Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Rok Hudobivnik

Analiza tveganja za samomor z uporabo globokih nevronskih mrež

diplomsko delo

univerzitetni študijski program
prve stopnje
računalništvo in informatika

Mentor: akad. prof. dr. Ivan Bratko

Ljubljana, 2017

Copyright. Rezultati diplomske naloge so intelektualna lastnina avtorja in Fakultete za računalništvo in informatiko Univerze v Ljubljani. Za objavo in koriščenje rezultatov diplomske naloge je potrebno pisno privoljenje avtorja, Fakultete za računalništvo in informatiko ter mentorja.

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Naloga je uporabiti dano podatkovno zbirko o samomorih za strojno učenje napovedovanja stopnje tveganja za samomor. Za učenje uporabite nevronske mreže in preizkusite razne metode za obravnavo manjkajočih vrednosti in normalizacije podatkov. Eksperimentirajte z raznimi nastavitvami nevronske mreže s ciljem doseganja čim večje napovedne točnosti. Analizirajte, kako število skritih slojev v mreži vpliva na napovedno točnost pri učenju iz danih podatkov.

Kazalo

Povzetek

Naslov: Analiza tveganja za samomor z uporabo globokih nevronskih mrež

Avtor: Rok Hudobivnik

Povzetek: Cilj diplomske naloge je na podlagi bioloških podatkov o ljudeh, ki so storili samomor, oz. ljudeh, ki ga niso, naučiti nevronsko mrežo ločevati med tema dvema skupinama. S tem bi lahko v nadaljevanju potencialno razvili način vnaprejšnjega preprečevanja samomorov. Rezultat analize podatkov je pokazal, da nevronska mreža ločuje med skupinama veliko bolj natančno, kot slepo ugibanje za ta primer, s povprečno točnostjo 71,4 % in standardnim odklonom 2,33 %. Tekom pisanja diplomskega dela sem reševal predvsem dva problema, prvi izmed dveh je bil problem manjkajočih vrednosti, ki se je izkazal za glavnega krivca pri omejitvah klasifikacijske točnosti podatkov. Drugi problem je bil problem iskanja prave konfiguracije nevronske mreže, ki bi pri določenem vhodu vrnila najboljšo možno klasifikacijsko točnost. Rezultati in zaključki diplomskega dela se skladajo s predhodnimi analizami teh podatkov.

Ključne besede: globoko učenje, nevronske mreže, samomor.

Abstract

Title: Suicide risk analysis using deep neural networks

Author: Rok Hudobivnik

Abstract: The goal of this thesis was to train a neural network to classify between two groups of people: those who have or have not committed suicide, based on the received biological data set. With the analysis of this data set further research could be performed with a goal of pre-emptive suicide prevention. In the experiments in this thesis, I achieved the average classification accuracy of 71,4 % and standard deviation of 2,33 %. The thesis deals with two distinct problems, first with the problem of missing values, that in the end proved to be the deciding factor for the limitations of the classification accuracy. Second, the problem of finding the optimal configuration of the neural network for this data set. The results and conclusions of this thesis are generally in agreement with other research done on this particular data set.

Keywords: neural nets, deep learning, suicide.

Poglavje 1

Uvod

Odkar obstajamo ljudje, obstaja tudi pojem samomora, skrajšanje lastnega življenja zaradi takih ali drugačnih razlogov. Nekateri ljudje se v zavetje smrti zatečejo zaradi ujetosti, ki jo predstavlja moderni svet ali njihova trenutna situacija, drugi zaradi osamljenosti, katere edini izhod je smrt, spet tretji zaradi lastnih duševnih motenj. Razlogov za to odločitev je izjemno veliko. S pravočasnim odkrivanjem nagnjenja osebe k samomoru, bi lahko tem ljudem priskrbeli primerno pomoč in podporo, ter jim s tem morda celo rešili življenje.

Odkrivanje stopnje nagnjenosti k samomoru, oziroma bolje rečeno, stopnje tveganja za samomor, je zapleten problem, saj je v enaki meri psihološki, kot tudi biološki problem. V svojem diplomskem delu se bom osredotočil predvsem na oprijemljive in izmerljive dele omenjenega problema. Zaradi kompleksnosti in povezanosti različnih atributov je računanje in ugotavljanje rezultata tega problema za vsako možno kombinacijo atributov praktično nemogoče, oziroma je časovno ali prostorsko neuresničljivo. Zaradi te problematike bom pri raziskovalnem delu uporabljal strojno učenje, bolj podrobno globoke nevronske mreže. Nevronske mreže se tako v preprostih, kot tudi bolj kompleksnih primerih izkažejo za bolj natančne in zanesljive, zaradi česar sem se odločil, da bom za analizo podatkov naučil globoko nevronsko mrežo. Na podlagi podatkov o bioloških meritvah vzorčne skupine ljudi, posredovanih

s strani dr. Alje Videtič Paska, bom poskušal naučiti nevronsko mrežo, ki bo v grobem z neko mero klasifikacijske točnosti lahko izračunala stopnjo tveganja za samomor za dano osebo.

Problem analize tveganja za samomor lahko razdelimo na dva dela. Vsi podatki, ki so na voljo o določeni osebi, niso enako pomembni za analizo tveganja za samomor, niso primerno predstavljeni v razmerju z drugimi podatki ali vsebujejo odstopanja, ki lahko v nadaljnjem prinesejo napake v izračunih. Zaradi tega je pred samo uporabo globoke nevronske mreže potrebno normalizirati prejete podatke, izločiti šum, ter izstopajoče podatke in preveriti povezanost različnih atributov z uporabo kovariance. Drugi del problema predstavlja izgradnja nevronske mreže, ter prilagajanje parametrov in učenje implementirane nevronske mreže, z namenom doseganja čim boljših rezultatov. Po končani implementaciji pa sledi tudi večkratno preverjanje implementiranega in statistika pridobljenih rezultatov, ter razlaga le teh.

Podobno analizo skoraj istih podatkov so leta 2016 v okviru seminarske naloge pod mentorstvom akad dr. Ivana Bratka opravili študenti Fakultete za računalništvo in informatiko [12]. Pri svojem delu so uporabili več različnih metod strojnega učenja, med drugim tudi nevronsko mrežo. V tem diplomskem delu bom poskušal nadaljevati njihovo raziskavo, ter razširiti analizo podatkov z uporabo nevronske mreže. Na koncu, bom pridobljene rezultate čim bolj podrobno primerjal z omenjeno seminarsko nalogo.

V prvem delu diplomskega dela na kratko predstavim domeno, na katero se navezujejo podatki. Drugo poglavje je namenjeno predstavitvi nevronskih mrež in globokega učenja, ter metod strojnega učenja, ki so bile uporabljene pri analizi podatkov. Sledi analiza bioloških podatkov, ter opis postopkov predprocesiranja teh podatkov. V nadaljevanju nato opišem različne pogoje, pod katerimi sem učil nevronsko mrežo, ter predstavim rezultate poskusov. V šestem poglavju opišem zanimive zaključke tega diplomskega dela, kaj vse sem se ob izdelavi naučil ter kako bi bilo možno v prihodnosti izboljšati rezultate, ki sem jih dobil. V zaključku povzamem glavne rezultate raziskovalnega dela, ter izoblikujem glavni sklep diplomskega dela.

Poglavje 2

Opis domene

Poglavje je namenjeno predstavitvi biološke domene, s katero se ukvarja to diplomsko delo.

2.1 Samomor

Kot je že iz same besede razvidno, je samomor nameren umor samega sebe. Samomor je prvinski del človeške kulture že od samega začetka [21]. Pogled na samomor se je skozi zgodovino in med različnimi kulturami močno spreminjal, od antike kjer je bil samomor obravnavan, kot edino logično dejanje ob določenih pogojih, npr. visoka starost, nepokretnost, duševni problemi itd., srednjega veka, kjer je samomor veljal za delo zlih sil, demonov itd., pa vse do dandanes, ko ni več stigmatiziran. S pomočjo znanosti skušamo preprečiti takšna dejanja in je zato primerno tej temi posvečeno veliko pozornosti v obliki študij in raziskav.

Velik vpliv na način obravnavanja samomora ima kultura sama, kako le ta interpretira samomor v povezavi z religijo, pomenom življenja in vrednostjo človeškega življenja, ter vplivom tega dejanja na druge člane kulture. Za primer je vredno vzeti japonski pogled na samomor v 17. stoletju, kjer je samomor, v obliki rituala imenovanega sepuku predstavljal častno smrt za vojake oziroma, kot oblika smrtne kazni [24]. Kot nasprotje japonski kulturi

lahko izpostavim kulturo 19. stoletja v Veliki Britaniji. Samomor je bil v tistem času velikokrat interpretiran, kot nesreča oziroma dejanje nore osebe, oziroma po presoji sodnikov tistega časa, osebe, ki je bila nora le v trenutku samomora (*non compos mentis*) [21].

Med razloge za samomor najpogosteje sodijo duševne motnje, ter zloraba substanc. Med druge manj pogoste razloge pa sodijo tudi impulzivna dejanja, ki so posledica stresa, kakršna koli oblika zlorabe, ali drugačnih posameznikovih življenjskih problemov. Pod najbolj pogoste oblike samomora spadajo samo-zastrupitev s pesticidi, samomor z uporabo strelnega orožja in samomor z obešenjem. V socialni skupini z največjim tveganjem za samomor sodijo ljudje med 15-im in 30-im letom, ter ljudje po 70-em letu starosti [22].

V zadnjih letih se po številu samomorov Slovenija uvršča v sam vrh Evrope. Statistično gledano se je leta 2015 Slovenija z oceno 21.4 oseb na 100,000 prebivalcev uvrstila na šesto mesto v Evropi [28].

Zaradi različnih razlogov za vsak samomor je tistim v stiski nudenih več možnosti, s katerimi se lahko spopadejo z mislimi na samomor. Onemogočanje dostopa do drog, ter drugih sredstev, ki bi lahko omogočala posameznikom, da storijo samomor, zdravljenje odvisnosti in kognitivne vedenjske terapije so le nekateri izmed pristopov k reševanju človeških življenj [22].

2.2 Serotonin

Kemijsko imenovana molekula 5-hidroksitripamin (5-HT) ali Serotonin, se primarno nahaja v prebavnem traktu, krvnih ploščicah (trombocitih) in v centralnem živčnem sistemu živali, vključno človeka. V prebavnem traktu Serotonin skrbi za uravnavanje gibanja črevesja [20]. Za diplomsko delo bolj pomembno, pa je delovanje Serotonina v centralnem živčnem sistemu, kjer molekule Serotonina delujejo, kot živčni prenašalec, katerega glavna vloga je sporočanje razpoložljivosti dobrin. Pod pojem dobrine v tem primeru sodijo tudi neoprijemljive dobrine, kot so na primer prijateljstvo, ljubezen in socialni status.

Serotonin modulira praktično skoraj celotno vedenje človeka z uravnavanjem količine te kemijske substance, poljudno ime za to molekulo je "hormon za srečo", saj je prisotnost večje količine serotonina v telesu velikokrat povezano z dobrim počutjem, srečo. Prav zaradi vpliva na človeško razpoloženje pa je količina serotonina v telesu povezana tudi z depresijo in drugimi razpoloženjskimi motnjami, ter posledično tudi s samomorom. Zaradi tega so geni, ki se ukvarjajo s prenašanjem in s sprejemanjem serotonina, glavna tarča študij o razpoloženjskih motnjah pri človeku.

Obravnavani primeri vzorčne skupine ljudi v tem diplomskem delu, pridobljeni s strani dr. Alja Videtič Paska, večinoma predstavljajo podatke o polimorfizmih na različnih genih ali delih genov, ki so povezani s sprejemanjem in prenašanjem Serotonina v možganih.

