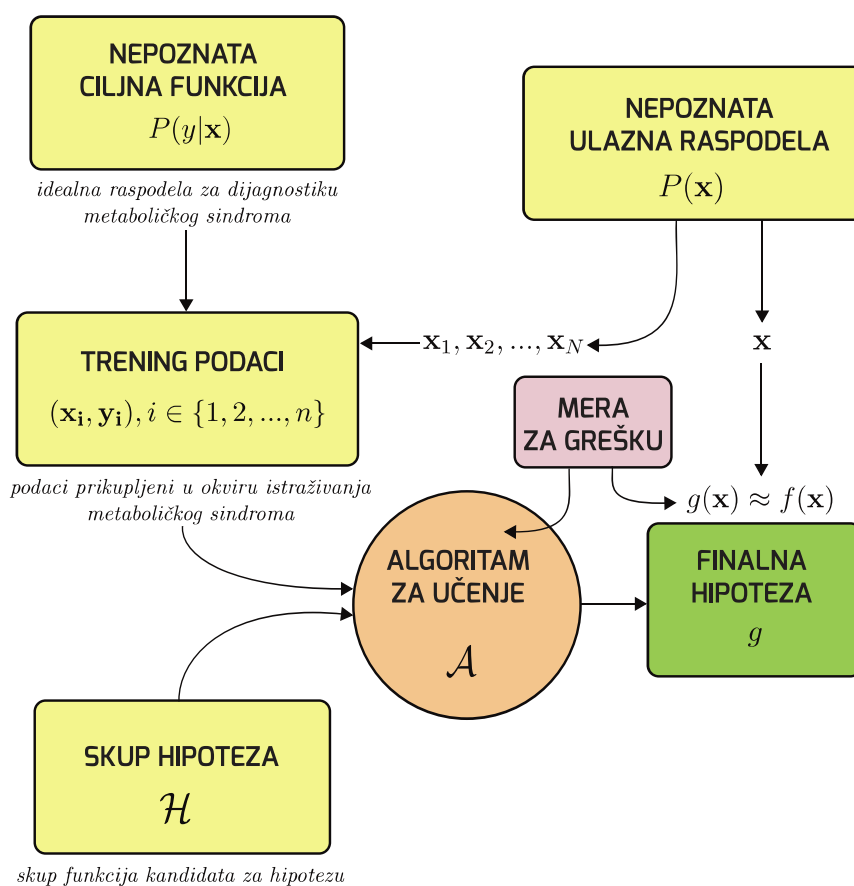
U praksi, idealan izbor mere za grešku nije uvek moguć. Ili zato što je tu meru teško definisati, ili zato što je izračunavanje te mere nedovoljno efikasno. Zbog toga se često koriste mere koje se biraju u skladu sa praktičnim ograničenjima koje nam problem koji rešavamo nameće.

3.2.5 Šum

U praksi, podaci iz kojih učimo često nisu rezultat determinističke ciljne funkcije. Umesto toga, podaci koje dobijamo često sadrže uzorke koji za iste vrednosti ulaznih parametara imaju različite izlaze. Dakle, formalno umesto prethodno definisane determinističke ciljne funkcije $y = f(\mathbf{x})$ sada imamo *ciljnu raspodelu* $P(y|\mathbf{x})$. Tako da su uzorci iz trening skupa (\mathbf{x}, y) sada generisani sa raspodelom $P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$.

Ovakva predstava cilja učenja, je analogna determinističkoj funkciji cilja $f(x)$ sa dodatim šumom. Ukoliko uzmemo da je šum nula ostaje nam samo $f(x)$ i imamo slučaj sa početka gde nam je cilj deterministička funkcija. Vidimo da uvođenjem ciljne raspodele ne gubimo na opštosti modela učenja.

Izvodljivost učenja važi i za model učenja sa ciljnom raspodelom, ali je samo učenje teže. Na slici u nastavku dat je uopšten model nadgledanog učenja.



Slika 6 – Uopšteni model nadgledanog učenja

3.3 Generalizacija

Pre ispita, profesor daje studentima skup zadataka kako bi studente što bolje pripremio za polaganje ispita. Iako to nisu zadaci koji će se pojaviti na ispitu, oni će pomoći studentima da nauče potrebno gradivo. Ti zadaci čine trening skup.

Ocena na ispitu je mera koliko je student dobro naučio gradivo. Ukoliko bi zadaci na ispitu bili iz skupa zadataka za pripremu, ocena ne bi bila mera koliko je student dobro naučio gradivo, nego koliko je dobro memorisao zadatke za pripremu.

Slično je i sa mašinskim učenjem, proveru kvaliteta učenja ne možemo da vršimo na uzorcima iz trening skupa, nego moramo da upotrebimo druge uzorke koje čine *test skup*.

Prethodno definisana *out-of-sample* greška E_{out} predstavlja meru koliko je učenje bilo uspešno na trening skupu \mathcal{D} , tj. koliko smo uspeli da *generalizujemo* na podatke koje nismo videli do sada. Sa druge strane, *in-sample* greška E_{in} nam govori koliko dobro smo naučili uzorke iz trening skupa.

Iz prethodno izvedene *Hoeffding*-ove nejednakosti može da se izvede sledeća nejednakost [58]:

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

Nejednakosti ovog tipa nazivamo *granica generalizacije*, u smislu da ograničava E_{out} u zavisnosti od E_{in} . Ono što je zanimljivo za gore dat izraz jeste njegova zavisnost od M za koje smo rekli da je mera veličine *skupa hipoteza* \mathcal{H} . Razlog za ovo je što su u praksi skoro svi zanimljivi skupovi hipoteza \mathcal{H} beskonačni. Kako bismo mogli da izučavamo *generalizaciju* na beskonačnim skupovima hipoteza, potrebno je M u gornjoj jednačini zameniti sa nečim konačnim.

Iako su tipični skupovi hipoteza beskonačni, oni dele osobinu da sadrže veliki broj veoma sličnih hipoteza. Kada uzmemo u obzir ova preklapanja sličnih hipoteza, koje možemo smatrati efektivno jednom hipotezom, možemo da dođemo do konačnog broja M , iako je skup hipoteza beskonačan [58].

Izraz koji potvrđuje izvodljivost učenja na modelima sa beskonačnim skupom hipoteza zove se *Vapnik-Chervonenkis (VC) granica generalizacije*.

VC granica generalizacije. Za svaku toleranciju $\delta > 0$ važi:

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

sa verovatnoćom $\geq 1 - \delta$.

Ovaj izraz, pored toga što potvrđuje izvodljivost učenja nad modelima sa beskonačnim skupom hipoteza, daje nam i nekoliko bitnih praktičnih rezultata, koje navodimo u nastavku.

3.3.1 Kompleksnost uzorka

Kompleksnost uzorka nam govori koliko uzoraka N je potrebno da bismo postigli postigli određeni stepen generalizacije. Step generalizacije je određen sa dva parametra

ϵ i δ . Tolerancija za grešku ϵ određuje dozvoljenu grešku generalizacije, a *parametar pouzdanosti* određuje koliko često tolerancija za grešku može da se prekorači. Koliko brzo raste N kako se ϵ i δ smanjuju, pokazuje nam koliko je potrebno podataka da bismo postigli dobru generalizaciju.

Iz izraza za VC granicu generalizacije se može dobiti sledeća nejednakost koja nam omogućava da procenimo veličinu potrebnog uzorka N :

$$N \geq \frac{8}{\epsilon^2} \ln \frac{4((2N)^{d_{VC}} + 1)}{\delta}$$

Iz gorenjeg izraza moguće je iterativnim postupkom odrediti N .

3.3.2 Kazna za kompleksnost modela

Kompleksnost uzorka iz prethodnog odeljka nam omogućava da na osnovu traženih parametara ϵ i δ procenimo potreban broj uzoraka N . U praksi, međutim, obično imamo dat fiksni skup podataka \mathcal{D} . U ovom slučaju razmatramo koje performanse možemo da očekujemo za dato N . Iz VC granice generalizacije imamo da sa verovatnoćom od bar $1 - \delta$ važi

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

Ovo možemo da zapišemo kao

$$E_{out}(g) \leq E_{in}(g) + \Omega(N, \mathcal{H}, \delta)$$

gde je

$$\Omega(N, \mathcal{H}, \delta) = \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}} \leq \sqrt{\frac{8}{N} \ln \frac{4((2N)^{d_{VC}} + 1)}{\delta}}$$

Gore naveden izraz može da se tumači kao *kazna za kompleksnost modela*. Vidimo da se E_{out} pogoršava kako se složenost \mathcal{H} povećava. Isto se dešava ukoliko insistiramo na većoj pouzdanosti (manje δ), a poboljšanje ove vrednosti se dobija povećavanjem broja uzoraka u trening skupu N .

Iako se $\Omega(N, \mathcal{H}, \delta)$ povećava sa složenošću \mathcal{H} , povećava se verovatnoća da će E_{in} da se smanjuje zato što u složenijem \mathcal{H} imamo više izbora pa je lakše pronaći hipotezu koja bolje aproksimira *funkciju cilja*. Optimalan model minimizuje kombinaciju ova dva parametra.

3.3.3 Test skup

Kao što smo videli, granica generalizacije nam daje prilično opuštenu procenu E_{out} u zavisnosti od E_{in} . Ovakva procena je prihvatljiva kao smernica za proces učenja, ali kao stvarna procena za E_{out} je neupotrebljiva.

Da bismo dali tačniju procenu kako će se finalna hipoteza ponašati van skupa \mathcal{D} , uvodimo *test skup* koji se sastoji od uzoraka koji se nisu koristili u procesu učenja, a koji se nakon obuke evaluiraju u odnosu na dobijeno g . Grešku koju dobijemo u procesu evaluacije obeležićemo sa E_{test} .

Vrednost E_{test} možemo da uzmemo za procenu E_{out} . Postavlja se pitanje koliko dobro smo generalizovali u zavisnosti od E_{test} .

Kako razmatramo samo finalnu hipotezu g , kao granicu generalizacije možemo da uzmemo jednostavni oblik *Hoeffding*-ove nejednakosti. Ova granica je mnogo jača nego što je to slučaj sa *VC* granicom generalizacije, i možemo da dobijemo značajno kvalitetniju ocenu koliko dobro aproksimiramo ciljnu funkciju van \mathcal{D} .

Negativan aspekt test skupa je što moramo da ga izdvojimo iz skupa \mathcal{D} , čime umanjujemo kvalitet učenja. Zbog toga je bitno da se uzme optimalan uzorak, koji neće značajno uticati na obuku, a biće dovoljan za kvalitetnu procenu generalizacije. U praksi standardno se uzima 80% podataka za učenje, a 20% posto podataka za testiranje.

3.4 Overfitting

Overfitting je fenomen gde dobro aproksimiranje u okviru trening skupa \mathcal{D} ne podrazumeva da ćemo dobiti prihvatljivu *out-of-sample* grešku E_{out} . Drugim rečima, algoritam za učenje je odabrao finalnu hipotezu g sa manjom greškom E_{in} , ali kao rezultat smo dobili nezadovoljavajuće visoku grešku E_{out} .

Prvi značajan razlog koji dovodi do *overfitting*-a je postojanje visokog nivoa šuma u ciljnoj funkciji. Ukoliko na takvu funkciju primenimo model sa širokim skupom hipoteza \mathcal{H} , algoritam za učenje će iskoristiti te dodatne stepene slobode da se prilagodi šumu (u ulaznim podacima nije moguće razlikovati šum od ispravnog uzorka). Prilagođavanjem šumu, proces obuke se udaljava od ciljne funkcije i na kraju rezultuje nezadovoljavajućom finalnom hipotezom g . Što je model složeniji, više je podložan

uticaju šuma.

Drugi razlog za *overfitting* je visoka kompleksnost ciljne funkcije. Pretpostavimo da nam je poznata h_{best} iz skupa hipoteza, koja najbolje aproksimira datu ciljnu funkciju. Sve tačke iz trening skupa koje nisu obuhvaćene sa h_{best} se mogu smatrati šumom. Ovu vrstu šuma nazivamo *deterministički šum*. Kako proces učenja ne zna koja je h_{best} , deterministički šum ima sličan uticaj na rezultujuću finalnu hipotezu kao i *stohastički šum* iz prethodnog slučaja.

Osnovni alati koji su nam na raspolaganju za rešavanje problema *overfitting*-a su *regularizacija* i *validacija*. *Regularizacija* izbegava *overfitting* tako što usmerava algoritam da za aproksimaciju koristi što jednostavnije hipoteze. *Validacija* uvodi pojam *validacionog skupa* koji ima iste osobine kao i test skup, ali se greška dobijena evaluacijom posmatrane hipoteze koristi za upravljanje algoritmom obučavanja [58].

3.5 Linearni model

U okviru ovog odeljka predstavimo *linearne modele učenja*. Razmotrićemo dva osnovna modela u upotrebi: *linearnu klasifikaciju* i *linearnu regresiju*. Linearna klasifikacija se koristi za aproksimaciju funkcija kod kojih se prostor izlaznih parametara sastoji od samo dva elementa (npr. $\mathcal{Y} = \{-1, 1\}$, ili slično). Ovaj model za cilj ima da skup ulaznih podataka podeli u dve klase pronalazeći hiper-ravan u prostoru ulaznih parametara koja je u skladu sa datim izlazima. Sa druge strane, linearna regresija se koristi za aproksimaciju ciljnih funkcija koje su kontinualne.

3.5.1 Linearna klasifikacija

Razmotrimo skup hipoteza \mathcal{H} sa d ulaznih parametara $f(x_1, x_2, \dots, x_d)$ gde $x_i \in \mathbb{R}$, i jednim izlaznim parametrom $y_0 \in \{-1, 1\}$. Neka je zadat *prag odluke* b tako da je:

$$h(x_1, x_2, \dots, x_n) = \begin{cases} 1, & \text{ako je } \sum_{i=1}^d w_i x_i > b \\ -1, & \text{ako je } \sum_{i=1}^d w_i x_i \leq b \end{cases}$$

gde su w_i težinski faktori. Sažeto, ovaj izraz možemo da zapišemo na sledeći način:

$$h(\mathbf{x}) = \text{sign} \left(\left(\sum_{i=1}^d w_i x_i \right) + b \right)$$

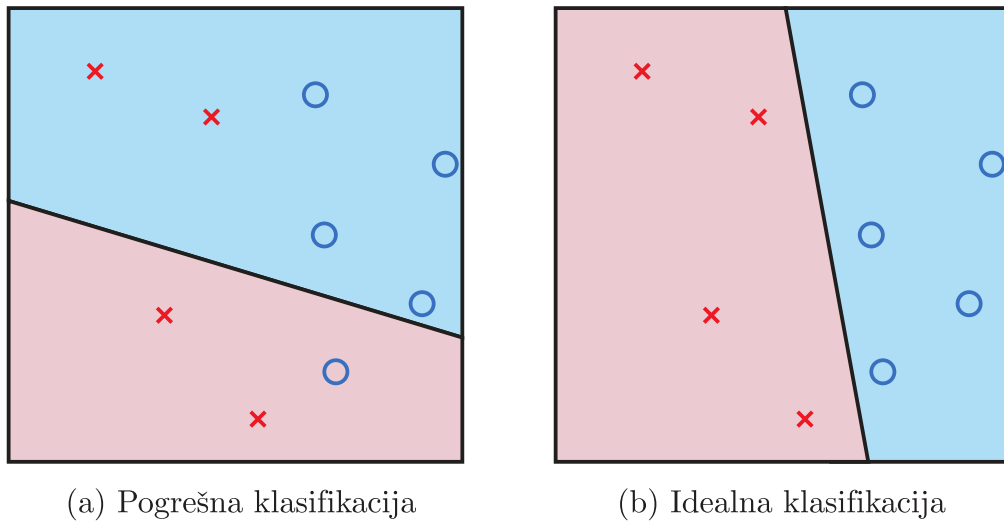
Da bismo dalje uprostiti gornji izraz, dodaćemo u prostor ulaznih parametara jedan parametar x_0 sa konstantnom vrednošću 1, koji odgovara težinskom faktoru $w_0 = b$, tako da sada možemo da pišemo:

$$h(\mathbf{x}) = \text{sign} \left(\sum_{i=0}^d w_i x_i \right)$$

Ukoliko uzmemo da je $\mathbf{x} = (x_0, x_1, \dots, x_d)$ i da je $\mathbf{w} = (w_0, w_1, \dots, w_d)$ dobijamo:

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x})$$

gde je $x_0 = 1$ za svako $\mathbf{x} \in 1 \times \mathbb{R}^d$, a vektor težinskih faktora $\mathbf{w} \in \mathbb{R}^{d+1}$. Ovako definisan linearni model učenja nazivamo *perceptron* i koristimo ga za klasifikaciju ulaznih podataka u dve klase. Na slici 7b data je grafička ilustracija rezultata perceptron modela za slučaj $d = 2$.



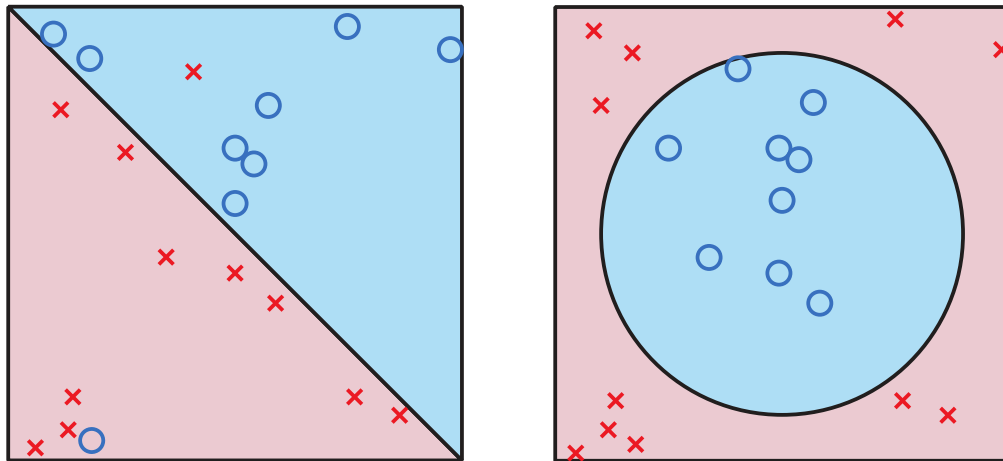
Slika 7 – Linearna klasifikacija podataka - perceptron

Kako bismo potvrdili da li je za slučaj *perceptrona* moguće da E_{out} bude dovoljno blizu E_{in} (prvi uslov izvodljivosti učenja), iskoristićemo VC granicu generalizacije [58]:

$$E_{out}(g) = E_{in}(g) + O\left(\sqrt{\frac{d}{N} \ln N}\right)$$

Dati izraz nam govori da će za dovoljno veliko N , E_{out} i E_{in} imati približnu vrednost. Drugi kriterijum izvodljivosti učenja, a to je da je $E_{in} \approx 0$ ćemo razmotriti nakon što prikažemo algoritam za učenje.

Algoritam za obuku perceptrona. Pretpostavimo da su podaci iz trening skupa *linearno separabilni*, što znači da postoji takav vektor \mathbf{w} da za svako $\mathbf{x}_i \in \mathcal{D}$ važi da je $h(\mathbf{x}_i) = y_i$. Drugim rečima, postoji takvo \mathbf{w} da je za njega $E_{in} = 0$ (slika 8a).



(a) Linearno separabilni podaci

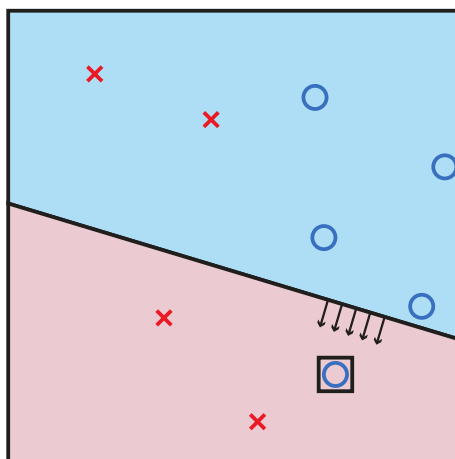
(b) Linearno neseparabilni podaci

Slika 8 – Linearna separabilnost podataka

Algoritam je iterativan. U svakoj iteraciji $t = 1, 2, \dots$ poznat je trenutni vektor težinskih faktora $\mathbf{w}(t)$. Sa $(\mathbf{x}(t), y(t))$ obeležimo uzorak koji u trenutnoj iteraciji nije dobro klasifikovan i iskoristićemo ga da promenimo trenutni vektor težinskih faktora po sledećem pravilu:

$$\mathbf{w}(t + 1) = \mathbf{w}(t) + y(t)\mathbf{x}(t)$$

Ovo pravilo pomera trenutnu granicu u pravcu dobre klasifikacije $\mathbf{x}(t)$ kao što je prikazano na slici 9. Algoritam nastavlja sa daljim iteracijama do momenta kada više nema uzoraka koji nisu ispravno klasifikovani. Može se dokazati da će ovakav algoritam sigurno konvergirati rešenju ukoliko su podaci linearno separabilni. Konvergencija ne zavisi od toga kojim redosledom biramo uzorke koji nisu dobro klasifikovani niti zavisi od toga kakve su inicijalne vrednosti težinskih faktora [58].



Slika 9 – Linearna klasifikacija podataka - perceptron

Pogledajmo sada slučaj kada podaci nisu linearno separabilni, što je prikazano na slici 8b. Algoritam za obuku perceptrona u ovom slučaju neće moći da konvergira i može da postane vrlo nestabilan prelazeći sa vrlo dobrog rešenja na veoma loše rešenje. Kvalitet rezultujućeg E_{in} nije moguće garantovati. U slučaju datom na slici 8b korišćenje linearnog modela nije opravdano, ali za slučaj prikazan na slici 8a gde nemamo idealnu linearnu separabilnost, nego vrlo približnu, ima smisla da nastavimo sa linearnim modelom i da kao rezultat damo hipotezu sa što manjim E_{in} .

Jedno od rešenja za ovaj problem je *pocket* varijanta algoritma za obučavanje perceptrona.

Algoritam 2 prikazuje *pocket* algoritam za obuku perceptrona koji pamti najbolji vektor težinskih faktora \hat{w} dobijen u prethodnim iteracijama do datog momenta. Na kraju, kao rezultat se vraća poslednja najbolja vrednost. Ukoliko podaci nisu linearno separabilni, nije moguće unapred predvideti E_{in} , ali je moguće naći hipotezu koja ima minimalni E_{in} . U ovom slučaju minimalno E_{in} zavisi od podataka u skupu \mathcal{D} i koliko su oni daleko od linearne separabilnosti.

Linearna klasifikacija po postavci odgovara problemu predikcije metaboličkog sindroma, ali neseparabilnost podataka sa kojim raspolažemo nam ne dozvoljava da primenimo gore pomenute algoritme. Zbog toga u ovoj tezi koristimo adekvatniji linearni

model koji predstavljamo u narednom odeljku.

Rezultat: vrednost *pocket težinskih faktora* $\hat{\mathbf{w}}$

$\hat{\mathbf{w}} = \mathbf{w}(0)$

$E_{in}(\hat{\mathbf{w}}) = \infty$

for $t = 0$ **to** $T - 1$ **do**

$\mathbf{w}(t + 1) = \mathbf{w}(t) + y(t)\mathbf{x}(t)$

 Izračunaj $E_{in}(\mathbf{w}(t))$

if $E_{in}(\mathbf{w}(t)) < E_{in}(\hat{\mathbf{w}})$ **then**

$\hat{\mathbf{w}} = \mathbf{w}(t + 1)$

$E_{in}(\hat{\mathbf{w}}) = E_{in}(\mathbf{w}(t))$

end

end

Algoritam 2: *Pocket* algoritam za obuku *perceptrona*

3.5.2 Linearna regresija

Linearna regresija se u odnosu na linearnu klasifikaciju razlikuje samo u prostoru izlaznih parametara, koji je u ovom slučaju kontinualan. Skup hipoteza modela linearne regresije ima sledeći oblik:

$$h(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$$

gde je $x_0 = 1$ za svako $\mathbf{x} \in 1 \times \mathbb{R}^d$, a vektor *težinskih faktora* $\mathbf{w} \in \mathbb{R}^{d+1}$, kao i kod prethodno razmatrane linearne klasifikacije.

