Statistična analiza podatkov v časovnem zaporedju

Jan Škorjanc
UP FAMNIT, Univerza na Primorskem
Glagoljaška 8, 6000 Koper
Koper, Slovenija
89172028@student.upr.si

***Povzetek* — V članku primerjamo možgane zdravih ljudi z možgani ljudi, ki se spopadajo z epilepsijo. Podatkovna zbirka sestoji iz podatkov elektroencefalograma (EEG), uporabljene pa so statistične metode kot so SVD (singular value decomposition), PCA (Principal component analysis), tSNE (t-Distributed Stochastic Neighbor Embedding)... Cilj postopka je statistični opis podatkov pri katerem si pomagamo tudi s klasifikacijo. Kot klasifikatorja sta uporabljena k-NN in logistična regresija. Statistična analiza poteka v programu Weka in programskem okolju R.**

Ključne besede: EEG, klasifikacija, analiza

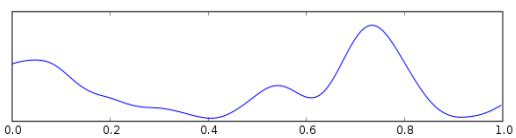
I. UVOD V ELEKTROENCEFALOGRAFIJO

Elektroencefalografija je postopek s katerim z elektrodami merimo možgansko električno aktivnost. Izmerjeni signali so predstavljeni kot vsota aktivacijskih in inhibicijskih postsinaptičnih električnih potencialov, kot tudi akcijskih potencialov živčnih celic. Potenciali so lahko spontani ali izzvani (utripanje luči), izmerjeni pa so v mikrovoltih. Pri merjenju se uporabljajo elektrode, ki so nameščene po celotni glavi pacienta. Preiskava poteka nekaj več kot 20 minut, izvaja pa se lahko tako med spanjem kot v stanju budnosti pacienta. Rezultat, ki ga dobimo po opravljeni elektroencefalografiji se imenuje elektroencefalogram (EEG). S takim postopkom se lahko diagnosticira in spremlja epilepsija, možganska kap, možganski tumor, predoziranje z drogami, poškodbe glave in tudi mentalno zaostalost. Z njim lahko dokažemo možgansko smrt in ugotavljamo globino kome. Za razliko od CT, PET, MRI je pri EEG možno spreminjanje stanja zaznati v milisekundah, zato je še vedno ena izmed najbolj uporabljenih tehnik. Bistven pomen EEG je razlikovati splošen možganski vzorec od vzorca, ki ga povzročajo motnje. Poznamo več vrst valov, ki se jih lahko opazi na EEG. To so alfa, beta, gama, delta in theta [10], [9].

- Delta (0.1 - 4Hz)

Delta val ima največjo amplitudo in predstavlja najbolj počasno možgansko aktivnost. Je normalno prisoten med spanjem in se pojavlja pri globokem dihanju, pri budnem stanju pa lahko nakazuje na bolezen možganov.

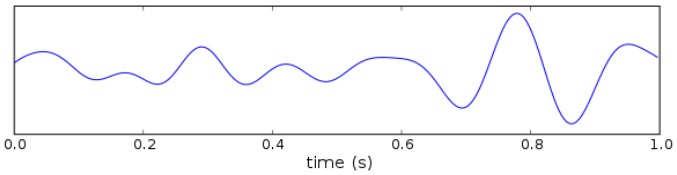
Slika 1: Prikaz delta valovanja



- Theta (4 - 8Hz)

Značilni za mladostnike in starejše ljudi. Pri odraslih pa se pojavljajo v rahlem spanju. Pri otrocih se pojavljajo tako v budnosti, kot v spanju.

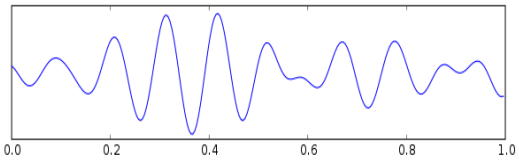
Slika 2: Prikaz Theta valovanja



- Alfa (8 - 15Hz)

Pojavljajo se pri odraslih ljudeh v procesu mirovanja. Pri usmerjanju pozornosti v določen objekt, val izgine.

