ORI E2

Kolokvijum 2

Zadatak na ovom kolokvijumu će biti iz oblasti medicinskih nauka i analiziraćemo podatke o pacijentima i procenjivaćemo rizika od srčanog udara. Skup podataka ne sadrži informacije o krvnoj slici, drugim bolestima kod istog pacijenta i slično (što bi sigurno bio mnogo bolji skup podataka jer i ovi faktori utiču na rizik od srčanog udara), ali je i dalje reprezentativan.

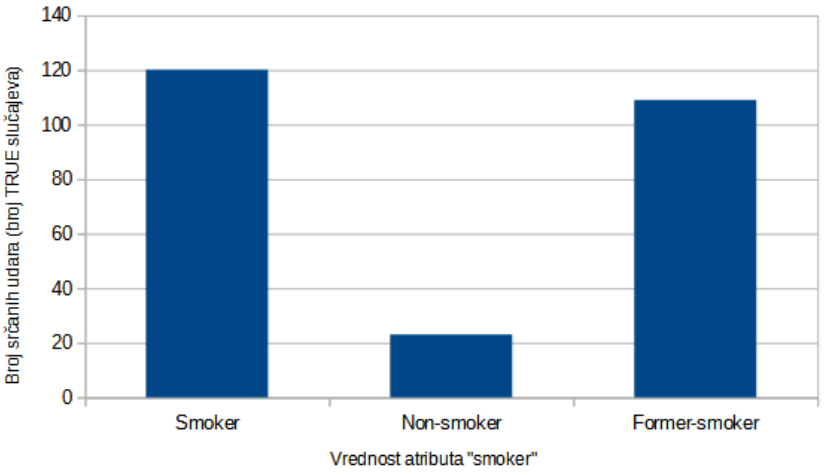
Napomena (za sve zadatke): Sve transformacije nad podacima treba raditi kroz *source* kod rešenja. Nije dozvoljeno menjanje skupa podataka direktno u csv fajlu. Svi modeli (regresija, klasterovanje, vnm) se ne smeju koristiti iz scikit-learn biblioteke i treba da se rade kao na vežbama (naravno, uz potrebne izmene da biste rešili ovaj zadatak). Za neke osnovne transformacije nad podacima je dozvoljeno koristiti scikit-learn i pandas, a po potrebi i pickle i joblib ako serijalizujete objekte modela.

1. Lak zadatak: Izvršiti aproksimaciju visine prosečne količine šećera u krvi (atribut *avg_glucose_level*) na osnovu broja godina pacijenta (atribut *age*). Iskomentarisati da li postoji zavisnost između ove dve veličine ili ne, na osnovu iscrtanog grafika. Nakon toga proveriti kolika bi bila prosečna količina šećera u krvi za osobe koje imaju 25, 45 i 65 godina starosti.

2. Srednji zadatak: Izvršiti klasterovanje pacijenata na osnovu njihovog broja godina (atribut age), informacije da li inače imaju problem sa povišenim krvnim pritiskom (atribut *hypertension*), na osnovu prosečne količine šećera u krvi (atribut *avg_glucose_level*), indeksa telesne mase (atribut *bmi*) i pola osobe (atribut *gender*). Naći optimalan broj klastera/grupa na osnovu elbow metode. Nakon toga analizirati klastere i proveriti koliki je odnos ljudi koji su u visokom i niskom riziku od srčanog udara (atribut *stroke*) u svakom od klastera. Ispisati procenat ili odnos za svaki klaster.
Ideja iza ovoga je da ako detektujete klaster sa ekstremnom vrednošću odnosa (bliže 0 ili 1), to je grupa/klaster ljudi koja je visoko ili nisko rizična i može nam koristiti za dalje analize faktora rizika. Obično se nakon toga posmatraju podaci u ovim interesantnim klasterima iz prethodnog koraka i traže se neke zajedničke osobine u njima (ovo nije deo zadatka nego samo komentar, da znate zašto se ovo inače radi).