Algoritam linearne regresije se bazira na minimizovanju kvadratne greške između $h(\mathbf{x})$ i y :

$$E_{out}(h) = \mathbb{E}[(h(\mathbf{x}) - y)^2]$$

gde je očekivana vrednost u skladu sa raspodelom $P(\mathbf{x}, y)$. Cilj je pronaći hipotezu koja ima malo $E_{out}(h)$. Kako je raspodela $P(\mathbf{x}, y)$ nepoznata, $E_{out}(h)$ nije moguće izračunati. Zbog toga koristimo *in-sample* grešku $E_{in}(h)$:

$$E_{in}(h) = \frac{1}{N} \sum_{n=1}^N (h(\mathbf{x}_n) - y_n)^2$$

Definišimo sada matricu $\mathbf{X} \in \mathbb{R}^{N \times (d+1)}$ dimenzija $N \times (d+1)$ čiji su redovi dati ulazi \mathbf{X}_n i vektor kolonu $\mathbf{y} \in \mathbb{R}$ čije su komponente dati izlazi y_n . Sada možemo da napišemo

izraz za grešku E_{in} u zavisnosti od \mathbf{w} , \mathbf{X} i \mathbf{y} :

$$\begin{aligned} E_{in}(\mathbf{w}) &= \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^\top \mathbf{x}_n - y_n)^2 \\ &= \frac{1}{N} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \\ &= \frac{1}{N} (\mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{y}^\top \mathbf{y}) \end{aligned} \tag{1}$$

gde je $\|\cdot\|$ euklidova norma vektora. Algoritam linearne regresije se zasniva na traženju minimuma $E_{in}(\mathbf{w})$ nad celim skupom mogućih $\mathbf{w} \in \mathbb{R}^{d+1}$, što formalno možemo zapisati kao:

$$\mathbf{w}_{\text{lin}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{d+1}} E_{in}(\mathbf{w})$$

Slika 10 ilustruje rešenje za jednu i dve dimenzije. Kako je $E_{in}(\mathbf{w})$ diferencijabilna, uslov za pronalaženje minimuma je $\nabla E_{in}(\mathbf{w}) = 0$. Iz prethodnih izraza dobijamo da je

$$\nabla E_{in}(\mathbf{w}) = \frac{2}{N} (\mathbf{X}^\top \mathbf{X} \mathbf{w} - \mathbf{X}^\top \mathbf{y})$$

Iz gornjeg izraza i uslova da je $\nabla E_{in}(\mathbf{w}) = 0$ dobijamo:

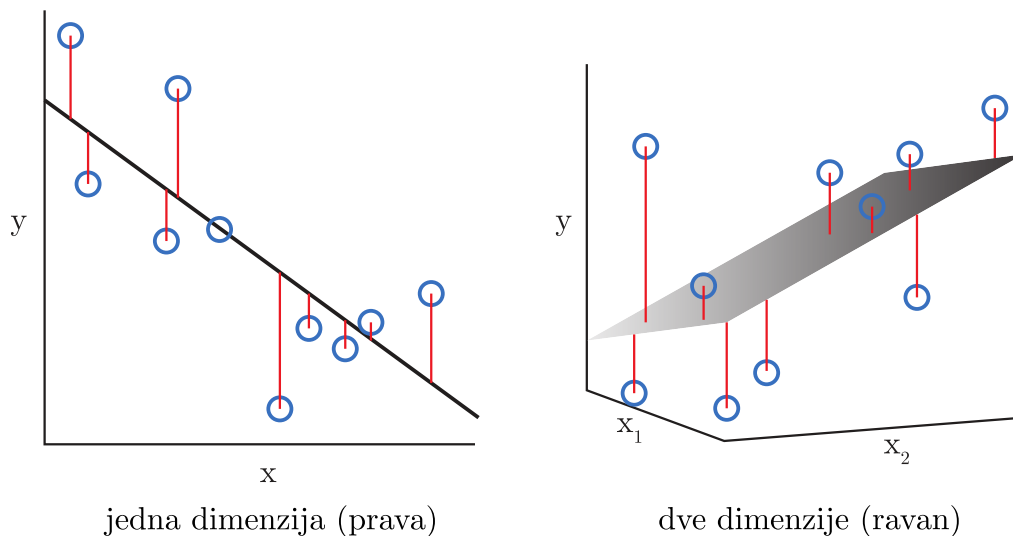
$$\mathbf{X}^\top \mathbf{X} \mathbf{w} = \mathbf{X}^\top \mathbf{y}$$

Za slučaj da $\mathbf{X}^\top \mathbf{X}$ ima svoju inverznu matricu možemo da izračunamo \mathbf{w} kao:

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

gde se $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ naziva *pseudo inverzna* matrica od \mathbf{X} koju obeležavamo sa \mathbf{X}^\dagger . Rezultujuće \mathbf{w} je jedinstveno optimalno rešenje koje tražimo. Ukoliko $\mathbf{X}^\top \mathbf{X}$ nema inverznu matricu, *pseudo inverzna* matrica postoji, ali rešenje nije jedinstveno. Zbog toga, kao konačnu formulu za \mathbf{w} dajemo:

$$\mathbf{w} = \mathbf{X}^\dagger \mathbf{y}$$



Slika 10 – Linearna regresija

Algoritam najmanjih kvadrata. Koristeći rezultate iz prethodnog teksta možemo jednostavno da izvedemo algoritam za linearnu regresiju:

Rezultat: \mathbf{w}_{lin}

Konstruisati matricu $\mathbf{X} \in \mathbb{R}^{N \times (d+1)}$

Izračunati *pseudo inverznu* matricu \mathbf{X}^\dagger od \mathbf{X}

Izračunati \mathbf{w}_{lin} pomoću izraza $\mathbf{w}_{\text{lin}} = \mathbf{X}^\dagger \mathbf{y}$

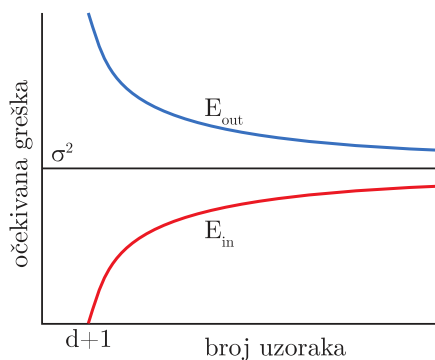
Algoritam 3: Algoritam najmanjih kvadrata

Vidimo da za razliku od algoritma za obučavanje perceptrona, algoritam najmanjih kvadrata nije iterativan po prirodi. Linearna regresija je redak slučaj algoritama za učenje, gde se rešenje dobija upotrebom analitičke formule. Ovo je razlog zašto je ova metoda u širokoj upotrebi.

Na kraju, osvrnimo se još na drugi uslov izvodljivosti učenja $E_{in} \approx 0$. Bez ulaženja u detalje [58], izraz za očekivanu *in-sample* grešku ima sledeću formu:

$$\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] = \sigma^2 \left(1 - \frac{d+1}{N}\right)$$

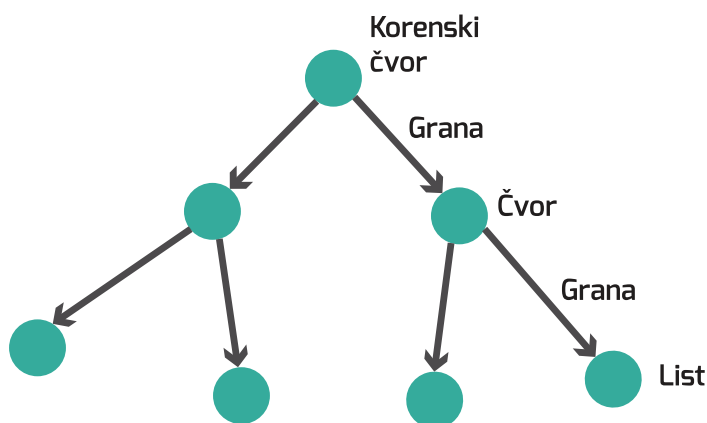
gde je σ^2 varijansa ciljne funkcije. Na slici 11 prikazan je odnos E_{in} i E_{out} u odnosu na broj uzoraka N .



Slika 11 – Kriva učenja za linearnu regresija (E_{in} i E_{out} u odnosu na broj uzoraka N)

3.6 Stabla odluke

Stabla odluke (*Decision Trees (DT)*) su poznata i često korišćena tehnika mašinskog učenja sa ciljem nalaženja unutrašnjih relacija u podacima. Stablo se formira u procesu obuke, a njegove glavne komponente su koren, čvorovi, grane i listovi (slika 12). Koren ima samo izlazne grane, dok list samo ulazne. Kada je stablo formirano, tok odlučivanja se kreće kroz grane – počev od korena, preko čvorova, pa do listova.



Slika 12 – Stablo odluke

Na početku obuke, potrebno je uraditi particiju skupa podataka na dva disjunktna podskupa, za obuku i za testiranje, izabrati algoritam i odrediti parametre obuke. Proces obuke se rekurzivno ponavlja i stablo se grana sve dok se ne naiđe na čvor u kojem dalje grananje nema doprinosa. Na osnovu tipa izlazne promenljive, razlikujemo *klasifikaciona stabla* u čijim listovima je kategorička promenljiva (klasa) i *regresiona stabla* u čijim listovima je kontinualna promenljiva.

U ovom istraživanju za kreiranje klasifikacionog stabla koristi se *C4.5 algoritam*. Skup za obuku \mathcal{T} čine sedmodimenzionalni vektori ($GEN, AGE, BMI, WHtR, SBP, DBP, MetS$), gde su $GEN, AGE, BMI, WHtR, SBP, DBP$ ulazni atributi na osnovu kojih se vrši klasifikacija, a $MetS$ izlazna klasa koja označava prisustvo metaboličkog sindroma i može imati dva ishoda *true* ili *false*.

C4.5 algoritam nije ograničen samo na binarna grananja, već kreira posebnu granu za svaku vrednost atributa. Na osnovu skupa za obučavanje \mathcal{T} se računa početna entropija po sledećoj formuli:

$$H(\mathcal{T}) = - \sum_j p_j \log_2(p_j).$$

Svako potencijalno grananje B deli skup za obuku na k podskupova $\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k$, nakon čega računamo entropiju:

$$H_B(\mathcal{T}) = \sum_{i=1}^k P_i H_B(\mathcal{T}_i).$$

Za svako grananje B se dobijena entropija $H_B(\mathcal{T})$ poredi sa početnom entropijom $H(\mathcal{T})$ sa ciljem da se izabere grananje koje najviše smanjuje entropiju.

U tabeli 3 je dat segment našeg skupa podataka za deset osoba na osnovu kojeg ćemo prikazati računanje entropija. Od navedenih deset osoba, pet imaju prisustvo metaboličkog sindroma, dok pet nemaju, pa početna entropija iznosi:

$$H(\mathcal{T}) = -\frac{5}{10} \log_2\left(\frac{5}{10}\right) - \frac{5}{10} \log_2\left(\frac{5}{10}\right) = 1.$$

Sada ćemo razmotriti grananje na osnovu BMI . Na osnovu tabele dobijamo:

$$P_{BMI < 25} = \frac{3}{10},$$

$$P_{25 \geq BMI < 30} = \frac{4}{10},$$

<i>GEN</i>	<i>AGE</i>	<i>BMI</i>	<i>WhtR</i>	<i>SBP</i>	<i>DBP</i>	<i>MetS</i>
1	46	22,33	0,40	150	120	0
2	32	21,00	0,43	120	80	0
2	27	22,49	0,49	120	100	1
1	57	27,94	0,64	140	95	1
2	48	25,72	0,55	135	80	1
2	49	26,49	0,51	125	80	0
1	43	28,98	0,52	120	80	0
2	25	38,50	0,81	140	75	1
1	58	33,11	0,63	155	95	1
2	25	40,83	0,54	115	80	0

Tabela 3 – Primer grananja stabla odluke

$$P_{BMI \geq 30} = \frac{3}{10},$$

$$H_{BMI < 25} = -\frac{1}{3} \log_2\left(\frac{1}{3}\right) - \frac{2}{3} \log_2\left(\frac{2}{3}\right) = 0.9183,$$

$$H_{25 \geq BMI < 30} = -\frac{2}{4} \log_2\left(\frac{2}{4}\right) - \frac{2}{4} \log_2\left(\frac{2}{4}\right) = 1,$$

$$H_{BMI \geq 30} = -\frac{2}{3} \log_2\left(\frac{2}{3}\right) - \frac{1}{3} \log_2\left(\frac{1}{3}\right) = 0.9183,$$

pa entropija nakon grananja iznosi

$$H_{BMI}(\mathcal{T}) = \frac{3}{10} \cdot 0.9183 + \frac{4}{10} \cdot 1 + \frac{3}{10} \cdot 0.9183 = 0.9510,$$

odnosno grananje smanjuje entropiju za

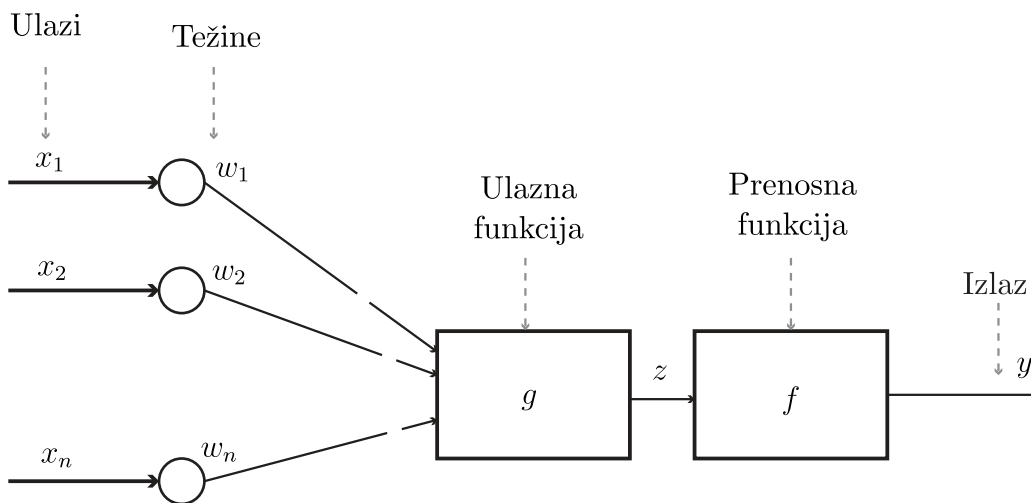
$$\Delta = H(\mathcal{T}) - H_{BMI}(\mathcal{T}) = 0.049.$$

3.7 Veštačke neuronske mreže

Veštačka neuronska mreža (Artificial Neural Network (ANN)) i algoritmi za njenu obuku spadaju u značajne metode mašinskog učenja. Veštačka neuronska mreža je nastala po uzoru na biološku neuronsku mrežu, i njenu strukturu čine povezani čvorovi, koje nazivamo *veštačkim neuronima*, koji predstavljaju osnovne jedinice obrade podataka.

Na slici 13 je prikazan model veštačkog neurona. Veštački neuron ima n ulaza $x_i, i = 1, 2, \dots, n$ i izlaz y . Svaki ulaz ima svoju težinu $w_i, i = 1, 2, \dots, n$ sa kojom se množi. Dobijeni proizvodi $x_i w_i, i = 1, 2, \dots, n$ su parametri ulazne funkcije neurona g , što izražavamo kao $z = g(x_1 w_1, x_2 w_2, \dots, x_n w_n)$. Najčešće u praksi funkcija g je suma u kojoj učestvuju i unutrašnje pobude $b_i, i = 1, 2, \dots, n$ (tzv. *bias*), odnosno važi da je:

$$z = \sum_{i=1}^n (x_i w_i + b_i).$$



Slika 13 – Šematski prikaz neurona

Izračunata suma je ulazna vrednost funkcije prenosna f , odnosno

$$y = f(z).$$

Najčešće prenosne funkcije su:

- sigmoidalna funkcija (na slici 14a)

$$y = \frac{1}{1 + e^{-z}},$$

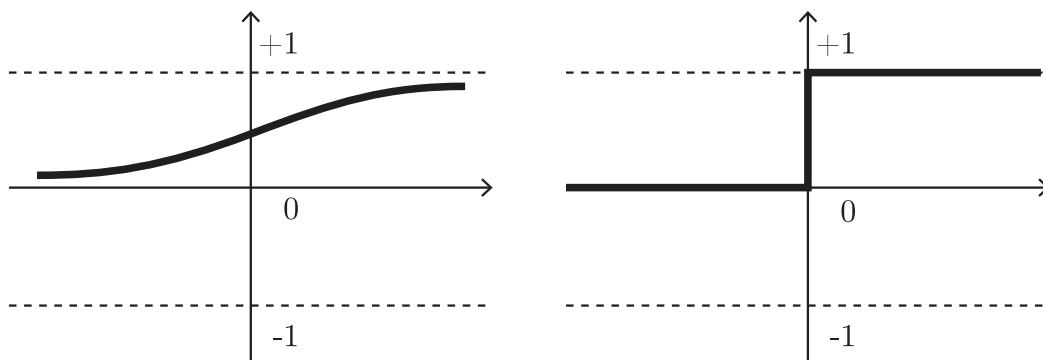
- odskočna funkcija (na slici 14b)

$$y = \begin{cases} 0, & z \leq 0 \\ 1, & z > 0 \end{cases},$$

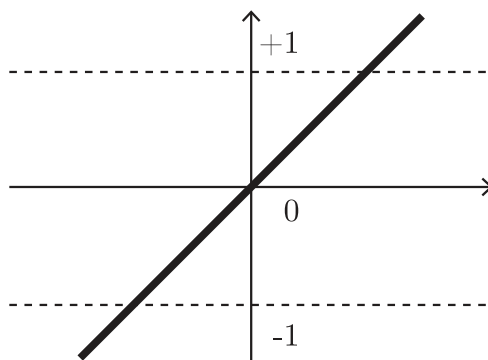
- linearna funkcija (na slici 15)

$$y = az,$$

gde je $z \in R$.



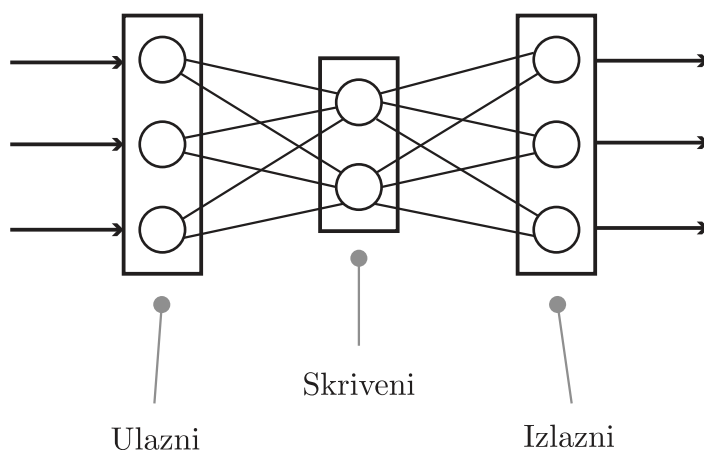
Slika 14 – Sigmoidalna (a) i odskočna (b) funkcija



Slika 15 – Linearna funkcija

Neuronska mreža ima slojevitú strukturu (slika 16), pa tako razlikujemo ulazni, skriveni i izlazni sloj. Ulazni sloj (*input layer*) prima ulazne podatke, a izlazni sloj (*output layer*) daje rezultat. Skriveni sloj (*hidden layer*) se nalazi između ulaznog i izlaznog sloja. Neuroni su raspoređeni po slojevima, pa se tako neuroni u skrivenom sloju nazivaju skrivenim neuronima (*hidden neurons*). Postoje dva tipa veza između slojeva:

- *feedforward*, gde signali putuju samo u jednom smeru (od ulaznog sloja, preko skrivenih slojeva, pa do izlaznog sloja),
- *feedback*, gde signali mogu putovati u oba smera.



Slika 16 – Veštačka neuronska mreža

U praksi se za obuku neuronske mreže najčešće koriste sledeće vrste učenja:

- Nadgledano učenje (*supervised learning*), gde su u trening skupu pored ulaznih podataka prisutni i tačno određeni izlazni podaci koje je prethodno odredio nadzornik (*supervisor*),
- Podržano učenje (*reinforcement learning*), gde su u obučavajućem skupu pored ulaznih podataka prisutni i određeni izlazni podaci sa ocenom koliko su oni dobri,
- Nenadgledano učenje (*unsupervised learning*), gde su u obučavajućem skupu prisutni samo ulazni podaci,
- Kompetitivno učenje (*competitive learning*), gde su u obučavajućem skupu prisutni samo ulazni podaci, a neuroni se takmiče i pobednik zauzima dominantnu poziciju kod donošenja odluka.

Početne vrednosti težina veza su slučajni brojevi. Nadgledano učenje se odvija iterativno po formuli

$$w_{ij}(k+1) = w_{ij}(k) + \delta_{ij}(k) = w_{ij}(k) + \mu \epsilon(y_j(k), y_j^*(k)),$$

gde su:

$w_{ij}(k)$ - težina veze od i do j u iteraciji k

$\delta_{ij}(k)$ - promena težine veze w_{ij} u iteraciji k

μ - koeficijent učenja

$\epsilon(y_j(k), y_j^*(k))$ - greška u iteraciji k koja se računa na osnovu dobijenog izlaza $y_j(k)$ i stvarnog izlaza $y_j^*(k)$ iz neurona j ,

dok se nenadgledano učenje bazira na formuli:

$$w_{ij}(k+1) = w_{ij}(k) + \delta_{ij}(k) = w_{ij}(k) + \mu y_j(k) x_j(k),$$

gde je:

$x_j(k)$ ulaz neurona j u iteraciji k .

Zbog efikasnog rešavanja izrazito složenih i nelinearnih problema, kao i u slučaju nekompletnosti ulaznih parametara, veštačke neuronske mreže imaju veliku primenu u kliničkoj medicini [60, 61]. Intenzivno i uspešno se koriste u algoritamskoj dijagnostici metaboličkog sindroma i kardiovaskularnih bolesti [60, 62, 63, 54].

Predikcija metaboličkog sindroma primenom *feedforward* veštačkih neuronskih mreža, obučenih algoritmom *Levenberg-Marquardt backpropagation*, je realizovana u MATLAB-u (*Neural Network Toolbox*) i prikazana u [28]. Poput *Gauss-Newton* algoritma, *Levenberg-Marquardt* algoritam, takođe, ne zahteva izračunavanje *Hessian* matrice:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 E}{\partial w_1^2} & \frac{\partial^2 E}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 E}{\partial w_1 \partial w_N} \\ \frac{\partial^2 E}{\partial w_2 \partial w_1} & \frac{\partial^2 E}{\partial w_2^2} & \cdots & \frac{\partial^2 E}{\partial w_2 \partial w_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 E}{\partial w_N \partial w_1} & \frac{\partial^2 E}{\partial w_N \partial w_2} & \cdots & \frac{\partial^2 E}{\partial w_N^2} \end{bmatrix},$$

koja u sebi sadrži druge izvode sume kvadrata grešaka

$$E(\mathbf{x}, \mathbf{w}) = \frac{1}{2} \sum_{p=1}^P \sum_{m=1}^M \mathbf{e}_{p,m}^2,$$

gde je

p - indeks paterna $p = 1, 2, \dots, P$

m - indeks izlaza $m = 1, 2, \dots, M$

\mathbf{x} - vektor ulaza

\mathbf{w} - vektor težina

$\mathbf{e}_{p,m}$ - vektor grešaka na izlazu m kada se primeni patern p , odnosno

$$\mathbf{e}_{p,m} = \mathbf{y}_{p,m} - \mathbf{y}_{p,m}^*$$

\mathbf{y} - dobijeni izlaz

\mathbf{y}^* - stvarni izlaz

Naime, *Levenberg-Marquardt* algoritam koristi sledeću aproksimaciju:

$$\mathbf{H} \approx \mathbf{J}^T \mathbf{J} + \mu \mathbf{I},$$

na osnovu koje se izvodi pravilo za obučavanje neuronske mreže:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - (\mathbf{J}_k^T \mathbf{J}_k + \mu \mathbf{I})^{-1} \mathbf{J}_k \mathbf{e}_k,$$

gde je k iteracija,
 μ koeficijent $\mu \geq 0$,
 \mathbf{I} jedinična matrica
i matrica

$$\mathbf{J} = \begin{bmatrix} \frac{\partial e_{1,1}}{\partial w_1} & \frac{\partial e_{1,1}}{\partial w_2} & \dots & \frac{\partial e_{1,1}}{\partial w_N} \\ \frac{\partial e_{1,2}}{\partial w_1} & \frac{\partial e_{1,2}}{\partial w_2} & \dots & \frac{\partial e_{1,2}}{\partial w_N} \\ \vdots & & \ddots & \vdots \\ \frac{\partial e_{1,M}}{\partial w_1} & \frac{\partial e_{1,M}}{\partial w_2} & \dots & \frac{\partial e_{1,M}}{\partial w_N} \\ \vdots & & \ddots & \vdots \\ \frac{\partial e_{P,1}}{\partial w_1} & \frac{\partial e_{P,1}}{\partial w_2} & \dots & \frac{\partial e_{P,1}}{\partial w_N} \\ \frac{\partial e_{P,2}}{\partial w_1} & \frac{\partial e_{P,2}}{\partial w_2} & \dots & \frac{\partial e_{P,2}}{\partial w_N} \\ \vdots & & \ddots & \vdots \\ \frac{\partial e_{P,M}}{\partial w_1} & \frac{\partial e_{P,M}}{\partial w_2} & \dots & \frac{\partial e_{P,M}}{\partial w_N} \end{bmatrix}.$$

U našem istraživanju ulazni vektori veštačke neuronske mreže su u formi:

$$\mathbf{x}_i = (GEN, AGE, BMI, WHtR, SBP, DBP)_i,$$

dok su na izlazu *MetS* koeficijenti, odnosno:

$$y_i = MetS_i,$$

gde je $i = 1, 2, \dots, 2928$. *MetS* koeficijenti imaju vrednost 1 ili 0 u zavisnosti od toga da li razmatrana osoba ima ili nema metabolički sindrom, a određeni su na osnovu *IDF* definicije koja je objašnjena u poglavlju 2.1.