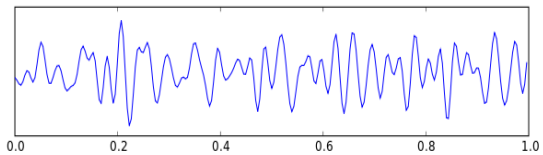
Slika 3: Prikaz Alfa valovanja



- Beta (15 - 31Hz)

Pojavljajo se pri ljudeh med govorom, reševanjem nalog, odločanju in raznih dejavnosti, ki vključujejo razmišljanje. Opazimo jih na delu možganov, ki je namenjen učenju in spominu.

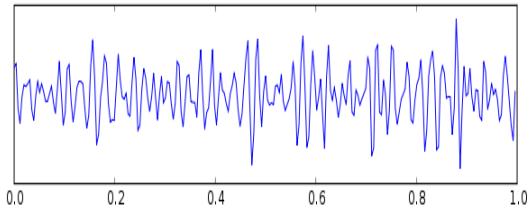
Slika 4: Prikaz Beta valovanja



- Gama (> 31Hz)

Pojavljajo se pri opravljanju zahtevnih nalog in stanja visoke budnosti in zavesti. Tudi med zaznavanjem več čutil (zvok,vid,vonj).

Slika 5: Prikaz Gama valovanja



II. IZBIRA PODATKOV

V originalni podatkovni zbirki je 5 množic (A-E), vsaka od njih pa vsebuje 100 EEG primerov v časovnem intervalu 23,6 sekunde. Segmenti so bili izrezani iz različnih stanj preiskovanja. Množici A in B vključujeta EEG signal zdravih prostovoljcev, vendar so v množici A med testiranjem imeli oči odprte, v množici B pa oči zaprte. Prostovoljci so bili sproščeni in umirjeni. Množici C in D vsebujeta paciente, katerim so diagnosticirali tumor v možganih. V skupino C spadajo merjenja zdravega dela možganov, v skupino D pa tistega dela možganov kjer se je tumor nahajal. Nihče od primerov pacientov v skupinah od A-D ni bil epileptik, množica E pa vsebuje paciente z diagnosticirano epilepsijo. Vsi EEG signali so bili posneti z enako napravo z uporabo digitalizacije 12 bitov. Frekvenca zapisa vzorcev je 173.61Hz, kar pomeni da približno 173 atributov predstavlja vrednosti v razmaku ene sekunde (razlika dveh vrednosti je v razmaku približno 5ms).

III. OPIS STATISTIČNIH METOD

A. *SVD*

SVD (singular value decomposition) je algoritem, ki poskuša zmanjšati rang matrike R na rang matrike K. Z drugimi besedami to pomeni, da lahko poiščemo aproksimacijo linearne kombinacije R vektorjev z novimi K vektorji. Tako matriko A razbijemo na zmnožek treh matrik UDV^T, kjer sta U in V ortonormirani, D pa diagonalna matrika. Ena izmed metod SVD je PCA metoda, njen cilj je zmanjšanje dimenzije podatkov. [3]

B. *PCA*

PCA (principal component analysis) je ena izmed najpogosteje uporabljenih metod. Osnovna ideja metode je opis razpršenosti n enot v m razsežnem prostoru (določenem z m spremenljivkami) z množico nekoreliranih spremenljivk (linearne kombinacije originalnih spremenljivk). Rezultat tako predstavljajo spremenljivke, ki so urejene po pomembnosti od najpomembnejše do najmanj pomembne. Pomembnost si lahko razlagamo kot razpršenost podatkov, ki jo spremenljivka vsebuje. Bolj kot je ta razpršena, bolj je spremenljivka pomembna. Cilj metode je poiskati nekaj prvih spremenljivk (komponent), ki vsebujejo največji del razpršenosti podatkov. Pri PCA podatki izgubijo določen del informacije, zato je največji problem predstavlja število novo narejenih spremenljivk. Če vzamemo teh spremenljivk premalo, se lahko zgodi, da izgubimo večji del informacije začetnih podatkov in tako so podatki neuporabni. V primeru da vzamemo preveč novih komponent, pa obstaja verjetnost, da so podatki enako razpršeni kot pred uporabo PCA. Algoritem deluje tako, da določi linearno kombinacijo spremenljivk tako, da je varianca te kombinacije največja [5]. PCA je definirana z enačbo:

$$Y_j = Xa_j ; \qquad j=1\ldots.m,$$

kjer je :

