Katedra za Signale i sisteme
Elektrotehnički fakultet
Univerzitet u Beogradu

Sistemi odlučivanja u medicini

Data mining u medicini

Student:
Aleksandar Đorđević 2023/3049

Mentor:
Marija Novičić

Sadržaj

1 Uvod

1.1 Uvod u data mining

Data mining je ključan deo *data science*-a koji predstavlja proces prolaženja kroz velike setove podataka u cilju identifikovanja obrazaca i veza između podataka. Ovo je posebno značajno u disciplinama u kojima je bitan uzrok neke pojava, a ne samo da li se ona manifestovala ili ne te *data mining* pronalazi široku primenu u biznisu i medicini. Ovde ćemo se fokusirati na njegove primene u medicini.

Prilikom dijagnostike je svakako bitno da li pacijent ima ili nema neku bolest kako bi znali na koji način treba da bude tretiran, međutim, ovo ne pomaže u prevenciji bolesti. Da bi se sprečio sam nastanak bolesti neophodno je da se zna njen uzrok te je očigledno veoma značajno da se zna koji je parametar najuticajniji na sam ishod testa na određenu bolest. Na ovaj način se, nakon redovne kontrole, pacijentu može skrenuti pažnja na potencijalnu opasnost koja se može sprečiti promenom životnih navika.

Iako *data mining* koristi slične tehnike kao i mašinsko učenje (logistička i linearna regresija, metod nosećih vektora, neuralne mreže itd.) ova dva pristupa su fundamentalno razilikuju. U mašinskom učenju je cilj u što većem broju slučaju dobiti predikciju koja odgovara realnosti, dok nam je u *data mining*-u od krucijalnog značaja da zaključimo koja su obeležja uticala da se takva odluka donese.

Osnovni koraci prilikom *data mining*-a su:

1. **Prikupljanje podataka** - korak u kome je potrebno shvatiti koji su nam podaci od značaja, prikupiti ih i skladišti ih
2. **Priprema podataka** - ovaj korak podrazumeva pretprocesiranje skupa podataka (ispravljanje grešaka kao što podaci koji nedostaju ili duplirani podaci, transformacija podataka itd.)
3. **Data mining** - korak u kome se nad pretprocesiranim podacima primenjuju odgovarajuće tehnike kao što su modeli mašinskog učenja, klasterizacija i druge
4. **Interpretacija** - korak u kome se na osnovu rezultata primenjenih tehnika formiraju zaključci o podacima

1.2 Opis skupa podataka

U ovom radu ćemo obraditi skup podataka koji govori o uticaju načina života na kvalitet sna. U skupu se nalaze 13 kolona za 400 osoba. Kolone su sledeće.

1. **ID osobe**
2. **Pol** - Muški/Ženski
3. **Godine starosti**
4. **Zanimanje**
5. **Trajanje sna** - broj sati koje osoba spava dnevno
6. **Kvalitet sna** - subjektivna ocena kvaliteta sna od 1 do 10
7. **Fizička aktivnost** - broj minuta koja osoba provodi trenirajući dnevno
8. **Nivo stresa** - subjektivna ocena nivoa stresa od 1 do 10
9. **BMI kategorija**
10. **Krvni pritisak** - gornji/donji krvni pritisak

11. **Otkucaji srca** - broj otkucaja srca u minutu
12. **Broj koraka** - dnevni broj koraka osobe
13. **Poremećaj sna** - Nema/Insomnia/Apnia sna

Dakle, cilj je utvrditi koji od navedenih faktora mogu dovesti do pojave poremećaja sna.
U skupu nije bilo nedostajućih podataka. Kolona krvni pritisak je razdvojena na dve od kojih jedna predstavlja gornji krvni pritisak (sistolni), a druga donji krvni pritisak (dijastolni). U BMI koloni postoje polja *Normal* i *Normal Weight* koji predstavljaju istu stvar te su polja *Normal Weight* zamenjena poljima *Normal*.

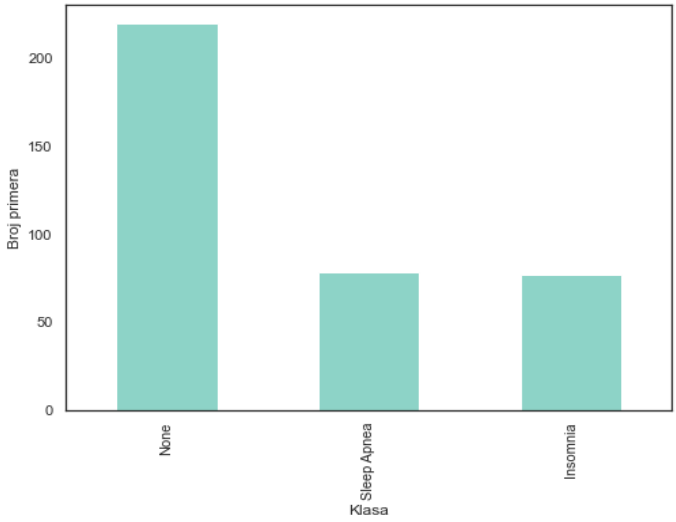
2 Eksplorativna analiza

Eksplorativna analiza podataka je veoma bitan korak u *data mining*-u. Ona nam omogućava da sagledamo odnose između obeležja i izlaza kao i između obeležja međusobno. Na ovaj način možemo napraviti pretpostavke koje kasnijemo možemo potvrditi ili odbaciti. Takođe moguće je odbaciti određena obeležja ako se ispostavi da su zavisna te zajedno ne daju nikakvu dodatnu informaciju u odnosu na korišćenje jednog obeležja.