4 Metodologija - implementacija algoritama

U okviru poglavlja posvećenog mašinskom učenju zaključili smo da ne postoji pouzdan način da unapred odredimo koji model učenja je najbolji za rešavanje datog problema. Ovo implicira da je potrebno izvršiti eksperimente sa različitim modelima i uporediti dobijene performanse, kako bismo odredili koji model daje najbolje rezultate. Pored zadovoljavajućih performansi, poželjno je da konačni model bude što jednostavniji.

Kako bi pronašli optimalan koji je u skladu sa gore navedenim kriterijumima, u ovoj tezi koristimo tri modela učenja različite složenosti. Za odabrane modele pretpostavljeno je da će dati rešenja zadovoljavajućih performansi adekvatnih za upotrebu algoritamske dijagnostike metaboličkog sindroma u praksi:

1. Stabla odluke
2. Linearna regresija
3. Veštačke neuronske mreže

U okviru ovog poglavlja pored detalja implementacije gore navedenih algoritama prikazujemo implementaciju sledećih bitnih funkcija:

- Izračunavanje kriterijuma performansi algoritma (*PPV* i *NPV*)
- Normalizovanje podataka
- Izračunavanje dijagnoze metaboličkog sindroma na osnovu *IDF* definicije

Algoritmi su implementirani u programskom jeziku *C#* [64, 65], korišćenjem biblioteke za mašinsko učenje *Accord.NET* [66, 67]. Izlazi algoritama broj 2 i broj 3 su kontinualne vrednosti u intervalu $[0, 1]$. Ovako dobijeni rezultati se konvertuju u konačan rezultat korišćenjem standardne *Math.Round* metode. Vrednost dobijena na ovaj način pripada skupu $\{0, 1\}$, gde nam vrednost 1 govori da je metabolički sindrom prisutan kod pacijenta, a vrednost 0 da nije. Izlaz algoritma broj 1 je vrednost iz skupa $\{0, 1\}$, i njegove vrednosti nije potrebno dodatno transformisati.

4.1 Kriterijum performanse algoritma

Kao kriterijum performanse algoritma primenjene su *pozitivna prediktivna vrednost PPV* (*positive predictive value*) i *negativna prediktivna vrednost NPV* (*predictive va-*

lue), koje se određuju primenom sledećih formula:

$$PPV = \frac{N_{TP}}{N_{TP} + N_{FP}} = \frac{N_{TP}}{N_P},$$
$$NPV = \frac{N_{TN}}{N_{TN} + N_{FN}} = \frac{N_{TN}}{N_N},$$

gde je

N_{TP} - broj tačnih pozitivnih predikcija (*number of true positives*)

N_{FP} - broj netačnih pozitivnih predikcija (*number of false positives*)

N_P - ukupan broj pozitivnih predikcija (*number of positives*)

N_{TN} - broj tačnih negativnih predikcija (*number of true negatives*)

N_{FN} - broj netačnih negativnih predikcija (*number of false negatives*)

N_N - ukupan broj negativnih predikcija (*number of negatives*)

U okviru našeg problema, pod pozitivnom predikcijom podrazumevamo da je metabolički sindrom prisutan, a pod negativnom predikcijom podrazumevamo da metabolički sindrom nije prisutan. Programski kod 1 u nastavku prikazuje implementaciju algoritma.

```
private static PredictiveValues CalculatePredictiveValues(  
    double[][] expectedOutputs, double[][] actualOutputs)  
{  
    var n_tp = 0;  
    var n_fp = 0;  
    var n_tn = 0;  
    var n_fn = 0;  
    var n = expectedOutputs.Length;  
  
    for (int i = 0; i < n; i++)  
    {  
        var expected = (int)Math.Round(expectedOutputs[i][0]);  
        var actual = (int)Math.Round(actualOutputs[i][0]);  
        if (actual == expected)  
        {  
            // u ovom slucaju imamo tacnu predikciju  
            // potrebno je odrediti da li pozitivna ili negativna  
            if (expected == 0)  
                n_tn++; // negativna  
            else  
                n_tp++; // pozitivna  
        }  
    }  
}
```

```

    }
    else
    {
        // u ovom slucaju imamo netacnu predikciju
        // potrebno je odrediti da li pozitivna ili negativna
        if (expected == 0)
            n_fn++; // negativna
        else
            n_fp++; // pozitivna
    }
}

var result = new PredictiveValues()
{
    N_TP = n_tp,
    N_TN = n_tn,
    N_FP = n_fp,
    N_FN = n_fn,
    PPV = (decimal)Math.Round((double)n_tp / ((double)n_tp + (double)n_fp), 6),
    NPV = (decimal)Math.Round((double)n_tn / ((double)n_tn + (double)n_fn), 6)
};

return result;
}

```

Programski kod 1 – Izračunavanje prediktivnih vredosti

4.2 Normalizacija podataka

Pod normalizacijom podataka podrazumevamo takvu transformaciju vrednosti ulaznih parametara koja za rezultat daje vrednosti parametara čija je srednja vrednost 0, a standardna devijacija 1. Ovo se postiže tako što se za svaki ulazni parametar x_1, x_2, \dots, x_n izračuna srednja vrednost $\mu_1, \mu_2, \dots, \mu_n$ i standardna devijacija $\sigma_1, \sigma_2, \dots, \sigma_n$. Zatim se nove transformisane vrednosti ulaznih parametara x_{nor_i} dobijaju po formuli:

$$x_{nor_i} = \frac{x_i - \mu_i}{\sigma_i}$$

Ovako definisana transformacija naziva se *Z-Scores*.

Iako su generalno, veštačke neuronske mreže invarijantne u odnosu na skaliranje ulaznih parametara, pokazano je da normalizacija podataka pozitivno utiče na ishod uče-

nja [68] u slučajevima manjih skupova podataka. Kako skup podataka raste, tako ovaj efekat opada.

U okviru ove teze, eksperimenti su izvođeni sa prethodno normalizovanim podacima. Programski kod 2 predstavlja implementaciju pomoću metoda koje su nam na raspolaganju u biblioteci *Accord.NET*.

```
inputs = new List<double[]>(inputs.ToArray().ZScores());
```

Programski kod 2 – Normalizacija podataka

4.3 Izračunavanje koeficijenta *MetS* na osnovu *IDF* definicije

Ovo istraživanje se oslanja na definiciju metaboličkog sindroma datu od strane *IDF* prikazanu u odeljku 2.1. Programski kod 3 prikazuje implementaciju algoritma 1 za izračunavanje *MetS* koeficijenta.

```
private const int MALE = 1;
private const int FEMALE = 2;
private static int CalculateMets(MetsDataRecord mr)
{
    var mets = 0;
    if ((mr.BMI > 30) || ((mr.GEN == FEMALE && mr.WC >= 80) || (mr.GEN == MALE &&
    ↪ mr.WC >= 94)))
    {
        var p = 0;
        if (mr.SBP >= 130 || mr.DBP >= 85)
            p += 1;
        if ((mr.GEN == FEMALE && mr.HDL < 1.29m) || (mr.GEN == MALE && mr.HDL <
        ↪ 1.03m))
            p += 1;
        if (mr.TG >= 1.7m)
            p += 1;
        if (mr.FPG >= 5.6m)
            p += 1;
        if (p >= 2)
            mets = 1;
    }
}
```

```
    return mets;
}
```

Programski kod 3 – Implementacija IDF definicije metaboličkog sindroma

4.4 Stablo odluke

Stabla odluke su razmatrana u odeljku 3.6, gde je opisan i algoritam za obuku C4.5. Biblioteka *Accord.NET* raspolaže implementacijom ovog algoritma, koja je data u vidu klase *C45Learning*. Pored toga, omogućeno je automatsko generisanje ulaznih varijabli na osnovu podataka iz *trening skupa* korišćenjem metode *DecisionVariable.FromData*. Implementacija je data u programskom kodu 4.

```
private static AlgoritamPredictiveValues DecisionTree(Data data)
{
    // konvertujemo date ulaze u oblik potreban za Learn metod
    var intLearnOutputs = data.LearnOutputs.Select(x => (int)Math.Round(x[0],
        ↪ 0)).ToArray();

    // promenjive automatski preuzimamo iz podataka
    var vars = DecisionVariable.FromData(data.LearnInputs);

    // kreiramo stablo odluke sa dve klase
    var decisionTree = new DecisionTree(vars, 2);

    // instanciramo algoritam obuke
    var c45learning = new C45Learning(decisionTree);

    // obuka stabla
    decisionTree = c45learning.Learn(data.LearnInputs, intLearnOutputs);

    // izracunavanje izlaza obucenog stabla odluke
    double[][] DecisionTreeCompute(double[][] inputs, double[][] outputs)
    {
        var res = new double[inputs.Length][];

        for (int i = 0; i < inputs.Length; i++)
        {
            var d3o = decisionTree.Decide(inputs[i]);
            var to = Math.Round(outputs[i][0], 0);
        }
    }
}
```

```

        res[i] = new double[] { d3o };
    }

    return res;
}

var learnOutputs = DecisionTreeCompute(data.LearnInputs, data.LearnOutputs);
var validateOutputs = DecisionTreeCompute(data.ValidateInputs,
    ↪ data.ValidateOutputs);
var testOutputs = DecisionTreeCompute(data.TestInputs, data.TestOutputs);

var learnPV = CalculatePredictiveValues(data.LearnOutputs, learnOutputs);
var validatePV = CalculatePredictiveValues(data.ValidateOutputs,
    ↪ validateOutputs);
var testPV = CalculatePredictiveValues(data.TestOutputs, testOutputs);

var result = new AlgoritamPredictiveValues()
{
    LearnNPV = learnPV.NPV,
    LearnPPV = learnPV.PPV,

    ValidateNPV = validatePV.NPV,
    ValidatePPV = validatePV.PPV,

    TestNPV = testPV.NPV,
    TestPPV = testPV.PPV
};

return result;

```

Programski kod 4 – Implementacija stabla odluke

4.5 Linearna regresija

Algoritam minimalnih kvadrata za *linearnu regresiju* opisan je u odeljku 3.5.2. Biblioteka *Accord.NET* raspolaže implementacijom ovog algoritma u vidu klase *OrdinaryLeastSquares*. Programski kod 5 prikazuje na koji način je upotrebljena u okviru ove teze.

```

private static AlgoritamPredictiveValues LinearRegression(Data data)
{

```

```

// instanciramo algoritam
var ols = new OrdinaryLeastSquares();

// konvertujemo ulazne vrednosti u oblik potreban za Learn metodu
var originalLearnOutputs = data.LearnOutputs.Select(x => x[0]).ToArray();

// izvršavamo obuku
var lr = ols.Learn(data.LearnInputs, originalLearnOutputs);

// izracunavanje izlaza
double[][] LRCompute(double[][] inputs, double[][] outputs)
{
    var res = new double[inputs.Length][];

    for (int i = 0; i < inputs.Length; i++)
    {
        var lrr = lr.Transform(inputs[i]);
        var lro = Math.Round(lrr);
        var to = Math.Round(outputs[i][0], 0);

        res[i] = new double[] { lro };
    }

    return res;
}

var learnOutputs = LRCompute(data.LearnInputs, data.LearnOutputs);
var validateOutputs = LRCompute(data.ValidateInputs, data.ValidateOutputs);
var testOutputs = LRCompute(data.TestInputs, data.TestOutputs);

var learnPV = CalculatePredictiveValues(data.LearnOutputs, learnOutputs);
var validatePV = CalculatePredictiveValues(data.ValidateOutputs,
↪ validateOutputs);
var testPV = CalculatePredictiveValues(data.TestOutputs, testOutputs);

var result = new AlgoritamPredictiveValues()
{
    LearnNPV = learnPV.NPV,
    LearnPPV = learnPV.PPV,

    ValidateNPV = validatePV.NPV,
    ValidatePPV = validatePV.PPV,

    TestNPV = testPV.NPV,

```

```

        TestPPV = testPV.PPV
    };

    return result;
}

```

Programski kod 5 – Implementacija linearne regresije

4.6 Veštačka neuronska mreža

Predikcija metaboličkog sindroma primenom *feedforward* veštačkih neuronskih mreža prethodno je prikazana u [28]. Korišćenjem MATLAB-ovog *Neural Network Toolbox*-a, neuronska mreža je obučena *backpropagation* algoritmom *Levenberg-Marquardt* (funkcija *trainlm*) sa standardnim parametrima i bez *regularizacije*. Ulazni podaci su prosleđeni algoritmu u izvornom obliku bez prethodne *normalizacije*. Razmotrene su jednoslojne veštačke neuronske mreže sa $HN = 1, 2, \dots, 100$ skrivenih neurona. Na osnovu dobijenih rezultata zaključeno je da je za pozitivnu predikciju metaboličkog sindroma optimalna ahritektura veštačke neuronske mreže sa $HN = 96$ skrivenih neurona (na osnovu maksimuma $PPV_{TS} = 0.8579$), dok je za negativnu predikciju optimalna sa $HN = 85$ skrivenih neurona ($NPV_{TS} = 0.8319$) [28].

Kako bismo pokušali da unapredimo prethodni rezultat, u ovoj tezi koristimo isti *backpropagation* algoritam *Levenberg-Marquardt*, ali ovaj put sa uključenom *bajesovskom regularizacijom* i prethodnom *normalizacijom* podataka. Pored toga, na kraju svake epohe obučavanja neuronske mreže vršimo validaciju pomoću validacionog skupa evaluiranjem kriterijuma performanse PPV i NPV . Proces učenja se prekida ukoliko je u poslednjih 10 epoha kriterijum performanse ostajao isti ili je opadao.

Programski kod 6 prikazuje implementaciju u C#-u. Za reprezentaciju neuronske mreže korišćena je klasa *ActivationNetwork*. Za inicijalizaciju početnih težinskih faktora koristimo klasu *NguyenWidrow*, a algoritam obučavanja je dat u vidu klase *LevenbergMarquardtLearning*. Da bi se u toku obučavanja koristila regularizacija potrebno je da se parametar *UseRegularization* postavi na vrednost **true**.

```

private static AlgoritamPredictiveValues Ann(Data data, int neuronCount)
{
    // za aktivacionu funkciju uzimamo standardnu sigmoidnu

```



```

var activationFunction = new SigmoidFunction();

// kreiramo neuronsku mrežu sa 6 neurona u ulaznom sloju i jednim u izlaznom
// broj skrivenih neurona je određen parametrom neuronCount
var ann = new ActivationNetwork(activationFunction, INPUT_COUNT, neuronCount,
↪ OUTPUT_COUNT);

// postavljanje inicijalnih težinski faktora
new NguyenWidrow(ann).Randomize();

// algoritam za obučavanje
var teacher = new LevenbergMarquardtLearning(ann);

teacher.LearningRate = 0.001;

// uključujemo regularizaciju
teacher.UseRegularization = true;

double err;
int count = 0;
int failCount = 0;
PredictiveValues lastPredictiveValues = null;
while (true)
{
    err = teacher.RunEpoch(data.LearnInputs, data.LearnOutputs);

    // izracunavamo prediktivne vrednosti
    var pv = CalculatePredictiveValues(ann, data);

    LogMessage($"Epoch #{count + 1:000}, PPV: {pv.PPV, 10:#0.000000}, NPV:
↪ {pv.NPV, 10:#0.000000}");

    if (lastPredictiveValues != null)
    {
        // proveravamo da li se performansa poboljsala ili ne
        if (lastPredictiveValues.PPV >= pv.PPV && lastPredictiveValues.NPV >=
↪ pv.NPV)
        {
            failCount++;
            // obuka se prekida ukoliko do poboljsanja nije doslo u poslednjih 10
            ↪ epoha
            if (failCount >= 10)
                break;
        }
    }
}

```

```

        else
            failCount = 0;
    }

    lastPredictiveValues = pv;

    count++;
    if (count >= 500)
        break;
}

// izracunavanje izlaza
double[][] AnnCompute(double[][] inputs, double[][] outputs)
{
    var res = new double[inputs.Length][];

    for (int i = 0; i < inputs.Length; i++)
    {
        var annr = ann.Compute(inputs[i])[0];
        var to = Math.Round(outputs[i][0], 0);
        var anno = Math.Round(annr, 0);

        res[i] = new double[] { anno };
    }

    return res;
}

var learnOutputs = AnnCompute(data.LearnInputs, data.LearnOutputs);
var validateOutputs = AnnCompute(data.ValidateInputs, data.ValidateOutputs);
var testOutputs = AnnCompute(data.TestInputs, data.TestOutputs);

var learnPV = CalculatePredictiveValues(data.LearnOutputs, learnOutputs);
var validatePV = CalculatePredictiveValues(data.ValidateOutputs,
↪ validateOutputs);
var testPV = CalculatePredictiveValues(data.TestOutputs, testOutputs);

var result = new AlgoritamPredictiveValues()
{
    LearnNPV = learnPV.NPV,
    LearnPPV = learnPV.PPV,

    ValidateNPV = validatePV.NPV,

```

```
    ValidatePPV = validatePV.PPV,  
  
    TestNPV = testPV.NPV,  
    TestPPV = testPV.PPV  
};  
  
return result;  
}
```

Programski kod 6 – Implementacija veštačke neuronske mreže

5 Eksperimenti i analiza rezultata

U okviru ovog istraživanja urađena su četiri eksperimenta:

1. Određivanje optimalne arhitekture neuronske mreže
2. Poređenje performansi algoritama
3. Obuka mreže optimalne arhitekture
4. Testiranje obučene neuronske mreže,

a zaključak o performansama se donosi na osnovu *PPV* i *NPV*, koje su opisane u poglavlju 4. Eksperimenti su implementirani u jeziku *C#* korišćenjem biblioteke za mašinsko učenje *Accord.NET*.

5.1 Određivanje optimalne arhitekture neuronske mreže

Sa ciljem da uporedimo rezultate sa [28], u ovom istraživanju korišćen je sličan skup podataka od 2928 osoba koji je na slučajan način deljen na tri particije u proporciji 80:10:10 – skup za obučavanje, skup za validaciju i skup za testiranje. Razmotrene su jednoslojne veštačke neuronske mreže tipa *feedforward* sa $HN = 1, 2, \dots, 100$ skrivenih neurona. Ulazni vektor veštačke neuronske mreže ima oblik:

$$\mathbf{x}_i = (GEN, AGE, BMI, WHtR, SBP, DBP)_i,$$

dok su na izlazu *MetS* koeficijenti, odnosno:

$$y_i = MetS_i,$$

gde je $i = 1, 2, \dots, 2928$. *MetS* koeficijenti imaju vrednost 1 ili 0 u zavisnosti od toga da li razmatrana osoba ima ili nema metabolički sindrom, a određeni su na osnovu *IDF* definicije koja je objašnjena u poglavlju 2.1.

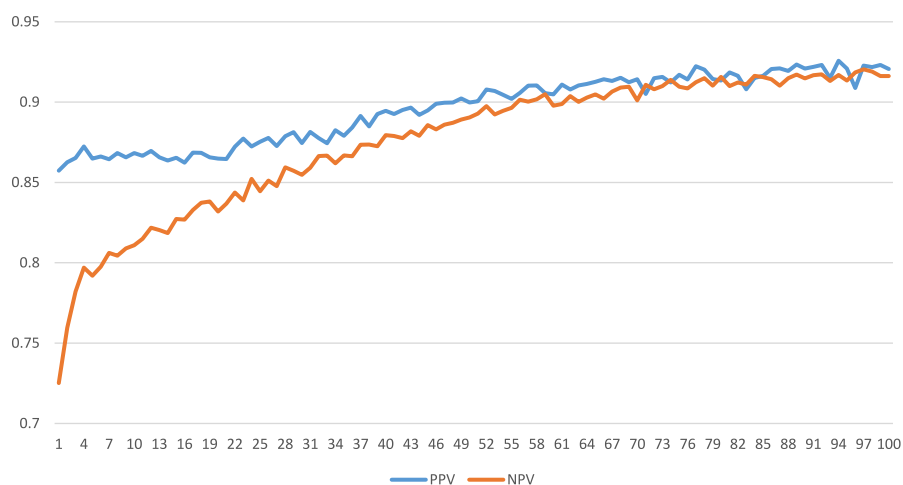
Za razliku od [28], gde je korišćen standardni *Levenberg-Marquardt backpropagation* algoritam za obuku, u okviru ove doktorske disertacije koristimo isti algoritam ali sa uključenom bajesovskom regularizacijom [69]. Ova varijanta algoritma naziva se i samo *bajesovska regularizacija*, i razmatrana je u različitim oblastima gde su postizani umereno do znatno bolji rezultati u odnosu na druge algoritme [70, 71, 72, 73].

Za osnovne parametre algoritma primenjene su uobičajne vrednosti definisane u biblioteci *Accord.NET*.

Pored toga u [28] ulazni podaci su prosleđeni algoritmu u izvornom obliku. Iako su generalno, veštačke neuronske mreže invarijantne u odnosu na skaliranje ulaznih parametara, pokazano je da normalizacija podataka pozitivno utiče na ishod učenja [68] u slučajevima manjih skupova podataka. Zbog toga je u okviru ove teze ulazni skup prethodno transformisan na način opisan u odeljku 4.2.

Optimalna arhitektura je određena primenom metode *repeated random subsampling validation* u okviru koje je za svako $HN = 1, 2, \dots, 100$ eksperiment 100 puta ponovljen i na kraju izračunate prosečne vrednosti PPV i NPV odvojeno na skupovima za obučavanje (PPV_{TR} i NPV_{TR}), validaciju (PPV_{VL} i NPV_{VL}) i testiranje (PPV_{TS} i NPV_{TS}). Dobijeni rezultati su navedeni u tabeli 9 i na osnovu njih je zaključeno da je za pozitivnu predikciju metaboličkog sindroma optimalna arhitektura veštačke neuronske mreže sa $HN = 94$ skrivenih neurona (na osnovu maksimuma $PPV_{TS} = 0.9257$), dok je za negativnu predikciju optimalna sa $HN = 97$ skrivenih neurona (na osnovu maksimuma $NPV_{TS} = 0.9204$).

Slika 17 prikazuje grafike prosečnih vrednosti PPV_{TS} i NPV_{TS} u zavisnosti od HN . Sa grafika se može primetiti da do $HN \approx 85$ vrednosti PPV_{TS} i NPV_{TS} imaju trend rasta, nakon čega ulaze u saturaciju. Ovaj zaključak je bitan jer u saturaciji PPV_{TS} i NPV_{TS} imaju zanemarljiv rast, što znači da se ostvaruje približna tačnost predikcije i sa manjim brojem skrivenih neurona (npr. umesto $HN = 97$ možemo koristiti $HN = 85$). Na taj način dobijamo slične performanse, ali uz značajno kraće vreme obučavanja. Dobijeni rezultati za PPV_{TS} i NPV_{TS} u ovom istraživanju su značajno bolji od rezultata koji su objavljeni u [28]. Dakle, pokazano je da se primenom *bajesovske regularizacije* i prethodnom normalizacijom ulaza postiže bolja generalizacija veštačke neuronske mreže za problem predikcije metaboličkog sindroma.