Poglavje 3

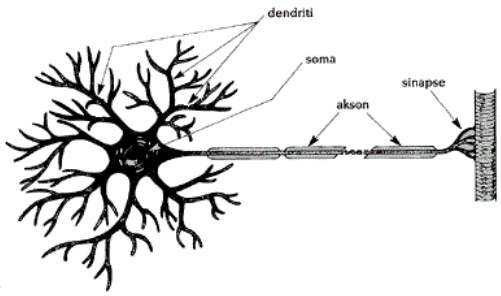
Uporabljene metode strojnega učenja

Poglavje je namenjeno predstavitvi metod strojnega učenja, ki so bile uporabljene pri analizi podatkov.

3.1 Nevron in aktivacijska funkcija

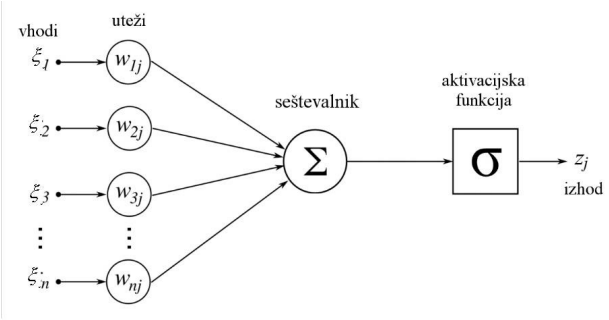
Umetne nevronske mreže so plod raziskav na področjih računalništva, matematike in nevrologije. Predstavljajo algoritmični ekvivalent povezovanja nevronskih celic v kompleksne mreže znotraj možganov. Osnovni del nevronskih mrež je živčna celica ali nevron. Živčne celice od drugih nevronskih celic prejemajo kemične in električne signale preko sinaps. V primeru, da je količina in moč prejetih signalov dovolj velika, da vzpodbudi živčno celico le ta posreduje signal preko dendritov v sinapse drugih živčnih celic, ter tako širi signal (Slika 3.1).

Osnovni del umetne nevronske mreže simulira obnašanje prave živčne celice in se zato imenuje nevron. V osnovi deluje zelo podobno pravi živčni celici. Vsak nevron sešteva utežene vrednosti izhodov nevronov iz predhodne plasti, ki so povezani z njim (podobno, kot živčna celica sprejema signale drugih živčnih celic). Vsoto vrednosti nato normalizira z uporabo vnaprej



Slika 3.1: Živčna celica.

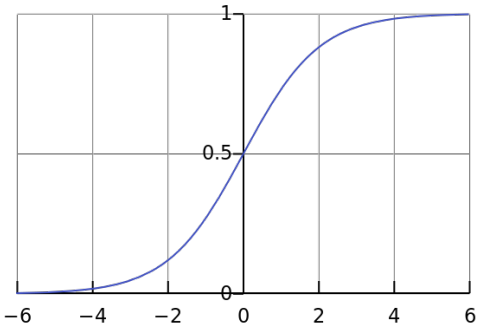
določene aktivacijske funkcije za sloj, v katerem se nevron nahaja, in nato na podlagi izhoda aktivacijske funkcije posreduje ustrezen izhod naslednjemu sloju nevronskih celic (Slika 3.2) [6].



Slika 3.2: Umetni nevron.

Aktivacijska funkcija je funkcija, ki normalizira vhodno spremenljivko (v tem primeru uteženo vsoto vhodov), bolje rečeno, preslika vhodno spremenljivko na vnaprej določen interval. Med najbolj prepoznane aktivacijske funkcije sodi tako imenovana sigmoidna funkcija (Enačba 3.1), ki preslika vhodno spremenljivko na interval med [0,1] (Slika 3.3).

$$f(z) = \frac{1}{1 + e^{-z}} \tag{3.1}$$



Slika 3.3: Krivulja sigmoidne funkcije

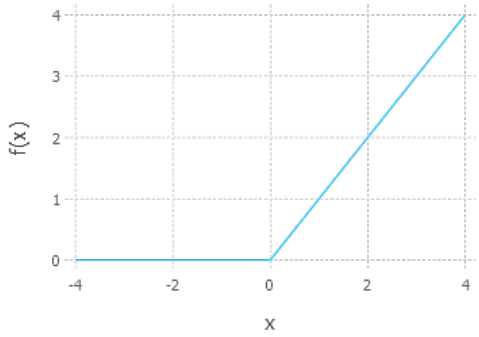
Sigmoidna funkcija kljub svoji prepoznavnosti ni popolna funkcija, razloga za to sta neničelni center funkcije, center sigmoidne funkcije ni v točki (0,0), ter problem izginjajočega gradienta (angl. *vanishing gradient*) pri majhnih vhodnih vrednostih (zelo majhne, skoraj ničelne izhodne vrednosti pri majhnih vhodnih vrednostih). Zaradi teh problematičnih aspektov sigmoidne funkcije, se v praksi pogosteje uporablja funkcija ReLu (rectified linear unit), ki odpravi problem izginjajočega gradienta. Funkcija ReLu (Enačba 3.2) preslika negativne vhode v 0 in ohrani pozitivne vhode (Slika 3.4). Pomanjkljivost funkcije ReLu se pokaže v problemu eksplozije gradienta (angl. *exploding gradient*), zelo velikih izhodnih vrednostih pri zelo velikih vhodnih vrednostih.

$$f(z) = \begin{cases} z & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \tag{3.2}$$

3.2 Arhitektura nevronskih mrež

Nevronske mreže so v svoji osnovi polni, usmerjeni, n-delni aciklični grafi, katerih vozlišča so nevroni. Tipično so nevroni povezani v obliki slojev (Slika 3.5).

Vhodne vrednosti potujejo od začetnega, vhodnega sloja preko skritih slojev do izhodnega sloja, katerega izhod predstavlja izhod nevronske mreže.

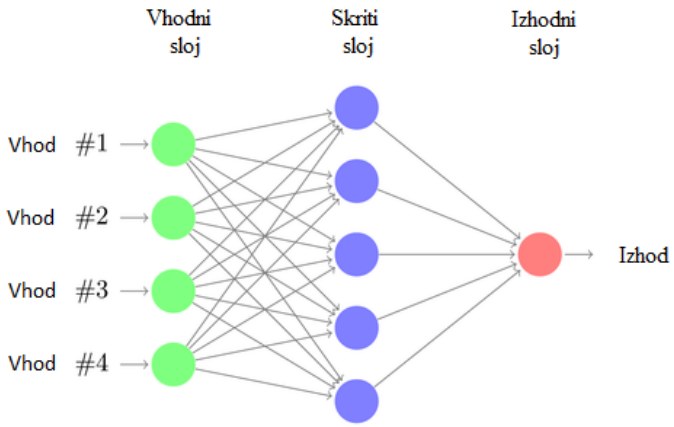


Slika 3.4: Krivulja ReLu funkcije

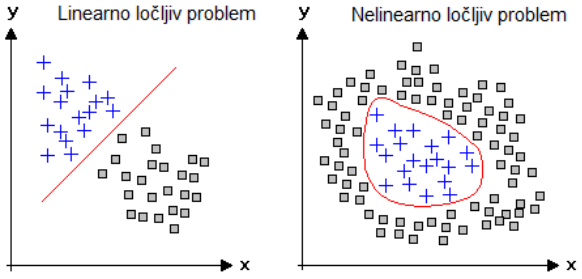
Pri tem se v vsakem nevronu nad vsoto vrednosti vhodov posameznega nevrona izvede aktivacijska funkcija, rezultat katere se nato prenese v naslednji sloj. Tak algoritem za računanje izhoda se imenuje algoritem usmerjene nevronske mreze (angl. *feedforward algorithm*). Sloji so med seboj povezani z uteženimi povezavami (angl. *weighted connections*), katerih število in razporejenost je odvisna od željne oblike nevronske mreže [9].

Globoke nevronske mreže so nevronske mreže z vsaj eno skrito plastjo. Globoke nevronske mreže omogočajo reševanje nelinearno ločljivih problemov, česar navadne nevronske mreže niso zmožne [8], kot je naprimer problem ločevanja dveh poljubnih skupin točk z uporabo funkcije. Pod linearno ločljive skupine točk sodijo tisti podatki, ki jih je možno ločiti z uporabo linearne funkcije. V nasprotju s tem pa so linearno neločljive skupine točk tiste, ki jih je možno ločiti le z uporabo nelinearnih funkcij (Slika 3.6). Z dodajanjem skritih slojev nevronski mreži omogočimo računanje nelinearnih funkcij.

Z večjo količino skritih nivojev povečamo abstrakcijo delovanja nevronske mreže. Večje število skritih slojev otežuje učenje, saj se s tem poveča število parametrov (uteži), ki jih je potrebno nastaviti, ter tudi čas potreben za računanje vrednosti posameznih nevronov in posodabljanja uteži. V primeru premajhnega števila podatkov tvegamo preveliko prilagojenost podanim po-



Slika 3.5: Nevronska mreža



Slika 3.6: Primer linearno ločljivih in ne ločljivih podatkov

datkom (angl. *overfitting*), kar pomeni, da se bo nevronska mreža izjemno izkazala pri računanju rezultatov za natanko te podatke, vendar pa bo pri vnosu drugačnih vhodnih podatkov učinek veliko slabši.

3.3 Gradientni spust

Gradientni spust je optimizacijski algoritem, ki na podlagi gradienta funkcije išče njen minimum. Da funkcija doseže svoj minimum, jo je potrebno proporcionalno spreminjati (narediti korak proti) v skladu z negativnim gradientom. Gradientni spust, ali tudi najbolj strm spust je v svoji osnovi

požrešni algoritem, katerega rezultat je v večini primerov lokalni optimum.

Gradientni spust se v povezavi s strojnim učenjem uporablja za minimizacijo cenovnih funkcij (angl. *cost function*). V primeru nevronskih mrež je to funkcija napake izhoda posameznega nevrona.

3.4 Učenje nevronske mreže

Učenje globoke nevronske mreže se prične po izračunu izhodne vrednosti. Izhodno vrednost nevronske mreže pridobimo z algoritmom širjenja vrednosti po usmerjeni nevronski mreži. Napaka določenega nevrona se izračuna na podlagi primerjave izhodne vrednosti s pričakovanim rezultatom nevrona. Tukaj je vredno omeniti, da glede na tip pričakovanih rezultatov, ki jih imamo na voljo ločimo učenje na tri vrste. Nadzorovano učenje (angl. *supervised learning*), množica pričakovanih rezultatov vsebuje točno vrednost, ki naj bi jo izhod nevronske mreže zasedel. Spodbujevano učenje (angl. *reinforcement learning*) priskrbi le namige o tem, ali je rezultat na pravi poti ali ne. V nasprotju z njima nenadzorovano učenje (angl. *unsupervised learning*) ne vsebuje nobenih pričakovanih rezultatov na podlagi katerih bi lahko učili nevronsko mrežo [9, 8].

Nevronska mreža se uči preko spreminjanja uteži na povezavah med nevroni v nevronski mreži. Spremembo posamezne uteži izračunamo z uporabo optimizacijskega algoritma, ki za cenovno funkcijo vzame funkcijo napake posameznega nevrona. Računanje napake in popravljanje uteži se izvaja v obratnem vrstnem redu, kot se izvaja računanje izhodne vrednosti nevronske mreže, saj se napaka posameznega nevrona računa z uporabo izračunane napake na kasnejših nevronih, v katere je omenjeni nevron posredoval izhodno vrednost svoje aktivacijske funkcije. Zaradi tega se algoritem za spreminjanje uteži in s tem učenje nevronske mreže imenuje algoritem vzvratnega širjenja napake (angl. *backpropagation algorithm*) [9].

