[1]

Detekcija epileptičnih napada iz EEG signala

Svetlana Baćina, BI 38/2019,
svetlana.bacina@best-eu.org

I. Uvod

Epilepsija je poremećaj mozga koji javno mnjenje najčešće dovodi u spoj sa napadima i padanjem u nesvest. Smatra se da u svetu ima oko 100 miliona osoba koje su obolele od epilepsije, dok se u Srbiji procenjuje da ih ima oko 75 000.

Mozak neprekidno generiše električne impulse reda milivolta [mV] koji se duž neurona prostiru u mozgu i kroz čitavo telo pomoću neurotransimeta - hemijskih veza.

Napad se obično definiše kao iznenadna promena ponašanja usled disbalansa hemije mozga koje uzrokuje iznenadni nalet električnih impulsa u mozak.

Normalni električni obrazac je poremećen iznenadnim i sinhronizovanim naletima električne energije koji mogu nakratko uticati na njihovu svest, pokrete ili senzacije.

Kao i kod drugih neuroloških poremećaja, procena i dijagnostika se vrše se elektroencefalogramom [EEG]. EEG je ključan za tačno prepoznavanje različitih oblika epilepsije.

Kako bismo kreirali modeč kojim bismo pomogli klasifikaciji stanja pacijenta, koristićemo dostupnu bazu snimanu na 500 snimaka. Rezultati će biti obrađeni u ovom projektu.

II. Baza podataka

Unutar ovog rada obrađujemo bazu podataka *"data_processed.csv"*, ali ćemo se u par navrata osvrnuti na bazu "*data.csv"*. Baza "*data.csv"* se sastoji od 500 EEG snimaka različitih ispitanika po 23.6s. Svaki snimak je podeljen na delove od po 1s, odnosno 178 odbiraka, te je svaki deo posmatran kao zaseban uzorak. Baza podataka "*data_processed.csv"* je bazirana na obrađenim signalima iz prethodno spomenute baze. Ona sadrži **11500 uzorka** snimljenih na 500 ispitanika za 23 sekunde.

Unutar te baze su izdvojena **četiri numerička obeležja**: *maksimalna amplituda signala, standardna devijacija, dinamički opseg amplitude i broj preseka signala sa nulom*. Poslednja kolona *y* predstavlja **kategoričko obeležje** predstavljeno numerički, brojevima od 1 do 5 (5 - otvorene oči tokom snimanja, 4 - zatvorene oči tokom snimanja, 3 - EEG sniman iz zdravog dela mozga (prethodno potvrđena tačna na lokacija tumora), 2 - EEG sniman sa lokacije gde se nalazi tumor, 1 - snimljeno tokom epilepticnog napada).

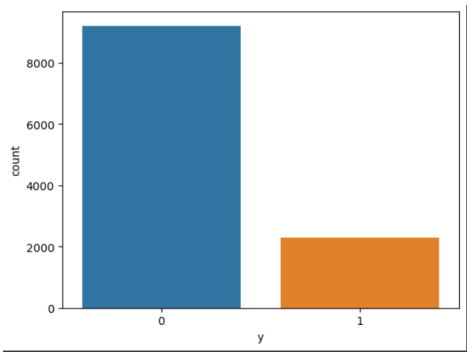
Potencijalni nedostatak podataka unutar baze smo proveravali upotrebom funkcije *isnull* i utvrdili smo da do istog nije došlo.

[1]

III. Analiza podataka

A. Vizuelizacija kategoričkog obeležja

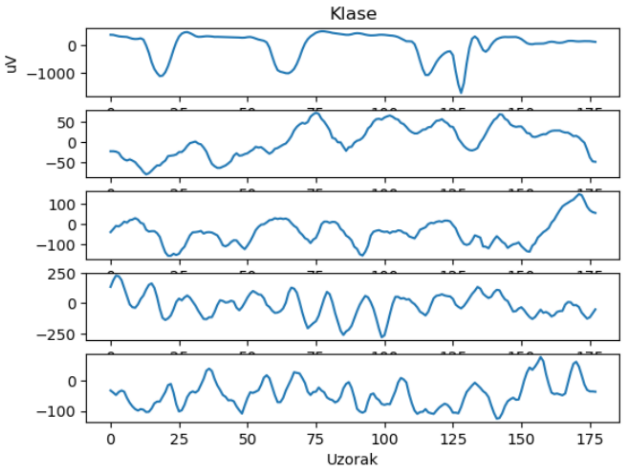
Korišćenjem funkcije *countplot* iz bibiloteke *seaborn* dobija se grafički prikaz raspodele kategoričkog obeležja *y*. Kako je isto bilo prikazano numerički, nije bilo potrebe za novim labelama. Sva stanja koja ne predstavljaju epileptični napad postavljena su na *0*. Ukupno imamo 9200 uzoraka koji ne sadrže epileptični napad i 2300 uzoraka u kojima se dogodio napad (Slika 1).