2.1 Balansiranost klasa

Prilikom primene modela mašinskog učenja i njihove evaluacije od velikog značaja je to da li imamo jednako primeraka iz svih klasa odnosno da li su klase balansirane. Ako klase nisu balansirane u nekim slučajevima je moguće postići visoke tačnosti nekim jednostavnim modelima npr. proglašavanjem svih primera za većinsku klasu. U medicini ovo je veoma čest slučaj jer skoro uvek postoji više testiranih ljudi koji nemaju neku bolest nego onih koji je imaju. Zbog ovoga je moguće napraviti test koji ima izuzetno visoku tačnost tako što sve pacijente svrstava u zdrave, međutim, očigledno je da je ovakav test beskoristan. Zbog ovoga prilikom evaluacije modela nije dovoljno koristiti prostu tačnost, već je neophodno koristiti parametre kao što su preciznost, osetljivost, f1-skor ili prikazati rezultate grafički pomoću matrice konfuzije.

Naš skup podataka je, kao i većina u medicini, nebalansiran sa najviše ljudi koji nemaju poremećaj spavanja. Raspored po klasama izgelda ovako:



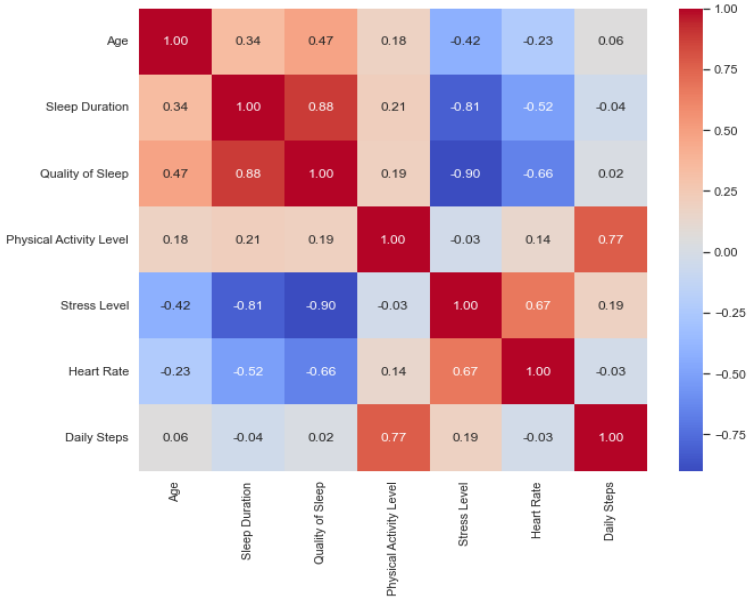
Slika 1: Raspored primerama po klasama

2.2 Matrica korelacije

Dobar pokazatelj linearne zavisnosti između dve slučajne promenljive je koeficijent korelacije koji za sa slučajne promenljive X i Y na računa na sledeći način.

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \tag{1}$$

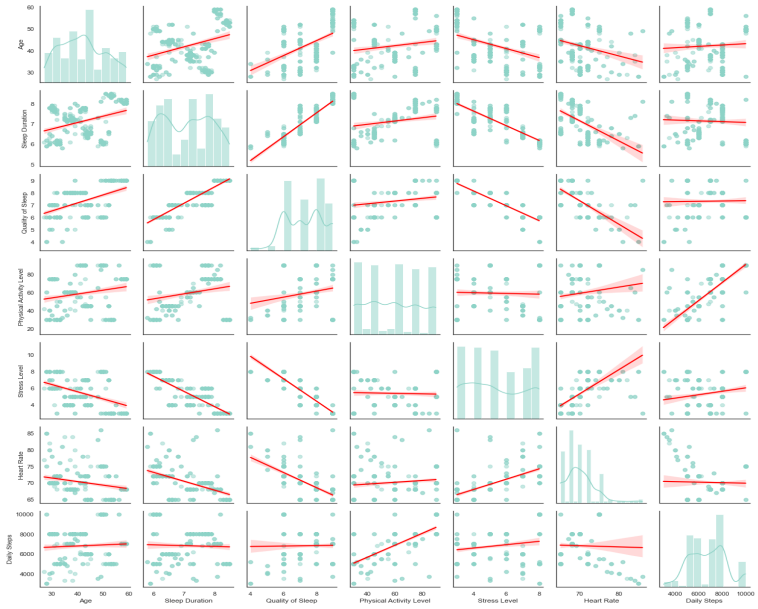
Ukoliko je vrednost koeficijenta korelacije između dve slučajne promenljive ± 1 to znači da su one linearno zavisne (+ indikuje direktnu srazmernost dok - obrnutu). Što je absolutna vrednost koeficijenta korelacije veća to je linearna zavisnost izraženija, ako je vrednost 0 tad ne postoji linearna zavisnost. Izračunavanjem korelacije između svih numeričkih kolona dobijamo matricu korelacije.



Slika 2: Matrica korelacije numeričkih obeležja

Nažalost, kako je izlaz kategorički nije moguće izračunati korelaciju sa njim. Ipak, iz matrice može zaključiti da li postoje veze između samih obeležja. Može se primetiti da je broj otkucaja srca izuzetno korelisan sa nivoom stresa što može značiti da je stres uzrok nekih srčanih oboljenja. Takođe, možemo primetiti da je nivo stresa izuzetno negativno korelisan sa kvalitetom i dužinom sna te on može biti jedan od glavnih uzroka problema sa spavanjem. Ovde se mogu videti mnoge zavisnosti, ali ova matrica pre svega treba da posluži za to da se otkriju potencijalne veze između obeležja koje se kasnije dokazuju ili opovrgavaju. Takođe, jedina velika mana ovog pristupa jeste ta što otkriva samo linearne zavisnosti te treba razumeti da koeficijent korelacije 0 ne znači da su obeležja nezavisna.

Lep način da se prikaže zavisnost numeričkih promenljivih jeste da se iscrtaju na 2D grafiku. Na sledećoj slici je upravo to prikazano s tim da se na glavnoj dijagonali nalaze histogrami obeležja.