Najbolj prepoznana oblika vzvratnega širjenja napake za optimizacijo funkcije napake uporablja stohastični gradientni spust, gradientni spust nad

vsemi primeri v množici podatkov, ter njihovimi razredi, ter na podlagi le tega popravlja uteži povezav (Enačba 3.3). V tej enačbi w_i prestavlja določeno utež, E_i predstavlja funkcijo napake nevrona pri določeni uteži in η predstavlja stopnjo učenja.

$$\begin{aligned} w_i &= w_i + \Delta w_i \\ \Delta w_i &= -\eta \frac{\partial E_i}{\partial w_i} \end{aligned} \tag{3.3}$$

3.5 ADAM

ADAM (adaptive moment estimation) je optimizacijski algoritem, ki je nesporna nadgradnja stohastičnega gradientnega spusta. ADAM se prilagaja parametrom (v tem primeru uteži povezav nevronske mreže), za vsak posamičen parameter izračuna svojo stopnjo učenja in svoj moment. Stopnja učenja je vrednost med 0 in 1, ki upočasni spreminjanje uteži glede na gradient z namenom hitrejše konvergence v lokalni optimum. Moment je tehnika spreminjanja uteži, kjer pri spreminjanju določene uteži prištejemo še del gradienta predhodne uteži. S tem pospešimo stohastični gradientni spust v relevantni smeri in zmanjšamo oscilacije v smeri proč od lokalnega optimuma (Enačba 3.4).

$$\begin{aligned} w_i &= \gamma w_{i-1} + \Delta w_i \\ (\gamma)&.....(moment) \end{aligned} \tag{3.4}$$

Optimizacijski algoritem ADAM adaptivno izračuna dva momenta (Enačbe 3.5, 3.6) za vsako utež (w), ter nato na podlagi le teh posodobi to utež (Enačba 3.7) [14].

$$\begin{aligned} m_w^{(i+1)} &= \beta_1 m_w^{(i)} + (1-\beta_1)\nabla_w L^{(i)} \\ v_w^{(i+1)} &= \beta_2 v_w^{(i)} + (1-\beta_2)(\nabla_w L^{(i)})^2 \end{aligned} \tag{3.5}$$

$$\begin{aligned} \hat{m}_w &= \frac{m_w^{(i+1)}}{1-\beta_1^i} \\ \hat{v}_w &= \frac{v_w^{(i+1)}}{1-\beta_2^i} \end{aligned} \tag{3.6}$$

$$w^{(i+1)} = w^{(i)} - \eta \frac{\hat{m}_w}{\sqrt{\hat{v}_w} + \epsilon} \tag{3.7}$$

V zgornjih enačbah predstavljata spremenljivki ${m^{(i)}}_w$ in ${v^{(i)}}_w$ prvi in drugi moment z vsebovano pristranskostjo (angl. *bias*), spremenljivki $\hat{m}_w$ in $\hat{v}_w$ pa popravljeni prvi in drugi moment, brez pristranskosti. ε predstavlja majhno neničelno število, ki služi le kot preventivni ukrep proti deljenju z 0, L pa predstavlja cenovno funkcijo. Spremenljivki β_1 in β_2 sta določeni s strani uporabnika.

Z uporabo optimizacijskega algoritma se nevronska mreža tako rekoč nauči (popravi uteži na povezavah), kakšne so pravilne izhodne vrednosti za podane vhodne vrednosti tako za učne primere, kot tudi za nove primere (primeri katerih izhodnih vrednosti ne poznamo, ali smo jih z namenom testiranja izvzeli).

Poglavje 4

Podatki

4.1 Analiza podatkov

Podatki uporabljeni v diplomski nalogi so posredovani s strani doc dr. Alje Videtič Paska. Podatki vsebujejo 1100 primerov s 23 atributi in klasifikacijskim razredom.

V spodnjem seznamu so v vrstnem redu, v kakršnem se nahajajo v tabeli, opisani atributi podatkovne množice:

1. (**ID**) Identifikacijska številka primera, 1100 različnih vrednosti atributa, za 1100 različnih primerov. Za atribut so podane vse vrednosti.

2. (**Vrsta Vzorca**) Opredelitev ali je oseba storila samomor ali ne. Atribut lahko zavzame 2 možni diskretni vrednosti (K ali S), kjer vrednost 'S' pomeni, da je oseba storila samomor. Za atribut so podane vse vrednosti.

3. (**Spol**) Spol osebe v posameznem primeru. Atribut lahko zavzame 2 možni diskretni vrednosti moški (M) ali ženska (Z). Za atribut manjka 157 (14,27 %) vrednosti.

4. (**Starost**) Starost osebe v času samomora, oz. v času odvzema meritev. Atribut zavzame vrednost zvezne spremenljivke med 2 in 94 let. Za atribut manjka 165 (15 %) vrednosti.

5. (**5 HT1A 1019**) 5-hydroxytryptamine receptor 1A (HTR1A), serotoninski receptor. Atribut vsebuje 3 možne diskretne vrednosti (GG, CC ali CG). Vrednosti atributa predstavljajo enojni nukleotidni polimorfizem (angl. *single nucleotid polymorphism*) na mestu 1019 v represivni regiji 5-HTR1A gena. [11, 25]. Možne oblike polimorfizma so Citozin na obeh kromosomih (CC), kombinacija Citozin in Gvanin (CG), ter Gvanin na obeh kromosomih (GG). Za atribut manjka 617 (56,09 %) vrednosti.

6. (**2A 1420**) 5-hydroxytryptamine receptor 2A (HTR2A), serotoninski receptor, ki se nahaja na 13. kromosomu [18]. Atribut zavzame vrednost izmed diskretne množice treh (3) vrednosti (CC, CT, TT). Vrednosti atributa predstavljajo enojni nukleotidni polimorfizem v genu 5-HT2A na mestu 1420. Možne oblike polimorfizma so Citozin na obeh kromosomih (CC), kombinacija Citozin in Timin (CT), ter Timin na obeh kromosomih (TT). Za atribut manjka 658 (59,81 %) vrednosti.

7. (**1F 78**) 5-hydroxytryptamine receptor 1F (HTR1F), serotoninski receptor. Atribut zavzame vrednost izmed diskretne množice dveh (2) vrednosti (CC, CT). Vrednosti atributa predstavljajo enojni nukleotidni polimorfizem v genu 5-HTR1F na mestu 78. Možne oblike polimorfizma so Citozin na obeh kromosomih (CC), kombinacija Citozin in Timin (CT), ter Timin na obeh kromosomih (TT). [25] Za atribut manjka 652 (59,27%) vrednosti.

8. (**1B G861C**) 5-hydroxytryptamine receptor 1B (HTR1B), seraotoninski receptor povezan z zlorabo substanc in močno depresijo [29]. Atribut zavzame vrednost izmed diskretne množice treh (3) vrednosti (CC, CT, TT). Vrednosti atributa predstavljajo enojni nukleotidni polimorfizem v genu 5-HTR1B na mestu 861. Možne oblike polimorfizma so Citozin na obeh kromosomih (CC), kombinacija Gvanin in Citozin (GC), ter Gvanin na obeh kromosomih (GG). Za atribut manjka 655 (59,45 %) vrednosti.

9. (**1B 161**) 5-hydroxytryptamine receptor 1B (HTR1B), serotoninski receptor povezan z zlorabo substanc in močno depresijo [29]. Atribut zavzame vrednost izmed diskretne množice treh (3) vrednosti (AT, AA, TT). Vrednosti atributa predstavljajo enojni nukleotidni polimorfizem v genu 5-HTR1B na mestu 161. Možne oblike polimorfizma so Adenin na obeh kromosomih (AA), kombinacija Adenin in Timin (AT), ter Timin na obeh kromosomih (TT). Za atribut manjka 670 (60,90 %) vrednosti.

10. (**LPR**) 5-hydroxytryptamine transporter (5-HTT ali SLC6A4), serotoninski transporter povezan z vplivom stresa na depresijo [3], polimorfizem LPR. Atribut zavzame vrednost izmed diskretne množice štirih (4) vrednosti (LL, LVL, LS in SS). Vrednosti predstavljajo polimorfizem v promotorski regiji gena 5-HTT. Možne oblike tega polimorfizma so dolga variacija na obeh kromosomih (LL), dolga variacija na enem in zelo dolga variacija na drugem kromosomu (LVL), dolga variacija na enem in kratka na drugem kromosomu (LS), ter kratka variacija na obeh kromosomih (SS). Slednja možna vrednost naj bi bila povezana z depresijo in samomorom [10]. Za atribut manjka 656 (59,63 %) vrednosti.

11. (**VNTR**) 5-hydroxytryptamine transporter (5-HTT ali SLC6A4), serotoninski transporter povezan z vplivom stresa na depresijo [3], polimorfizem VNTR. Atribut zavzame vrednost izmed diskretne množice petih (5) vrednosti (1,2,3,4,5). Vrednosti predstavljajo število tandemskih ponovitev polimorfizma na genu 5-HTT. Za atribut manjka 656 (59,72 %) vrednosti.

12. (**5 HT2C 995**) 5-hydroxytryptamine receptor 2C (HTR2C), serotoninski receptor povezan z depresijo in kliničnim odzivom na antidepresive [4]. Atribut zavzame vrednost izmed diskretne množice petih (5) vrednosti (A,GA,G,GG,AA). Vrednosti atributa predstavljajo enojni nukleotidni polimorfizem v genu 5-HTR2C na mestu 995. Možne oblike

polimorfizma za moške so Gvanin (G), oziroma Adenin (A) na X kromosomu, medtem, ko so možne oblike polimorfizma za ženske Gvanin na obeh X kromosomih (GG), Gvanina na enem in Adenin na drugem (GA) X kromosomu, ali pa Adenin na obeh X kromosomih. Za atribut manjka 587 (53,36 %) vrednosti.

13. (**5 HT2C G68C**) 5-hydroxytryptamine receptor 2C (HTR2C), serotoninski receptor povezan z depresijo in kliničnim odzivom na antidepresive [4]. Atribut zavzame vrednost izmed diskretne množice petih (5) vrednosti (GC, G, C, CC, GG). Vrednosti atributa predstavljajo enojni nukleotidni polimorfizem v genu 5-HTR2C na mestu 68. Možne oblike polimorfizma za moške so Gvanin (G), oziroma Citozin (C) na X kromosomu, medtem, ko so možne oblike polimorfizma za ženske Gvanin na obeh X kromosomih (GG), Gvanina na enem in Citozin na drugem (GC) X kromosomu, ali pa Citozin na obeh X kromosomih (CC). Za atribut manjka 613 (55,72 %) vrednosti.

14. (**Alkohol**) Alkoholizem. Atribut lahko zavzame 2 vrednosti (DA, NE), ki se nanašata na prisotnost alkoholizma pri osebi. Za atribut manjka 869 (79 %) vrednosti.

15. (**HTR2A**) 5-hydroxytryptamine receptor 2A (HTR2A), serotoninski receptor. Atribut zavzame vrednost izmed diskretne množice treh (3) vrednosti (TC, TT, CC). Vrednosti atributa predstavljajo enojni nukleotidni polimorfizem v genu 5-HTR2A na mestu 102, ki vpliva na odziv osebe na antipsihotična zdravila [5]. Možne oblike polimorfizma so Citozin na obeh kromosomih (CC), kombinacija Citozin in Timin (TC), ter Timin na obeh kromosomih (TT). Za atribut manjka 603 (54,81 %) vrednosti.

16. (**MAOBCC rsTC799836**) Polimorfizem rs1799836 na genu Monoamine oxidase B (krajše MAOB). Gen MAOB se nahaja na X kromosomu [17]. Atribut zavzame vrednost izmed diskretne množice treh

(3) vrednosti (TC, TT, CC). Možne oblike polimorfizma so Citozin na obeh kromosomih (CC), kombinacija Citozin in Timin (TC), ter Timin na obeh kromosomih (TT). Možni kombinaciji pri moškem spolu sta TT in CC na X kromosomu, medtem, ko je pri ženskem spolu možna tudi kombinacija TC X kromosomih. Za atribut manjka 434 (39,45 %) vrednosti.