Sl. 1. Grafički prikaz kategoričkog obeležja

B. Prikaz EEG signala po klasama

Na slici 2. su prikazani uzorci signala za svaku klasu. Uočavamo da se kod signala iz klase 1 - snimljeni tokom epileptičnog napada opseg kretanja amplitude znatno razlikuje u odnosu na ostale klase. Takođe, primećujemo i pikove (spikes) koji su karakteristični za ovo stanje.



Sl. 2. Uzorci signala za klase 1-5 od gore ka dole

C. Statističke veličine parametara

Korišćenjem funkcije *describe* dobijamo informacije o statističkim parametrima obeležja. Pomoću datih podataka, dobijen je uvid o parametrima: srednja vrednost (mean), standardna devijacija (std), minimalna (min) i maksimalna (max) vrednost, medijana (50%), donji kvartil (25%) i gornji kvartil(75%) opseg.

Na osnovu tabele (Tabela 1.) može se uočiti da postoji znatna razlika između gornji i maksimuma. Ako uzmemo u obzir da 20% klasa čini klasa koja sadrži snimak epileptičnog napada, ova razlika ima smisla.

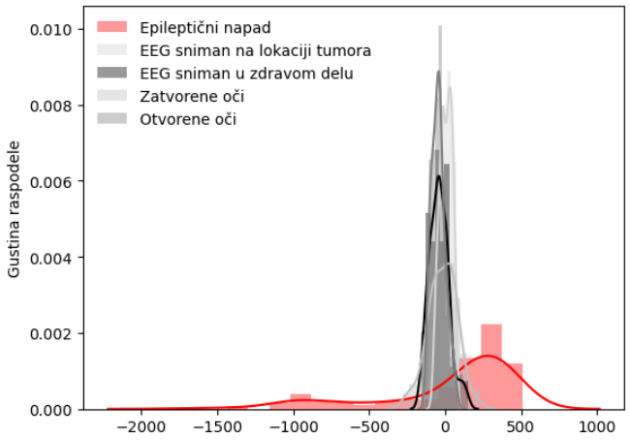
Tabela 1. Statistički parametri obeležja Range i Std

Veličina	Range	Std
Mean	471.01	101.92
Std	577.35	125.62
Min	49.00	10.71
25%	172.00	36.55
50%	243.00	51.05
75%	422.25	90.31
Max	3696.00	810.40

Visoka varijabilnost u aktivnosti mozga kao i velik dinamički opseg amplitude su karakteristički kod EEG signala tokom Epileptičnog napada.

D. Vizuelno poređenje raspodele različitih stanja

Pomoću funkcije *distplot* iz biblioteke *seaborn* dobijen je grafik na kom je iscrtana distribucija podataka, prikazana pomoću distplotova. Distplotovi predstavljaju jednodimenzionalnu distribuciju podataka, odnosno distribuciju promenljive u odnosu na gustinu.



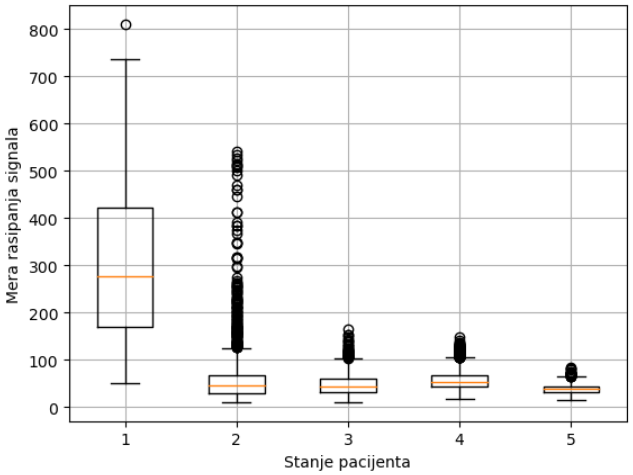
Sl. 3. Grafički prikaz zavisnosti koncentracije čestica PM2.5 od temperature, pritiska i vlage

Uočavamo da su histogram klasa u kojima se nije dogodio epileptični napad u obliku “peak”-a, dok je za klasu u kojoj se dogodio epileptični napad širi i značajno različit u odnosu na ostale, što se vidi na slici broj 3. Različitost u obliku histograma može biti olakšica prilikom kreiranja klasifikacionog modela.