X – matrika podatkov

a_j – vektor uteži

in

$$\mathrm{var}(Y_j) = \mathrm{var}(Xa_j) = \max$$

C. *T-SNE*

t-SNE je algoritem strojnega učenja, ki je namenjen zmanjšanju dimenzije podatkov. Je algoritem, ki visoko dimenzijske podatke zelo dobro pretvori v dve dimenziji, ki jih potem lahko prikažemo na grafu. To stori tako, da točkam, ki so si bolj podobne in bolj povezane priredi vrednosti, ki so si bližje v grafu, točkam, ki so si manj povezane pa priredi vrednosti, ki so si na grafu bolj oddaljene. Sestavljen je iz dveh glavnih faz. V prvi fazi naredi verjetnostno porazdelitev tako, da visoko dimenzionalnim objektom, ki so si podobni dodeli veliko verjetnost izbora, tisti ki si pa niso podobni pa majhno verjetnost izbora. V drugi fazi pa izbranim objektom definira podobno verjetnostno porazdelitev točkam v nizkih dimenzij in minimizira Kullback-Leiberevo divergenco med obema porazdelitvama. Verjetnost se izračuna po tej enačbi: [7]

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

D. *Logistična regresija*

Logistična regresija je različica regresije, kjer je odvisna spremenljivka kategorična, kar pomeni, da lahko zavzame zgolj določene vrednosti. Tak primer so recimo krvne skupine, binarne vrednosti itd. V osnovi je podobna linearni regresiji, ki je transformirana z logistično funkcijo (od tod tudi njeno ime). Deluje po enačbi: [1]

$$f(x) = \frac{L}{1 + e^{-k(x - x_0)}}$$

E. *k-NN*

k-NN algoritem deluje tako, da izračuna razdaljo do vseh instanc, ki imajo znano vrednost razreda. Klasifikacija poteka na podlagi k najbližjih instanc. Algoritem se ne uči, ampak za vsak nov primer, ki ga želi klasificirati na novo izračuna razdalje. Je časovno zahteven algoritem, ki zasede veliko prostora. Njegova prednost pa je njegova enostavnost [8]. Za izračun razdalje se lahko uporabi več različnih formul. V

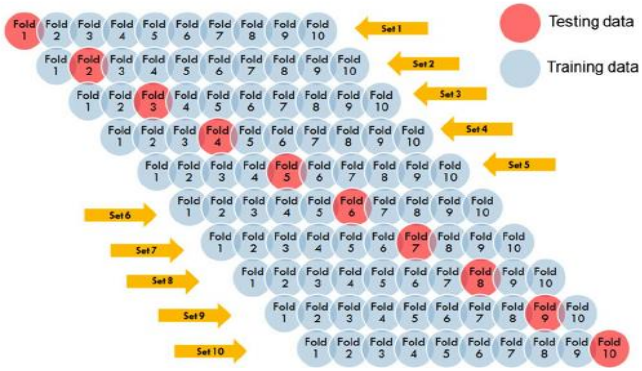
raziskavi je bila uporabljena Evklidska razdalja, ki se jo izračuna po formuli:

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + \cdots (a_n - b_n)^2}$$

F. *k-kratno prečno preverjanje*

K-kratno prečno preverjanje (angl. K-fold cross validation) je način vrednotenja algoritmov. Podatkovno zbirko razdeli na k enako velikih disjunktnih podmnožic. Nato uporabljen klasifikacijski algoritem zgradi učni model na k-1 podmnožicah, na zadnji pa ga testira. Postopek se ponovi k-krat, kjer se podmnožice vsakič zamenjajo, kar prikazuje slika. Natančnost algoritma izračuna tako, da število pravilno napovedanih instanc deli s številom vseh instanc v podmnožici [6]. V raziskavi je bilo uporabljeno 10-kratno prečno preverjanje, ki se je izvedlo 10-krat. Kot rezultat je izpisano povprečje vseh desetih iteracij.