Slika 3: 2D prikaz zavisnosti obeležja

3 Modeli mašinskog učenja

Pre nego što se primene modeli mašinskog učenja potrebno je obaviti određeno pretprocesiranje skupa podataka. Najpre ćemo kodovati kategoričke promenljive celim brojevima. Numerička obeležja su skalirana tako da je srednja vrednost 0, a varijansa jedinična na sledeći način:

$$x = \frac{x - mean(x)}{std(x)} \tag{2}$$

Ovo je urađeno kako red veličine samog obeležja ne uticao na njegovu važnost.
U ovom poglavlju ćemo opisati korišćene modele mašinskog učenja.

3.1 Logistička regresija

Logistička regresija je model koji se koristi za klasifikaciju primera u dve klase (videćemo kasnije kako se ovo može proširiti i na više klasa). Ona polazi od linerane hipoteze.

$$h_\theta(x) = \theta^T x \tag{3}$$

gde θ predstavlja parametre modela, a x ulaz. Ova hipoteza se zatim koristi kao argument logističke (sigmoid) funkcije te finalna hipoteza izgleda ovako:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{4}$$

Izlaz logističke funkcije se nalazi u intervalu od 0 do 1 te se može tumačiti kao verovatnoća da dati primer pripada prvoj klasi. Primer se klasifikuje tako što se izlaz poredi sa pragom (ako je veći od praga pripada prvoj klasi, ako ne pripada drugoj). Za prag se obično uzima 0.5, ali moguće je i menjati ga.

Parametri se uče maksimizacijom log-verodostojnosti na obučavajućem skupu.

$$l(\theta) = \ln \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) = \sum_{i=1}^{m} y^{(i)} \ln h_\theta(x^{(i)}) + (1 - y^{(i)}) \ln(1 - h_\theta(x^{(i)})) \tag{5}$$

Najčesći način da se izvrši ova maksimizacija jeste gradijentni spust. Pored toga moguće je korišćenje metoda zasnovanih na Hesijanu kao što je Newton-Raphson metoda.

Kako bi se sprečilo preobučavanje često se uvodi regularizacija. Za L1 regularizaciju se na optimizacionu funkciju dodaje suma apsolutnih vrednosti parametara pomnožena faktorom regularizacije, dok se prilikom L2 regularizacije dodaje suma kvadrata parametara pomnožena faktorom regularizacije. L1 regularizacija ima tendenciju da anulira određene parametre što je veoma zgodno jer je potrebno zapamtiti manje parametara prilikom implementacije modela. Međutim, velika mana ove regularizacije jeste ta što optimizaciona funkcija nije više diferencijabilna. L2 regularizacija ima tendenciju da drži parametre malim, a za razliku od L1 regularizacije optimizaciona funkcija ostaje diferencijabilna i to je najveći razlog za popularnost ove metode.

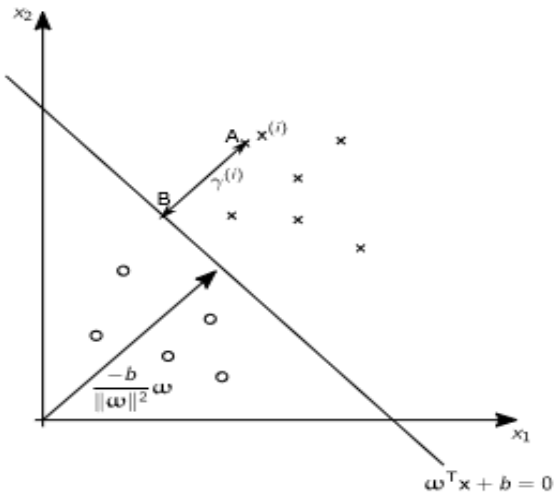
Takođe, moguće je izboriti se sa nebalansiranim klasama uvođenjem težinskih koeficijenata klasama. Najčešće su ovi koeficijenti obrnuto proporcionalni broju primera klase te su na taj način najmanje klase najviše otežinjene.

Još treba napomenuti da iako logistička regresija rešava problem klasifikacije između dve klase moguće je rešiti i problem višeklasne klasifikacije jednim od dva pristupa *one vs all* ili *one vs one*. U *one vs all* pristupu se projektuje onoliko klasifikatora koliko ima klasa i to tako što se posmatrana klasa posmatra kao jedna klasa, a sve ostale kao druga. Nako projektovanja odluka se donosi tako što se primer propusti kroz sve klasifikatore i svrsta se u onu klasu čiji je klasifikator da najveću verovatnoću da pripada toj klasi. *One vs one* pristup funkcioniše tako što se projektoje po jedan klasifikator za svaki par klasa. Prilikom odlučivanja primer se propusti kroz sve klasifikatore i smešta u onu klasu koja je dobila najviše glasova.

3.2 Metod nosećih vektora

Metod nosećih vektora (eng. *Support vector machine*) predstavlja model linearne klasifikacije između dve klase. Ovaj metod teži da maksimizuje marginu između klasifikacione prave i njoj najbližih primera i na ovaj način obezbedi generalizaciju.

Za geometrijsku marginu usvajamo rastojanje primera i od separacione prave $\gamma^{(i)}$. Za ispravno klasifikovane primere uzima se da je ono pozitivno, dok za neispravne je ono negativno. Klase se označavaju sa $y = \pm 1$.