17. (**MAOA3 rs3AAGG74AA7**) Polimorfizem rs3027407 na genu Monoamine oxidase A (krajše MAOA), ki naj bi bil povezan s povečano verjetnostjo za psihiatrične bolezni [16]. Gen se podobno, kot MAOB nahaja na X kromosomu, vendar na nasprotnem delu le tega. Atribut zavzame vrednost izmed diskretne množice treh (3) vrednosti (GG, AA, AG). Možne oblike polimorfizma so Gvanin na obeh kromosomih (GG), kombinacija Gvanin in Adenin (AG), ter Adenin na obeh kromosomih (AA). Možni kombinaciji pri moškem spolu sta AA ali GG na X kromosomu, pri ženskem spolu pa tudi AG na X kromosomih. Za atribut manjka 443 (40,27 %) vrednosti.

18. (**MAOACT rs9CC95TT5**) Polimorfizem rs909525 na genu Monoamine oxidase A. Atribut zavzame vrednost izmed diskretne množice treh (3) vrednosti (TT, CC, CT). Možne oblike polimorfizma so Citozin na obeh kromosomih (CC), kombinacija Citozin in Timin (TC), ter Timin na obeh kromosomih (TT). Možni kombinaciji pri moškem spolu sta CC ali TT na X kromosomu, pri ženskem spolu pa tudi CT na X kromosomih. Za atribut manjka 439 (39,9 %) vrednosti.

19. (**MAOA4 rsCTCT37CC7CC**) Polimorfizem rs1137070 na genu Monoamine oxidase A. Atribut zavzame vrednost izmed diskretne množice treh (3) vrednosti (TT, CC, CT). Možne oblike polimorfizma so Citozin na obeh kromosomih (CC), kombinacija Citozin in Timin (TC), ter Timin na obeh kromosomih (TT). Možni kombinaciji pri moškem spolu sta CC ali TT na X kromosomu, pri ženskem spolu pa tudi CT na X kromosomih. Za atribut manjka 436 (39,63 %) vrednosti.

20. (**MAOA uVNTR**) Polimorfizem uVNTR na genu Monoamine oxidase A. Polimorfizem te vrste je povezan z odvisnostjo od substanc in agresivnim/impulzivnim vedenjem [27]. Atribut zavzame vrednost izmed diskretne množice devetih (9) vrednosti (0, 3*2, 5, 3, 5*4, 3*5, 3*4, 3, 4, 2). Vrednosti atributa predstavljajo število tandemskih ponovitev 30-bp dolge regije MAOA gena, ki se pojavi 1.2-kb višje v toku od MAOA kodirnega zaporedja. Ker se MAOA gen nahaja na X kromosomu, se pri moških in ženskah pojavijo različne oblike polimorfizma. Pri moškem spolu je možnih 0, 2, 3, 4 ali 5 tandemskih ponovitev regije na X kromosomu. Pri ženskem spolu pa je možnih 0,2,3,4 ali 5 tandemskih ponovitev na obeh X kromosomih hkrati ali pa 3*2 (3 ponovitve na enem in 2 ponovitvi na drugem kromosomu), 5*4 (5 ponovitve na enem in 4 ponovitev na drugem kromosomu), 3*5 (3 ponovitve na enem in 5 ponovitev na drugem kromosomu) ali pa 3*4 (3 ponovitve na enem in 4 ponovitve na drugem kromosomu). Za atribut manjka 577 (52,45 %) vrednosti.

21. (**Assay2**) Triptofan hidroksilaza 2 (TPH2), polimorfizem rs1843809, ki naj bi bil povezan z depresijo in samomorilnostjo [23, 26]. Atribut zavzame vrednost izmed diskretne množice treh (3) vrednosti (GG, TT, GT). Možne oblike polimorfizma so Gvanin na obeh kromosomih (GG), kombinacija Gvanin in Timin (GT), ter Timin na obeh kromosomih (TT). Za atribut manjka 556 (50,54 %) vrednosti.

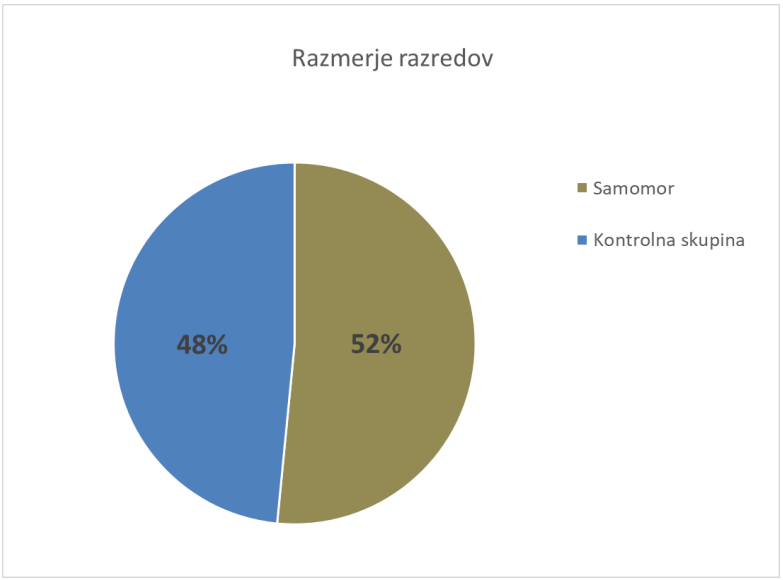
22. (**Assay3**) Triptofan hidroksilaza 2 (TPH2), polimorfizem rs1386493, ki naj bi bil povezan z depresijo in samomorilnostjo [23, 26]. Atribut zavzame vrednost izmed diskretne množice treh (3) vrednosti (CC, CT, TT). Možne oblike polimorfizma so Citozin na obeh kromosomih (CC), kombinacija Citozin in Timin (CT), ter Timin na obeh kromosomih (TT). Za atribut manjka 554 (50,36 %) vrednosti.

23. (**Assay4**) Triptofan hidroksilaza 2 (TPH2), polimorfizem rs4131348, ki naj bi bil povezan z bipolarno motnjo [2]. Atribut zavzame vrednost iz-

med diskretne množice treh (3) vrednosti (CC, CT, TT). Možne oblike polimorfizma so Citozin na obeh kromosomih (CC), kombinacija Citozin in Timin (CT), ter Timin na obeh kromosomih (TT). Za atribut manjka 556 (50,54 %) vrednosti.

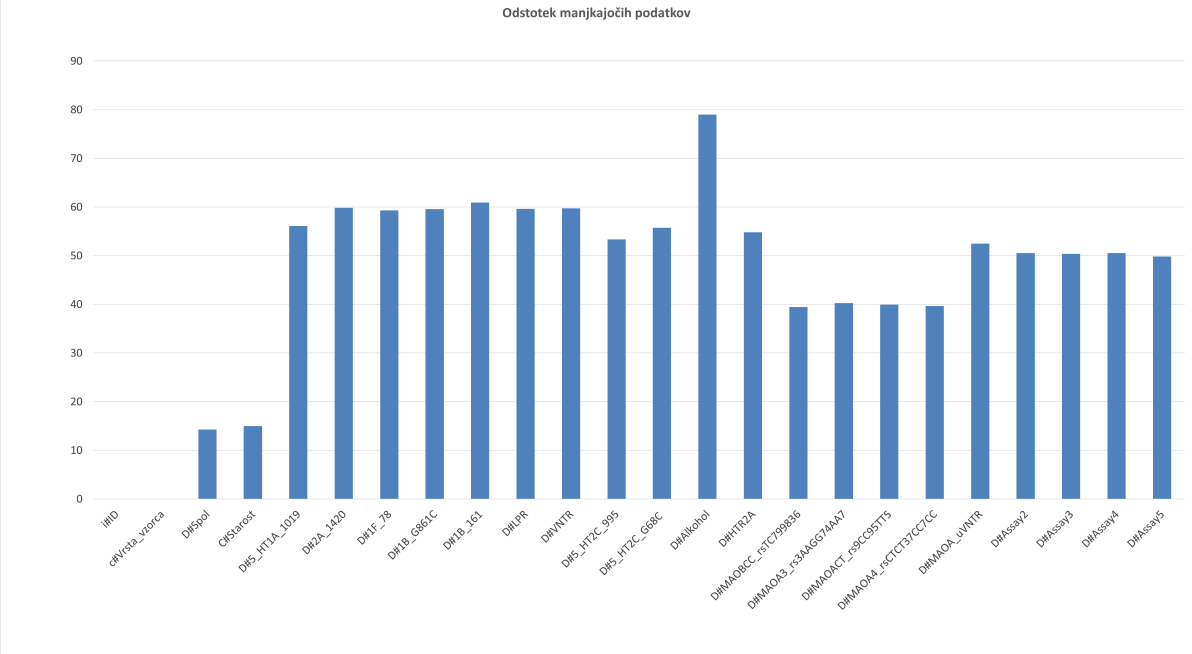
24. (**Assay5**) Triptofan hidroksilaza 2 (TPH2), polimorfizem rs11178997. Atribut zavzame vrednost izmed diskretne množice treh (3) vrednosti (AT, AA, TT). Vrednosti atributa predstavljajo enojni nukleotidni polimorfizem v genu 5-HTR1B na mestu 161. Možne oblike polimorfizma so Adenin na obeh kromosomih (AA), kombinacija Adenin in Timin (AT), ter Timin na obeh kromosomih (TT). Za atribut manjka 548 (49,81 %) vrednosti.

Podatki vsebujejo razredno spremenljivko "vrsta vzorca", ki primere loči v kontrolno skupino in skupino, ki je storila samomor. V kontrolni skupini se izmed 1100 primerov nahaja 533 primerov (48 %), medtem, ko se v skupini, ki je storila samomor nahaja 567 primerov (52 %) (Slika 4.1).



Slika 4.1: Razmerje vrednosti v razredni spremenljivki

Opisana podatkovna množica vsebuje veliko manjkajočih vrednosti (Slika 4.2), bolj podrobno, v podatkih manjka 45,84 % vseh vrednosti. Od vseh atributov le atributa ID in Vrsta vzorca ne vsebujeta prav nobene manjkajoče vrednosti. Z natanko 79 % manjkajočimi vrednostmi je Alkohol atribut z največ manjkajočimi vrednostmi. Vrednosti v podatkovni množici manjkajo popolnoma naključno.



Slika 4.2: Odstotek manjkajočih vrednosti po atributih

4.2 Predprocesiranje podatkov

Manjkajoči vrednosti predstavljajo velik problem pri podatkovnem rudarjenju in uporabi podatkovnih množic na splošno. V izogib temu obstaja veliko načinov za predprocesiranje podatkovnih množic, ki vsebujejo manjkajoče vrednosti, od najbolj preprostih, ki le prezrejo primere z manjkajočimi vrednostmi, pa do bolj kompleksnih, ki manjkajoče vrednosti izračunajo z uporabo strojnega učenja in/ali statistike.

V okviru praktičnega dela diplomskega dela sem poleg originalne podatkovne množice ustvaril tri dodatne podatkovne množice. Z naslednjimi štirimi podatkovnimi množicami sem v nadaljevanju izvedel generiranje manjkajočih vrednosti:

1. Začetna podatkovna množica brez dodatnih sprememb.

2. Začetna podatkovna množica v kateri so manjkajoče vrednosti atributa Alkohol zapolnjene z dodatno vrednostjo, ki pomeni "VERJETNO NE"(Poleg možnih vrednosti "DA"in "NE"). Razlog za uporabo te podatkovne množice je primerjava rezultatov s predhodnim seminarskim delom, v katerem se je uporabil enak pristop [12].