Slika 6: 10-kratno prečno preverjanje



IV. OPIS PROGRAMSKE OPREME

A. *Weka*

Weka (Waikato Environment for Knowledge Analysis) je programska oprema, ki vsebuje zbirko algoritmov za strojno učenje in podatkovno rudarjenje. Program je v celoti napisan v programskem jeziku Java in je odprtokoden, saj je na voljo pod GNU licenco. Ime je dobil po vrsti ptice, ki živi na Novi Zelandiji, saj je bil tam tudi program narejen [2]. Podpira splošne algoritme za klasifikacijo, regresijo, razvrščanje v skupine in ostale metode podatkovnega rudarjenja. Weko lahko uporabljamo na več načinov, zato je razdeljena na 5 delov. To so raziskovalec (angl. explorer), preizkuševalec (angl. experimenter), tok znanja (angl. knowledgeFlow), delovno okolje (angl. workbench) in preprost vmesnik z ukazno vrstico (angl. simple CLI) Algoritmi se izvajajo na podatkih, ki morajo biti v program naloženi. Weka podpira več formatov datotek, kot so CSV, LibSMV in C4.5, vendar vse pretvori v poseben format imenovan ARFF. Ta vsebuje glavo, v kateri podamo ime relacije in opišemo atribute.

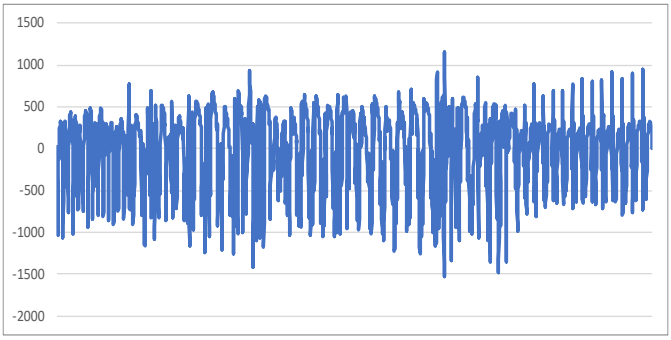
B. *R*

R je programski jezik in brezplačno programsko okolje namenjeno statističnem programiranju in vizualizaciji podatkov. Je eden izmed najbolj pogostih uporabljenih jezikov med statistiki in uporabniki podatkovnega rudarjenja. Na voljo je pod GNU licenco, deluje pa že od leta 1993. Omogoča linearno in nelinearno modeliranje, vizualizacijo, statistične teste, časovno analizo podatkov, klasifikacijo, razvrščanje v gruče. Ima kar nekaj paketov, ki nudijo široko paleto algoritmov [4].

V. POSTOPEK DELA IN REZULTATI

V podatkovni zbirki so instance razdeljene na 5 razredov (množic) kot je predstavljeno v poglavju "Izbira podatkov". Za potrebe raziskave, v kateri nas zanima zgolj ali je oseba epileptik ali ne, sem te razrede spremenil. Tako sem vrednosti od A do D spremenil v vrednost N (ni epileptik), množico E pa v vrednost Y (je epileptik). Ker je podatkov zelo veliko, bi bilo vseh 23 sekund (več kot 4000 meritev) težko predstaviti, saj bi bili podatki preveč zgoščeni (slika 7), zato sem se odločil, da bom podatke skrčil in se osredotočal na manjši del podatkov. Tako bo naša podatkovna zbirka vsebovala zgolj eno sekundo meritve (173 atributov). Če bi nam uspelo podatke dobro klasificirati, bi tako že po zgolj eni sekundi EEG ugotovili ali je oseba epileptik ali ne in s tem proces merjenja zelo skrajšali.

Slika 7: Prikaz EEG vrednosti vseh 23 sekund enega pacienta



Tako bo podatkovna zbirka v splošnem vsebovala 500 instanc (100 epileptikov, 400 neepileptikov) in 173 atributov.

A. *Osnovna statistična analiza*

Tabela 1: rezultati osnovne statistične analize nad podatki

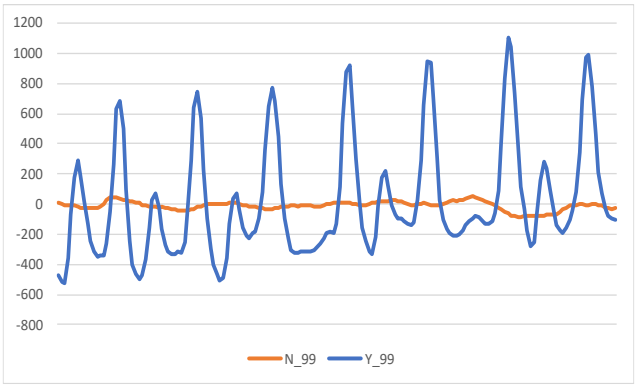
	N	Y
Minimum	-889	-1885
Maksimum	2047	1624
Mediana	-8	-5
Povprečje	-9,25	-7,92
Standardni odklon	70,70	347,14

Iz osnovnih statističnih meritev lahko ugotovimo, da imajo vsi primeri v povprečju podobne rezultate, saj je razlika med epileptiki in neepileptiki zgolj za 1,33 mikrovolta, kar je zanemarljivo malo. Lahko pa ugotovimo, da imajo epileptiki v večini primerov večji impulz možganov, kar nam pove tudi maksimum in minimum vseh meritev. Mediana nam potrdi enako tezo.