Slika 17 – Grafik vrednosti PPV i NPV u odnosu na broj skrivenih neurona

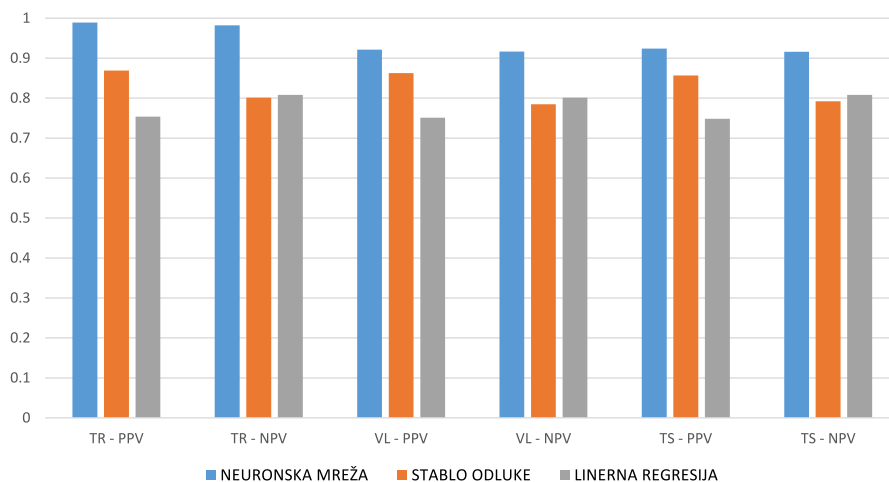
5.2 Poređenje performansi algoritama

U okviru ovog eksperimenta poređene su performanse stabla odluke, linearne regresije i veštačke neuronske mreže optimalne strukture. Implementacije algoritama date su u odeljcima 4.4, 4.5 i 4.6. Stablo odluke je formirano pomoću algoritma $C4.5$, a linearna regresija je urađena metodom najmanjih kvadrata.

U poglavlju 5.1 je zaključeno da do $HN \approx 85$ vrednosti PPV_{TS} i NPV_{TS} imaju trend rasta, nakon čega ulaze u saturaciju i imaju zanemarljiv rast. Zbog toga možemo koristiti neuronsku mrežu i sa manjim brojem skrivenih neurona, pa u ovom eksperimentu koristimo neuronsku mrežu sa $HN = 90$.

Performanse razmatranih metoda mašinskog učenja (stablo odluke, linearna regresija, i veštačka neuronska mreža) upoređene su na 100 *random* primera i dobijeni rezultati su prikazani u tabeli 10. U svakom primeru je izvršena *random* particija skupa podataka 80:10:10 i dobijeni su skupovi za obučavanje, validaciju i testiranje. Na istom obučavajućem skupu je obučeno stablo odluke i neuronska mreža, a zatim određeni parametri linearne regresije. Nakon toga, na istom skupu su testirane sve tri metode i izračunate PPV i NPV . U prilogu je data tabela 10 koja detaljno prikazuje PPV i NPV za svih 100 primera. Prosečne vrednosti PPV i NPV za svih 100 primera su date u tabeli 4, koje su ujedno grafički prikazane na slici 18. Na osnovu ovih rezultata

zaključujemo da veštačke neuronske mreže tačnije dijagnostikuju metabolički sindrom u odnosu na stabla odluke i linearnu regresiju.



Slika 18 – Poređenje ANN90, DT i LR na osnovu PPV i NPV

Metoda	PPV_{TR}	NPV_{TR}	PPV_{VL}	NPV_{VL}	PPV_{TS}	NPV_{TS}
ANN90	0,9890	0,9820	0,9216	0,9165	0,9239	0,9162
DT	0,8693	0,8013	0,8629	0,7848	0,8567	0,7920
LR	0,7537	0,8082	0,7511	0,8013	0,7484	0,8083

Tabela 4 – Grafik vrednosti PPV i NPV za ANN90, DT i LR

Tabela 5 prikazuje rezultate Wilcoxon-Mann-Whitney testa na osnovu kojih zaključujemo da postoji statistički značajna razlika između srednjih vrednosti PPV i NPV prikazanih u tabeli 4.

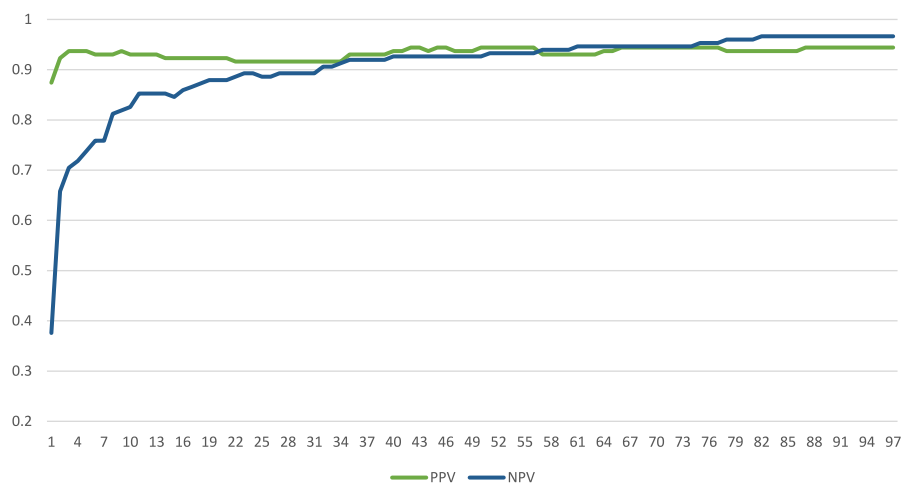
Metode	W	p -value
ANN90 - DT	9324,5	<0,0001*
ANN90 - LR	9923,5	<0,0001*
DT - LR	9647,0	<0,0001*

Tabela 5 – Rezultati Wilcoxon-Mann-Whitney testa

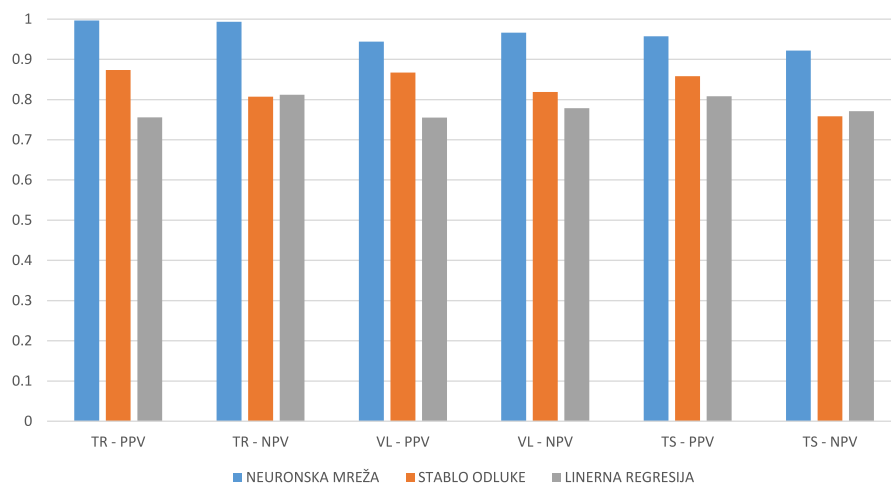
5.3 Obuka mreže optimalne arhitekture

Ovaj eksperiment razmatra obuku veštačke neuronske mreže sa 90 skrivenih neurona. Tabela 11, koja je data u prilogu, prikazuje PPV i NPV na skupu za validaciju u zavisnosti kako se broj epoha povećava koji je dat u prvoj koloni tabele. Od druge do pete kolone su date vrednosti na osnovu kojih se računa PPV i NPV , a to su redom N_{TP} (broj tačnih pozitivnih predikcija), N_{TN} (broj tačnih negativnih predikcija), N_{FP} (broj netačnih pozitivnih predikcija) i N_{FN} (broj netačnih negativnih predikcija). Kretanje PPV i NPV na skupu za validaciju u zavisnosti od broja epoha je prikazano na slici 19.

Nakon toga, ponovo je izvršena *random* particija skupa podataka 80:10:10 i dobijeni su skupovi za obučavanje, validaciju i testiranje. Na istom obučavajućem skupu je obučeno stablo odluke i neuronska mreža, a zatim određeni parametri linearne regresije. Nakon toga, na istom skupu su testirane sve tri metode i izračunate PPV i NPV . Dobijeni rezultati su prikazani u tabeli 6 i na slici 20.



Slika 19 – PPV i NPV za validacioni skup u odnosu na broj epohe



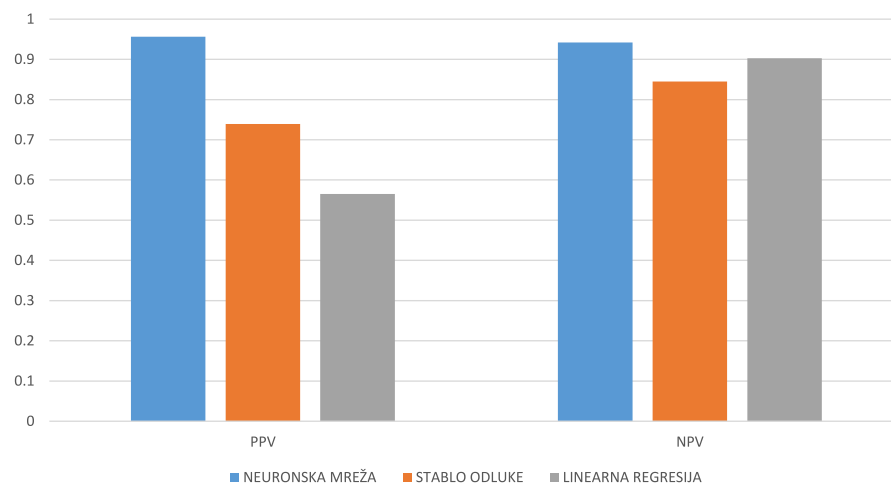
Slika 20 – PPV i NPV za trening, validacioni i test skup

Metoda	PPV_{TR}	NPV_{TR}	PPV_{VL}	NPV_{VL}	PPV_{TS}	NPV_{TS}
$ANN90$	0,9910	0,9894	0,9583	0,9054	0,9710	0,9359
DT	0,8737	0,7953	0,8750	0,7297	0,8768	0,7885
LR	0,7572	0,8157	0,7569	0,7770	0,7101	0,7885

Tabela 6 – PPV i NPV za trening, validacioni i test skup

5.4 Testiranje obučene mreže

Četvrti eksperiment demonstrira primenu neuronske mreže, stabla odluke i linearne regresije. Na 100 nepoznatih primera je testiran rad već obučene metode mašinskog učenja $ANN90$, DT i LR , a u tabeli 12 su prikazni dobijeni odgovori, kao i tačan odgovor. Određene su vrednosti PPV i NPV koje su prikazane u tabeli 7. Na slici 21 su upoređene performanse razmatranih metoda, odvojeno za PPV i NPV .



Slika 21 – PPV i NPV za obučenu mrežu nad test skupom

Metoda	PPV_{TS}	NPV_{TS}
<i>ANN</i>	0,9565	0,9417
<i>DT</i>	0,7391	0,8447
<i>LR</i>	0,5652	0,9029

Tabela 7 – Poređenje PPV i NPV po algoritmima za test skup

6 Funkcionalni opis i arhitektura softverskog sistema

Rezultati prikazani u poglavlju 5 nam daju osnov da implementaciju prikazanu u odeljku 4.6 iskoristimo u stvarnom sistemu za algoritamsku dijagnostiku metaboličkog sindroma. Kao što je razmatrano u poglavlju 2, metabolički sindrom je skup međusobno povezanih metaboličkih poremećaja koji povećavaju rizik za razvoj brojnih hroničnih bolesti, gde se na prvom mestu nalaze kardiometaboličke bolesti – aterosklerotske kardiovaskularne bolesti i dijabetes tipa 2. Pravovremenom dijagnostikom ovog sindroma moguće je uticati na faktore rizika kod pacijenata i smanjiti rizik od razvoja ozbiljnijih oboljenja.

Kao što je pokazano u poglavlju 5, moguće je obučiti veštačku neuronsku mrežu tako da se postigne visok nivo prediktivnih vrednosti (maksimalno dobijeni $PPV = 0,9257$, a maksimalno dobijeni $NPV = 0,9204$). Pri tome, ulazni parametri koji se koriste su izuzetno lako merljivi. Sve potrebne podatke moguće je prikupiti elementarnim pregledom u lekarskoj ordinaciji od strane medicinskih tehničara, ili, što je veoma važno, pacijenti mogu i sami to da urade bez odlaska kod lekara.

Gore navedeno nam daje osnov i motivaciju za implementaciju inteligentnog softverskog sistema za dijagnostiku metaboličkog sindroma, koji opisujemo u ovom poglavlju.

6.1 Konceptualni opis sistema

Opis inteligentnog softverskog sistema za dijagnostiku metaboličkog sindroma dajemo u odnosu na osnovne ciljeve koje očekujemo da ispuni:

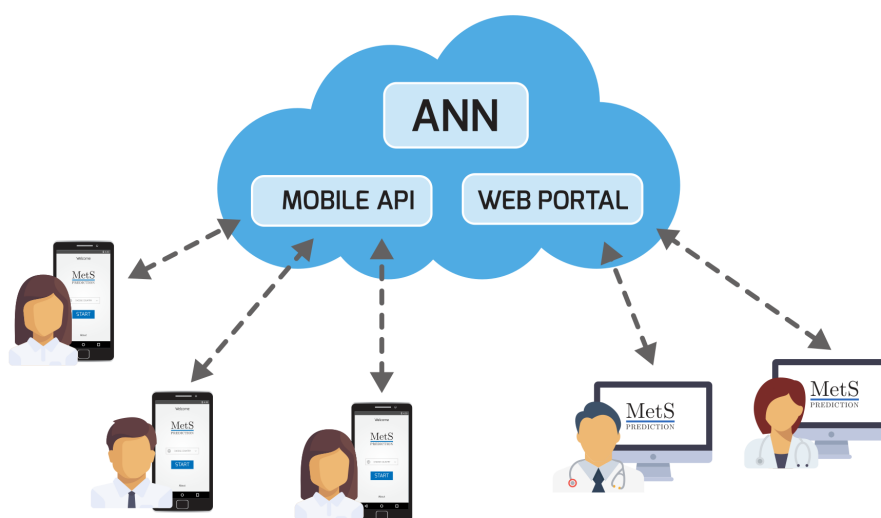
1. Laka dostupnost širokoj populaciji
2. Mogućnost dijagnostike od strane stručnjaka
3. Mogućnost samodijagnostike
4. Globalno rešenje (različiti skupovi podataka po državama)
5. Prikupljanje proširenog skupa podataka za dalja istraživanja
6. Efikasna razmena podataka

Trenutna tehnološka rešenja u vidu *cloud sistema* i *mobilnih aplikacija* nam omogućavaju da rezultate prikazane u poglavlju 5 napravimo dostupnim širokom krugu ljudi.

Predviđeno je da stručne osobe - lekari sistem koriste u svakodnevnom radu za usmeravanje dijagnostičke procedure. Pozitivan rezultat testa daje jaku indikaciju da se urade dalja ispitivanja i da se definitivno potvrdi ili opovrgne prisustvo metaboličkog sindroma kod pacijenta. U suprotnom, ako je test negativan, lekar treba da na osnovu svog iskustva i pregleda pacijenta odluči o tome da li je dalja dijagnostika potrebna.

Pacijentima, s druge strane, sistem omogućava da na jednostavan način, bez odlaska lekaru, utvrde da li postoji mogućnost da imaju metabolički sindrom. Sve što treba jeste da izmere težinu, visinu, obim struka i krvni pritisak i da dobijene podatke unesu u mobilnu aplikaciju koja im prikazuje rezultat testa i preporučuje im dalje korake koje bi trebalo da preduzmu.

Gore opisana funkcionalnost je postignuta implementacijom *web* portala sa naprednim funkcijama sistema za lekare i istraživače, dok je za pacijete predviđena jednostavna i intuitivna mobilna aplikacija. Konceptualna šema sistema data je na slici 22.



Slika 22 – Konceptualni dijagram sistema za dijagnostiku metaboličkog sindroma

Važno je napomenuti da ovako opisan sistem koji koristi algoritamsku dijagnostiku ne može da zameni lekara i stručno određivanje dijagnoze.

Skup podataka korišćen u ovom radu za obuku neuronske mreže je dobijen iz istraživanja na populaciji u Srbiji. Poznato je da metabolički sindrom varira u zavisnosti od geografske regije. Prikupiti kvalitetne podatke za globalnu populaciju je veoma teško.

Zbog toga nije moguće obučiti jedinstvenu neuronsku mrežu za svetsku populaciju. Globalno rešenje koje je predloženo predviđa različite skupove podataka za posebne države, i za njih posebno obučene neuronske mreže.

Performanse obučene neuronske mreže zavise od kvalitetnog trening skupa. Kako bi se omogućila zamena trening skupa boljim ili dodavanje trening skupova za druge države, u odeljku 6.5 je definisan format za uvoz podataka za obuku. Sistem treba da omogući korisniku da testira novi skup podataka i ukoliko je zadovoljan rezultatima, da zameni stari.

Sistem koji je ovde predložen nije zamišljen isključivo kao dijagnostičko sredstvo, nego i kao alat koji će da omogući dalji napredak oblasti. Poznato je da postoje veze između metaboličkog sindroma i drugih parametara kao što su zanimanje, nivo fizičke aktivnosti, dijetetski režim i sl. Tako da, pored svega navedenog, sistem omogućava lekarima i istraživačima da za pojedinačne testove koje su izvodili, nakon sprovedene dalje dijagnostike evidentiraju te dodatne parametre, zajedno sa rezultatima laboratorijskih analiza i konačno postavljenom dijagnozom. Na ovaj način se prikupljaju podaci koji mogu da se upotrebe kao osnova za nova istraživanja. Date podatke je moguće na jednostavan način preuzeti iz sistema u formatu definisanom u odeljku 6.5.

6.2 Web portal za lekare

Pristup osnovnim funkcijama sistema implementiran je u vidu *web* portala. Portal je koncipiran kao dijagnostičko sredstvo i kao platforma za istraživanje metaboličkog sindroma. Oslanja se na *backend* u čijoj osnovi je veštačka neuronska mreža razmatrana u prethodnim poglavljima. Funkcije sistema koje su dostupne kroz portal su:

1. Obučavanje neuronske mreže različitim trening skupovima za različite države
2. Probno obučavanje neuronske mreže za dati trening skup sa prikazom rezultata
3. Predikcija metaboličkog sindroma na osnovu datih parametara (*test*)
4. Unos dodatnih podataka i konačne dijagnoze nakon dobijenih laboratorijskih nalaza
5. Pregled izvedenih testova i eksport podataka

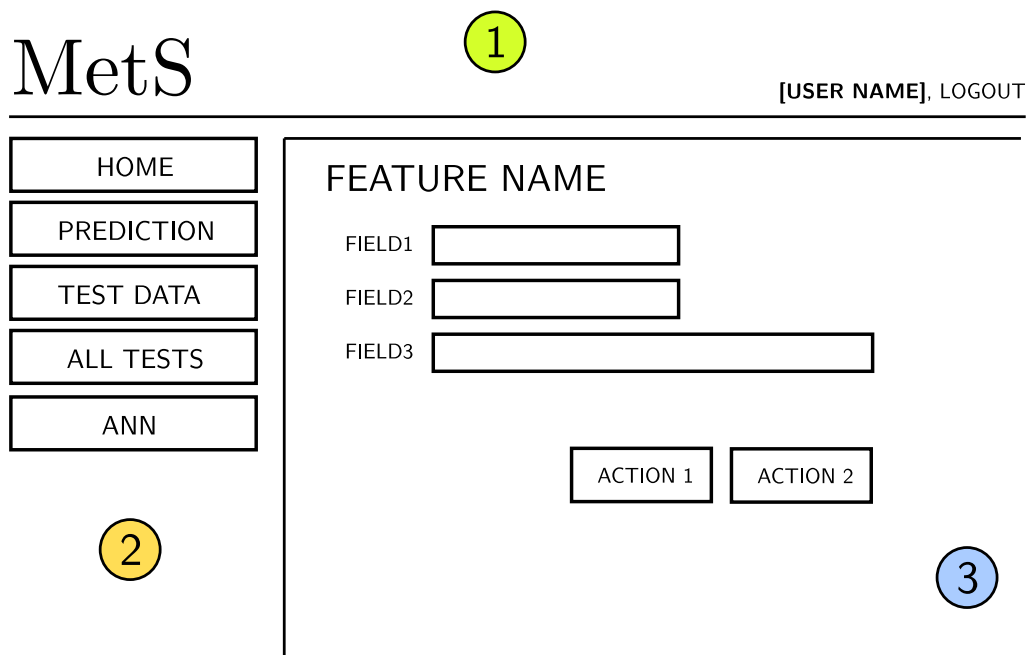
Portal je realizovan kao *single page* aplikacija u jezicima *TypeScript* [74] i *HTML* korišćenjem biblioteke *Angular* [75]. Narednih nekoliko odeljaka su posvećeni detaljnom opisu njegove funkcionalnosti i prikazu njegovog dizajna.

6.2.1 Organizacija korisničkog interfejsa

Dizajn korisničkog interfejsa je standardan za portale ovog tipa i sastoji se iz tri osnovna dela (slika 23):

1. Zaglavlje
2. Navigacioni panel
3. Radni prostor

Zaglavlje (1) sadrži logo aplikacije, ime korisnika i link za odjavu iz sistema. Linkovi koji korisnika vode do pogleda za pojedinačne funkcije portala se nalaze na navigacionom panelu (2). Radni prostor (3) prikazuje elemente potrebne za realizovanje konkretnih funkcija sistema. Interakcija sa korisnikom je intezivnija u ovom delu, gde korisnik unosi potrebne podatke i inicira akcije.

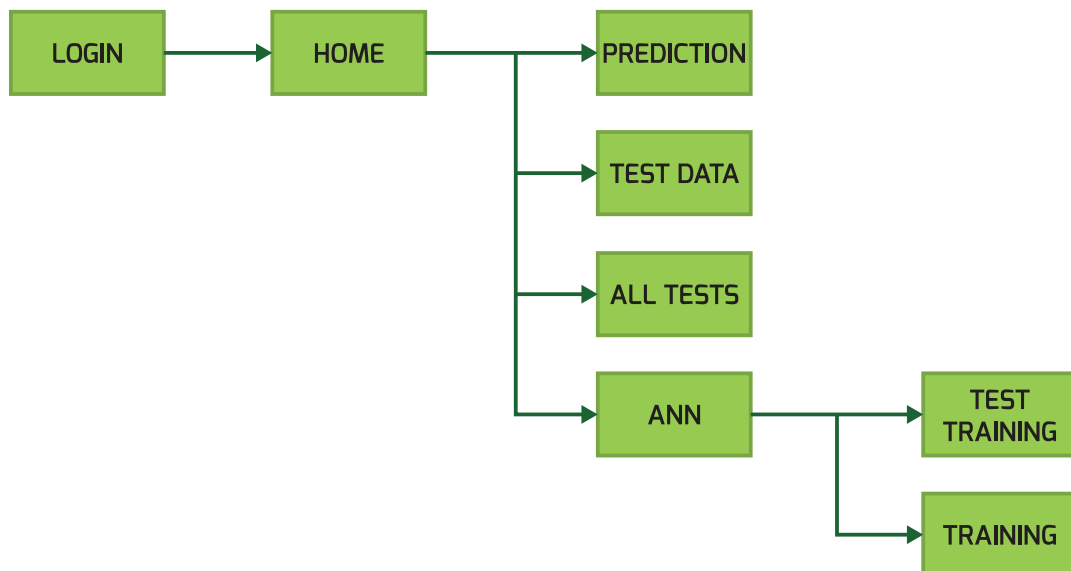


Slika 23 – Koncept korisničkog interfejsa

6.2.2 Hijerarhija web stranica

Na slici 24 je data hijerarhija stranica web portala. Portalu se pristupa isključivo kroz *LOGIN* stranicu nakon izvršene uspješne identifikacije korisnika. Stranica *HOME* sa leve strane sadrži navigacioni panel pomoću koga korisnik pristupa različitim funkcijama portala.

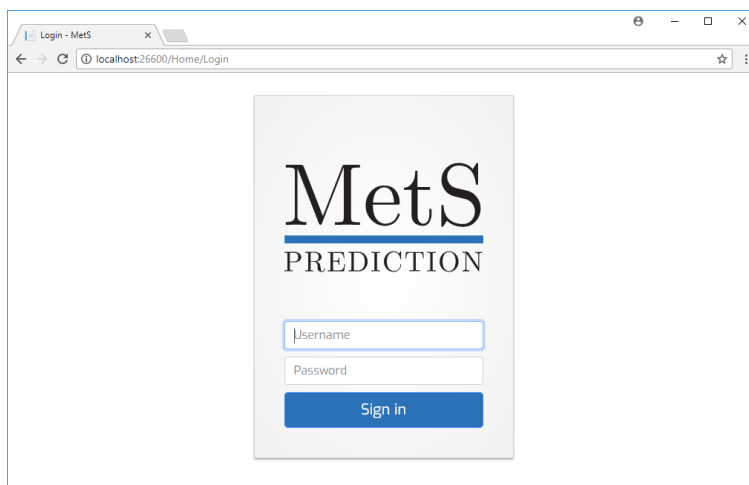
Osim *LOGIN* stranice, sve ostale imaju istu osnovnu strukturu opisanu u prethodnom odeljku. Za raspoređivanje elemenata na stranicama i njihov dizajn upotrebljena je biblioteka *Bootstrap*. U narednih nekoliko odeljaka se upoznajemo sa sadržajem pojedinačnih stranica i funkcijama za koje su zadužene.