E. Uporedni prikaz mere rasipanja raličitih klasa

Na slici 4. prikazani su boxplotovi na osnovu standardnih devijacija pet različitih grupa. Uočavamo da je srednja vrednost standardne devijanse kod klase u kojoj se dogodio epileptični napad značajno veći u odnosu na druge klase, što svedoči o većoj meri rasipanja signala, odnosno da su vrednosti podataka daleko razmaknute u odnosu na srednju vrednost.

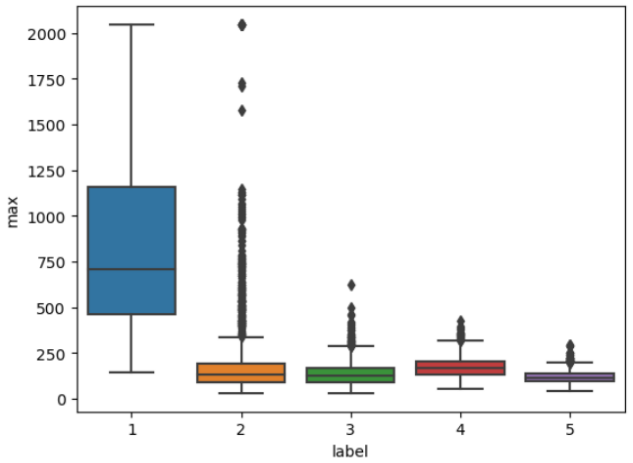
Kod EEG signala koji je snimljen na lokaciji tumora uočavamo da ima dosta autlajera



Sl. 4. Grafički prikaz standardne devijanse postojećih klasa

F. Vizuelni prikaz zavisnosti maksimalne amplitude signala od stanja pacijenta

Maksimalna amplituda signala se često koristi u analizi EEG signala za detekciju artefakta. Korišćenjem funkcije *boxplot* prikazan je grafik na slici broj 5.



Sl. 5. Grafički prikaz zavisnosti pojave maksimalne amplitude od stanja pacijenta

Posmatranjem prikazanog grafika možemo zaključiti da se najveće amplitude pojavljuju kod pacijenta koji doživljava epileptični napad. Kad je EEG sniman na lokaciji na kojoj se nalazi tumor uočavamo veliki broj autlajera.

IV. KNN KLASIFIKATOR

Metoda k najbližih suseda (engl. *k nearest neghbors* - *k*NN) predstavlja neparametarsku metodu klasifikacije. Ova metoda spada u grupu metoda kasnog učenja (engl: lazy learning), koje podrazumevaju odlaganje obrade uzoraka za obuku do trenutka kada treba klasifikovati neobeleženi uzorak.

A. Priprema

Kolona *label* je pomoću funkcije *labela_napad* svedena na

binarni problem, gde su sve vrednosti u kojima se nije dogodio epileptični napad svedene na 0.
Pomoću funkcije *iloc* u promenjivu *X* su smeštena obeležja, a u promenjivu *y* je smeštena poslednja kolona koja predstavlja klase .
Pomoću funkcije *train_test_split* izvršena je podela uzoraka na 20% za test finalnog klasifikatora i 80% ostalih uzoraka na kojima će se vršiti metoda unakrsne validacije sa 10 podskupova.

B. *Evaluacija klasifikatora*

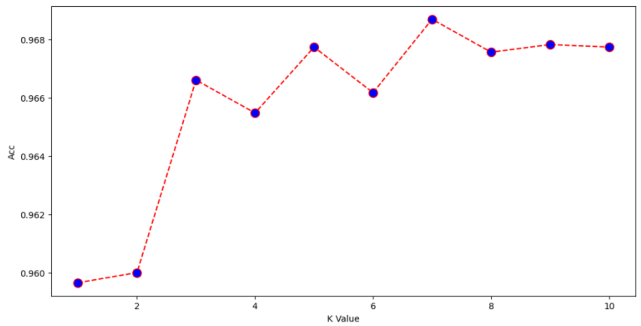
U cilju određivanja mera uspešnosti za svaku klasu, napisana je funkcija *Evaluation_Classifier*. Kao ulaz u funkciju unosi se matrica konfuzije. U matrici konfuzije pojedine vrste odgovaraju stvarnim vrednostima labela, dok pojedine kolone odgovaraju predviđenim vrednostima (oznaka klase koju je uzorku dodelio klasifikator). Na osnovu matrice konfuzije računaju se mere uspešnosti klasifikatora: tačnost, preciznost, osetljivost , specifičnost i F-mera.