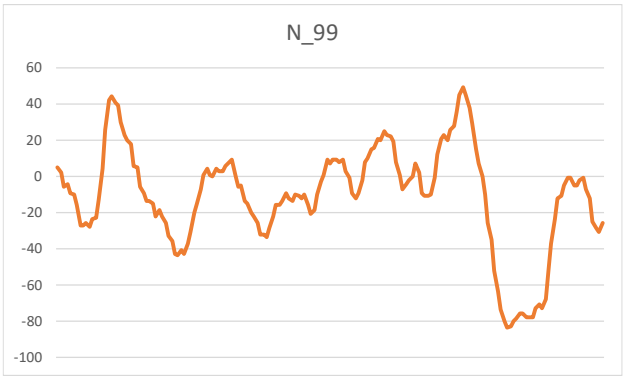
Največjo razliko med obema skupinama primerov pa nam da standardni odklon, ki nam pove za koliko se v povprečju posamezna meritev razlikuje od povprečne vrednosti meritve. Z drugimi besedami nam prikaže razpršenost meritve, kar v primeru EEG pomeni večjo amplitudo vala. Vidimo, da imajo epileptiki v eni sekundi kar 5x večje nihanje kot neepileptiki, kar nam prikazuje tudi slika 8, ki predstavlja primer vsakega primera skupine v eni sekundi.

Slika 8 nam prikazuje naključni izbrani primer epileptika in neepileptika v času ene sekunde. Vidimo, da je v primeru epileptika nihanje zelo izrazito, kar potrjuje tudi standardni odklon. Zaradi tako izrazitega nihanja se zdi, da impulz v možganih neepileptika sploh ne niha, vendar nam standardni odklon pove, da tudi impulz v zdravih možganih niha, le da je to nihanje v primerjavi z epileptikom skoraj neopazno. Za lažjo predstavo nihanja zdravih možganov je na sliki 9 predstavljen enak primer nepileptika v svojem grafu. Iz slik razberemo, da je v tem primeru pri epileptiku prisoten alfa val, pri neepileptiku pa theta val. Potrebno pa je poudariti, da je to zgolj primer ene sekunde. Posledično se je potrebno zavedati, da je morda valovanje v naslednjih sekundah drugačno in da prisotnost enega tipa valovanja ne nujno diagnosticira bolezni v eni sekundi časa.

Slika 8: Prikaz EEG obeh primerov v eni sekundi

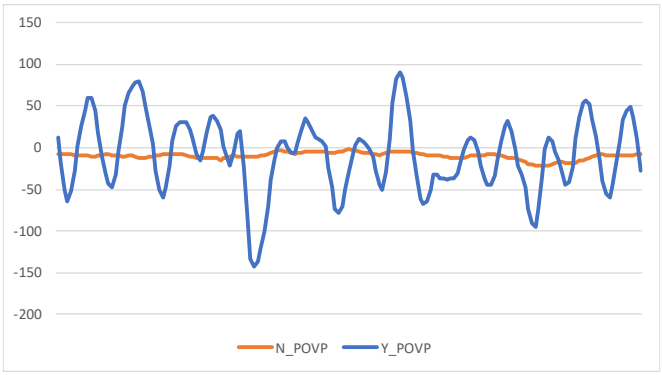


Slika 9: Prikaz EEG neepileptika v eni sekundi



Poleg prikaza posameznega primera EEG nam slika 10 prikazuje povprečne vrednosti primerov epileptikov in neepileptikov. Tudi povprečne vrednosti nam potrdijo to, da je nihanje možganske aktivnosti pri epileptikih bolj izrazito v primerjavi z neepileptiki.