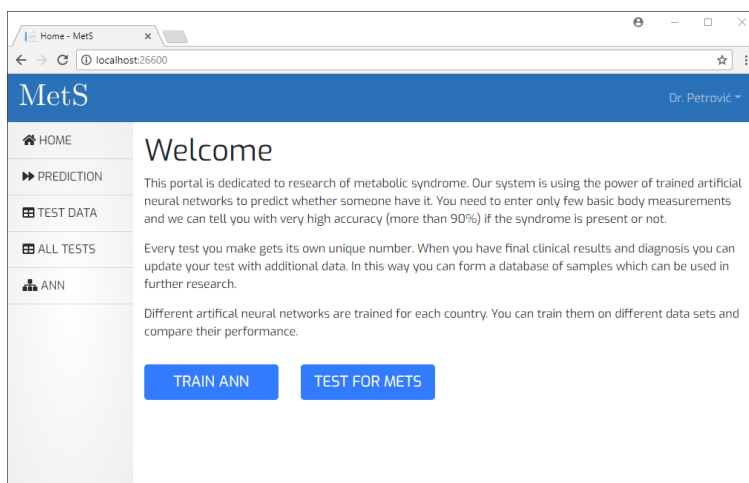
Slika 24 – Hijerarhija pogleda za web portal

6.2.3 Stranice *LOGIN* i *HOME*

Stranica za prijavu u sistem (*LOGIN*) je prvi pogled (slika 25) koji se prikazuje korisniku. Korisnike registruje administrator na zahtev. Kako bi ga sistem identifikovao potrebno je da unese korisničko ime (*user name*) i lozinku (*password*). Nakon uspešne prijave na sistem, korisniku se prikazuje osnovni pogled portala prikazan na slici 26. Na ovom pogledu korisniku je upućen pozdrav i prikazan kratak tekst o tome čemu je posvećen portal i koje su njegove funkcije. Ponuđene su mu prečice do najbitnijih funkcija sistema.



Slika 25 – Web stranica *LOGIN*



Slika 26 – Web stranica *HOME*

6.2.4 Web stranica PREDICTION

Stranica *PREDICTION* prikazana na slici 27 posvećena je predikciji metaboličkog sindroma. Njena osnovna uloga je unos podataka potrebnih za test i prikaz rezultata.

Kao što je razmatrano u prethodnim poglavljima ulazni parametri za neuronsku mrežu su: pol (*GEN*), starosna dob (*AGE*), indeks telesne mase (*BMI*), odnos visine i obima struka (*WHtR*), sistolni krvni pritisak (*SBP*) i dijastolni krvni pritisak (*DBP*). Kako bi se olakšala sama procedura, parametri *BMI* i *WHtR* su zamenjeni jednostavnijim parametrima: telesna visina (*HT*), telesna težina (*WT*) i obim struka (*WC*). Iz ovih parametara jednostavno se izračunavaju *BMI* i *WHtR* po sledećim formulama:

$$BMI = \frac{WT}{HT^2}, WHtR = \frac{WC}{HT}$$

Prilikom unosa, podaci se validiraju i korisnik se upozorava ukoliko nisu u dozvoljenim granicama. Nakon što unese sve podatke, korisnik pritiskom na dugme "GO" inicira predikciju i dobija rezultat testa.

The screenshot shows a web browser window with the URL `localhost:26600/Home/Prediction`. The page title is "MetS" and the user is logged in as "Dr. Petrović". The main content area is titled "Prediction" and contains the following form elements:

- Location: Serbia (dropdown menu)
- Gender: Male (selected), Female
- Age: 42
- Height: 175
- Weight: 101
- Waist circ.: 91.2
- Systolic BP: 135
- Diastolic BP: 110
- Go! button

Below the form, a green box displays the result: "Test result is POSITIVE" and "Test id: 12478".

Slika 27 – Web stranica PREDICTION

6.2.5 Web stranica *PREDICTION*

Na ovoj stranici je korisniku omogućeno da prethodno izvedene testove može da dopuni sa dodatnim podacima. Parametri koji su predviđeni za evidentiranje uključuju rezultate laboratorijskih analiza, dodatne podatke o pacijentu i konačnu klinički određenu dijagnozu metaboličkog sindroma.

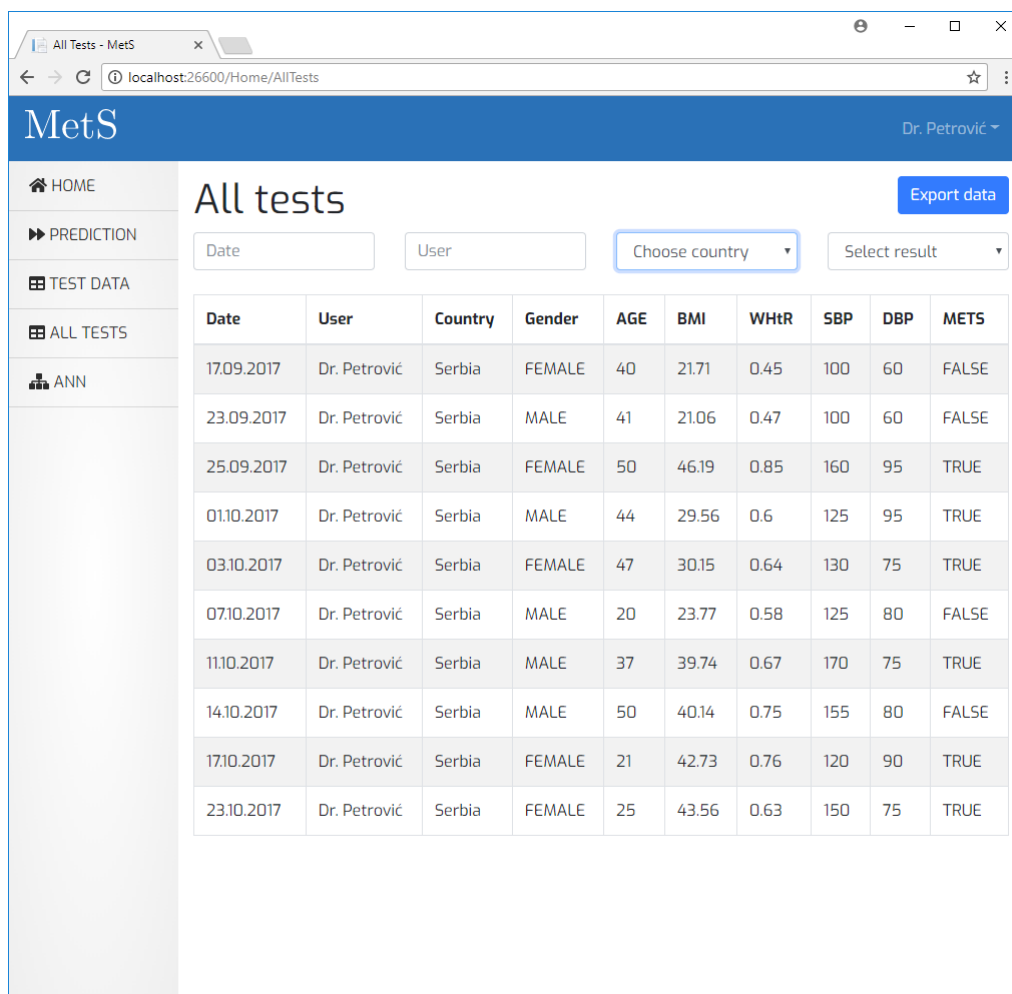
Od laboratorijskih analiza moguće je uneti rezultate za nivo triglicerida, nivo *HDL* holesterola i nivo glukoze u krvi našte. Dodatni opšti podaci o pacijentu zanimljivi sa aspekta metaboličkog sindroma, a koji mogu da se evidentiraju su: zanimanje pacijenta, nivo fizičke aktivnosti i dijeta. Svi ovi podaci se prikupljaju sa ciljem da se s jedne strane proveri tačnost predikcije obučene veštačke neuronske mreže, a s druge strane da se formira novi skup podataka koji može da se upotrebi kao osnova za dalja istraživanja.

The screenshot shows a web browser window with the URL `localhost:26600/Home/TestData`. The page title is "MetS" and the user is logged in as "Dr. Petrović". The left sidebar contains navigation links: HOME, PREDICTION, TEST DATA, ALL TESTS, and ANN. The main content area is titled "Test number" and shows a search bar with "12478" and a "Find" button. Below this, there are two columns of input fields. The left column contains: Country (Serbia), Gender (Male selected), Age (42), Height (175), Weight (101), Waist circ. (121), Systolic BP (135), and Diastolic BP (110). The right column contains: Diet (NATIONAL), Occupation (SEDENTARY), Exercise (NONE), HDL level (0.85), Fasting glucose (4.7), Triglycerides (1.68), and METS diagnosis (TRUE). A blue "Save" button is located below the right column. At the bottom left, a green notification box displays "Test result is POSITIVE" and "Test id: 12478".

Slika 28 – Web stranica *TEST DATA*

6.2.6 Web stranica ALL TESTS

Osnovna uloga ove stranice je da korisniku omogući pregled podataka o urađenim testovima. Podatke je moguće filtrirati i sortirati po različitim parametrima kao što su datum urađenog testa, korisnik, država ili rezultat predikcije. Korisnik može jednostavnim klikom na dugme “Export data” da izveze sve podatke o urađenim testovima u formatu koji je definisan u odeljku 6.5.



The screenshot displays the 'All tests' page in the MetS application. The page includes a navigation sidebar with options: HOME, PREDICTION, TEST DATA, ALL TESTS (selected), and ANN. The main content area shows a table of test results with the following data:

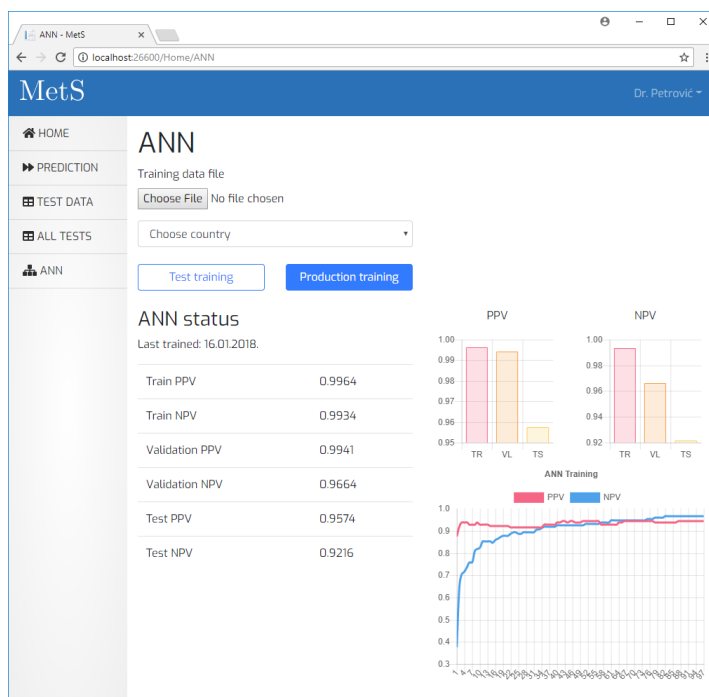
Date	User	Country	Gender	AGE	BMI	WHtR	SBP	DBP	METS
17.09.2017	Dr. Petrović	Serbia	FEMALE	40	21.71	0.45	100	60	FALSE
23.09.2017	Dr. Petrović	Serbia	MALE	41	21.06	0.47	100	60	FALSE
25.09.2017	Dr. Petrović	Serbia	FEMALE	50	46.19	0.85	160	95	TRUE
01.10.2017	Dr. Petrović	Serbia	MALE	44	29.56	0.6	125	95	TRUE
03.10.2017	Dr. Petrović	Serbia	FEMALE	47	30.15	0.64	130	75	TRUE
07.10.2017	Dr. Petrović	Serbia	MALE	20	23.77	0.58	125	80	FALSE
11.10.2017	Dr. Petrović	Serbia	MALE	37	39.74	0.67	170	75	TRUE
14.10.2017	Dr. Petrović	Serbia	MALE	50	40.14	0.75	155	80	FALSE
17.10.2017	Dr. Petrović	Serbia	FEMALE	21	42.73	0.76	120	90	TRUE
23.10.2017	Dr. Petrović	Serbia	FEMALE	25	43.56	0.63	150	75	TRUE

Slika 29 – Web stranica ALL TESTS

6.2.7 Web stranica ANN

Ova stranica je posvećena veštačkoj neuronskoj mreži koja čini centralni deo sistema. Kao što je u prethodnim poglavljima opisano, koristi se veštačka neuronska mreža sa 90 skrivenih neurona, koja se obučava *backpropagation* algoritmom sa bajesovskom regulacijom i prethodnom normalizacijom ulaza.

Ovaj pogled omogućava korisniku da uveze u sistem skup podataka za obuku veštačke neuronske mreže i da nakon toga inicira njen trening. Korisniku je ponuđeno da izvede probnu obuku (klik na dugme "Test training") ili produkcionu obuku (klik na dugme "Production training"). Test obuka će kao rezultat prikazati performanse koje mreža postiže kada se obuči sa datim podacima. Produkciona obuka će zameniti novom trenutnu veštačku neuronsku mrežu koja se koristi za predikciju. U oba slučaja na stranici se prikazuju vrednosti *PPV* i *NPV* za trening, validacioni i test skup podataka. Pored toga, prikazan je grafik obuke neuronske mreže i poređenje *PPV* i *NPV* u zavisnosti od skupa podataka.



Slika 30 – Web stranica ANN

6.3 Mobilna aplikacija za pacijente

Široka upotreba mobilnih uređaja i jednostavan postupak dijagnoze pružaju nam mogućnost da u vidu mobilne aplikaciju namenjene pacijentima dijagnostičku proceduru učinimo dostupnom široj populaciji. Njenom upotrebom, bez posete lekaru, pacijent može sa velikom verovatnoćom da utvrdi da li ima metabolički sindrom, ili ne.

Primenljivost ovakvog rešenja u praksi direktno zavisi od njegove ergonomičnosti. Aplikacija mora biti jednostavna i lako upotrebljiva za običnog korisnika. Prilikom njenog dizajniranja posebnu pažnju treba posvetiti komunikaciji sa korisnikom. Funkcionalnost aplikacije ima veze sa zdravstvenim stanjem neke osobe, što je samo za sebe osetljiva tema.

U skladu sa gore navedenim, na početku je potrebno jasno objasniti funkcionalnost aplikacije i njena ograničenja. Zatim, prilikom unosa parametara neophodnih za dijagnozu smanjiti mogućnost greške. Na kraju, rezultat testa treba prikazati na način da ne izazove nepotrebnu zabrinutost kod pacijenta, i predložiti mu korake koje u skladu sa rezultatom treba da preduzme.

Kako aplikacija nije u mogućnosti da postavi konačnu dijagnozu, tj. da sa stoprocentnom sigurnošću potvrdi ili opovrgne postojanje sindroma, ovu činjenicu treba dodatno jasno naglasiti.

Globalno rešenje podrazumeva lokalizaciju aplikacije na različite jezike. Aplikacija se automatski prilagođava prema jeziku koji se koristi na samom uređaju i prikazuje poruke na jeziku razumljivom korisniku. Prototip prikazan u ovoj doktorskoj disertaciji koristi srpski i engleski jezik.

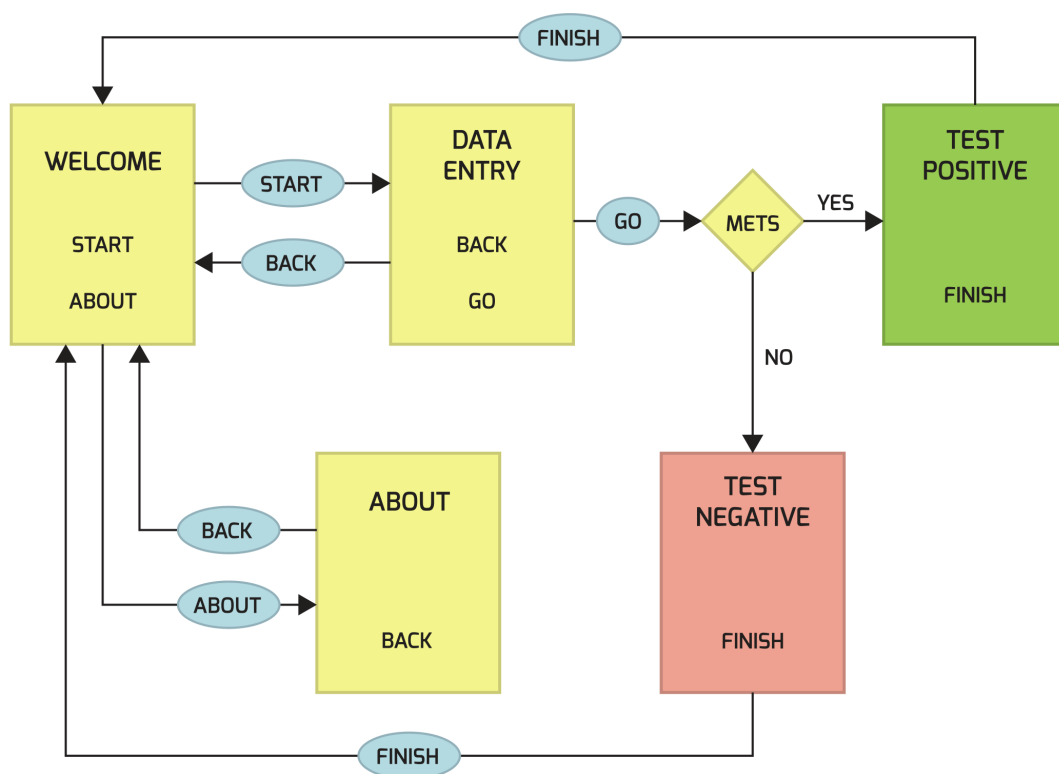
Na osnovu gore razmotrenog, implementirana je mobilna aplikacija za pacijente, čiji su dizajn i funkcionalnost predstavljeni u narednim odeljcima. Aplikacija je implementirana korišćenjem *Xamarin* platforme pomoću biblioteke *Xamarin.Forms* u jeziku *C#*. Kao i *web* portal oslanja se na zajednički *backend* sa kojim komunicira putem *REST API*-ja.

6.3.1 Hijerarhija pogleda sa tranzicijama

Aplikacija se sastoji iz pet pogleda:

- *WELCOME* - početni ekran aplikacije
- *ABOUT* - objašnjenje funkcionalnosti aplikacije
- *DATA ENTRY* - unos parametara potrebnih za predikciju
- *TEST POSITIVE* - prikaz pozitivnog rezultata testa
- *TEST NEGATIVE* - prikaz negativnog rezultata testa

Hijerarhija ovih pogleda sa tranzicijama prikazana je na slici 31, a detaljni opisi i sadržaj, zajedno sa stvarnim slikama dati su u narednim odeljcima.



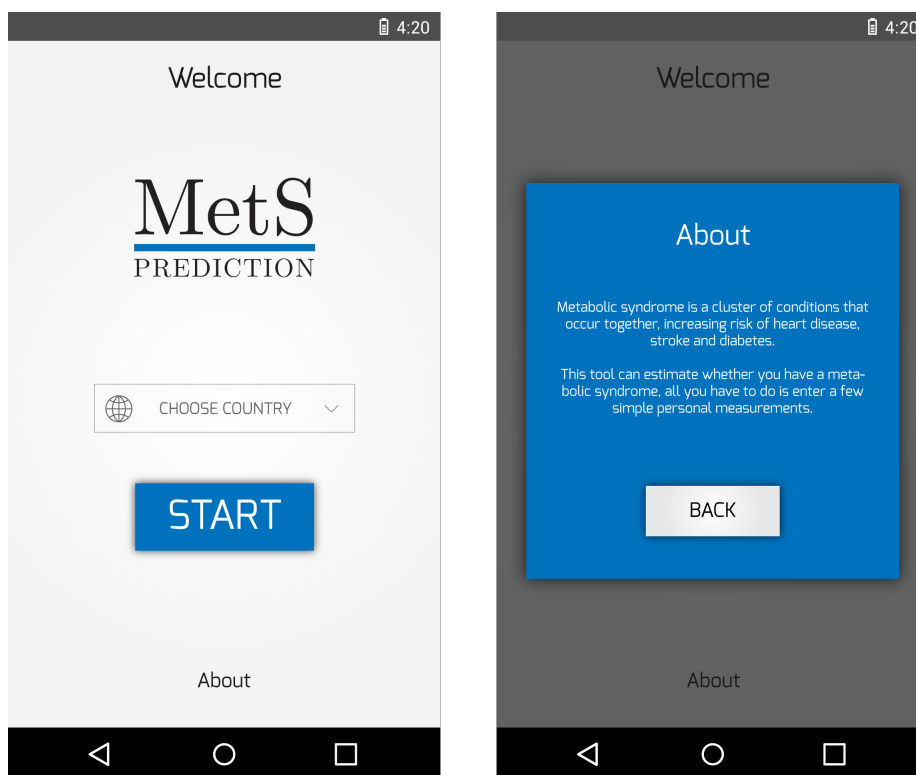
Slika 31 – Pogledi za mobilnu aplikaciju sa tranzicijama

6.3.2 Pogled *WELCOME*

Inicijalni pogled nakon pokretanja aplikacije je dat na slici 32a. Sadrži pozdravnu poruku i logo aplikacije. Korisniku je omogućeno da izabere državu i da se upozna sa funkcionalnošću aplikacije pritiskom na dugme "About". Nakon što korisnik odabere državu, pritiskom na dugme "START" prelazi na unos podataka potrebnih za predikciju.

6.3.3 Pogled *ABOUT*

Pogled sadrži tekst koji na jednostavan način opisuje šta je to metabolički sindrom i objašnjava korisniku funkcionalnost aplikacije. Nakon što korisnik pročita tekst pritiskom na dugme "BACK" vraća se na inicijalni pogled. Metabolički sindrom je predstavljen kao skup zdravstvenih stanja koja se javljaju zajedno i povećavaju rizik od srčanih bolesti, infarkta i dijabetesa. Aplikacija je opisana kao alat koji može da proceni da li neko ima metabolički sindrom ili ne, samo je potrebno uneti jednostavne lične mere.



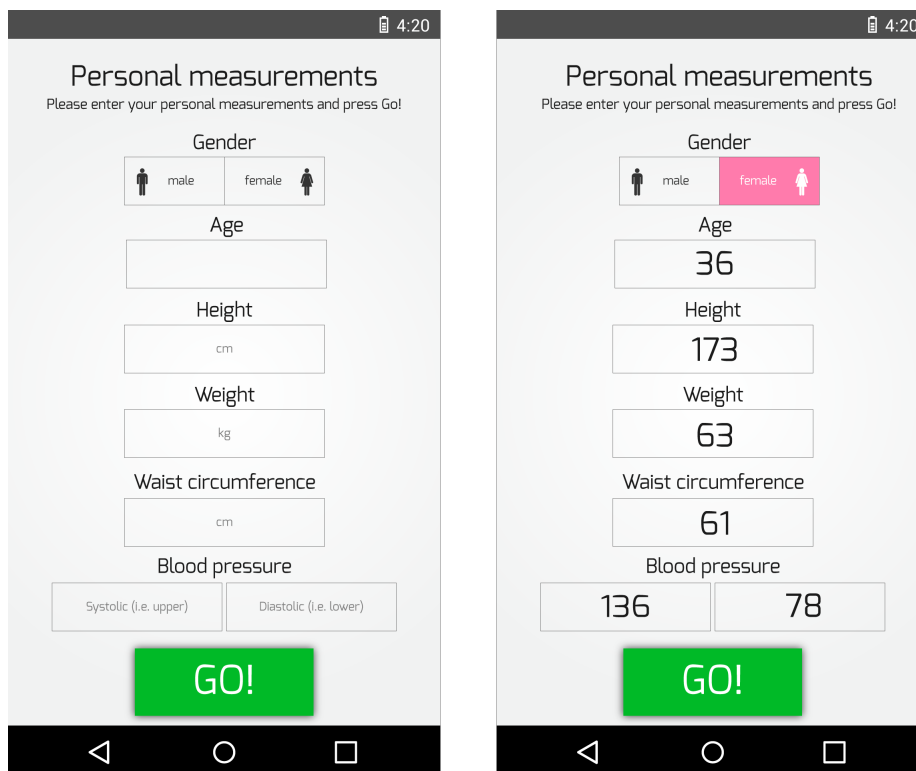
Slika 32 – Pogled *WELCOME* (a) i Pogled *ABOUT* (b)

6.3.4 Pogled DATA ENTRY

Osnovna uloga pogleda prikazanog na slici 33 je unos podataka potrebnih za predikciju metaboličkog sindroma. Isto kao što je urađeno u slučaju *web* portala (odjeljak 6.2.4) parametri koji su korišćeni za obuku neuronske mreže: indeks telesne mase (*BMI*), i odnos visine i obima struka (*WHtR*), zamenjeni su jednostavnijim parametrima: telesna visina (*HT*), telesna težina (*WT*) i obim struka (*WC*). Kao što je u pomenutom odeljku navedeno, iz ovih parametara jednostavno se izračunavaju *BMI* i *WHtR* po formulama:

$$BMI = \frac{WT}{HT^2}, WHtR = \frac{WC}{HT}$$

Prilikom unosa, podaci se validiraju i korisnik se upozorava ukoliko nisu u dozvoljenim granicama. Nakon što unese sve podatke korisnik pritiskom na dugme “GO” inicira predikciju i dobija rezultat testa.