3.Težak zadatak: Izvršiti predikciju da li je pacijent u riziku od srčanog udara (atribut *stroke*). Na raspolaganju su vam svi ostali atributi u skupu podataka (pol - *gender*, da li pacijent ima problem sa hipertenzijom - *hypertension*, da li pacijent ima neku od bolesti srca - *heart_disease*, da li je ikad bio/la u braku - *ever_married*, tipa zaposlenja - *work_type*, tipa sredine u kojoj živi - *residence_type*, prosečne visine šećera u krvi - *avg_glucose_level*, indeksa telesne mase - *bmi*, da li je pacijent pušač - *smoking_status*). Zadatak se sastoji iz dva dela:

A. **Eksplorativna analiza. Sami izaberite koje atribute ćete koristiti kao ulaz u model** (ne morate sve), da biste dobili što bolje rezultate. Za kategoričke atribute se često u praksi radi provera zavisnosti svakog od njih u odnosu na target/label atribut (atribut *stroke*). Ovo možete uraditi i tako što ćete iscrtavati grafike (bar chart) za svaki kategorički atribut u odnosu na labelu. Npr, naredni dijagram pokazuje primer zavisnosti informacije da li je pacijent pušač i broja srčanih udara (dijagram je proizvoljan i nije iz skupa podataka). Jasno se vidi da pušači imaju manje srčanih udara, pa je ovaj atribut interesantan. Za ovaj deo posla ne treba model, nego je dovoljno samo proći kroz ceo skup podataka i prebrojati slučajeve. Ovo ne treba raditi za numeričke atribute, nego samo za kategoričke.



Primer iscrtavanja ovakvih grafika imate na sledećem linku. Potrebno je samo napuniti dve liste i proslediti biblioteci matplotlib, koju smo i koristili na vežbama za iscrtavanje.

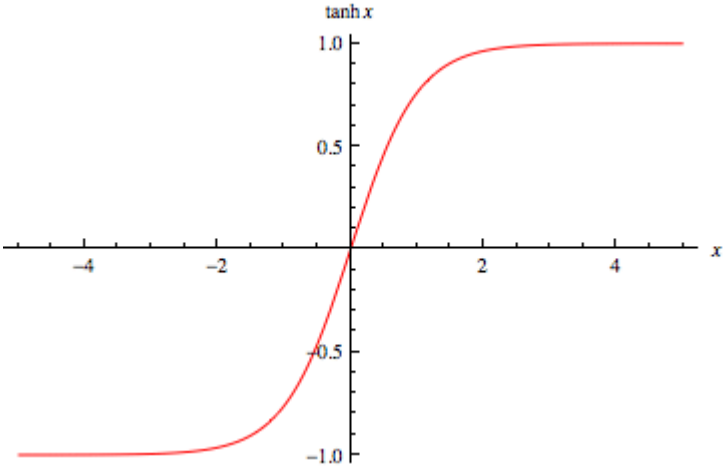
B. **Istrenirati model veštačke neuronske mreže** i izvršiti predikciju rizika od srčanog udara, na osnovu atributa izabranih u prethodnom koraku. Kao aktivacionu funkciju u skrivenim slojevima koristiti tangens hiperbolični (tanh), čija se osnovna jednačina i izvod nalaze ispod, a u izlaznom sloju koristiti sigmoidnu aktivacionu funkciju. Skup podataka podeliti u odnosu od 70:30 (trening:test) uz posebnu pažnju na to da raspodela klasa bude ravnomerna u train i test skupu. Evaluirati model nad test podacima i izračunati *precision*, *recall i F1* meru modela, a nakon toga ih ispisati na konzoli.

$$precision = \frac{TP}{TP+FP}$$

$$recall = \frac{TP}{TP+FN}$$

$$F1\ score = \frac{precision*recall}{precision+recall}$$

TP (true positive) = predviđeno True i stvarno je trebalo biti True
TN (true negative) = predviđeno False i svarno je trebalo biti False
FP (false positive) = predviđeno True, a trebalo je biti False
FN (false negative) = predviđeno False, a trebalo je biti True



$$tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$$

$$\frac{\partial tanh(x)}{\partial x} = 1 - tanh(x)^{2}$$