Slika 4: SVM

Tačka A predstavlja primer $x^{(i)}$. $\overrightarrow{AB}$ je normalno na separacionu pravu te je paralelno sa vektorom ω. Tačka B se može predstaviti kao: $x^{(i)} - \gamma^{(i)} \frac{\omega}{||\omega||}$. Takođe, tačka B je na separacionoj pravi pa ispunjava njenu jednačinu te važi: $\omega^T(x^{(i)} - \gamma^{(i)} \frac{\omega}{||\omega||}) + b = 0$. Konačno geometrijska margina je data sledećim izrazom.

$$\gamma^{(i)} = y^{(i)}(\frac{\omega^T}{||\omega||}x + \frac{b}{||\omega||}) \tag{6}$$

Uslov za ispravno klasifikovan primer se može zapisati ovako:

$$y^{(i)}(\omega^T x^{(i)} + b) > 0 \tag{7}$$

Zbog toga se uvodi funkcionalna margina koja se definiše na sledeći način.

$$\hat{\gamma}^{(i)} = y^{(i)}(\omega^T x^{(i)} + b) \tag{8}$$

Na osnovu ovoga veza između funkcionalne i geometrijske margine je $\gamma^{(i)} = \frac{\hat{\gamma}}{||\omega||}$. Na obučavajućem skupu želimo da maksimizujemo minimalnu geometrijsku marginu odnosno rastojanje do najbližeg primera. Takođe, želimo što veću tačnost. Ako maksimizujemo funkcionalnu marginu imamo problem da ako ikad nadjemo pravu koja perfektno deli klase možemo povećati funkcionalnu marginu skaliranjem parametara ω i b bez uticaja na samu pravu. Zbog ovoga se ona ne maksimizuje već se koristi samo kao uslov za dobru klasifikaciju, a margine se povećavaju minimizacijom $||\omega||$. U praksi se, zbog toga da bi se optimizacija rešavala kvadratnim programiranjem, minimizuje kvadrat norme vektora ω odnosno problem nalaženja optimalne separacione krive izgleda ovako.

$$\min_{\omega,b} \frac{1}{2}||\omega||^2 \text{ pod uslovom } 1 - y^{(i)}(\omega^T x^{(i)} + b) \leq 0 \tag{9}$$

Problem sa ovom optimizacijom jeste da će raditi samo na linearno separabilnom skupu podataka. Zbog toga je potrebno dozvoliti da neki primeri budu pogrešno klasifikovani. To je učinjeno uvođenjem šarka gubitaka. Sada optimizacioni problem izgleda ovako:

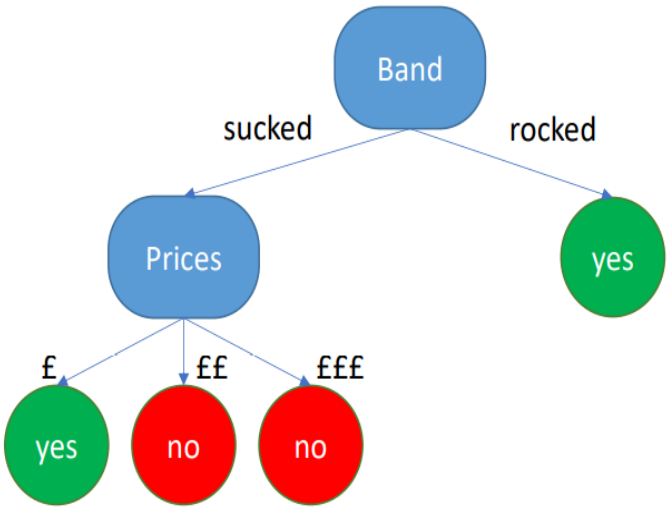
$$\min_{\omega,b,\xi} \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{m} \xi_i \text{ pod uslovom } \hat{\gamma} \geq 1 - \xi_i,\ \xi_i \geq 0 \tag{10}$$

Parametar C ima smisao faktora regularizacije i što je on veći više se penalizuju pogrešne klasifikacije te je regularizacija manje izražena. Smanjivanjem C regularizacija postaje izraženija. Ovaj optimizacioni problem se rešava algoritmom uspona po koordinatama ili algoritmom sekvencijalne minimalne optimizacije.

Ovako definisan SVM može rešiti samo probleme u kojem je prirodna separaciona kriva između klasa linearna. Ovo je moguće prevazići drugčijom predstavom ulaznih podataka. Kako ovo može biti veoma zahtevno za velike skupove podataka koristi se takozvani kernel-trik. Kernel-trik omogućava da se problem implicitno prebaci u drugi optimizacioni prostor izračunavanjem funkcije na osnovu ulaznih podataka koja ima smisao unutrašnjeg proizvoda transofrmisanih ulaznih podataka. Na ovaj način je moguće projektovati hiper-ravan u višedimenzionalnom prostoru koja će predstavljati nelinearnu separacionu hiper-ravan kada se projektuje u polazni prostor.

3.3 Stabla odlučivanja

Stabla odlučivanja su modeli mašinskog učenja koji se sastoje od čvorova koji predstavljaju obeležja i dele skup podataka na podskupove u odnosu na to obeležje. Jedna od glavnih prednosti stabala odlučivanja jesu njihova interpretabilnost i lakoća obrađivanja kako kategoričkih tako i numeričkih obeležja. Postavlja se pitanje kako odabrati obeležja i rasporediti ih po stablu. Ideja je da se dobiju što uređeniji podskupovi (idealno podskupovi koji su sačinjeni samo od pripadnika jedne klase).



Slika 5: Stablo odlučivanja

Kao mera neuređenosti može se koristit entropija koja se definiše ovako:

$$H(X) = -\sum_i p_i \log p_i \tag{11}$$

gde X predstavlja obeležje koje posmatramo. Kako je ovo mera neuređenosti najbolje obeležje je ono koje minimizuje entropiju. Često se koristi i mera pod nazivom Džinijeva nečistoća (eng. *Gini impurity*) čija definicija izgleda ovako:

$$I(X) = \sum_i p_i(1 - p_i) = 1 - \sum_i p_i^2 \tag{12}$$

I ova metrika se takođe minimizuje. Na osnovu ovih metrika stablo se gradi na sledeći način. Prođe se kroz sva obeležja i izračuna se metrika na osnovu podskupova koje to obeležje pravi. Izabere se ono obeležje na osnovu kog je dobijena minimalna metrika. Ovo možemo raditi sve dok nam svi podskupovi ne budu u potpunosti sačinjeni od pripadnike iste klase. Ovo nam, naravno, nije u interesu jer će ovakav model najverovatnije biti preobučen. Načini regularizacije stabala odlučivanja jeste njegovo skraćivanje statističkim testovima nakon što je formirana u potpunosti. Drugi načini jesu da se unapred odredi dubina stabla (na ovaj način se smanjuje broj podela te se sprečava preobučavanje) ili minimalni broj primera po listu (podskup na kraju stabla).