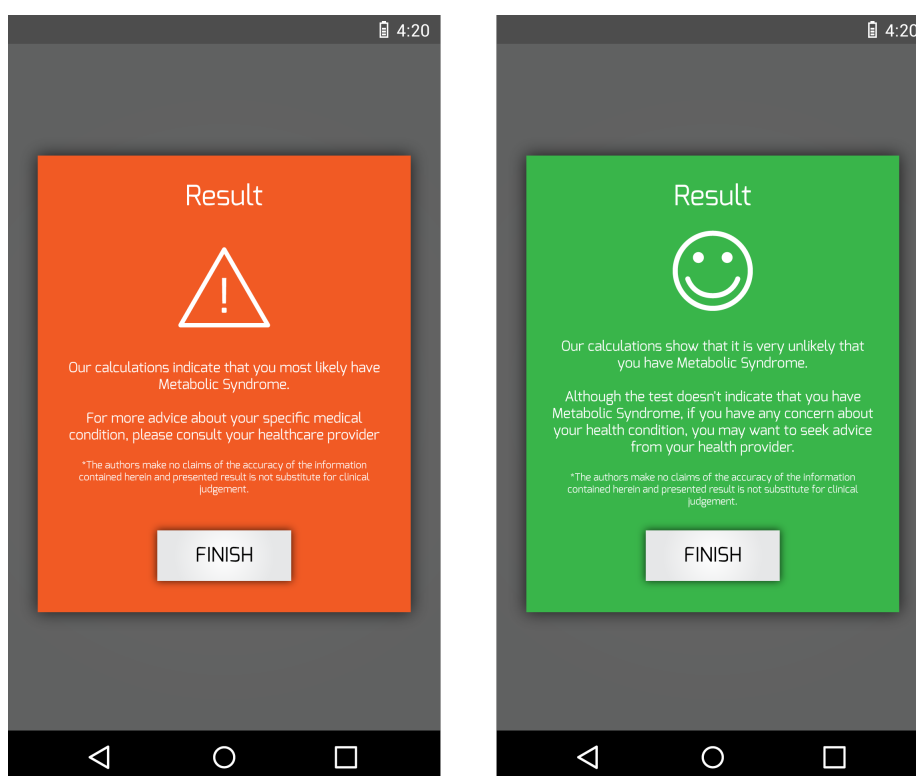
Slika 33 – Pogled DATA ENTRY

6.3.5 Pogled *POSITIVE TEST*

Kada je rezultat predikcije pozitivan, tj. kada osoba čiji su podaci uneti verovatno ima metabolički sindrom, prikazuje se pogled dat na slici 34a. Korisniku je prikazan pogled sa upozoravajućom bojom i odgovarajućim simbolom, kao i tekstom o rezultatu testa.

Pošto ne možemo sa apsolutnom sigurnošću da tvrdimo da osoba ima metabolički sindrom ($PPV \approx 0.92$), saopštavamo da je to vrlo verovatno. Pored toga dajemo joj savet da se javi svom lekaru kako bi proverila svoje zdravstveno stanje. Kako naš sistem nije u mogućnosti da zameni stručno lice, na kraju još jednom napominjemo da ne tvrdimo da je dati rezultat tačan, i da on nije zamena za kliničku procenu.

Nakon pročitanih poruka korisnik se pritiskom na dugme "FINISH" vraća na inicijalni pogled aplikacije.



Slika 34 – Pogledi *POSITIVE TEST* (a) i *NEGATIVE TEST* (b)

6.3.6 Pogled *NEGATIVE TEST*

Kada je rezultat predikcije negativan, tj. kada osoba čiji su podaci uneti verovatno nema metabolički sindrom, prikazuje se pogled dat na slici 34b. Korisniku je prikazan pogled sa umirujućom bojom i simbolom i tekstom o rezultatu testa.

S obzirom da takođe ne možemo sa apsolutnom sigurnošću da tvrdimo da osoba nema metabolički sindrom ($NPV \approx 0.92$), saopštavamo da je to vrlo verovatno. Pored toga dajemo joj savet da ukoliko je zabrinuta za svoje zdravstveno stanje, treba da se javi svome lekaru. Kako naš sistem nije u mogućnosti da zameni stručno lice, na kraju još jednom napominjemo da ne tvrdimo dati rezultat, i da on nije zamena za kliničku procenu.

Nakon pročitanih poruka korisnik se pritiskom na dugme “FINISH” vraća na inicijalni pogled aplikacije.

6.4 Softverska arhitektura sistema

Predloženi sistem je realizovan u obliku standardne više-slojne arhitekture [76] prikazane na slici 35.

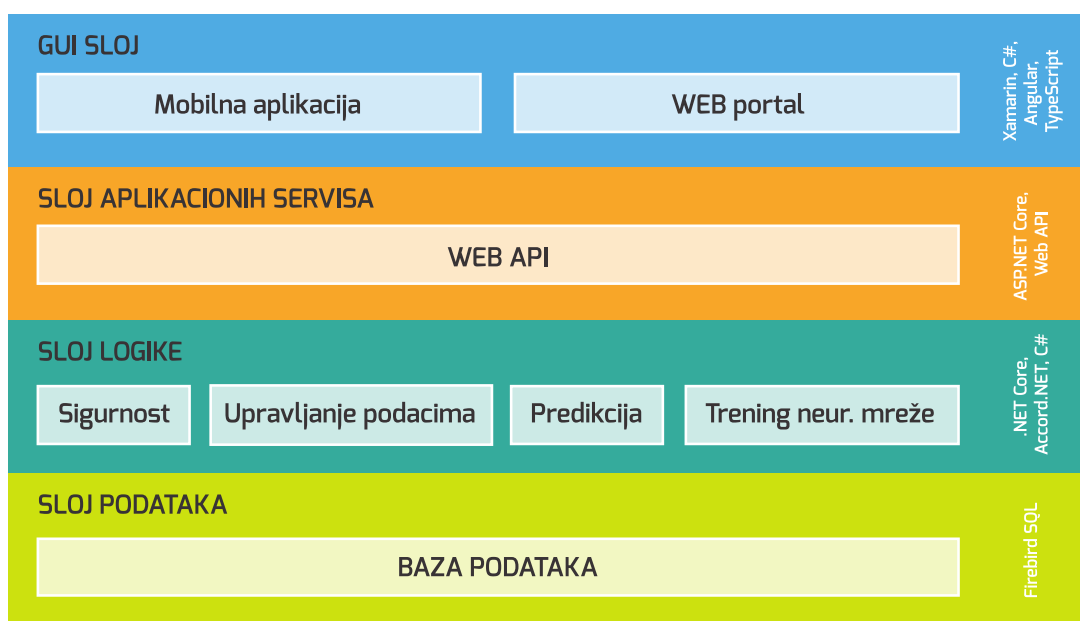
Sloj korisničkog interfejsa (GUI sloj) čine mobilna aplikacija za pacijente i *web* portal. Mobilna aplikacija je implementirana pomoću biblioteke *Xamarin.Forms* [77] u jeziku C#, koja podrazumeva upotrebu *MVVM* [78] (*Model - View - View Model*) paterna i *data-binding*-a. Isti patern je korišćen i za implementaciju *web* portala koji je realizovan kao *single page* aplikacija zasnovana na biblioteci *Angular* i jezicima *TypeScript* i *HTML*. Raspored elemenata na pogledima i osnovni dizajn aplikacije baziran je na *Bootstrap* [79] biblioteci.

Sloj aplikacionih servisa je realizovan u vidu *REST API*-ja [80] koji predstavlja komunikacioni sloj između *GUI* sloja i nižih slojeva arhitekture. Baziran je na *ASP.NET Core* [81] platformi i *Web API*-ju [82].

Sloj logike je zadužen za konkretnu implementaciju funkcija sistema. Pored nezaobilazne sigurnosti sistema u ovom sloju je implementiran deo za upravljanje podacima korišćenjem *Entity Framework Core* [83] biblioteke.

Sloj podataka baziran je na RDBMS-u (*Relational Database Management System*) Firebird SQL [84].

Na kraju, napomenimo da je za razvojno okruženje korišćen *Visual Studio 2017* i da su sve upotrebljene biblioteke *open-source*.



Slika 35 – Softverska arhitektura sistema

6.5 Format za razmenu podataka

Na kraju ovog poglavlja dajemo specifikaciju formata za razmenu podataka. Podaci se razmenjuju u oba smera. Uzorci za probnu i produkcionu obuku se uvoze u sistem, dok je prošireni skup podataka moguće izvesti iz sistema. Za osnovni format izabran je veoma jednostavan *CSV (Comma Separated Value)* format, koji je široko prihvaćen i podržan. Specifikacija za ovaj format data je u *RFC 4180* [85].

Podrazumeva se da je prvi red u datoteci sa podacima zaglavlje sa nazivima parametara koji su uključeni. Podaci koji se prosleđuju sistemu ne moraju da sadrže sve podržane kolone iz tabele. Za obuku veštačke neuronske mreže dovoljno je specificirati jedan od sledeća dva skupa parametara:

1. Osnovni parametri sa dijagnozom (*GEN, AGE, BMI, WHTR, SBP, DBP, METS*)
2. Prošireni parametri bez dijagnoze (*GEN, AGE, WT, HT, WC, SBP, DBP, TG, FPG, HDL*)

U drugom slučaju sistem će sam odrediti dijagnozu *METS* na osnovu implementirane *IDF* definicije metaboličkog sindroma date u 4.3.

Prilikom eksporta podataka iz sistema, koriste se sve definisane kolone.

Parametar	Tip	Opis
<i>GEN</i>	INTEGER	Pol (<i>Gender</i>), 1 - muški, 2 - ženski
<i>AGE</i>	INTEGER	Starosna dob (<i>Age</i>)
<i>WT</i>	DECIMAL	Telesna težina (<i>Body weight</i>)
<i>HT</i>	DECIMAL	Telesna visina (<i>Body height</i>)
<i>WC</i>	DECIMAL	Obim struka (<i>Waist Circumference</i>)
<i>BMI</i>	DECIMAL	Indeks telesne mase (<i>Body Mass Index</i>)
<i>WHTR</i>	DECIMAL	Odnos obima struka i visine (<i>Waist circumference to height ratio</i>)
<i>SBP</i>	INTEGER	Sistolni krvni pritisak (<i>Systolic Blood Pressure</i>)
<i>DBP</i>	INTEGER	Dijastolni krvni pritisak (<i>Diastolic Blood Pressure</i>)
<i>TG</i>	DECIMAL	Nivo triglicerida (<i>Triglycerides level</i>)
<i>FPG</i>	DECIMAL	Nivo glukoze našte (<i>Fasting Plasma Glucose</i>)
<i>HDL</i>	DECIMAL	HDL holesterol (HDL cholesterol)
<i>METS</i>	INTEGER	Klinička dijagnoza 0 - sindrom nije prisutan, 1 - sindrom prisutan
<i>METSP</i>	INTEGER	Procenjena dijagnoza 0 - sindrom nije prisutan, 1 - sindrom prisutan
<i>OCCP</i>	STRING[100]	Zanimanje (<i>Occupation</i>)
<i>EXCR</i>	STRING[100]	Fizička aktivnost (<i>Excercise</i>)
<i>DIET</i>	STRING[100]	Dijeta (<i>Diet</i>)

Tabela 8 – Format za razmenu podataka

7 Zaključak

U ovoj doktorskoj disertaciji je razmotren problem algoritamske dijagnostike metaboličkog sindroma na osnovu lako merljivih parametara: pol (GEN), starosna dob (AGE), indeks telesne mase (BMI), odnos obima struka i visine ($WHtR$), sistolni krvni pritisak (SBP) i dijastolni krvni pritisak (DBP). Istraživanjem su dobijeni sledeći konkretni zaključci:

- Između dijagnoze metaboličkog sindroma i lako dostupnih parametara (pol, starost, indeks telesne mase (BMI), odnos obima struka i telesne visine ($WHtR$), sistolni i dijastolni krvni pritisak) postoji veoma složena matematička relacija i najbolje sredstvo za njeno otkrivanje jesu metode veštačke inteligencije (neuronske mreže, stabla odluke i sl.).
- Poređenjem rezultata, utvrđeno je da veštačke neuronske mreže tačnije dijagnostikuju prisustvo metaboličkog sindroma u odnosu na stabla odluke i linearnu regresiju.
- Shodno tome, ispitane su jednoslojne *feed-forward* veštačke neuronske mreže sa 1-100 skrivenih neurona i maksimum pozitivne prediktivne tačnosti $PPV = 0.9257$ je postignut sa 94 skrivena neurona, a maksimum negativne prediktivne tačnosti $NPV = 0.9204$ sa 97 neurona.
- Do $HN \approx 85$ vrednosti PPV_{TS} i NPV_{TS} imaju trend rasta, nakon čega ulaze u saturaciju i imaju zanemarljiv rast, što znači da se ostvaruje približna tačnost predikcije i sa manjim brojem skrivenih neurona, zbog čega smo usvojili da u daljim eksperimentima koristimo $HN = 90$.
- Primenom *bajesovske regularizacije* sa prethodnom normalizacijom ulaza postiže se bolja generalizacija veštačke neuronske mreže za problem predikcije metaboličkog sindroma.

Na osnovu prethodnih zaključaka, u doktorskoj disertaciji je definisan i implementiran inteligentni softverski sistem koji u praksi omogućava jednostavnu i ekonomski povoljnu dijagnostiku metaboličkog sindroma. Mogućnost da se dijagnoza metaboličkog sindroma može predvideti u svakodnevnoj kliničkoj praksi ima veliki značaj za zdravlje cele populacije.

Buduća istraživanja na ovu temu mogu da se usmere na ispitivanje mogućnosti primene i drugih metoda mašinskog učenja, kao i drugačijih arhitektura i algoritama obuke koji bi mogli dodatno da povećaju tačnost predikcije.

Razvijeni softverski sistem bi trebalo dalje unapređivati u skladu sa potrebama prakse i nauke. Takođe, bilo bi značajno da se prikazano rešenje primeni na domaćem i ako je to moguće, međunarodnom nivou, kako bi se formirao značajan skup podataka za nova istraživanja u oblasti.

Literatura

- [1] Mottillo S, Filion KB, and Genest J. The metabolic syndrome and cardiovascular risk a systematic review and meta-analysis. *J Am Coll Cardiol*, 56:1113–1132, 2010.
- [2] Ford ES, Li C, and Sattar N. Metabolic syndrome and incident diabetes: current state of the evidence. *Diabetes Care*, 31:1898–1904, 2008.
- [3] Global health risks: mortality and burden of disease attributable to selected major risks. Technical report, WHO, Geneva, Dec 2009.
- [4] Strategija za prevenciju i kontrolu hroničnih nezaraznih bolesti Republike Srbije. Technical report, Vlada Republike Srbije, 2006.
- [5] Hivert MF, Grant RW, Shrader P, and Meigs JB. Identifying primary care patients at risk for future diabetes and cardiovascular disease using electronic health records. *BMC Health Serv Res.*, 2009.
- [6] Chen J, Muntner P, and Hamm LL. The metabolic syndrome and chronic kidney disease in us adults. *Ann Intern Med*, 140:167–174, 2004.
- [7] Esposito K, Chiodini P, Colao A, Lenzi A, and Giugliano D. Metabolic syndrome and risk of cancer: a systematic review and metaanalysis. *Diabetes Care*, 35:2402–2411, 2012.
- [8] Rao DP, Dai S, Lagace C, and Krewski D. Metabolic syndrome and chronic disease. *Chronic Diseases and Injuries in Canada*, 2014.
- [9] World Health Organization. Definition, diagnosis and classification of diabetes mellitus and its complications. *Report of a WHO consultation*, 1999.
- [10] Balkau B and Charles MA. Comment on the provisional report from the who consultation. european group for the study of insulin resistance (egir). *Diabet Med*, 16(5):442–443, May 1999.
- [11] Expert Panel on Detection Evaluation and Treatment of High Blood Cholesterol in Adults. Executive summary of the third report of the national cholesterol education program (ncep) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii). *JAMA*, 285(19):2486–2497, 2001.

- [12] International Diabetes Federation. The idf consensus worldwide definition of the metabolic syndrome. 2006.
- [13] Wang X, Yang F, Bots ML, Guo WY, Zhao B, Hoes AW, and Vaartjes I. Prevalence of the metabolic syndrome among employees in northeast china. *Chin Med J*, 128:1989–1993, 2015.
- [14] Škrbić D. *METABOLIČKI SINDROM KOD PACIJENATA SA HRONIČNOM OP-STRUKTIVNOM BOLESTI PLUĆA I BRONHIEKTAZIJAMA*. PhD thesis, Univerzitet u Novom Sadu, Medicinski fakultet, 2015.
- [15] Alberti KG and Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. part 1: diagnosis and classification of diabetes mellitus provisional report of a who consultation. *Diabet Med*, 15(7):539–553, May 1998.
- [16] Appel SJ, Jones ED, and Kennedy-Malone L. Central obesity and the metabolic syndrome: implications for primary care providers. *J Am Acad Nurse Pract*, 16(8):335–342, 2004.
- [17] Rutter MK, Meigs JB, Sullivan LM, D’Agostino RB, and Wilson PW. Insulin resistance, the metabolic syndrome, and incident cardiovascular events in the framingham offspring study. *Diabetes*, 54(11):3252–3257, 2005.
- [18] Berg AH and Scherer PE. Adipose tissue, inflammation, and cardiovascular disease. *Circ Res*, 96:939–949, 2005.
- [19] Misra A and Vikram NK. Clinical and pathophysiological consequences of abdominal adiposity and abdominal adipose tissue depots. *Nutrition*, 19:457–466, 2003.
- [20] Stokić E, Tomić-Naglić D, Derić M, and Jorga J. Therapeutic options for treatment of cardiometabolic risk. *Med Pregl*, 62 Suppl 3(54–58), 2009.
- [21] Kupusinac A, Doroslovački R, Malbaški D, Srdić B, and Stokić E. A primary estimation of the cardiometabolic risk by using artificial neural networks. *Comput Biol Med*, 43(6):751–757, 2013.
- [22] Rosolova H and Nussbaumerova B. Cardio-metabolic risk prediction should be superior to cardiovascular risk assessment in primary prevention of cardiovascular diseases. *EPMA Journal*, 2:15–26, 2011.

- [23] World Health Organization. Obesity: preventing and managing the global epidemic. report of a who consultation. *World Health Organ Tech Rep Ser*, 894:1–253, 2000.
- [24] Ashwell M, Gunn P, and Gibson S. Waist-to height ratio is a better screening tool than waist circumference and bmi for adult cardiometabolic risk factors: systematic review and meta-analysis. *Obes Rev*, 13(3):275–286, Mar 2012.
- [25] Kupusinac A, Stokić E, and Srdić B. Determination of whtr limit for predicting hyperglycemia in obese persons by using artificial neural networks. *TEM ž*, 1:270–272, 2012.
- [26] Grujić V, Martinov-Cvejin M, Ač-Nikolić E, and Nićiforović Šurković O. Epidemiologija gojaznosti odraslog stanovništva vojvodine. *Med Pregl*, 58(5-6):292–295, 2005.
- [27] Friedewald WT, Levy RI, and Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem*, 18(6):499–502, Jun 1972.
- [28] Ivanović D, Kupusinac A, Stokić E, Doroslovački R, and Ivetić D. Ann prediction of metabolic syndrome: a complex puzzle that will be completed. *Journal of medical systems*, 40(12):264, December 2016.
- [29] Flach P. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press, New York, NY, USA, 2012.
- [30] Mitchell TM. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.
- [31] Kononenko I and Kukar M. Machine learning for medical diagnosis. *Proc. Workshop on Computer-Aided Data Analysis in Medicine*, November 1995.
- [32] Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, Bellazzi R, and Abu-Hanna A. The coming of age of artificial intelligence in medicine. *Artificial Intelligence in Medicine*, 46(1):5 – 17, 2009. Artificial Intelligence in Medicine AIME' 07.
- [33] Foster KR, Koprowski R, and Skufca JD. Machine learning, medical diagnosis, and biomedical engineering research - commentary. *BioMedical Engineering OnLine*, 13(1):94, Jul 2014.
- [34] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, and Wang

- Yongjun. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2017.
- [35] Vinay A, Shekhar VS, Rituparna J, Aggrawal T, Balasubramanya Murthy KN, and Natarajan S. Cloud based big data analytics framework for face recognition in social networks using machine learning. *Procedia Computer Science*, 50:623 – 630, 2015. Big Data, Cloud and Computing Challenges.
- [36] Gondhi NK and Gupta A. Survey on machine learning based scheduling in cloud computing. In *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, ISMSI '17, pages 57–61, New York, NY, USA, 2017. ACM.
- [37] Li KL, Gibson C, Ho DGE, Zhou Q, Kim J, Buhisi O, Brown DE, and Gerber MS. Assessment of machine learning algorithms in cloud computing frameworks. *2013 IEEE Systems and Information Engineering Design Symposium*, pages 98–103, 2013.
- [38] Augustin Tataru. *Metrics for Evaluating Machine Learning Cloud Services*. PhD thesis, 2017.
- [39] Rao RB, Bi J, Fung G, Salganicoff M, Obuchowski N, and Naidich D. Lungcad: A clinically approved, machine learning system for lung cancer detection. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 1033–1037, New York, NY, USA, 2007. ACM.
- [40] Chang CL and Hsu MY. The study that applies artificial intelligence and logistic regression for assistance in differential diagnostic of pancreatic cancer. *Expert Systems with Applications*, 36(7):10663 – 10672, 2009.
- [41] Chiu JS, Wang YF, Su YC, Wei LH, Liao JG, and Li YC. Artificial neural network to predict skeletal metastasis in patients with prostate cancer. *Journal of Medical Systems*, 33(2):91, May 2008.
- [42] Catto JWF, Linkens DA, Abbod MF, Chen M, Burton JL, Feeley KM, and Hamdy FC. Artificial intelligence in predicting bladder cancer outcome. *Clinical Cancer Research*, 9(11):4172–4177, 2003.
- [43] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, and Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, October 2000. Evaluation Studies.

- [44] Gertych A, Ing N, Ma Z, Fuchs TJ, Salman S, Mohanty S, Bhele S, Velásquez-Vacca A, Amin MB, and Knudsen BS. Machine learning approaches to analyze histological images of tissues from radical prostatectomies. *Computerized Medical Imaging and Graphics*, 46:197 – 208, 2015. Information Technologies in Biomedicine.
- [45] Narula S, Shameer K, Omar AMS, Dudley JT, and Sengupta PP. Machine-learning algorithms to automate morphological and functional assessments in 2d echocardiography. *Journal of the American College of Cardiology*, 68(21):2287 – 2295, 2016.
- [46] Seera M, Lim CP, Liew WS, Lim E, and Loo CK. Classification of electrocardiogram and auscultatory blood pressure signals using machine learning models. *Expert Systems with Applications*, 42(7):3643 – 3652, 2015.
- [47] Li Q, Rajagopalan C, and Clifford GD. A machine learning approach to multi-level ecg signal quality classification. *Computer Methods and Programs in Biomedicine*, 117(3):435 – 447, 2014.
- [48] Clayton RH, Murray A, and Campbell RWF. Recognition of ventricular fibrillation using neural networks. *Medical and Biological Engineering and Computing*, 32(2):217–220, Mar 1994.
- [49] Liu N, Koh ZX, Chua ECP, Tan LML, Lin Z, Mirza B, and Ong MEH. Risk scoring for prediction of acute cardiac complications from imbalanced clinical data. *IEEE Journal of Biomedical and Health Informatics*, 18(6):1894–1902, Nov 2014.
- [50] Lee BJ and Kim JY. Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. *IEEE Journal of Biomedical and Health Informatics*, 20(1):39–46, Jan 2016.
- [51] Choi S, Jae Kim W, Keun Yoo T, Park JS, Won Chung J, Lee YH, Kang ES, and Kim DW. Screening for prediabetes using machine learning models. 2014:618976, 07 2014.
- [52] Sarwar A and Sharma V. Comparative analysis of machine learning techniques in prognosis of type ii diabetes. *AI & SOCIETY*, 29(1):123–129, Feb 2014.
- [53] Yu W, Liu T, Valdez R, Gwinn M, and Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inform Decis Mak*, 10:16–16, Mar 2010. 1472-6947-10-16[PII].