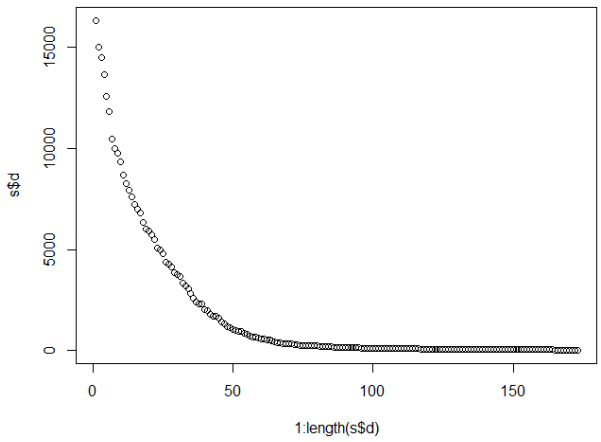
Slika 10: Povprečna vrednost EEG



B. *SVD*

Po splošnem postopku SVD, ki je bil narejen v R-ju dobimo tri matrike, iz katerih lahko rekonstruiramo podatke. Ena izmed teh je matrika D, ki nam določi pomembnost pridobljenih podatkov, kar je prikazano na grafu na sliki 11. Iz tega grafa lahko vidimo, da se pri številki 50 pomembnost atributov ustavi, kar pomeni, da lahko z zgolj 50-imi atributi opišemo večji del podatkovne množice. Z drugimi besedami to pomeni, da je vpliv preostalih atributov skoraj zanemarljiv. Glede na to ugotovitev so bili konstruirani podatki na podlagi zgolj 50 atributov.

Slika 11: Graf D matrike



C. *PCA*

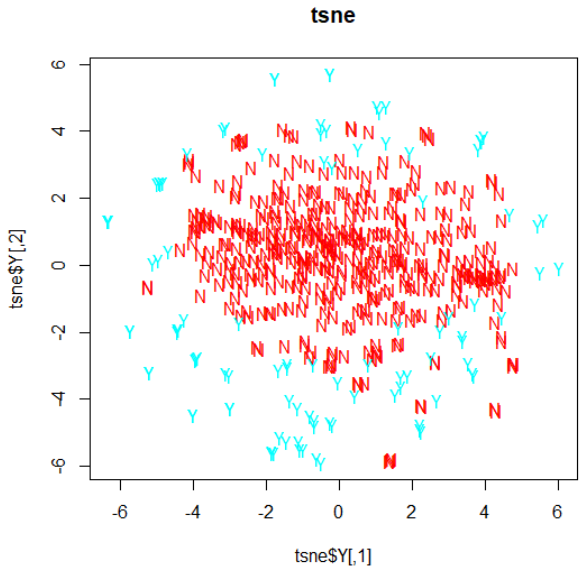
PCA je edini algoritem statistične obdelave, ki je bil implementiran v Weki. Ta nam obdela vhodne podatke tako, da določi nove atribute na podlagi linearne kombinacije prejšnjih atributov in s tem zmanjša dimenzijo podatkov. Kot vhodni parameter sem določil 95% vsebovanost informacije starih podatkov v novih podatkih. Algoritem je kot rezultat vrnil 30 novih atributov. To pomeni, da so bili podatki tako preurejeni,

da je za njihov opis 95% informacije namesto 173 potrebnih zgolj 30 atributov.

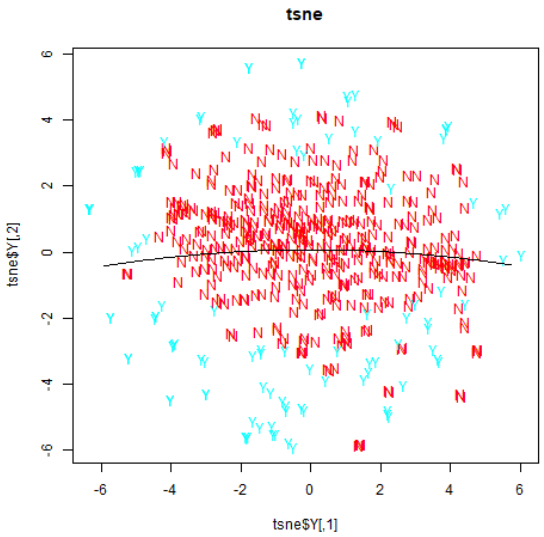
D. T-SNE

Algoritem t-SNE je zmanjšal dimenzijo podatkov iz 173 na zgolj dva atributa. Tako je mogoče vrednosti prikazati tudi na grafu (slika 12). Z rdečo barvo so označeni primeri zdravih možganov, z modro pa možgani epileptikov. Tako je že iz grafa lepo vidno, da so podatki razdeljeni na dve skupini. Primerov epileptikov okoli točke (0,0) ni in so zelo razpršeni po ravnini. Ravno nasprotno pa je pri večini primerov zdravih možganov, ki se nahajajo okoli točke (0,0). Grafu t-SNE podatkov je bila dodana še nelinearna least squares regresija, ki je prikazana na sliki 13.