3. Začetna podatkovna množica, iz katere so odstranjeni primeri, ki vsebujejo 17 ali več manjkajočih vrednosti. Odstranitev primerov, ki vsebujejo manjkajoče podatke, sodi med najbolj preproste in tudi najpogosteje uporabljene metode za obdelavo podatkovnih množic, ki vsebujejo manjkajoče vrednosti [1]. Zaradi tega sem se tudi odločil, da do neke mere vključim to metodo v svojem diplomskem delu. Meja 17-ih manjkajočih vrednosti je postavljena na podlagi analize podatkov, saj tako odstranim zgornjih 10 % primerov z največ manjkajočimi vrednostmi (pri višji oz. nižji meji bi se s tem odstranilo preveč oz. premalo primerov, glede na število vseh primerov v podatkovni množici).

4. Podatkovna množica, ki vključuje kombinacijo pristopov omenjenih v točkah 2 in 3, katere namen je testiranje skupka predhodnih dveh množic.

Vse vrednosti v podatkovnih množicah so bile pred nadaljnjimi spremembami preslikane iz diskretnih tekstovnih vrednosti v diskretne numerične vrednosti z namenom olajšanja generiranja vrednosti v nadaljevanju. Dodatni razlog za spremembo tipa podatkov je tudi razlog, da nevronska mreža v tem primeru lahko, kot tip vhodnih podatkov, sprejme le numerične vrednosti. V diplomskem delu sem z namenom primerjave različnih pristopov k

problemu generiranja manjkajočih podatkov uporabil tri različne pristope z vsako izmed podatkovnih množic omenjenih zgoraj:

1. **Zamenjava manjkajočih podatkov z najpogostejšo vrednostjo (Modus)**: Ena izmed bolj preprostih metod generiranja manjkajočih vrednosti, pri kateri se vse manjkajoče vrednosti nekega atributa nadomestijo z najbolj pogosto vrednostjo tega atributa. Kljub temu, da se proti koncu predprocesiranja podatkov vse vrednosti prevedejo v numerične vrednosti, ta metoda ohrani diskretnost le teh. Z uporabo metode zamenjave manjkajočih vrednosti, z npr. povprečno vrednostjo, bi na tem mestu ustvarili dodatno možno vrednost atributa, kar pa pri uporabljeni metodi ni problem. Metoda služi, kot osnova za primerjavo z bolj kompleksnimi metodami, ki so uporabljene v tem diplomskem delu.

2. **Metoda K-najbližjih sosedov**: Metoda uporabi algoritem K-najbližjih sosedov za izračun najprimernejše vrednosti za manjkajoče vrednosti. Metoda za vsako manjkajočo vrednost določi K tej manjkajoči vrednosti najbližjih vrednosti gleda na ostale atribute v določenem primeru, zatem zamenja manjkajočo vrednost z najpogostejšo vrednostjo v teh K primerih. Enaka metoda je bila uporabljena pri izdelavi predhodne seminarske naloge, ter služi po eni strani kot nekoliko bolj kompleksna metoda, katere rezultati se lahko primerjajo z bolj prepresto metodo, ter po drugi strani kot metoda, katere rezultati se lahko primerjajo z rezultati predhodne seminarske naloge [12].

3. **Metoda MICE (Multivariate imputation by chained equations)**: Metoda je veliko bolj kompleksna kot ostali dve uporabljeni metodi. Metoda lahko z dokaj veliko klasifikacijsko točnostjo izračunava približke manjkajočim vrednostim v podatkih, zaradi česar je bila tudi uporabljena v tem diplomskem delu. MICE sprva pretvori vse manjkajoče vrednosti v začasne vrednosti, ki so izračunane z uporabo preprostih algoritmov, kot sta npr. zamenjava s srednjo vrednostjo ali

zamenjava z najpogostejšo vrednostjo. Za tem eno po eno začasne vrednosti spremeni nazaj v manjkajoče vrednosti, ter jih ponovno izračuna na podlagi porazdelitve vrednosti v trenutnem atributu, oziroma modela, ki za vhodne vrednosti vzame vrednosti drugih atributov v tem primeru. Izračun vseh manjkajočih vrednosti na ta način predstavlja en cikel algoritma. Tekom večih ciklov se imputacije manjkajočih vrednosti posodabljajo. Priporočljivih je med 10 in 40 ciklov za generariranje vrednosti [19].

Zaradi velikih razlik v vrednostih in porazdelitvah med različnimi atributi, sem po generiranju podatkov v okviru pripravljanja na učenje nevronske mreže in zbiranje rezultatov preizkusil tudi nekaj različnih načinov normalizacije in njihov vpliv na končno klasifikacijsko točnost nevronske mreže. Tekom teh poskusov sem testiral tri različne pristope, podatkovna množica brez kakršnekoli normalizacije vrednosti, skaliranje vrednosti na interval med 0 in 1, ter skaliranje vrednosti na način, da je srednja vrednost posameznega atributa enaka 0, standardni odklon pa enak 1. Rezultati testov, izvedenih z različnimi konfiguracijami nevronske mreže in več različnimi permutacijami učnih in testnih primerov, ter različnih začetnih vrednosti uteži, so pokazali, da se je najbolje odrezala zadnja vrsta normalizacije, saj je nasproti drugima metodama dosegla v povprečju 10 - 11 % večjo klasifikacijsko točnost izhodnih vrednosti. Glavni razlog za razlike v klasifikacijski točnosti leži predvsem v dejstvu, da so si po uporabi različnih vrst normalizacije podatkovne množice zelo različne. V tretjem primeru npr. so podatki skalirani na način, da je njihova povprečna vrednost enaka 0, standardni odklon pa enak 1. To predstavlja veliko razliko z drugima pristopoma, katerih srednja vrednost je odvisna od vrednosti atributov, razpon vrednosti je med atributi zelo različen. Normalizacija podatkov pred uporabo le teh v nevronski mreži je potrebna zaradi različnih razponov vrednosti, ki se lahko pojavijo, saj veliko večje vrednosti nekega vhoda lahko popolnoma zasenčijo majhne vrednosti drugega vhoda, ter tako posredno vplivajo na učenje in s tem na rezultate nevronske mreže. Na podlagi rezultatov omenjenih testiranj sem

v nadaljevanju za učenje nevronske mreže uporabljal najbolj uspešno vrsto normalizacije (srednja vrednost 0, ter standardni odklon 1). Za predprocesiranje podatkov je bil uporabljen programski jezik R, zaradi enostavnosti uporabe pri obdelavi podatkov in možnosti implementacije algoritma MICE.

Poglavje 5

Rezultati

5.1 Implementacija

V postopku učenja nevronske mreže je bilo testiranih skupaj 16 različnih konfiguracij, števila skritih slojev in nevronov v teh skritih slojih, nevronske mreže:

- 1 skriti sloj s 5, 11, 12 ali 30 nevroni v vseh slojih (skupaj 4 različnih oblik nevronske mreže),
- 2 skrita sloja s 5, 11, 12 ali 30 nevroni v vseh slojih (skupaj 4 različnih oblik nevronske mreže),
- 3 skriti sloji s 5, 11, 12 ali 30 nevroni v vseh slojih (skupaj 4 različnih oblik nevronske mreže),
- 4 skriti sloji s 5, 11, 12 ali 30 nevroni v vseh slojih (skupaj 4 različnih oblik nevronske mreže),

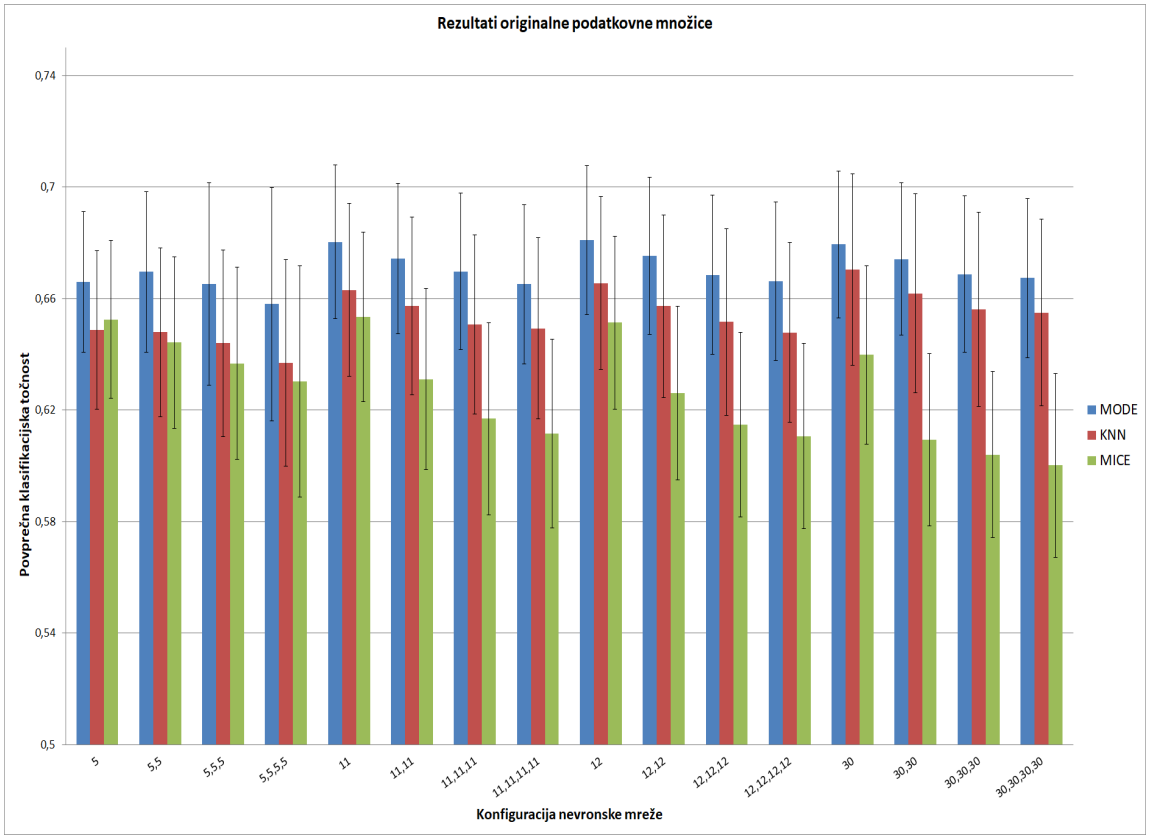
Izbira števila skritih slojev v nevronski mreži, ter števila nevronov v posameznem sloju vpliva na klasifikacijsko točnost izhodov naučene mreže. Zaradi težavnosti določanja [13] teh dve parametrov zgolj iz števila vhodnih in izhodnih nevronov sem v diplomski nalogi testiral več različnih kombinacij parametrov, omenjenih v začetku poglavja [7].

Z vsako obliko nevronske mreže je bilo izvedeno učenje za vseh 12 različnih podatkovnih množic (4 različne začetne podatkovne množice, na vsaki izmed njih so bile izvedene 3 oblike generiranja manjkajočih podatkov). Tekom učenja nevronske mreže je bilo izvedenih 1000 iteracij s podano vhodno podatkovno množico. Vsaka oblika nevronske mreže je bila z isto podatkovno množico naučena pri 100 različnih naključnih začetnih postavitvah (inicializacijah) uteži povezav v tej nevronski mreži. Izbira števila 100 temelji na zagotavljanju dovolj velikega števila rezultatov pri določeni obliki in vhodu nevronske mreže za pravilen in čim bolj statistično natančen izračun povprečne klasifikacijske točnosti nevronske mreže. Vsak proces učenja s temi 100 različnimi začetnimi postavitvami uteži je bil nato ponovljen 10 krat z naključno izbranimi permutacijami dveh podmnožic podatkovne množice, ki sta bili namenjeni procesu učenja in procesu testiranja izhodov naučene nevronske mreže. Podmnožici sta bili v razmerju 80:20, kjer je bila večja podmnožica uporabljena za učenje nevronske mreže. Razmerje 80:20 je bilo izbrano na podlagi Paretovega principa, z dodatnim argumentom, da, glede na relativno majhno število primerov, večje število testnih primerov pomeni nekoliko slabšo naučenost nevronske mreže, manjše število testnih primerov pa bi pomenilo manj natančno testiranje izhodov naučene nevronske mreže [15]. Za namene testiranja delovanja različnih konfiguracij nevronskih mrež je bil uporabljen programski jezik python zaradi izjemno priročnih programskih knjižnic numpy in scipy.