- [54] Hirose H, Takayama T, Hozawa S, Hibi T, and Saito I. Prediction of metabolic syndrome using artificial neural network system based on clinical data including insulin resistance index and serum adiponectin. *Computers in Biology and Medicine*, 41(11):1051 – 1056, 2011.
- [55] Worachartcheewan Ak, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V, and Nantasenamat C. Predicting metabolic syndrome using the random forest method. page 581501, 06 2015.
- [56] Miller B and Fridline M. Development and validation of metabolic syndrome prediction and classification-pathways using decision trees. *Journal of Metabolic Syndrome*, 4(1):-, 2015.
- [57] Karimi-Alavijeh F, Jalili S, and Sadeghi M. Predicting metabolic syndrome using decision tree and support vector machine methods. *ARYA Atheroscler*, 12(3):146–152, May 2016. ARYA-12-146[PII].
- [58] Abu-Mostafa YS, Magdon-Ismail M, and Lin HT. *Learning From Data*. AMLBook, 2012.
- [59] Hoeffding W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
- [60] Baxt WG. Application of artificial neural networks to clinical medicine. *Lancet*, 346:1135–1138, 1995.
- [61] Anselmino M, Malmberg K, Rydén L, and Öhrvik J. A gluco-metabolic risk index with cardiovascular risk stratification potential in patients with coronary artery disease. *Diabetes & Vascular Disease Research*, 6(2):62–70, 2009.
- [62] Patil SB and Kumaraswamy YS. Intelligent and effective heart attack prediction system using data mining and artificial neural network. *European Journal of Scientific Research*, 31(4):642–656, 2009.
- [63] Lin CC, Bai YM, Chen JY, Hwang TJ, Chen TT, Chiu HW, and Li YC. Easy and low-cost identification of metabolic syndrome in patients treated with second-generation antipsychotics: Artificial neural network and logistic regression models. *J Clin Psychiatry*, 71(3):225–234, 2010.
- [64] Microsoft. *C# Guide*. <https://docs.microsoft.com/en-us/dotnet/csharp/index>.
- [65] Michael McMillan. *Data Structures and Algorithms Using C#*. Cambridge University Press, 2007.

- [66] *Accord.NET*. <http://accord-framework.net/>.
- [67] Mathias Brandewinder. *Machine Learning Projects for .NET Developers*. APress.
- [68] Shanker M, Hu MY, and Hung MS. Effect of data standardization on neural network training. *Elsevier*, 24:385–397, 1996.
- [69] MacKay DJC. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.
- [70] Saini LM. Peak load forecasting using bayesian regularization, resilient and adaptive backpropagation learning based artificial neural networks. *Electric Power Systems Research*, 78(7):1302 – 1310, 2008.
- [71] Lauret P, Fock E, Randrianarivony RN, and Manicom-Ramsamy JF. Bayesian neural network approach to short time load forecasting. *Energy Conversion and Management*, 49(5):1156 – 1166, 2008.
- [72] Ticknor JL. A bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications*, 40(14):5501 – 5506, 2013.
- [73] Bruneau P and NR McElroy. logd 7.4 modeling using bayesian regularized neural networks. assessment and correction of the errors of prediction. 46:1379–87, 05 2006.
- [74] *TypeScript*. <https://www.typescriptlang.org/>.
- [75] *Angular*. <https://angularjs.org/>.
- [76] Fowler M. *Patterns of Enterprise Application Architecture*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.
- [77] Hancock E. *Xamarin Forms for IT Students*. CreateSpace Independent Publishing Platform, USA, 2016.
- [78] Smith J. *Advanced Mvvm*. Josh Smith, 2013.
- [79] Oliver HA. *An Introduction to Twitter Bootstrap 3*. CreateSpace Independent Publishing Platform, USA, 2015.
- [80] Richardson L, Amundsen M, and Ruby S. *RESTful Web APIs*. O’Reilly Media, Inc., 2013.
- [81] Chambers J, Paquette D, and Timms S. *ASP.NET Core Application Development: Building an Application in Four Sprints*. Microsoft Press, 1st edition, 2016.

- [82] Kanjilal J. *ASP.NET Web API: Build RESTful Web Applications and Services on the .NET Framework*. Packt Publishing, 2013.
- [83] Troelsen A and Japikse P. Introducing entity framework core, 11 2017.
- [84] *Firebird SQL*. <https://www.firebirdsql.org/en/start/>.
- [85] *Common Format and MIME Type for Comma-Separated Values (CSV) Files*. <https://tools.ietf.org/html/rfc4180>.

A Rezultati eksperimenta br. 1

HN	PPV _{TR}	NPV _{TR}	PPV _{VL}	NPV _{VL}	PPV _{TS}	NPV _{TS}
1	0,8580	0,7327	0,8526	0,7302	0,8573	0,7251
2	0,8673	0,7694	0,8588	0,7654	0,8626	0,7595
3	0,8695	0,7881	0,8626	0,7846	0,8652	0,7821
4	0,8747	0,7975	0,8666	0,7868	0,8724	0,7970
5	0,8760	0,8045	0,8698	0,7917	0,8648	0,7920
6	0,8805	0,8108	0,8703	0,8026	0,8661	0,7976
7	0,8836	0,8199	0,8661	0,8093	0,8644	0,8062
8	0,8860	0,8248	0,8681	0,8069	0,8683	0,8044
9	0,8861	0,8270	0,8688	0,8193	0,8656	0,8090
10	0,8907	0,8321	0,8692	0,8122	0,8683	0,8109
11	0,8939	0,8374	0,8681	0,8158	0,8666	0,8149
12	0,8943	0,8416	0,8681	0,8134	0,8696	0,8218
13	0,8968	0,8454	0,8685	0,8189	0,8657	0,8204
14	0,9006	0,8491	0,8617	0,8183	0,8637	0,8185
15	0,9028	0,8585	0,8635	0,8268	0,8653	0,8273
16	0,9026	0,8581	0,8671	0,8280	0,8624	0,8268
17	0,9045	0,8636	0,8688	0,8309	0,8686	0,8328
18	0,9123	0,8709	0,8650	0,8335	0,8684	0,8373
19	0,9138	0,8744	0,8728	0,8341	0,8657	0,8382
20	0,9159	0,8783	0,8712	0,8330	0,8648	0,8319
21	0,9171	0,8821	0,8670	0,8373	0,8646	0,8368
22	0,9204	0,8869	0,8688	0,8422	0,8723	0,8437
23	0,9242	0,8914	0,8708	0,8473	0,8774	0,8388
24	0,9252	0,8940	0,8701	0,8458	0,8724	0,8522
25	0,9283	0,8978	0,8699	0,8486	0,8753	0,8446
26	0,9300	0,9029	0,8736	0,8530	0,8778	0,8512
27	0,9299	0,9008	0,8687	0,8506	0,8727	0,8477
28	0,9370	0,9122	0,8716	0,8558	0,8788	0,8593
29	0,9373	0,9113	0,8761	0,8517	0,8813	0,8572
30	0,9385	0,9121	0,8791	0,8531	0,8746	0,8548
31	0,9432	0,9184	0,8824	0,8580	0,8814	0,8591
32	0,9452	0,9251	0,8817	0,8688	0,8777	0,8665
33	0,9488	0,9276	0,8793	0,8657	0,8744	0,8667
34	0,9486	0,9249	0,8845	0,8645	0,8824	0,8620
35	0,9513	0,9310	0,8845	0,8714	0,8790	0,8668
36	0,9533	0,9314	0,8822	0,8711	0,8843	0,8663
37	0,9580	0,9382	0,8896	0,8749	0,8915	0,8735
38	0,9593	0,9368	0,8878	0,8667	0,8849	0,8736

HN	PPV _{TR}	NPV _{TR}	PPV _{VL}	NPV _{VL}	PPV _{TS}	NPV _{TS}
39	0,9584	0,9398	0,8909	0,8720	0,8926	0,8726
40	0,9614	0,9436	0,8935	0,8780	0,8946	0,8795
41	0,9614	0,9451	0,8893	0,8786	0,8926	0,8789
42	0,9669	0,9490	0,8945	0,8770	0,8951	0,8776
43	0,9671	0,9528	0,8974	0,8818	0,8966	0,8818
44	0,9649	0,9493	0,8932	0,8858	0,8920	0,8791
45	0,9665	0,9503	0,8946	0,8790	0,8949	0,8857
46	0,9690	0,9563	0,8986	0,8859	0,8989	0,8831
47	0,9690	0,9548	0,8980	0,8921	0,8997	0,8860
48	0,9693	0,9536	0,8996	0,8824	0,8998	0,8871
49	0,9701	0,9560	0,8992	0,8850	0,9023	0,8892
50	0,9718	0,9589	0,9002	0,8918	0,8998	0,8904
51	0,9746	0,9602	0,9054	0,8870	0,9006	0,8929
52	0,9763	0,9637	0,9062	0,8935	0,9079	0,8975
53	0,9738	0,9615	0,9003	0,8911	0,9070	0,8923
54	0,9761	0,9645	0,9068	0,8902	0,9045	0,8945
55	0,9750	0,9667	0,9040	0,8985	0,9020	0,8965
56	0,9756	0,9644	0,9043	0,8942	0,9057	0,9015
57	0,9792	0,9657	0,9067	0,8925	0,9102	0,9003
58	0,9780	0,9673	0,9070	0,8962	0,9104	0,9018
59	0,9803	0,9713	0,9062	0,8978	0,9056	0,9050
60	0,9751	0,9695	0,8999	0,9031	0,9048	0,8978
61	0,9797	0,9696	0,9102	0,8979	0,9109	0,8988
62	0,9817	0,9720	0,9111	0,9024	0,9078	0,9037
63	0,9820	0,9718	0,9128	0,9060	0,9104	0,9002
64	0,9823	0,9715	0,9132	0,9027	0,9113	0,9028
65	0,9828	0,9732	0,9130	0,9065	0,9127	0,9048
66	0,9818	0,9717	0,9086	0,8991	0,9143	0,9022
67	0,9822	0,9729	0,9098	0,9046	0,9131	0,9065
68	0,9827	0,9741	0,9147	0,9037	0,9151	0,9091
69	0,9838	0,9765	0,9150	0,9076	0,9124	0,9096
70	0,9843	0,9749	0,9132	0,9048	0,9143	0,9011
71	0,9749	0,9757	0,9043	0,9117	0,9051	0,9108
72	0,9841	0,9740	0,9152	0,9024	0,9148	0,9080
73	0,9833	0,9746	0,9116	0,9033	0,9158	0,9100
74	0,9832	0,9767	0,9141	0,9114	0,9123	0,9138
75	0,9840	0,9760	0,9176	0,9104	0,9171	0,9096
76	0,9847	0,9764	0,9159	0,9125	0,9141	0,9085
77	0,9858	0,9764	0,9224	0,9101	0,9223	0,9124
78	0,9856	0,9784	0,9182	0,9098	0,9203	0,9149
79	0,9852	0,9775	0,9175	0,9083	0,9144	0,9103

HN	PPV _{TR}	NPV _{TR}	PPV _{VL}	NPV _{VL}	PPV _{TS}	NPV _{TS}
80	0,9853	0,9784	0,9188	0,9158	0,9135	0,9158
81	0,9878	0,9778	0,9167	0,9091	0,9185	0,9100
82	0,9846	0,9796	0,9170	0,9167	0,9164	0,9122
83	0,9773	0,9786	0,9079	0,9138	0,9080	0,9110
84	0,9870	0,9789	0,9198	0,9144	0,9151	0,9164
85	0,9840	0,9815	0,9173	0,9175	0,9164	0,9156
86	0,9857	0,9788	0,9159	0,9104	0,9206	0,9143
87	0,9855	0,9770	0,9158	0,9152	0,9210	0,9103
88	0,9869	0,9806	0,9190	0,9214	0,9194	0,9149
89	0,9883	0,9806	0,9221	0,9156	0,9234	0,9171
90	0,9893	0,9820	0,9226	0,9178	0,9209	0,9148
91	0,9872	0,9805	0,9221	0,9127	0,9219	0,9168
92	0,9875	0,9806	0,9230	0,9126	0,9231	0,9173
93	0,9869	0,9794	0,9193	0,9171	0,9151	0,9131
94	0,9890	0,9838	0,9230	0,9191	0,9257	0,9170
95	0,9876	0,9800	0,9203	0,9160	0,9208	0,9134
96	0,9775	0,9789	0,9135	0,9178	0,9088	0,9185
97	0,9893	0,9827	0,9194	0,9194	0,9228	0,9204
98	0,9880	0,9851	0,9230	0,9197	0,9219	0,9190
99	0,9871	0,9807	0,9280	0,9167	0,9231	0,9162
100	0,9884	0,9811	0,9206	0,9180	0,9205	0,9162

Tabela 9 – *PPV* i *NPV* na skupu za obuku (*TR*), validaciju (*VL*) i testiranje (*TS*) za broj skrivenih neurona $HN = 1, 2, \dots, 100$.

B Rezultati eksperimenta br. 2

NEURONSKA MREZA						STABLO ODLUKE						LINEARNA REGRESIJA					
PPV _{TR}	NPV _{TR}	PPV _{VL}	NPV _{VL}	PPV _{TS}	NPV _{TS}	PPV _{TR}	NPV _{TR}	PPV _{VL}	NPV _{VL}	PPV _{TS}	NPV _{TS}	PPV _{TR}	NPV _{TR}	PPV _{VL}	NPV _{VL}	PPV _{TS}	NPV _{TS}
0,9928	0,9919	0,9379	0,9320	0,8873	0,9539	0,8659	0,7880	0,8828	0,7959	0,8310	0,7632	0,7552	0,8099	0,7586	0,8367	0,6761	0,8224
0,9928	0,9870	0,9530	0,9301	0,8613	0,8917	0,8705	0,7992	0,8859	0,8182	0,8175	0,7452	0,7545	0,8106	0,7919	0,8042	0,6788	0,7771
0,9946	0,9902	0,9091	0,9063	0,9324	0,9247	0,8408	0,8292	0,8258	0,7813	0,7973	0,8356	0,7648	0,8113	0,7197	0,7813	0,7230	0,8219
0,9875	0,9722	0,9098	0,8553	0,9388	0,9184	0,8712	0,7925	0,8195	0,7799	0,8980	0,7959	0,7478	0,8056	0,7368	0,8113	0,8095	0,8435
0,9946	0,9919	0,9203	0,9221	0,9091	0,9571	0,8635	0,7937	0,8841	0,7922	0,8442	0,8000	0,7477	0,8123	0,7826	0,7727	0,7403	0,8071
0,9839	0,9779	0,9412	0,9423	0,9510	0,9404	0,8615	0,7809	0,8676	0,8077	0,8671	0,8079	0,7426	0,8070	0,7721	0,8718	0,7133	0,7947
0,9876	0,9769	0,9323	0,9371	0,8971	0,8924	0,8645	0,7997	0,8947	0,7673	0,8088	0,7848	0,7520	0,8054	0,7970	0,7925	0,7500	0,8165
0,9965	0,9893	0,9070	0,9387	0,9437	0,9079	0,8651	0,8189	0,8372	0,8282	0,8732	0,8553	0,7374	0,7959	0,7519	0,8405	0,8239	0,8618
0,9928	0,9951	0,9658	0,8973	0,9362	0,9608	0,8605	0,8262	0,8562	0,7534	0,8440	0,8693	0,7381	0,8132	0,7397	0,7329	0,7163	0,8497
0,9912	0,9942	0,9225	0,9755	0,9098	0,9503	0,8856	0,7662	0,8527	0,8098	0,8947	0,7640	0,7570	0,8010	0,7597	0,8466	0,7444	0,8137
0,9736	0,9702	0,9034	0,9252	0,8636	0,9214	0,8772	0,8182	0,8069	0,8435	0,8571	0,8071	0,7425	0,8142	0,6897	0,8707	0,6494	0,8357
0,9893	0,9943	0,8849	0,9608	0,9225	0,9342	0,8926	0,7771	0,8273	0,8170	0,8310	0,8289	0,7592	0,8065	0,6691	0,8235	0,6831	0,8618
0,9937	0,9927	0,9098	0,9182	0,9470	0,9161	0,8734	0,7891	0,8647	0,7673	0,8609	0,7483	0,7513	0,8070	0,7669	0,8113	0,6954	0,8531
0,9734	0,9597	0,8824	0,8846	0,9630	0,8868	0,8838	0,7811	0,8456	0,7821	0,8815	0,7862	0,7551	0,8025	0,7574	0,8013	0,7852	0,8113
0,9891	0,9887	0,9108	0,9333	0,9085	0,9539	0,8690	0,8109	0,9045	0,8000	0,8732	0,8158	0,7443	0,8142	0,7834	0,8148	0,7042	0,8026
0,9965	0,9827	0,9385	0,9012	0,9708	0,9490	0,8744	0,7993	0,8538	0,7346	0,9051	0,8153	0,7622	0,8010	0,7154	0,7901	0,8102	0,8025
0,9910	0,9716	0,9416	0,9203	0,9407	0,8679	0,8170	0,8459	0,8247	0,8551	0,8444	0,7862	0,7583	0,8086	0,7727	0,8551	0,6889	0,7547
0,9973	0,9902	0,9021	0,9128	0,9149	0,9346	0,7675	0,8665	0,8462	0,8255	0,7731	0,8889	0,7496	0,8062	0,7692	0,8054	0,8156	0,8105
0,9874	0,9854	0,9500	0,9276	0,9252	0,8912	0,8776	0,8172	0,8857	0,8092	0,8844	0,7279	0,7561	0,8164	0,7286	0,8355	0,7551	0,6939
0,9964	0,9846	0,9124	0,9290	0,9295	0,9058	0,8697	0,7906	0,8540	0,7677	0,8910	0,7754	0,7502	0,8181	0,6715	0,7935	0,8141	0,7609
0,9973	0,9766	0,9172	0,9259	0,8993	0,9226	0,8675	0,7839	0,9045	0,8370	0,8417	0,8065	0,7378	0,8137	0,7516	0,8741	0,7482	0,7935
0,9882	0,9782	0,9272	0,8794	0,9375	0,9600	0,8749	0,8047	0,8676	0,7518	0,8403	0,7467	0,7434	0,8257	0,8079	0,7872	0,7500	0,7600
0,9893	0,9910	0,9323	0,9308	0,9583	0,9067	0,8662	0,8215	0,8346	0,8365	0,8611	0,7733	0,7583	0,8100	0,7293	0,8491	0,7222	0,7533
0,9875	0,9804	0,9504	0,9272	0,9000	0,8961	0,8854	0,7690	0,9149	0,8079	0,8857	0,7403	0,7556	0,8033	0,8014	0,8411	0,7929	0,7857
0,9946	0,9854	0,9231	0,9396	0,9296	0,9474	0,8652	0,7901	0,8671	0,7919	0,8592	0,8092	0,7484	0,8129	0,7343	0,7651	0,7606	0,8421
0,9920	0,9885	0,9433	0,9404	0,8881	0,9500	0,8718	0,7916	0,8582	0,7815	0,8358	0,7125	0,7614	0,7998	0,7447	0,8079	0,7537	0,7625
0,9921	0,9859	0,8880	0,8922	0,9559	0,9114	0,8690	0,7983	0,8880	0,7665	0,8676	0,7658	0,7669	0,8066	0,7280	0,8024	0,7426	0,7848
0,9838	0,9756	0,9103	0,8571	0,9362	0,8954	0,8669	0,7911	0,8828	0,7687	0,8723	0,8235	0,7473	0,8081	0,7586	0,7755	0,7943	0,8039
0,9919	0,9862	0,9333	0,9225	0,9191	0,9241	0,8705	0,7935	0,8600	0,8310	0,8676	0,7405	0,7554	0,8106	0,7600	0,8169	0,7868	0,7658
0,9875	0,9910	0,9214	0,9474	0,8500	0,9026	0,8605	0,8015	0,8714	0,8355	0,8214	0,7403	0,7621	0,8105	0,7571	0,8158	0,7000	0,7792
0,9864	0,9782	0,9306	0,9054	0,8824	0,9078	0,8837	0,8187	0,8542	0,8108	0,8105	0,7943	0,7502	0,8131	0,7083	0,8243	0,7255	0,8440
0,9856	0,9781	0,9448	0,9116	0,9155	0,8816	0,8461	0,8310	0,8552	0,8027	0,8239	0,7566	0,7588	0,8148	0,7793	0,7959	0,6972	0,8026
0,9955	0,9673	0,9366	0,9133	0,8986	0,9359	0,8784	0,8162	0,8451	0,7800	0,8841	0,8205	0,7460	0,8113	0,7394	0,7800	0,7609	0,8590
0,9483	0,9303	0,9248	0,8805	0,9161	0,9139	0,8645	0,7910	0,8722	0,8113	0,8392	0,8013	0,7442	0,8082	0,7895	0,8176	0,7413	0,8013
0,9841	0,9851	0,8905	0,9355	0,9225	0,9333	0,8958	0,7570	0,8978	0,6452	0,8992	0,7758	0,7597	0,8099	0,7664	0,7548	0,8140	0,8182
0,9847	0,9691	0,8690	0,9524	0,9424	0,9226	0,8806	0,7875	0,8966	0,8231	0,8345	0,8000	0,7531	0,8127	0,6690	0,8163	0,7482	0,8000
0,9839	0,9828	0,8731	0,9557	0,9091	0,9272	0,8707	0,7805	0,8433	0,8038	0,8951	0,7815	0,7591	0,8002	0,7388	0,8418	0,7762	0,8079
0,9874	0,9748	0,9366	0,9067	0,9091	0,8742	0,8589	0,7966	0,9366	0,7800	0,8531	0,7616	0,7538	0,8047	0,7535	0,7933	0,7413	0,8411
0,9825	0,9701	0,8978	0,9097	0,8760	0,9538	0,8719	0,7962	0,8832	0,7484	0,8430	0,7919	0,7658	0,8062	0,7445	0,7935	0,7355	0,8208
0,9913	0,9942	0,9524	0,9458	0,9535	0,8848	0,8661	0,8173	0,8889	0,8193	0,8992	0,8121	0,7559	0,8015	0,7698	0,8253	0,7752	0,7939
0,9910	0,9731	0,9225	0,8867	0,9220	0,9085	0,8780	0,8036	0,8944	0,7133	0,8582	0,7843	0,7480	0,8109	0,7465	0,7800	0,7943	0,8039
0,9946	0,9667	0,9467	0,8944	0,9209	0,8839	0,8557	0,8313	0,8733	0,7465	0,8705	0,8323	0,7448	0,8151	0,7667	0,7113	0,7626	0,8258
0,9900	0,9903	0,9329	0,9161	0,9600	0,9444	0,8772	0,7796	0,8926	0,8042	0,8667	0,7639	0,7425	0,8085	0,7383	0,7972	0,7733	0,8403
0,9927	0,9912	0,8844	0,9034	0,8981	0,9489	0,8812	0,7829	0,8503	0,7103	0,8790	0,8102	0,7431	0,8229	0,6395	0,8069	0,7389	0,7883
0,9937	0,9919	0,8971	0,9038	0,9470	0,9371	0,8848	0,7994	0,8529	0,7628	0,8609	0,7692	0,7516	0,8083	0,7059	0,8077	0,7550	0,8112
0,9894	0,9595	0,9085	0,9200	0,9360	0,8876	0,8762	0,7944	0,8592	0,7933	0,8720	0,7811	0,7692	0,8101	0,7324	0,7533	0,7360	0,7929
0,9928	0,9878	0,9220	0,9272	0,9371	0,9139	0,8878	0,7850	0,8511	0,8079	0,9301	0,8079	0,7487	0,7997	0,7518	0,8278	0,7133	0,8411
0,9806	0,9818	0,9333	0,8981	0,9231	0,9451	0,8438	0,8230	0,8148	0,8153	0,8846	0,7988	0,7529	0,8015	0,7778	0,8280	0,7769	0,8171
0,9946	0,9822	0,9371	0,8859	0,9195	0,9034	0,8544	0,8341	0,8881	0,7852	0,8255	0,7862	0,7432	0,8155	0,7692	0,7852	0,7785	0,7793
0,9965	0,9909	0,8978	0,9161	0,9394	0,9012	0,8627	0,7997	0,8321	0,7548	0,8636	0,8210	0,7591	0,8071	0,7372	0,7677	0,7955	0,8148
0,9849	0,9802	0,9603	0,9036	0,9021	0,8940	0,8760	0,7947	0,8333	0,8072	0,8392	0,7152	0,7600	0,8120	0,7222	0,8494	0,7622	0,7285
0,9885	0,9810	0,8955	0,8861	0,9333	0,9120	0,8654	0,8129	0,8507	0,7911	0,8593	0,8428	0,7511	0,7923	0,7761	0,8228	0,7111	0,8868
0,9823	0,9785	0,9433	0,9338	0,8968	0,9048	0,8603	0,8208	0,8794	0,7815	0,8571	0,8095	0,7560	0,8059	0,7589	0,8079	0,6905	0,8036
0,9820	0,9772	0,8571	0,9803	0,8472	0,9200	0,8689	0,8217	0,8643	0,8618	0,7917	0,7933	0,7513	0,8054	0,7357	0,8750	0,7292	0,7933
0,9813	0,9844	0,9429	0,9737	0,9399	0,9130	0,8756	0,8143	0,8929	0,8487	0,8195	0,7764	0,7582	0,8110	0,7571	0,8289	0,6992	0,8012
0,9929	0,9869	0,9007	0,9205	0,9265	0,9304	0,8421	0,8362	0,8156	0,7550	0,8162	0,8797	0,7449	0,8084	0,7447	0,7550	0,7426	0,8544
0,9929	0,9712	0,9535	0,8773	0,9296	0,8882	0,8660	0,8214	0,8682	0,7914	0,8662	0,8092	0,7622	0,8165	0,7907	0,7239	0,7394	0,7763
0,9973	0,9901	0,9200	0,8802	0,8912	0,9524	0,8766	0,7854	0,8800	0,7365	0,8503	0,7823	0,7496	0,8100	0,8000	0,7425	0,7687	0,8435
0,9874	0,9699	0,9357	0,9145	0,9041	0,8986	0,8534	0,8016	0,8929	0,8224	0,8630	0,7297	0,7500	0,8089	0,8071	0,8224	0,7603	0,7905