C. *Određivanje optimalnih parametara metodom unakrsne validacije*

Metoda unakrsne validacije skup podataka za obuku se deli na približno 10 jednakih podskupova. Ono po čemu se odlikuje ova metoda je to što se svaki od uzoraka jednom javi kao test uzorak. Skup podataka bude bolje iskorišćen. Prosek ponašanja ovih 10 klasifikatora daje očekivano ponašanje na neočekivanim uzorcima.
Podelu na 10 podskupova smo izvršili pomoću funkcije *StratifiedKFold*. Vrednost parametra *k* označava koliko će suseda biti uzeto u obzir, te kroz petlje isprobavamo vrednosti od 1 do 10, kako bismo konačan model mogli da obučimo pomoću klasifikatora koji daje najbolje rezultate. Za metriku su izabrane *manhattan* i *minkowski*.

D. *Rezultati klasifikacije metodom Knn*

Nakon izvršenih iteracija najveću tačnost je pokazala metrika *manhattan*, za k=9, gde je tačnost 97%. (slika 6.)



Sl. 6 Grafik uspešnosti Knn klasifikacije za Manhattan metriku

Matrica konfuzije je prikazana na tabeli 2. Vrste prikazuju prave vrednosti, dok se u kolonama nalaze vrednosti koje su predviđene.

	Predviđene vrednosti	
Stvarne vrednosti	9009	191
	179	2121

Tabela 2. Matrica konfuzije za model klasifikacije Knn metodom

Na glavnoj dijagonali matrice konfuzije nalaze se ispravno klasifikovani uzorci. Može se zaključiti da model ima dobru sposobnost prepoznavanja.

V. Logistička regresija

Logistička regresija predstavlja metodu nadgledanog učenja koja koristi se za predviđanje vrednosti binarne promenljive. Logistička regresija definiše se kao generalizovan linearni regresioni model koji za dati uzorak na ulazu predviđa verovatnoću da uzorak pripada klasi y = 1 koristeći pogodnu nelinearnu funkciju ulaznih promenljivih (x1, x2, . . . , xD) = x. Nelinearna funkcija koja se koristi u modelu logističke regresije je sigmoidalna ili logistička funkcija.

A. *Podela podataka*

Kako bismo testirali podatke na skupu koji nije skup obuke, potrebno je bilo izvršiti podelu na test skup koji čini 10% baze podataka, 10% validacioni skup i ostatak baze koji čine skup za obuku modela. Ukoliko ne bismo prethodno podelili podatke pre samog obučavanja, testirali bi sa već viđenim podacima i procena ne bi bila relevantna. Funkcijom *train_test_split* smo izvršili podelu.

B. *Mere performansi*

U cilju lakšeg određivanja najboljih parametara za model logističke regresije koristićemo se evaluacionim parametrom tačnosti koji dobijamo pomoću ugrađene funkcije *accuracy_score,* kao i *recall* gde ćemo otkriti osetljivost. Osetljivost će pokazati koliko dobro model uspeva da identifikuje stvarne pozitivne instance.

C. *Evaluacija performansi logističke regresije*

Kao maksimalni broj iteracija *max_iteration* smo postavili 100, 200, 500, 1000 i 2000. Solveri koji su korišćeni su *netwon-cg, lbfgs, liblinear* i *saga.* Svaki solver pokušava da pronađe vrednosti parametara koji minimiziraju funkciju troškova.
Najveća tačnost je postignuta pomoću solvera *newton-cg* i *lbfgs*, dok je maksimalni broj iteracija 2000. Preciznost modela sa ovim parametrima je 82.6%, a tačnost 95.8%.

VI. SVM Klasifikator

Klasifikator pod nazivom "mašina na bazi vektora nosača" (engl. *Support Vector Machine* - SVM) se zasniva na *klasifikatoru maksimalne margine*, čiji je cilj da podeli prostor na dva dela tako da se u jednom delu nađu samo uzorci jedne klase, a u drugom druge, koristeći hiperravan. Da bi ovo bilo moguće, potrebno je da klase budu linearno separabilne, što kod brojnih klasifikacionih problema nije slučaj. U tim slučajevima se ne može koristiti klasifikator maksimalne margine, jer se dozvoljava da uzorci jedne klase budu na pogrešnoj strani hiperravni razdvajanja. Tada se koristi *klasifikator meke margine*.

Jedan od podesljivih parametara je parametar *C* koji predstavlja regularizacioni parametar i određuje toleranciju na grešku klasifikacije.

Postoje problemi za koje linearne granice nisu odgovarajuće čak ni uz korišćenje meke margine i promenljivog parametra C. Tada se primenjuje tzv. *kernel trik,* tj. uzorci se preslikavaju u višedimenzioni prostor u kom jesu linearno separabilni, a tome u linearnom prostoru odgovara nelinearna granica odlučivanja.