5.2 Opis rezultatov

Naučena nevronska mreža z začetno originalno podatkovno množico (Slika 5.1) je dosegla najboljšo povprečno klasifikacijsko točnost 68,09 %, s standardnim odklonom 2,66 % pri generiranju manjkajočih podatkov z uporabo zamenjave manjkajoči podatkov z modusom, ter konfiguraciji nevronske mreže, ki je vsebovala 1 skriti sloj z 12 nevroni (Slika 5.5). Najslabša povprečna klasifikacijska točnost 60 % s standardnim odklonom 3,3 % je bila dosežena pri

konfiguraciji nevronske mreže, ki je vsebovala 4 skrite sloje s po 30 nevroni. Po pregledu rezultatov je razvidno, da se v primeru te začetne podatkovne množice najboljši rezultati pojavijo po generiranju manjkajočih podatkov z zamenjavo manjkajočih vrednosti z modusom, po drugi strani pa se povprečno najslabši rezultati pojavijo pri bolj kompleksni metodi MICE. Razlog za to bi bilo možno pripisati velikemu številu manjkajočih vrednosti v prvotni podatkovni množici, saj se, napram le tej, metoda MICE odreže veliko bolje v okviru sledečih podatkovnih množic (Slika 5.6).

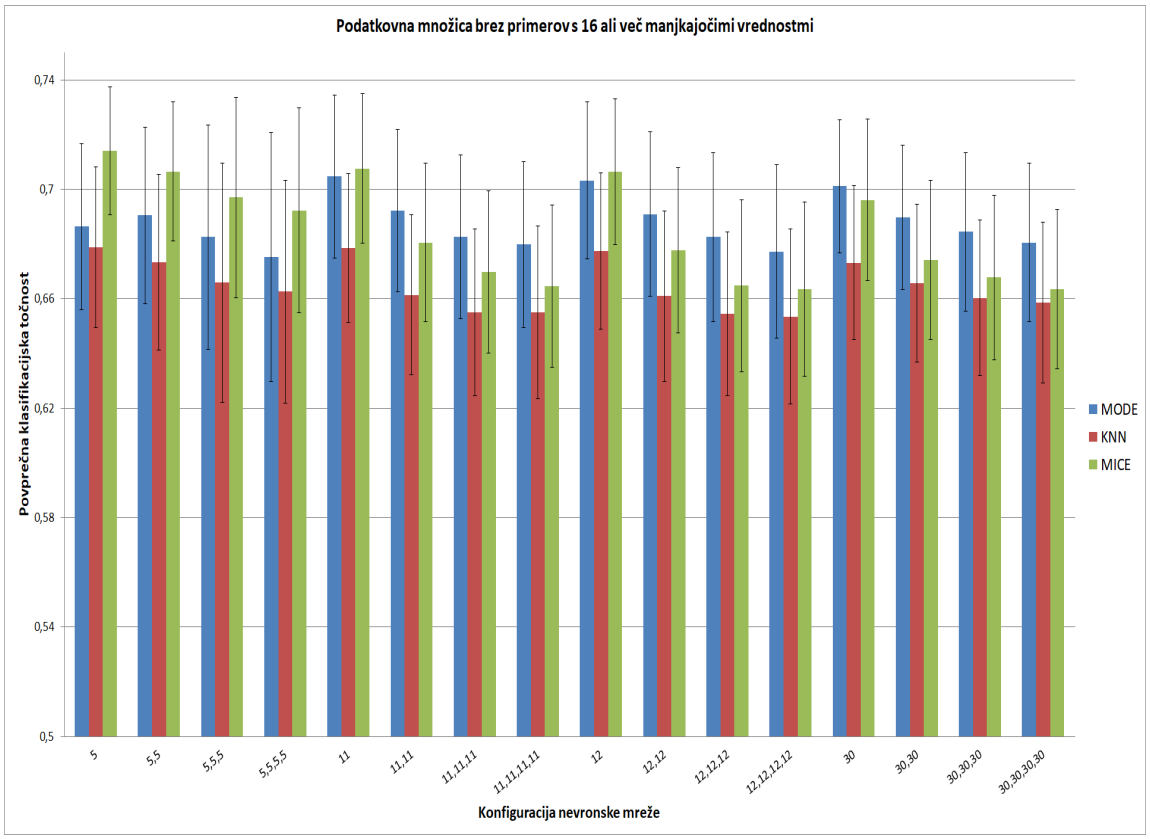


Slika 5.1: Povprečni rezultati klasifikacijskih točnosti nevronskih mrež za originalno podatkovno množico

Nevronska mreža, ki je bila naučena s podatki iz začetne množice, kateri so bili odstranjeni primeri z več kot 16 manjkajočimi vrednostmi (Slika 5.2),

je med različnimi konfiguracijami nevronske mreže dosegla najboljšo povprečno vrednost 71,40 % s standardnim odklonom 2,33 % z uporabo kompleksne metode MICE, pri konfiguraciji nevronske mreže, ki je vsebovala 1 skriti sloj s 30 nevroni. Najslabša povprečna klasifikacijska točnost 65,34 % s standardnim odklonom 3,2 % je bila dosežena pri konfiguraciji nevronske mreže, ki je vsebovala 4 skrite sloje s po 12 nevroni. Učenje in testiranje sta bila v tem primeru izvedena z uporabo nekoliko manjše množice (zaradi odstranitve primerov z več kot 16 manjkajočimi vrednostmi), vendar še vedno z enako porazdelitvijo (80:20) učne in testne podmnožice. Povprečno najboljša metoda generiranja manjkajočih vrednosti za to začetno podatkovno množico je bil način zamenjave manjkajočih vrednosti z modusom. V nasprotju s predhodno začetno podatkovno množico je bila v tem primeru povprečno najslabša metoda K-najbližjih sosedov. Razlog za doseganje slabše klasifikacijske točnosti v primeru uporabe te metode bi morda lahko pripisali manjšemu številu učnih primerov v tej podatkovni množici, saj se metoda K-najbližjih sosedov zanaša na podobnost med primeri, za računanje manjkajočih vrednosti. Kot zanimivost bi tu omenil še delovanje funkcije MICE, ki pri konfiguraciji nevronske mreže z manj nevroni v skritih slojih doseže večjo klasifikacijsko točnost od metode zamenjave manjkajočih vrednosti z modusom. V nasprotju s tem pa v konfiguracijah z večjim številom nevronov v skritih slojih doseže nekoliko slabšo klasifikacijsko točnost od metode zamenjave manjkajočih vrednosti z modusom. Razlog za takšne rezultate morda leži v preveliki prilagojenosti učnim primerom.

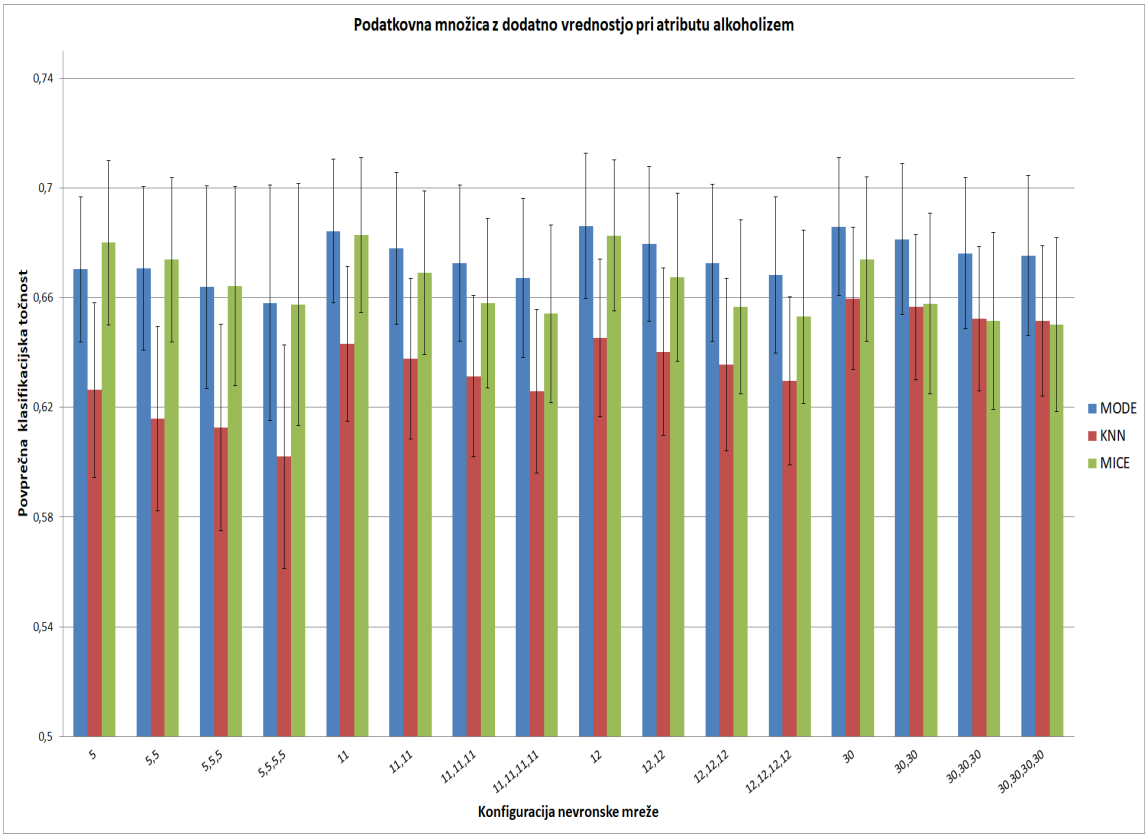
Nevronska mreža naučena z začetno podatkovno množico, kateri je bila pri atributu alkoholizem dodana tretja vrednost (Slika 5.3), je dosegla povprečno najboljši rezultat 68,62 % s standardnim odklonom 2,65 %. Rezultat je bil dosežen z uporabo zamenjave manjkajočih podatkov z modusom, pri konfiguraciji nevronske mreže, ki je vsebovala 1 skriti sloj z 12 nevroni. Najslabša povprečna klasifikacijska točnost 60,2 % s standardnim odklonom 4,07 % je bila dosežena pri konfiguraciji nevronske mreže, ki je vsebovala 4 skrite sloje s po 5 nevroni. Splošno najboljša metoda generiranja manj-



Slika 5.2: Povprečni rezultati klasifikacijskih točnosti nevronskih mrež za podatkovno množico, kateri so bili odstranjeni primeri z več kot 16 manjkajočimi vrednostmi