3.4 Slučajna šuma

Primećeno je da više slabih učenika (modeli koji imaju tačnost više od 50%) mogu ostvariti veće finalne tačnosti nego jaki učenik (standardni modeli mašinskog učenja). Zbog ovoga su razvijeni ansambl metodi koji su jedni od najmoćnijih modela današnjice. Jedan od primera ansambl metoda jeste slučajne šuma. Slučajne šume se sastoje od više stabla odlučivanja (broj stabla odlučivanja je hiper-parametar modela). Naravno, bilo koji broj stabla nam ne pomaže ukoliko su ona trenirana na isti način. Potrebno je na neki način napraviti stabla koja su naučila različite stvari i na taj način dopunjuju jedno drugo. Ovo se radi bootstraping-om. Bootstraping predstavlja odabir primera koji će biti dati kao obučavajući skup svakom od slabih učenika. Nasumično biramo onoliko puta koliko ima primera u obučavajućem skupu, ali sa ponavljanjem. Na ovaj način svaki slabi učenik ima oko 70% jedinstvenih primera. Pored broja slabih

učenika, hiper-parametre slabih šuma predstavlju svi hiper-parametri stabla odlučivanja (vrsta metrike, maksimalna dubina, minimalan broj primera u listovima). Velika prednost slučajnih šuma jeste što se svaki slabi učenik obučava nezavisno te je moguće trenirati ih potpuno paralelno što u doba modernih grafičkih kartica dovodi do značajnog ubrzavanja proces obučavanja. Treba napomenuti da slučajne šume imaju mogućnost da daju važnost obeležja na osnovu toga koliko se ona često koriste. Ovo je veoma značajno za primene u *data mining*-u gde je bitno znati koja obeležja su donela do konkretne odluke.

3.5 Gradient Boosting Trees

Gradient boosting Trees je još jedan ansambl metod koji se zasniva na stablima. Međutim, za razliku od slučajnih šuma ovde se koriste regresiona stabla. Ona su u mnogome slična stablima odlučivanja s tim da umesto klasifikacije rade regresiju te se listovima ne dodeljuje klasa, već srednja vrednost primera u njima.

Model se inicijalizuje konstantom koja minimizuje funkciju gubitka.

$$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma) \tag{13}$$

Nakon toga se proizvoljan broj puta (M) ponavlja sledeća sekvenca. Prvo se sračunaju takozvani pseudo-reziduali.

$$r_{im} = -\left(\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right)_{F(x)=F_{m-1}(x)} \tag{14}$$

Sada se slabi učenik (stablo) trenira na rezidualima što rezultuje u funkciji $h_m(x)$. Sada minimizacijom funkcije gubitka dobijamo multiplikativni faktor na sledeći način.

$$\gamma_m = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{15}$$

Ažurirani model izgleda ovako.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{16}$$

Kako bi se sprečilo preobučavanje mogu se koristiti standarne metode ograničavanja dubine stabla ili minimalnog broj primera u listovima. Takođe, često se uvodi parametar stope obučavanja koji množi koeficijent γ_m prilikom ažuriranja modela. Kao i slučajne šume i ovaj model može da prikaže važnost obeležja te je i on dosta pogodan za *data mining*.

4 Određivanje hiper-parametara modela

Proces treniranja služi da se odrede parametri modela, ali pre toga je neophodno odrediti hiper-parametre. Najpre se skup podataka podeli na trening i test skup (urađeno je u odnosu 70:30). Standardan metod određivanja hiper-parametara je izdvajanje određenog dela trening skup kao validacioni skup koji koristimo za procenu kvaliteta modela prilikom menjanja hiper-parametara. Kako bi se izbeglo dodatno smanjivanje trening skupa ovde je korišćena k-fold krosvalidacija.

4.1 K-fold krosvalidacija

Trening skup se podeli nasumično na k delova (u našem slučaju 5). U svakoj iteraciji se 1 deo uzme kao validacioni skup, a ostatak kao trening skup. Sračuna se greška na validacionom skupu i kao procena greške generalizacije se uzima srednja vrednost ovih grešaka kroz iteracije. Uzimamo one hiper-parametre koji su doveli do minimalne greške generalizacije. Za izbor hiper-parametra korišćen je *grid search* odnosno za svaki hiper-parametar koji se optimizuje se izaberu nekoliko vrednosti i onda se isprobaju sve kombinacije

zadatih vrednosti. Ovakvo podešavanje hiper-parametara je dalo mnogo bolje rezultate nego korišćenje *default*-nih podešavanja te će u sledećem poglavlju biti prikazani samo rezultati sa podešenim hiper-parametrima.

5 Rezultati

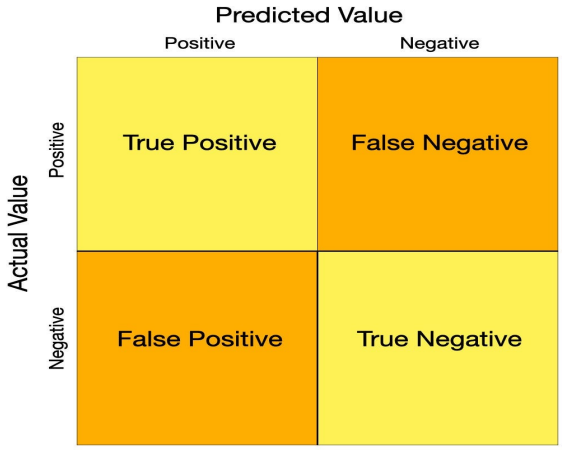
U ovom poglavlju će za sve prethodno navedene modele biti prikazano koji su hiper-parametri podešavani, koje su njihove vrednosti ispitane, koje su izabrane vrednosti kao i konačni rezultati istreniranog modela.