A. Priprema

Za problem klasifikacije u ovoj bazi biće korišćene 10, 20, 30, 40 i 50 kao vrednosti parametra C, kao i linearni, radijalni i polinomijalni kernel. Kako je u pitanju binarna klasifikacija, nije potrebno definisati *decision_function_shape* , te će se uzeti podrazumevana vrednost *ovr (one versus rest)*.

Pri kreiranju ovog modela će se koristiti unakrsna validacija sa 10 podskupova.

B. Rezultati klasifikacije metodom SVM

Najbolje rezultate je dao SVM klasifikator sa parametrima C= 50, za *rbf* kernel. Udeo uspešno klasifikovanih uzoraka je 96.62%.
Matrica konfuzije je data u Tabeli 3.

	Predviđene vrednosti	
Stvarne vrednosti	9012	188
	200	2100

Tabela 3. Matrica konfuzije za model SVM klasifikatora sa parametrima C=50 i kernel =rbf

VII. Smanjenje dimenzionalnosti metodom PCA

Smanjenje dimenzionalnosti prostora ima veliki značaj kada je u pitanju ubrzanje algoritama i pojednostavljenje modela. PCA je jedna od najčešće korišćenih metoda za linearnu redukciju obeležja.

PCA (eng. *Principal Component Analysis*) ima za cilj da što vernije prikaže uzorke iz visokodimenzionalnog prostora u prostoru sa manjim brojem dimenzija. Spada u nenadgledanu metodu učenja. Izborom najinformativnijih komponenata obrazuje se novi prostor sa manjim brojem dimenzija i uzorci se projektuju oko njega. Najinformativnije PCA komponente biće one koje odgovaraju pravcima najvećeg rasipanja uzoraka.

A. Standardizacija obeležja

Izvršena je standardizacija obeležja pomoću funkcije *StandardScaler*. Normalizacijom svodimo srednju vrednost svakog obeležja na 0, a standardnu devijuansu na 1. Ovim se sprečava pristrastnost PCA prema obeležjima koja imaju veći opseg vrednosti.

B. Redukcija dimenzionalnosti

U biblioteci *sklearn.decomposition* postoji implementirana klasa *PCA* pomoću koje ćemo izvršiti redukciju dimenzionalnosti. Glavni parametar klase je *n_components* koji definiše udeo varijanse. Za ovaj slučaj je definisana vrednost 0.9, kako bi se sačuvao dovoljan broj glavnih komponenti kako bi se sačuvalno 90% varijabilnosti originalnih podataka.

Redukovani prostor ima dimenziju 11500,2. ranije je imao 11500,4.

XIII. Modeli nakon izvršene redukcije dimenzionalnosti

Pokrećemo modele iz prethodno urađenih algoritama koji su imali najbolji procenat tačnosti i koristimo parametre za date slučajeve. Koristićemo redukovane *X_train_r* i *X_test_r* kako bismo mogli uporedili performanse oba modela.

A. Model SVM nakon PCA

Kako je model koji je pokazao najbolji procenat tačnosti bio sa parametrima C = 50 i kernelom = rbf, na njemu ćemo primeniti nov skup podataka. Nakon izvršenja algoritma procenat tačnosti je 97.48%. Što je više nego bez redukcije dimenzionalnosti.

B. Model Knn nakon PCA

Parametri koji su nakon izvršenih iteracija pokazali najveću tačnost su bili: metrika *manhattan*, za k=9, gde je tačnost bila 97%.
Koristeći skupove sa redukovanom dimenzionalnošću tačnost je 97.04%.

C. Model logističke regresije nakon PCA

Nakon izvršenja logističke regresije na skupu podataka manjih dimenzija, dogodilo se da je tačnost za sve parametre identična, osnosno 95.2%.
Zaključujem da je model dostigao optimalnu performansu.

XIX. Poređenje rezultata primenjenih algoritama

Kod KNN Klasifikatora je preciznost klase 0 (Bez epileptičnog napada) 98%, a klase 1 (Desio se epileptični napad) je 92%. SVM klasifikator je za klasu 0 izračunao preciznost 98%, a za klasu 1 dao preciznost 93%. Preciznost kod logističke regresije je 96.2%, a tačnost 95.2% što je u odnosu na druga dva modela slabije.
Razlika između KNN i SVM modela je neznatna, što se uočava najviše na razlici prosečne tačnosti od 0.43% .
Uspešnost datog SVM modela klasifikacije u prepoznavanju epileptičnih napada je najveća, ali je uspešnost KNN klasifikacije približna.