Slika 12: Prikaz razpršenosti podatkov po t-SNE algoritmu



Slika 13: Dodana nelinearna least squares regresija



.

VI. REZULTATI KLASIFIKACIJE

Za klasifikacijo sta bila uporabljena algoritma k-NN in Logistic regression. Pri k-NN je bil k=1. Klasifikatorja sta bila naključno izbrana in sta implementirana že v Weki. Rezultati klasifikacije so predstavljeni v tabeli 2.

Tabela 2: Prikaz uspešnosti klasifikacije v odstotkih

	k-NN	Logistic regression
Originalni podatki	93,4	85,4
SVD	93.4	87.4
PCA	89,4	88
t-SNE	94,2	83,2

Iz tabele lahko vidimo, da je k-NN v splošnem bolje klasificiral podatke kot logistična regresija. Očitno je, da je znotraj ene sekunde hitro mogoče ugotoviti razliko med epileptiki in neepileptiki, saj je uspešnost klasifikatorjev (okoli 90%) zelo visoka in dobra. Iz rezultatov je razvidno, da je SVD v katerem je bilo uporabljenih zgolj 50 najboljših atributov enako dober kot originalni podatki oz. pri logistični regresiji celo boljši. Pri PCA postopku padec uspešnosti k-NN algoritmu ne preseneča, saj je z manjšo količino podatkov težko pričakovati boljši rezultat. Pozitivno pa se je odrezal t-SNE, ki edini izboljša uspešnost k-NN klasifikatorja. Iz rezultatov tudi vidimo, da je pri vseh postopkih uspešnost klasifikacije sorazmerno porazdeljena glede na več klasifikatorjev. Če pri enem klasifikatorju izgubimo uspešnost, jo pri drugem pridobimo.

VII. ZAKLJUČEK

Med postopkom analize podatkov EEG ugotovimo, da je razlika med epileptiki in neepileptiki v procesu merjenja jasno razvidna brez dodatnih statističnih postopkov. To nam potrdijo tudi statistične meritve, ki nam razkrijejo veliko razliko v nihanju možganske aktivnosti med obema skupinama. Tudi klasifikacija nad podatki je bila nadpovprečno uspešna z okoli 90% pravilno klasificiranih podatkov. S tem je bilo pokazano, da je epilepsija bolezen, ki jo lahko hitro diagnosticiramo in za diagnozo lahko poskrbi kar računalnik.

VIRI IN LITERATURA

[1] J. S. Cramer, *The origins of logistic regression*, 119. Tinbergen Institute, 2002

[2] E. Frank, M. A. Hall, in I. H. Witten, *The WEKA Workbench Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 4. izd, 2016.

[3] G. H. Golub in C. Reinsch, *Singular Value Decomposition and Least Squares Solutions*, Heidelberg, 1971

[4] K. Hornik (4. oktober 2017). R FAQ [Online]. Dosegljivo: https://cran.r-project.org/doc/FAQ/R-FAQ.html#What-is-R_003f. [Dostopano: 5. 5. 2019].

[5] I. T. Jolliffe, *Principal Component Analysis, Series: Springer Series in Statistics*, 3. izd. Springer NY, 2002

[6] R. Kohavi, *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, 2016.

[7] L. J. P. van der Maaten in G. E. Hinton, *Visualizing Data Using t-SNE*, Journal of Machine Learning Research. 9, 2008, str. 2579–2605

[8] T. Mitchell, *Instance Based Learning*, Mach. Learn., 1997, str. 199–214

[9] E. Niedermeyer in F. L. da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams & Wilkins, 2004

[10] R. Sucholeiki. (6. oktober 2017) Normal EEG Waveforms [Online]. Dosegljivo:https://emedicine.medscape.com/article/1139332-overview#a1. [Dostopano: 19. 4. 2019].