5.1 Metrike za evaluaciju modela

Najintuitivnija metrika za procenu kvaliteta modela jeste njegova tačnost odnosno broj tačnih klasifikacija u odnosu na ukupan broj klasifikacija ili matematički ovako:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

gde TP(*True Positive*) i TN(*True Negative*) predstavljaju prave pozitive odnosno negative, FP(*False Positive*) lažne alarme, a FN(*False Negative*) propuštene detekcije. Značenje ovih promenljivih se lepo predstavlja tabelarno.



Međutim, ova metrika često ne daje realnu sliku valjanosti modela pogotovo kod nebalansiranih skupova podataka (što je slučaj kod nas). Zamislimo skup podataka gde od 100 pacijenata 5 ima neko oboljenje, a ostalih 95 nemaju. Lako je napraviti model koji sve pacijente proglašava za zdrave i dostiže tačnost od 95%. Naravno, ovakav model je potpuno beskoristan i zbog toga je neophodno uvesti naprednije metrike za evaluaciju. Dve najpopularnije metrike za evaluaciju testova u medicini su preciznost (*precision*) i osetljivost (*sensitivity*). Osetljivost nam govori o tome koliko često će naš test prepoznati da je pacijent bolestan ukoliko on to zaista jeste ili matematički ovako:

$$Sensitivity = \frac{TP}{TP + FN} \tag{18}$$

Ona je veoma važna jer je veoma bitno da se što veći procenat pacijenata koji je bolesno prepozna kao takvo kako bi se lečili. Preciznost govori o tome koji procenat pacijenata koji je proglašen za bolesne to zaista i bio ili matematički ovako:

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

Iako nije krucijalan kao osetljivost i ovaj parametar je od velike važnosti jer lažne detekcije mogu dovesti da dalje ispitivanje ode u pogrešnom pravcu i na taj način se izgubi dragoceno vreme u lečenju pacijenta.

Metrika koja kombinuje prethodne dve i koja se najčešće koristi za evaluaciju modela na nebalansiranim skupovima podataka jeste F1-skor. Ona se računa na sledeći način.

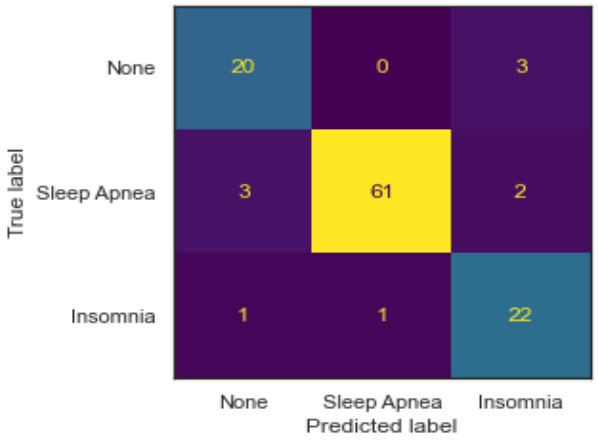
$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Sensitivity}} \tag{20}$$

Iz formule se vidi da se F1-skor kreće između 0 i 1 i da može biti blizak jedinici samo ako su i preciznost i osetljivost bliske jedinici tako da nam dobar F1-skor garantuje dobru i preciznost i osetljivost. Zbog ovoga je F1-skor korišćen prilikom određivanja hiper-parametara umesto standardne greške generalizacije.

5.2 Logistička regresija

Ovaj model je treniran sa maksimumom od 10000 iteracija i klase su otežinjene faktorima obrnuto proporcionalnim broju primera koji njima pripadaju. Korišćena je L2 regularizacija i biran je hiper-parameter C koji je jednak recipročnoj vrednosti faktora regularizacije λ. Vrednosti parametra C koje su ispitivane su $[0.001, 0.01, 0.1, 1, 10, 100]$. Takođe, je birano između dva optimizaciona algoritma LBFGS i SAGA. Krosvalidacijom je izabrano $C = 1$ i LBFGS algoritam.

Lep grafički način da se predstave rezultati jeste konfuziona matrica. Po redovima su stvarne klase primera, a po kolonama klase koje je dao model. Vidimo da su rezultati veoma dobri i da model radi



Slika 6: Rezultati logističke regresije

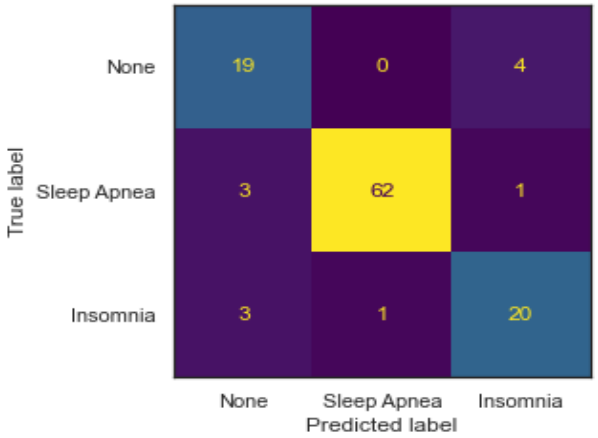
dobro na svim klasama. Postignuta tačnost je 91%. Kako je ovo problem 3 klase napredne metrike će se razlikovati u zavisnosti od toga koju klasu smatramo pozitivom odnosno imaćemo tri vrednosti za preciznost, osetljivost i F1-skor. One su predstavljene u sledećoj tabeli.