NEURONSKA MREZA						STABLO ODLUKE						LINEARNA REGRESIJA					
PPV _{TR}	NPV _{TR}	PPV _{VL}	NPV _{VL}	PPV _{TS}	NPV _{TS}	PPV _{TR}	NPV _{TR}	PPV _{VL}	NPV _{VL}	PPV _{TS}	NPV _{TS}	PPV _{TR}	NPV _{TR}	PPV _{VL}	NPV _{VL}	PPV _{TS}	NPV _{TS}
0,9884	0,9820	0,9318	0,8875	0,9726	0,8986	0,8768	0,7971	0,8561	0,7625	0,8356	0,8041	0,7607	0,8069	0,7803	0,7750	0,7192	0,8446
0,9838	0,9683	0,9114	0,8881	0,9219	0,9157	0,8786	0,8000	0,8608	0,8433	0,8672	0,8133	0,7473	0,8122	0,7595	0,7910	0,7656	0,8373
0,9875	0,9796	0,9524	0,9172	0,9091	0,9321	0,8803	0,7768	0,8639	0,7793	0,8333	0,8086	0,7435	0,7980	0,8027	0,8138	0,7576	0,8580
0,9901	0,9894	0,8832	0,9355	0,9252	0,8912	0,8636	0,8062	0,8613	0,7613	0,8639	0,7483	0,7460	0,8168	0,7591	0,7355	0,7483	0,8027
0,9794	0,9763	0,9160	0,8820	0,9133	0,8819	0,8809	0,7878	0,8626	0,7950	0,8000	0,7639	0,7574	0,8082	0,7328	0,8323	0,6867	0,7708
0,9920	0,9902	0,9281	0,8627	0,9130	0,9423	0,8707	0,8018	0,8561	0,7190	0,8841	0,7949	0,7493	0,8092	0,7266	0,7843	0,8261	0,8397
0,9901	0,9878	0,9155	0,9000	0,9586	0,9530	0,8767	0,7725	0,8803	0,7600	0,8897	0,7852	0,7462	0,8091	0,7606	0,7867	0,7586	0,8523
0,9955	0,9821	0,9489	0,8968	0,9456	0,8844	0,8689	0,8054	0,8613	0,7161	0,9048	0,7891	0,7540	0,8151	0,7153	0,7677	0,7619	0,8163
0,9900	0,9895	0,8881	0,9262	0,9097	0,9209	0,8882	0,8003	0,8392	0,7383	0,8258	0,7986	0,7455	0,8124	0,7692	0,7785	0,7032	0,8273
0,9955	0,9844	0,9306	0,9730	0,9173	0,9379	0,8626	0,8100	0,8472	0,8041	0,8872	0,7329	0,7609	0,8133	0,7222	0,8243	0,7293	0,7267
0,9910	0,9910	0,9058	0,9221	0,9441	0,9272	0,8693	0,7910	0,8478	0,7662	0,8741	0,8411	0,7440	0,8041	0,8116	0,7922	0,7762	0,8543
0,9875	0,9796	0,9297	0,8780	0,9211	0,8944	0,8766	0,7933	0,8906	0,7317	0,8289	0,7465	0,7621	0,8105	0,7578	0,7866	0,7434	0,7535
0,9964	0,9926	0,9111	0,9363	0,9197	0,9427	0,8623	0,7878	0,8741	0,7580	0,9051	0,8344	0,7531	0,8026	0,7481	0,8025	0,7153	0,8408
0,9802	0,9740	0,9091	0,9128	0,9103	0,8993	0,8721	0,8239	0,8252	0,7718	0,8207	0,7987	0,7532	0,8019	0,7762	0,8121	0,7241	0,8389
0,9964	0,9926	0,9856	0,9412	0,9333	0,9245	0,8817	0,8038	0,9065	0,8105	0,8593	0,7799	0,7518	0,8030	0,7842	0,8170	0,7556	0,8239
0,9947	0,9901	0,9118	0,9487	0,9531	0,9096	0,8607	0,8262	0,8529	0,8269	0,8438	0,8193	0,7584	0,7997	0,8162	0,7949	0,7109	0,8193
0,9620	0,9488	0,8947	0,8742	0,8955	0,8875	0,8771	0,8126	0,8797	0,7736	0,8955	0,8063	0,7622	0,8059	0,6992	0,7925	0,7239	0,8250
0,9876	0,9876	0,9197	0,9161	0,9399	0,9130	0,8865	0,8072	0,8175	0,7613	0,8947	0,8075	0,7544	0,8089	0,7518	0,8065	0,7744	0,7764
0,9849	0,9779	0,9184	0,9103	0,9609	0,9337	0,8807	0,7785	0,8912	0,7517	0,9141	0,8193	0,7587	0,8031	0,7551	0,8000	0,7734	0,7952
0,9965	0,9885	0,9111	0,9108	0,9254	0,9125	0,8654	0,7964	0,8519	0,7452	0,8731	0,8313	0,7653	0,8054	0,7185	0,7452	0,8582	0,7938
0,9955	0,9829	0,9030	0,8924	0,9392	0,8973	0,8737	0,7993	0,8284	0,7975	0,8581	0,7603	0,7527	0,8100	0,7463	0,8038	0,8311	0,7808
0,9892	0,9853	0,9310	0,9660	0,8273	0,8903	0,8761	0,8184	0,8966	0,8027	0,7698	0,7935	0,7504	0,8160	0,7586	0,8095	0,7410	0,8000
0,9867	0,9803	0,9574	0,9139	0,9621	0,9198	0,8880	0,7946	0,8440	0,7152	0,8333	0,7963	0,7556	0,7987	0,8085	0,7616	0,7727	0,8333
0,9830	0,9780	0,8980	0,8690	0,9179	0,8938	0,8711	0,8073	0,8503	0,7310	0,8731	0,7625	0,7547	0,8139	0,7755	0,7862	0,7463	0,7750
0,9903	0,9884	0,9231	0,9136	0,9399	0,9317	0,8758	0,7945	0,8308	0,7531	0,8496	0,8323	0,7656	0,8020	0,6923	0,7593	0,8045	0,8571
0,9929	0,9917	0,9259	0,9108	0,9690	0,9455	0,8660	0,7839	0,8222	0,8217	0,8760	0,8424	0,7601	0,8030	0,7037	0,8344	0,7442	0,8364
0,9887	0,9815	0,9453	0,9268	0,9333	0,8448	0,8722	0,8121	0,8594	0,7561	0,8500	0,7356	0,7713	0,8020	0,7734	0,7683	0,7583	0,7701
0,9801	0,9733	0,8777	0,9020	0,9205	0,9231	0,8565	0,7958	0,8993	0,8039	0,8411	0,7972	0,7374	0,8088	0,7482	0,8366	0,7417	0,8252
0,9883	0,9789	0,9306	0,9324	0,9514	0,9600	0,8613	0,7873	0,9583	0,8176	0,8819	0,8267	0,7496	0,8084	0,7569	0,7905	0,7222	0,8067
0,9831	0,9893	0,8993	0,9216	0,9270	0,9618	0,8788	0,7820	0,8417	0,7386	0,8321	0,8344	0,7540	0,8090	0,7698	0,7582	0,7299	0,8280
0,9973	0,9919	0,9437	0,9800	0,9580	0,9007	0,8679	0,7714	0,8662	0,8067	0,8462	0,8079	0,7367	0,8080	0,7746	0,8333	0,7413	0,8411
0,9956	0,9909	0,9542	0,9006	0,9015	0,8827	0,8599	0,8020	0,8779	0,7640	0,8485	0,7654	0,7674	0,8020	0,7634	0,7764	0,7348	0,7778

Tabela 10 – *PPV* i *NPV* na skupu za obuku (*TR*), validaciju (*VL*) i testiranje (*TS*) za *ANN90*, *DT* i *LR*.

C Rezultati eksperimenta br. 3

EN	N _{TP}	N _{TN}	N _{FP}	N _{FN}	PPV	NPV
1	63	104	66	59	0,4884	0,6380
2	73	126	56	37	0,5659	0,7730
3	78	139	51	24	0,6047	0,8528
4	85	141	44	22	0,6589	0,8650
5	90	144	39	19	0,6977	0,8834
6	96	147	33	16	0,7442	0,9018
7	98	148	31	15	0,7597	0,9080
8	100	150	29	13	0,7752	0,9202
9	100	150	29	13	0,7752	0,9202
10	103	150	26	13	0,7985	0,9202
11	105	149	24	14	0,8140	0,9141
12	105	147	24	16	0,8140	0,9018
13	105	148	24	15	0,8140	0,9080
14	103	149	26	14	0,7985	0,9141
15	104	149	25	14	0,8062	0,9141
16	104	150	25	13	0,8062	0,9202
17	104	150	25	13	0,8062	0,9202
18	104	151	25	12	0,8062	0,9264
19	105	151	24	12	0,8140	0,9264
20	105	149	24	14	0,8140	0,9141
21	105	149	24	14	0,8140	0,9141
22	105	149	24	14	0,8140	0,9141
23	104	151	25	12	0,8062	0,9264
24	105	151	24	12	0,8140	0,9264
25	105	151	24	12	0,8140	0,9264
26	105	151	24	12	0,8140	0,9264
27	105	151	24	12	0,8140	0,9264
28	106	151	23	12	0,8217	0,9264
29	107	151	22	12	0,8295	0,9264
30	108	151	21	12	0,8372	0,9264
31	108	151	21	12	0,8372	0,9264
32	108	151	21	12	0,8372	0,9264
33	108	151	21	12	0,8372	0,9264
34	109	151	20	12	0,8450	0,9264
35	110	152	19	11	0,8527	0,9325
36	109	152	20	11	0,8450	0,9325
37	110	152	19	11	0,8527	0,9325
38	111	152	18	11	0,8605	0,9325

EN	N _{TP}	N _{TN}	N _{FP}	N _{FN}	PPV	NPV
39	111	152	18	11	0,8605	0,9325
40	111	152	18	11	0,8605	0,9325
41	111	152	18	11	0,8605	0,9325
42	112	152	17	11	0,8682	0,9325
43	112	153	17	10	0,8682	0,9387
44	112	153	17	10	0,8682	0,9387
45	112	153	17	10	0,8682	0,9387
46	113	153	16	10	0,8760	0,9387
47	113	153	16	10	0,8760	0,9387
48	112	153	17	10	0,8682	0,9387
49	112	153	17	10	0,8682	0,9387
50	111	153	18	10	0,8605	0,9387
51	111	154	18	9	0,8605	0,9448
52	111	154	18	9	0,8605	0,9448
53	112	154	17	9	0,8682	0,9448
54	112	156	17	7	0,8682	0,9571
55	112	156	17	7	0,8682	0,9571
56	112	156	17	7	0,8682	0,9571
57	113	156	16	7	0,8760	0,9571
58	115	156	14	7	0,8915	0,9571
59	115	156	14	7	0,8915	0,9571
60	115	157	14	6	0,8915	0,9632
61	115	155	14	8	0,8915	0,9509
62	115	155	14	8	0,8915	0,9509
63	115	155	14	8	0,8915	0,9509
64	115	155	14	8	0,8915	0,9509
65	115	155	14	8	0,8915	0,9509
66	114	155	15	8	0,8837	0,9509
67	114	155	15	8	0,8837	0,9509
68	114	155	15	8	0,8837	0,9509
69	116	154	13	9	0,8992	0,9448
70	117	153	12	10	0,9070	0,9387
71	118	153	11	10	0,9147	0,9387
72	118	153	11	10	0,9147	0,9387
73	118	153	11	10	0,9147	0,9387
74	118	153	11	10	0,9147	0,9387
75	118	153	11	10	0,9147	0,9387
76	118	153	11	10	0,9147	0,9387
77	119	153	10	10	0,9225	0,9387
78	119	153	10	10	0,9225	0,9387
79	119	153	10	10	0,9225	0,9387

EN	N _{TP}	N _{TN}	N _{FP}	N _{FN}	PPV	NPV
80	119	153	10	10	0,9225	0,9387
81	119	153	10	10	0,9225	0,9387
82	119	153	10	10	0,9225	0,9387
83	120	153	9	10	0,9302	0,9387
84	122	153	7	10	0,9457	0,9387
85	122	153	7	10	0,9457	0,9387
86	122	153	7	10	0,9457	0,9387
87	122	153	7	10	0,9457	0,9387
88	122	153	7	10	0,9457	0,9387
89	122	153	7	10	0,9457	0,9387
90	122	153	7	10	0,9457	0,9387
91	122	153	7	10	0,9457	0,9387
92	122	153	7	10	0,9457	0,9387
93	122	153	7	10	0,9457	0,9387
94	122	153	7	10	0,9457	0,9387

Tabela 11 – Primer obuke veštačke neuronske mreže *ANN90*.

D Rezultati eksperimenta br. 4

											METS predicted		
GEN	AGE	BMI	WC	WHtR	SBP	DBP	TG	HDL	FPG	METS	ANN	DT	LR
2	36	24,32	76,00	0,4674	115	75	1,35	1,43	4,2	0	0	0	0
2	46	27,50	89,00	0,5463	140	85	3,7	1,1	4,1	1	1	1	1
2	38	43,00	116,00	0,7306	140	70	1,76	0,99	5,3	1	1	1	1
2	48	36,30	97,00	0,5844	140	100	2,4	1,2	5,4	1	1	1	1
2	42	26,04	77,00	0,4650	110	70	2,26	1,36	4,5	0	0	0	0
1	60	29,11	97,20	0,5468	145	80	2,24	1	3,5	1	1	1	0
1	45	24,30	99,00	0,5470	120	80	1,19	1,57	4,81	0	0	0	0
1	46	26,03	91,60	0,5614	120	80	2,7	1,2	5,7	0	0	0	0
1	60	27,79	94,20	0,5324	140	100	1,9	1,2	5,4	1	0	1	1
2	37	23,68	70,00	0,4152	110	60	1,07	1,43	5,1	0	0	0	0
2	21	33,50	87,00	0,5180	110	60	2,46	0,78	3,6	1	1	0	0
1	50	29,00	103,60	0,5754	150	100	2,5	1,05	5,4	1	1	1	1
2	38	37,50	100,00	0,6130	120	65	2,2	1,18	4,7	1	1	0	0
1	52	29,40	104,80	0,5545	130	75	2,74	1,19	4,18	1	1	1	0
1	36	24,36	85,90	0,4832	145	90	1,9	1,4	4,8	0	0	0	0
2	31	23,13	79,00	0,4825	130	80	2,36	0,66	6,3	0	0	0	0
1	58	33,11	109,30	0,6321	155	95	1,26	1,37	6,2	1	1	1	1
2	35	36,00	101,00	0,6612	120	80	1,2	1,31	3,7	0	0	0	1
1	21	25,47	89,00	0,5115	120	110	1,74	0,97	4,1	0	0	0	0
2	47	37,90	126,00	0,7081	160	70	1,2	1,2	4,3	1	1	1	1
2	54	28,39	91,80	0,5888	140	100	2,3	0,8	5,8	1	1	1	1
2	39	28,05	93,50	0,5844	135	85	0,54	1,21	4,9	1	1	1	1
2	41	20,14	75,20	0,4510	120	80	0,89	1,38	4,6	0	0	0	0
1	41	20,81	77,10	0,4234	110	90	0,94	1,02	5,7	0	0	0	0
2	34	21,80	84,00	0,4941	130	80	0,74	1,53	4,79	0	1	0	0
2	46	28,99	88,50	0,5395	130	60	1,68	0,91	4,4	1	1	1	0
2	47	44,90	107,00	0,6362	140	95	2,48	1,01	5,6	1	1	1	1
2	48	28,45	82,40	0,5253	120	100	0,87	1,14	4,5	1	1	1	0
2	49	28,10	96,50	0,5684	140	80	1,8	0,92	5,8	1	1	1	1
1	53	27,45	96,20	0,5499	125	75	2,8	0,9	4,9	1	1	0	0
1	55	21,79	81,20	0,4377	125	70	1,4	1,35	4,1	0	0	0	0
1	43	29,15	99,20	0,5775	160	95	1,59	1,1	5,25	0	0	1	1
2	54	28,18	83,80	0,5207	150	80	2,4	0,9	6,2	1	1	1	1
1	56	32,28	111,50	0,6605	165	120	2,6	0,9	9	1	1	1	1
1	53	33,21	95,20	0,5599	155	100	1,7	0,9	4,4	1	1	1	1
1	49	26,25	90,30	0,5048	140	65	2,1	1,1	5,3	0	0	0	0
2	25	40,83	92,00	0,5412	115	80	1,2	1,45	4,1	0	0	0	0

											METS predicted		
GEN	AGE	BMI	WC	WHtR	SBP	DBP	TG	HDL	FPG	METS	ANN	DT	LR
2	27	22,49	83,00	0,4882	120	100	3,7	1,22	4	1	1	0	0
2	23	41,70	105,00	0,6608	120	110	1,74	0,75	5,2	1	1	1	1
1	21	22,92	80,50	0,4447	130	100	3,1	0,9	11,4	0	0	0	0
1	26	23,26	91,00	0,4876	120	80	0,76	1	4,8	0	0	0	0
2	50	46,10	136,00	0,8429	160	100	1,2	1,2	5,3	1	1	1	1
1	51	30,84	96,10	0,5594	156	110	1,81	1,02	6,4	1	1	1	1
2	53	39,10	116,00	0,6823	160	85	1,1	1,87	4,8	0	0	1	1
2	51	46,50	130,00	0,7745	140	95	2	1,64	5,3	1	1	1	1
2	49	26,49	83,10	0,5076	125	80	1,52	1,12	5,6	1	1	0	0
2	20	47,00	116,00	0,7603	140	95	2	0,9	5,2	1	1	1	1
2	22	38,20	97,00	0,6025	115	70	0,63	0,84	3,9	0	0	0	0
2	35	39,50	105,00	0,6880	130	75	1,8	1,2	4,1	1	1	1	1
1	52	27,45	96,40	0,5511	135	100	2	1,54	4,7	1	0	1	1
2	53	29,16	93,80	0,6054	130	75	2,07	1	4,8	1	1	1	1
2	28	45,70	140,00	0,7955	90	60	1,34	1,11	4,4	0	0	1	1
1	49	35,14	113,20	0,6398	175	90	0,9	1,24	7,2	1	1	1	1
2	43	27,38	77,90	0,4868	110	80	0,82	1,11	3,7	0	0	0	0
1	56	22,49	76,30	0,4488	125	80	0,74	1,86	3,3	0	0	0	0
2	29	23,00	92,00	0,5603	120	90	3,1	0,9	5,7	1	1	1	0
1	21	23,52	98,00	0,5601	120	100	1,7	1,07	9,9	1	1	0	0
1	52	26,40	101,60	0,5597	125	120	1,26	0,52	5,2	1	1	0	1
1	46	22,33	70,60	0,4016	150	120	2,1	1,11	6,7	0	0	0	0
2	32	21,00	75,00	0,4330	120	80	0,62	1,28	4,1	0	0	0	0
1	57	27,94	105,30	0,6385	140	95	2,8	1,24	5,8	1	1	1	1
1	40	23,09	78,40	0,4673	125	60	2,94	0,9	4,8	0	0	0	0
1	46	22,33	70,60	0,4016	145	86	2,1	1,04	3,3	0	0	0	0
2	47	27,17	76,00	0,4735	110	65	0,97	1,38	4	0	0	0	0
1	46	28,10	92,30	0,5276	120	90	2,4	0,97	4,4	0	0	1	0
1	46	25,66	86,20	0,4976	130	75	2,3	1,1	4,8	0	0	0	0
2	62	25,50	75,50	0,4402	100	65	1,25	1,36	4,5	0	0	0	0
1	57	34,31	112,30	0,6419	120	120	2,5	0,94	5,7	1	1	1	1
1	43	28,98	96,20	0,5179	120	80	1,45	0,96	4,8	0	0	0	0
2	48	25,72	97,30	0,5483	135	80	2,2	1	5,4	1	1	1	0
2	30	34,70	98,00	0,6318	140	105	1,9	0,9	4,7	1	1	1	1
2	25	38,50	128,00	0,8064	140	75	3,7	1,88	3,9	1	1	1	1
2	21	32,80	93,00	0,5633	140	80	0,85	0,82	5,7	1	1	1	1
2	52	39,80	112,00	0,6665	110	80	1,01	1,31	4,4	0	0	0	1
2	46	48,00	112,00	0,7372	140	85	2,2	0,89	5,7	1	1	1	1
1	42	26,26	98,60	0,5546	125	85	2,64	0,84	4,5	1	1	0	0
2	28	39,20	110,00	0,6450	140	75	2,7	1,2	4,9	1	1	1	1

											METS predicted		
GEN	AGE	BMI	WC	WHtR	SBP	DBP	TG	HDL	FPG	METS	ANN	DT	LR
2	56	26,17	98,00	0,6129	110	70	1,12	1,58	5,3	0	0	0	0
1	37	24,83	94,80	0,5455	100	75	0,66	1,31	4,5	0	0	0	0
2	46	39,40	99,00	0,6245	125	75	2,2	1,2	6,5	1	1	0	1
2	49	26,95	102,70	0,6418	130	80	3,9	1,95	6,1	1	1	1	1
2	50	35,22	92,30	0,5873	145	95	1	0,9	4,7	1	1	1	1
2	54	45,00	117,00	0,6863	180	75	1,5	0,8	6,5	1	1	1	1
1	21	23,28	77,50	0,4259	130	75	1,6	1,23	4,5	0	0	0	0
1	56	28,67	98,40	0,5966	135	80	0,9	1,35	5,1	0	0	1	0
2	46	34,00	102,00	0,6181	110	80	0,88	1,22	3,9	0	0	0	0
1	47	31,30	112,60	0,5953	120	100	2,6	0,9	5,8	1	1	1	1
1	47	21,19	81,30	0,4237	165	85	2,3	1,04	4,5	0	0	0	0
2	34	23,51	73,00	0,4459	120	80	1,53	1,4	4,4	0	0	0	0
2	28	17,50	74,00	0,4229	90	70	0,67	1,45	4,74	0	0	0	0
1	47	26,14	84,30	0,4819	130	80	2,45	1,2	5,6	0	0	0	0
1	47	33,95	113,20	0,6289	145	85	2	1,2	5,1	1	1	1	1
2	49	33,14	102,30	0,5889	140	80	2,4	1,32	5,9	1	1	1	1
2	61	30,08	94,00	0,5875	170	100	2,74	1,17	5	1	1	1	1
2	30	19,10	73,00	0,4148	120	85	0,37	1,4	4	0	0	0	0
1	20	36,52	111,00	0,6000	140	80	1,8	1,15	3,3	1	1	1	1
2	37	24,92	73,00	0,4449	100	70	0,75	1,26	4,9	0	0	0	0
2	27	21,00	71,00	0,4329	120	70	1,06	1,82	4,89	0	0	0	0
1	54	29,10	99,00	0,6047	160	100	2,2	0,98	6	1	1	1	1

Tabela 12 – Primeri primene *ANN90*, *DT* i *LR* u predikciji metaboličkog sindroma.