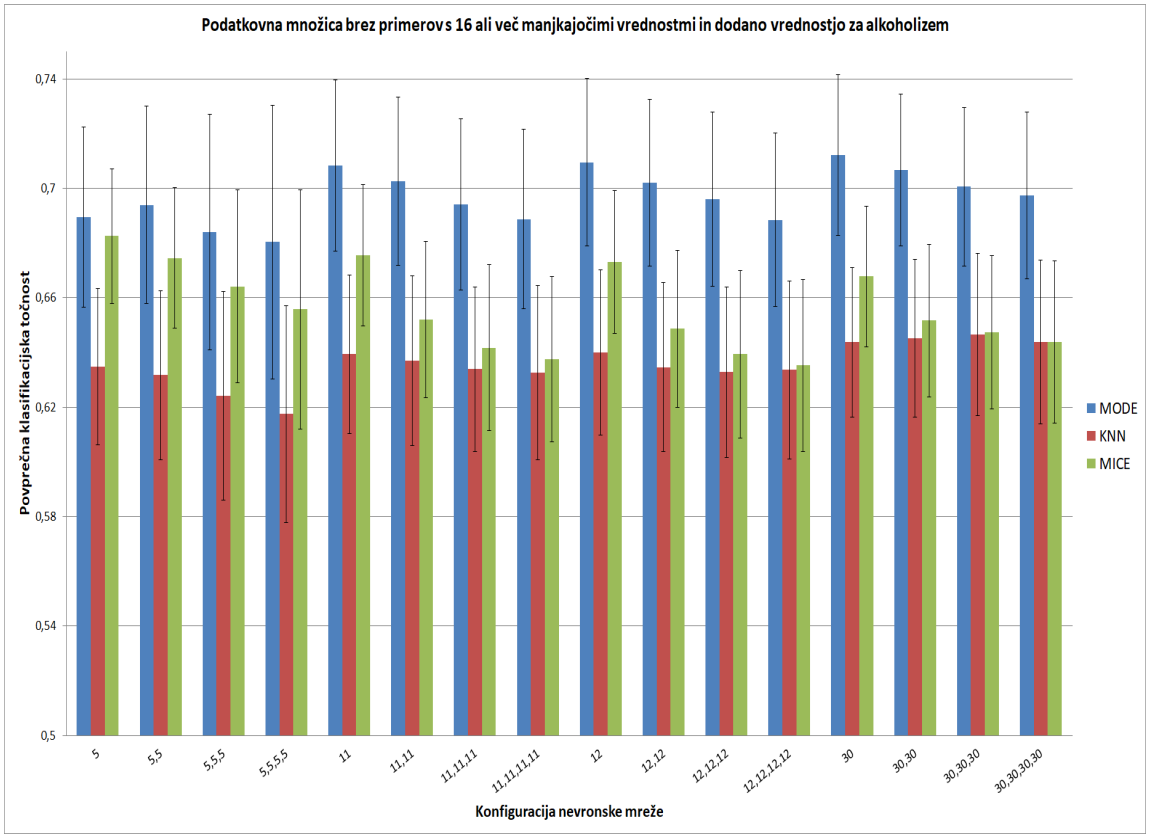
kajočih vrednosti za to začetno podatkovno množico je bil način zamenjave manjkajočih vrednosti z modusom. Izjema temu se pojavi pri konfiguraciji nevronske mreže z le 5 nevroni v skritih slojih, ko ta metoda doseže nekoliko slabšo klasifikacijsko točnost kot metoda MICE.

Začetna podatkovna množica, ki je uporabljala kombinacijo dodatne vrednosti pri atributu alkoholizem, ter odstranitve primerov z več kot 16 manjkajočimi vrednostmi (Slika 5.4), je dosegla povprečno klasifikacijsko točnost 71,06 % s standardnim odklonom 2,93 %. Najslabša povprečna klasifikacijska točnost 61,47 % s standardnim odklonom 3,97 % je bila dosežena pri konfigu-

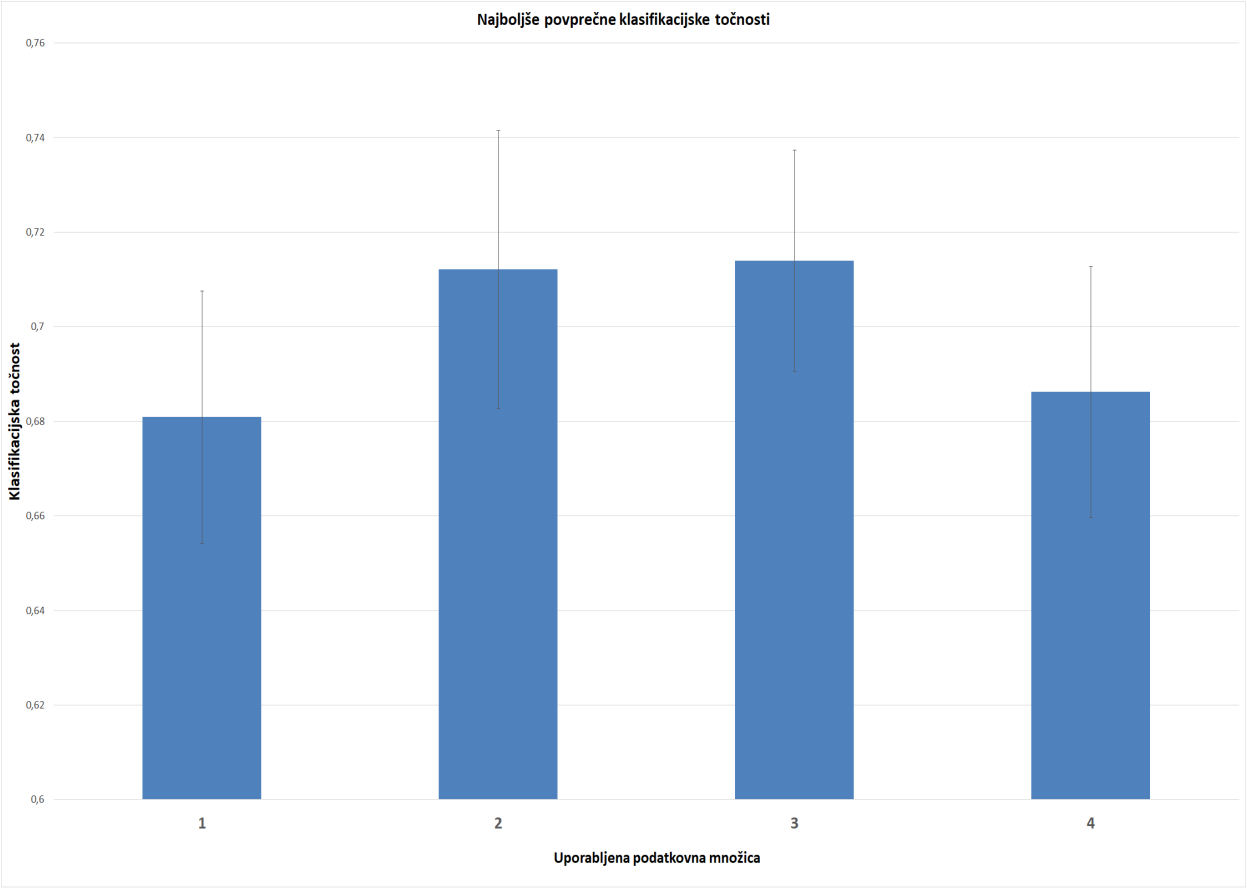


Slika 5.3: Povprečni rezultati klasifikacijske točnosti nevronskih mrež z podatkovno množico, kateri je bila dodana dodatna vrednost za atribut alkoholizem

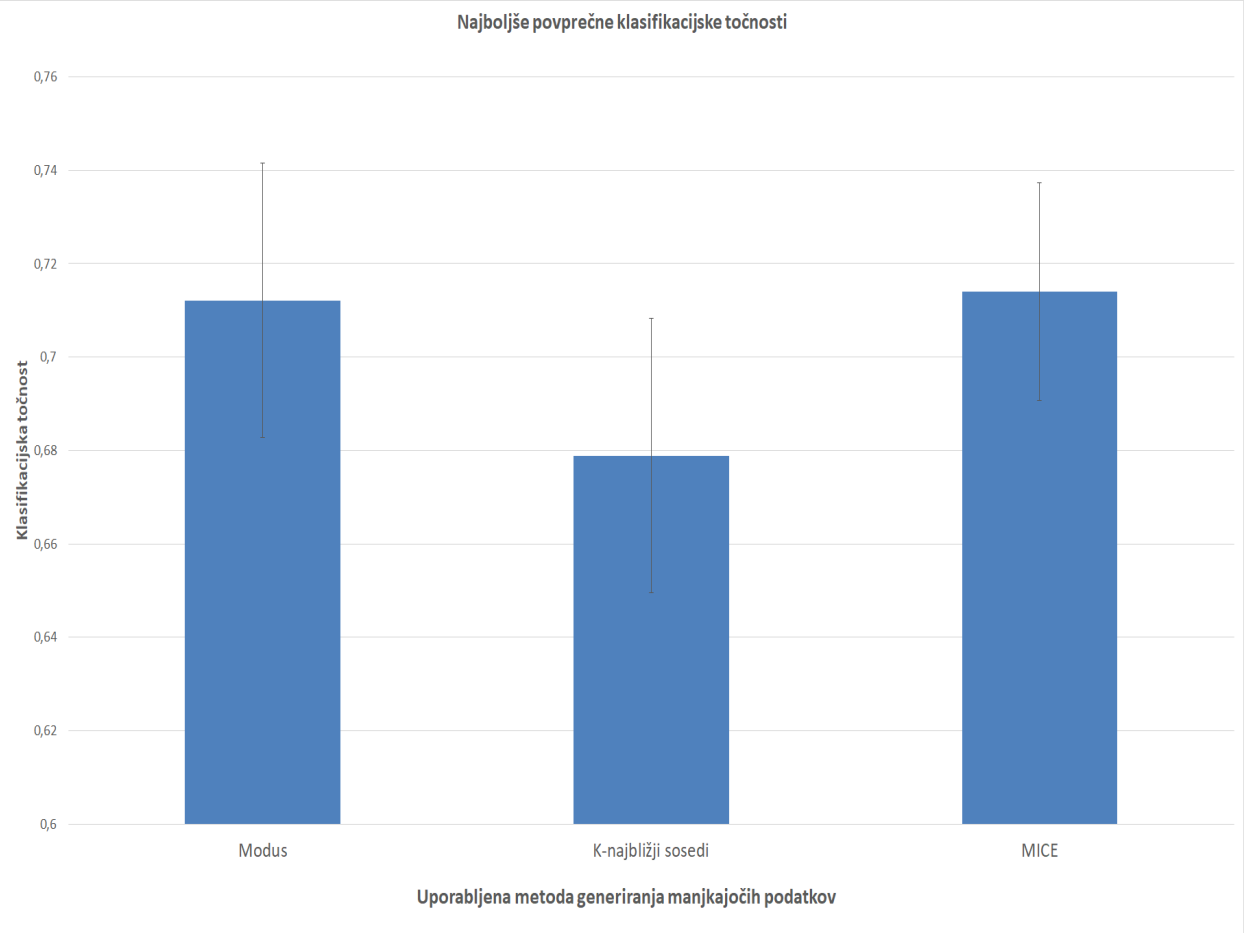
raciji nevronske mreže, ki je vsebovala 4 skrite sloje s po 5 nevroni. Rezultat je bil dosežen pri uporabi zamenjave manjkajoči podatkov z modusom, ter konfiguraciji nevronske mreže, ki je vsebovala 1 skriti sloj s 30 nevroni.



Slika 5.4: Povprečni rezultati klasifikacijskih točnosti nevronskih mrež za podatkovno množico, kateri je bila dodana dodatna vrednost za atribut alkoholizem, ter odstranjeni primeri z več kot 16 manjkajočimi vrednostmi



Slika 5.5: Najboljši povprečni rezultati glede na uporabljeno podatkovno množico



Slika 5.6: Najboljši povprečni rezultati glede na uporabljeno metodo zamenjave manjkajočih vrednosti v podatkovnih množicah

Poglavje 6

Diskusija

V primerjavi s predhodnim projektom, ki je analiziral isto podatkovno množico z različnimi metodami strojnega učenja, so rezultati tega diplomskega dela primerljivi z zelo podobnimi zaključki [12]. V predhodnem projektu so študenti interdisciplinarnega študija Kognitivna znanost, z uporabo nevronskih mrež in algoritmov naključnega gozda (angl. *Random forest*), ter K-najbližjih sosedov za generiranje manjkajočih podatkov, dosegli klasifikacijsko točnost 69,88 %, z uporabo klasifikacijskega drevesa (angl. *Classification tree*) pa točnost 71,27 %. Za primerjavo je najboljši rezultat, ki je bil dosežen z uporabo K-najbližjih sosedov za generacijo manjkajočih vrednosti, v sklopu tega diplomskega dela, dosegel povprečno klasifikacijsko točnost 67,89 % s standardnim odklonom 3,8 %.