	Preciznost	Osetljivost	F1-skor
None	0.83	0.87	0.85
Sleep Apnea	0.98	0.92	0.95
Insomnia	0.81	0.92	0.86

Vidimo da su najlošiji rezultati kad je pozitivna klasa klasa da pacijent nema poremećaj sna što je uredu jer u praksi to nikada neće biti slučaj. Rezultati su najbolji kada je pozitivna klasa Sleep Apnea, ali bi rezultati za insomniju mogli biti malo bolji. Takođe, ovde je teško reći koja obeležja su najviše doprinela, ali ovo služi kao dobra osova da se vidi koje tačnosti je moguće dobiti iz datog skupa podataka.

5.3 Metod nosećih vektora

Za ovaj model je bira parametar regularizacije C iz skupa $[0.001, 0.01, 0.1, 1, 10, 100]$. Takođe, biran je tip korišćenog kernela iz skupa linearnog, polinomijalnog i Gausovog. Izabrano je $C = 100$ i polinomijalni kernel. Matrica konfuzije izgleda ovako:



Slika 7: Rezultati metoda nosećih vektora

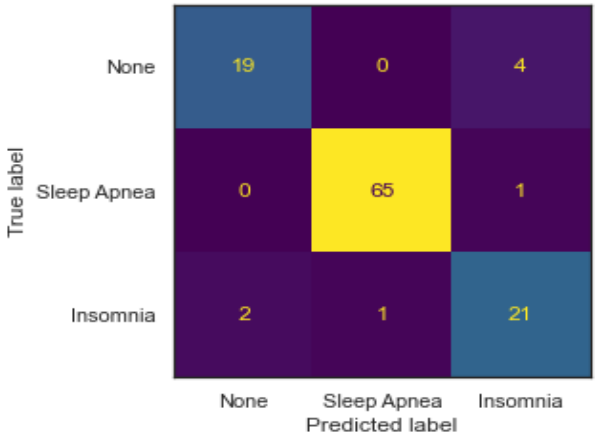
Dobijena tačnost je nešto manji nego kod logističke regresije i iznosi 89%. Napredne metrike izgledaju ovako.

	Preciznost	Osetljivost	F1-skor
None	0.76	0.83	0.79
Sleep Apnea	0.98	0.94	0.96
Insomnia	0.80	0.89	0.82

Opet su najlošiji rezultati kad je pozitivna klasa klasa bez poremećaja što je opet uredu. Najbolje za klasu Sleep Apnea, a za insomniju su dobijeni nešto lošiji rezultati te je opšti utisak da se ovaj model pokazao nešto lošije nego logistička regresija. Ni ovaj model ne može dati važnost obeležja te ćemo sada preći na ansambl metode koji imaju tu mogućnost.

5.4 Slučajna šuma

Hiper-parametri koji su podešavani za ovaj model su broj slabih učenika iz skupa $[50, 100, 200]$, maksimalna dubina stabla jednog slabog učenika iz skupa $[10, 20, 30, \infty]$, minimalni broj primera u listovima iz skupa $[1, 2, 4]$ i da li se koristi *bootstraping* ili ne. Izabrani hiper-parametri su 100 slabih učenika neograničene dubine i minimalno 4 primera u listu sa korišćenjem *bootstraping*-a.

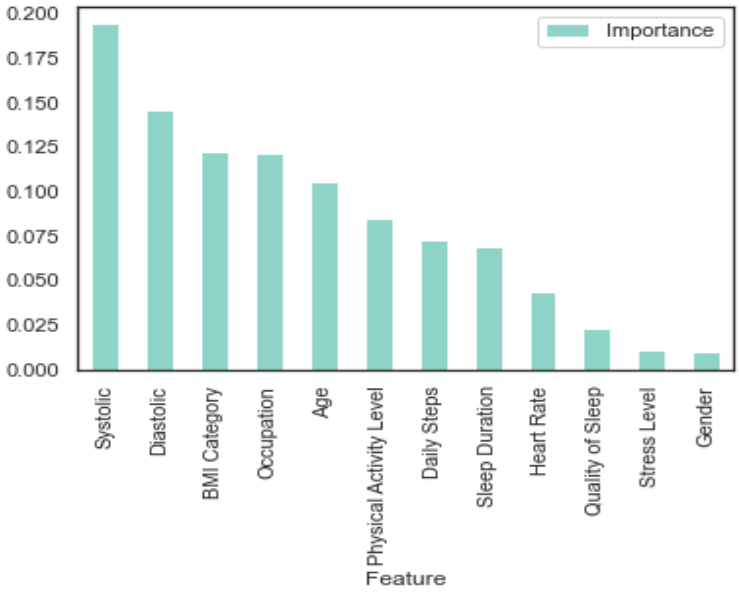


Slika 8: Rezultati slučajne šume

Postiguta je tačnost od 93%. Napredne metrike izgledaju ovako:

	Preciznost	Osetljivost	F1-skor
None	0.90	0.83	0.86
Sleep Apnea	0.98	0.98	0.98
Insomnia	0.81	0.88	0.84

Rezultati su bolji nego sa prethodnim modelima, ali i dalje ostaje tačno da su najbolji kada je pozitivna klasa Sleep Apnea, a manje dobri kada su to None ili Insomnia. Kako model ima visoku tačnost i dobre napredne metrike ima smisla pogledati koja su obeležja najvažnija za ovaj model jer su ona dovela do dobrih rezultata. Slučajne šume imaju mogućnost rangiranja obeležja na osnovu toga koliko se ona često koriste u stablima i toga koliko uspšno razbijaju skup na homogene podskupove. Zbog ovoga su, kako je već napomenuto, one veoma zgodne za primene u *data mining*-u. Lista važnosti obeležja izgleda ovako:

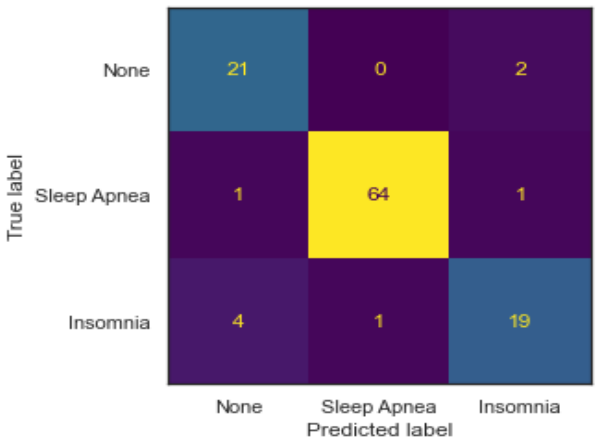


Slika 9: Važnost obeležja na osnovu slučajne šume

Vidimo da su najznačajnija obeležja gornji i donji krvni pritisak, kao i BMI kategorija te bi se moglo zaključiti da ljudi koji ne vode računa o ishrani i ne bave se fizičkom aktivnošću mogu dobiti neki poremećaj sna. Fizička aktivnost je nešto niže na listi ali i dalje predstavlja značajno obeležje te bi dobra prevencija poremećaja sna bila zdrava ishrana i bavljenje sportom. Naravno, i godine imaju značajan uticaj ali na to se ne može uticati te ovde nije od interesa.

5.5 Gradient Boosting Trees

Hiper-parametri koji su podešvani su broj slabih učenika (regresionih stabla) iz skupa $[50, 100, 200]$, maksimalna dubina jednog stabla iz skupa $[10, 20, 30, \infty]$ i minimalni broj primera u listovima iz skupa $[1, 2, 4]$. Izabrani hiper-parametri su 100 slabih učenika, maksimalna dubina od 10 nivoa i minimalno 2 primera u listu.

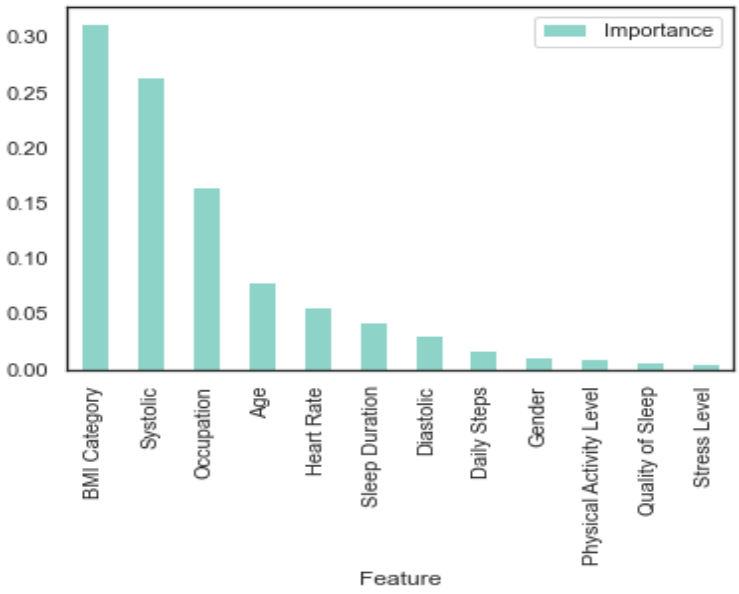


Slika 10: Rezultati *gradient boosting*-a

Postignuta tačnost je 92%. Napredne metrike izgledaju ovako.

	Preciznost	Osetljivost	F1-skor
None	0.81	0.91	0.86
Sleep Apnea	0.98	0.97	0.98
Insomnia	0.86	0.79	0.83

Napredne metrike su sličnog oblika kao i kod slučajne šume. Kao i kod slučajne šume i ovde se na sličan način formira lista važnosti obeležja i ona izgleda ovako.



Slika 11: Važnost obeležja na osnovu *gradient boosting*-a

Ovde je BMI kategorija najznačajnija, a drugi je gornji krvni pritisak. Zanimljivo je da je ovde donji krvni pritisak dosta nisko rangiran. Zanimanje i godine su opet među značajnijim obeležjima. Takođe, zanimljivo je da je fižička aktivnost gotovo irelevantna.

6 Zaključak

Cilj ovog rada je bio da se pretpostave uzroci nastajanja poremećaja sna. Prvo je sprovedena eksplorativna analiza u kojoj su analizirane korelacije obeležja kako sa izlazom tako i međusobno. Zatim su opisani metodi mašinskog učenja koji su kasnije korišćeni za projektovanje klasifikatora. Od posebnog značaja su ansambl metode koju imaju opciju da formiraju listu važnosti obeležja i na taj način daju do znanja koja su obeležja najviše uticala na donošenje odluki. Takođe, kako bi se potvrdilo da modeli dobro rade te je važnost njihovih obeležja relevantna korišćene su metrike kao što su tačnost, preciznost, osetljivost i F1-skor.

Na osonvu prethodne analiza može se pretpostaviti da bi zdrava ishrana bila dobar vid prevencije poremećaja sna. Naravno, ovo je potrebno potvrditi eksperimentalno u kontrolisanim uslovima, ali ovakav pristup predstavlja dobar prvi korak pogotovo u manje istraženim oblastima.

Literatura

[1] Paul R. Harper, *A review and comparison of classification algorithms for medical decision making*

[2] Umar Sidiq and dr. Rafi Ahmad Khan, *Data Mining for diagnosis in Healthcare Sector-A Review*

[3] Parvez Ahmad, Saqib Qamar and Syed Qasim Afser Rizvi, *Techniques of Data Mining In Healthcare: A Review*

[4] dr. Predrag Tadić, *Beleške sa predavanja iz Mašinskog učenja*