Po analizi povprečnih klasifikacijskih točnosti za določeno konfiguracijo nevronske mreže je iz rezultatov razviden trend padanja z dodajanjem več kot ene skrite plasti. Ob tem se rezultati testiranj istega števila skritih slojev in različnega števila nevronov v posameznem sloju le malenkostno razlikujejo. Zaradi tega bi v primeru nadaljnje analize te domene predlagal uporabo enega skritega sloja ter bolj obsežno testiranje primernega števila nevronov v tem sloju.

Rezultati klasifikacijske točnosti nevronske mreže so veliko boljši od slepega ugibanja v primeru te podatkovne množice z najboljšo doseženo pov-

prečno klasifikacijsko točnostjo 71,4 % in standardnim odklonom 2,33 %. Največji vpliv na stanje rezultatov imajo nedvoumno vhodni podatki, katerih je relativno malo. Za količino primerov, ki so na voljo, le ti vsebujejo presenetljivo veliko število manjkajočih vrednosti. Zaradi dejstva, da je bilo te manjkajoče podatke mogoče izračunati na podlagi njihove porazdelitve in drugih atributov, so rezultati te diplomske naloge le približek potencialnim rezultatom, ki bi jih dobili z uporabo primerov brez manjkajočih vrednosti.

Od uporabljenih treh načinov generiranja manjkajočih vrednosti (srednja vrednost, K-najbližjih sosedov, ter MICE) v uporabljeni podatkovni množici so v veliki večini primerov najboljši rezultati izhajali iz podatkovnih množic katerih manjkajoči podatki so bili preprosto zapolnjeni s srednjo vrednostjo. Tu se zato pojavi vprašanje, kakšne so mejne količine (odstotek) manjkajočih podatkov v podatkovni množici, pri katerih kompleksne metode (npr. MICE) delujejo bolje od najbolj preprostih metod. Temu pa bi lahko tudi dodal vprašanje, ali v okviru uporabljene podatkovne množice kompleksne metode zamenjave manjkajočih vrednosti sploh lahko delujejo bolje od preproste metode menjave manjkajočih vrednosti s srednjo vrednostjo. To vprašanje je podprto s pridobljenimi rezultati, saj se MICE le redko izkaže bolje od metode menjave manjkajočih vrednosti z modusom.

Kot zanimivost je potrebno omeniti še vpliv normalizacije na klasifikacijsko točnost naučenih nevronskih mrež. Iz kratkega testiranja normalizacijskih tehnik s podatkovno množico so bile opazne ogromne razlike med točnostjo pri različnih normalizacijskih metodah, kar spodbudi dodatno analizo te podatkovne množice z namenom izboljšave rezultatov klasifikacije.

Poglavje 7

Zaključek

Tekom praktičnega dela diplomske naloge je bilo uporabljenih več načinov predprocesiranja podatkov, kot tudi več različnih oblik nevronske mreže. Izmed vseh je imela največjo povprečno klasifikacijsko točnost 71,19 % nevronska mreža z enim skritim slojem, ki je vseboval 5 nevronov, katere vhodna množica podatkov je bila predprocesirana z uporabo algoritma MICE. Preko vseh konfiguracij nevronske mreže in začetnih vhodnih podatkovnih množic se pojavlja trend padanja klasifikacijske točnosti z dodajanjem skritih slojev, kar kaže na povečevanje pretirane prilagojenosti podatkom nevronske mreže s povečevanjem števila skritih slojev. Z uporabo nevronske mreže je bila v primerjavi z naključnim ugibanjem (52 %) dosežena večja napovedna točnost (71,4 %).

Povprečno najboljši rezultati so nastali pri uporabi podatkovne množice, kateri so bili odstranjeni primeri z več kot 16 manjkajočimi vrednostmi, ter dodana dodatna vrednost za atribut alkoholizem. Kar se tiče algoritma za generiranje manjkajočih vrednosti je splošno najboljše rezultate dosegal algoritem zamenjave manjkajočih vrednosti z modusom.

Pridobljeni rezultati niso popolni odsev realnega sveta, ter tudi skoraj nikoli ne bodo, vendar pa je ta razlika še veliko večja zaradi velike količine manjkajočih vrednosti v podatkovni množici, ki so bili nato izpeljani (izračunani) iz drugih vrednosti v podatkovni množici. Zaradi tega bi bila pred nadalj-

njo analizo tega področja, z namenom čim večjega približevanja rezultatov realnemu svetu, potrebna dopolnitev manjkajočih vrednosti v podatkovni množici.

Literatura

[1] Alan C Acock. Working with missing values. *Journal of Marriage and Family*, 67(4):1012–1028, 2005.

[2] Ann Van Den Bogaert, Kristel Sleegers, Sonia De Zutter, Lien Heyrman, Karl-Fredrik Norrback, Rolf Adolfsson, Christine Van Broeckhoven and Jurgen Del-Favero. Association of brain-specific tryptophan hydroxylase, tph2, with unipolar and bipolar disorder in a northern swedish, isolated population. *Archives of General Psychiatry*, 63(10):1103–1110, 2006.

[3] Avshalom Caspi, Karen Sugden, Terrie E Moffitt, Alan Taylor, Ian W Craig, Honalee Harrington, Joseph McClay, Jonathan Mill, Judy Martin, Antony Braithwaite and others. Influence of life stress on depression: Moderation by a polymorphism in the 5-htt gene. *Science*, 301(5631):386–389, 2003.

[4] Beverly H Brummett, Michael A Babyak, Redford B Williams, Kathleen Mullan Harris, Rong Jiang, William E Kraus, Abanish Singh, Paul T Costa, Anastasia Georgiades, Ilene C Siegler. A putatively functional polymorphism in the htr2c gene is associated with depressive symptoms in white females reporting significant life stress. Dosegljivo: `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114451`. [Dostopano: 14. 7. 2017].

[5] Engelborghs, Sebastiaan and Holmes, Clive, McCulley, Michelle, De Deyn, Peter P. 5-ht˜ 2˜ a receptor polymorphism may modulate antip-

sychotic treatment response in alzheimer's disease. *International Journal of Geriatric Psychiatry*, 19(11):1108–1109, 2004.

[6] Kevin Gurney. *An Introduction to Neural Networks*. Routledge, 1997.

[7] Jeff Heaton. *ntroduction to neural networks with Java*. Heaton Research, Inc., 2008.

[8] Igor Kononenko. *Strojno učenje*. Fakulteta za računalništvo in informatiko, 1997.

[9] Igor Kononenko, Marko Robnik Šikonja. *Inteligentni sistemi*. Založba FE in FRI, 2010.

[10] John J Mann, Yung-yu Huang, Mark D Underwood, et al. A serotonin transporter gene promoter polymorphism (5-httlpr) and prefrontal cortical binding in major depression and suicide. Dosegljivo: `http://jamanetwork.com/journals/jamapsychiatry/fullarticle/481646`. [Dostopano: 14. 7. 2017].

[11] Joshua Kaufman, Christine DeLorenzo, Sunia Choudhury, Ramin V Parsey. The 5-ht 1a receptor in major depressive disorder. *European Neuropsychopharmacology*, 26(3):397–410, 2016.

[12] Jure Fabjan, Iva Ilioska, Miha Medved, Monika Pirc and Maja Zupančič. Machine learning of suicide data of the slovene population. Report for subject: Umetna inteligenca 1, Interdisciplinarni študij Kognitivna znanost, Univerza v Ljubljani, 2016.

[13] Saurabh Karsoliya. Approximating number of hidden layer neurons in multiple hidden layer bpnn architecture. *International Journal of Engineering Trends and Technology*, 3(6):714–717, 2012.

[14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Ankunda R Kiremire. The application of the pareto principle in software engineering. *Consulted January*, 13:2016, 2011.

[16] Kwon Ho Jang, Jin Han Jun and Lim Myung Ho. Association between monoamine oxidase gene polymorphisms and attention deficit hyperactivity disorder in korean children. *Genetic Testing and Molecular Biomarkers*, 18(7):505–509, 2014.

[17] LM Kochersperger, EL Parker, M Siciliano, GJ Darlington and RM Denney. Assignment of genes for human monoamine oxidases a and b to the x chromosome. *Journal of Neuroscience Research*, 16(4):601–616, 1986.

[18] Małgorzata Wrzosek, Jacek Łukaszkiewicz, Michał Wrzosek, Piotr Serafin, Andrzej Jakubczyk, Anna Klimkiewicz, Halina Matsumoto, Kirk J Brower and Marcin Wojnarc. Association of polymorphisms in htr2a, htr1a and tph2 genes with suicide attempts in alcohol dependence: a preliminary report. Dosegljivo: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3169703/`. [Dostopano: 13. 7. 2017].

[19] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis and Philip J Leaf. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011.

[20] Miles Berger, John A Gray and Brian L Roth. The expanded biology of serotonin. *Annual review of medicine*, 60:355–366, 2009.

[21] Society of Old Age Rational Suicide. A brief history of suicide. Dosegljivo: `http://www.soars.org.uk/index.php/about/2014-06-06-18-57-53`. [Dostopano: 18. 5. 2017].

[22] World Health Organisation. Suicide. Dosegljivo: `http://www.who.int/mediacentre/factsheets/fs398/en/`. [Dostopano: 18. 5. 2017].

[23] P Zill, TC Baghai, P Zwanzger, C Schüle, D Eser, R Rupprecht, HJ Möller, B Bondy and MSNP Ackenheil. Snp and haplotype analysis of a novel tryptophan hydroxylase isoform (tph2) gene provide evidence for association with major depression. *Molecular Psychiatry*, 9(11):1030, 2004.

[24] Andrew Rankin. *Seppuku: A History of Samurai Suicide*. Kodansha International, 2011.

[25] Tomaž Zupanc, Peter Pregelj, Martina Tomori, Radovan Komel and Alja Videtič Paska. No association between polymorphisms in four serotonin receptor genes, serotonin transporter gene and alcohol-related suicide. Technical report, Institute of Forensic Medicine, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia, University Psychiatric Hospital Ljubljana, Ljubljana Polje, Slovenia, Institute of Biochemistry, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia, 2010.

[26] Tomaž Zupanc, Peter Pregelj, Martina Tomori, Radovan Komel and Alja Videtič Paska. Tph2 polymorphisms and alcohol-related suicide. *Neuroscience letters*, 490(1):78–81, 2011.

[27] Verônica Contini, Francine ZC Marques, Carlos ED Garcia, Mara H Hutz, Claiton HD Bau. MAOA-uVNTR Polymorphism in a Brazilian Sample: Further Support for the Association With Impulsive Behaviors and Alcohol Dependence. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 141(3):305–308, 2006.

[28] Word Health Organisation. World health organisation suicide mortality rate. Dosegljivo: `http://www.who.int/gho/mental_health/suicide_rates_crude/en/`. [Dostopano: 15. 5. 2017].

[29] Yung-yu Huang, Maria A Oquendo, Jill M Harkavy Friedman, Lawrence L Greenhill, Beth Brodsky, Kevin M Malone, Vadim Khait and John J Mann. Substance abuse disorder and major depression are associated

with the human 5-ht1b receptor gene (htr1b) g861c polymorphism. Technical report, Department of Neuroscience, New York State Psychiatric Institute, Columbia University College of Physicians and Surgeons, New York, USA, 2003.