PROJEKAT IZ PREPOZNAVANJA OBLIKA

U zavisnosti od odabrane baze, potrebno je **rešavati problem klasifikacije, regresije ili klasterizacije**. Pre svega, uraditi analizu baze što podrazumeva i rešavanje nedostajućih vrednosti i po potrebi prevođenje kategorija kategoričkih varijabli u numeričke vrednosti. Potom odabrati **bar 3 algoritma** i unakrsnom validacijom uporediti njihove performanse na odabranoj bazi. **Pokušati i redukciju dimenzionalnosti** nekom od metoda, pa ponoviti odabrane algoritme i uporediti performanse. Konačno, u izveštaju (do 4 str.) predstaviti najvažnije detalje iz analize podataka, kao i uporedne analize performansi modela, uz kratke teorijske opise korišćenih algoritama. Podrazumeva se da izveštaj treba da ima i uvod u problem, cilj istraživanja, kao i zaključke na samom kraju i preporučljivo je navesti korišćenu literaturu. U okviru pojedinih foldera za projektne fajlove možete naći dodatne dokumente koji mogu da posluže za inspiraciju ili koji detaljnije opisuju bazu.

Projekti se mogu raditi samostalno ili u parovima. Najviše dva tima (do 4 studenta) mogu da rade na istoj temi, pritom se metodološki projekti moraju razlikovati.

Rok za odabir projekta je 28.02.2022. putem moodle ankete.

Rok za predaju projekta (kod+izveštaj) je 30.04.2022.

BAZE

1. **Detekcija karcinoma dojke**
   Baza sadrži podatke o 116 pacijenata, njihove godine, BMI i podatke dobijene analizom krvi, kao i podatak o prisustvu/odsustvu karcinoma dojke. Potrebno je napraviti algoritam koji će na osnovu datih podataka odlučivati da li ispitanik boluje od karcinoma dojke ili ne.
2. **Terapija lečenja bradavica**
   Baza sadrži 180 pacijenata. Na 90 je primenjena imunoterapija, a na preostalih 90 krioterapija. Za svakog pacijenta poznata je informacija o polu, starosti, vremenu trajanja bradavice, broju bradavica, vrsti bradavice, površini koju zauzimaju, kao i podatak da li je pacijent odreagovao na terapiju ili nije. Treba ispitati ima li neki od parametara (ili njihova kombinacija) uticaj na ishod terapije i može li se utvrditi prema datim parametrima koju terapiju je bolje primeniti i može li se uopšte ishod terapije predvideti korišćenjem metoda mašinskog učenja.
3. **Satelitski snimci njiva**
   Podaci se sastoje iz 4 Landsat-8 satelitska snimka koji su snimljeni iznad Bačke: 13.06., 31.07., 16.08. i 01.09.2013. godine, a iz kojih je za svaki od piksela izdvojeno po 10 spektralnih merenja sa dva multispektralna merna instrumenta: OLI i TIRS senzora, koji se nalaze na satelitu. Treba napraviti algoritam na nivou piksela koji će na osnovu spektralnih merenja moći da odluči koja vrsta žitarice je zasađena na određenom području (obratiti pažnju da se pikseli koji pripadaju jednoj njivi ne smeju istovremeno naći i u trening i u test skupu).
4. **Detekcija prisustva zaposlenih**
   Baza se sastoji od nekoliko csv fajlova koji sadrže varijable. Neke varijable, kao npr. unutrašnja temperatura, date su po prostoriji, dok su neke varijable, kao npr. spoljašnja temperatura date globalno. Pored navedenih tu su i drugi parametri izmereni unutar poslovnog prostora, kao i drugi meteorološki parametri i vremenski trenuci snimanja (datum i vreme). Treba napraviti algoritam koji na osnovu (nekih od) podataka iz baze može da zaključi da li u poslovnom prostoru ima nekog

od zaposlenih ili ne. Studentu se ostavlja izbor da li će raditi procenu prisustva zaposelnih u celom prostoru ili nekoj konkretnoj prostoriji, kao i da li će raditi klasifikacioni ili regresioni problem.

5. **Predviđanje jačine betona**

   Beton predstavlja jedan od najvažnijih materijala u građevinarstvu. Kompresivna snaga betona predstavlja jednu od njegovih najvažnijih karakteristika, a koja zavisi od starosti betona i sastojaka od kojih je napravljen. Na osnovu 1030 uzoraka betona i datih karakteristika, kreirati model koji će na osnovu datih karakteristika predviđati kompresivnu snagu betona.

6. **Milano telekomunikacioni podaci**

   Baza sadrži vremenske serije agregiranih mobilnih aktivnosti korisnika po jedinici površine. U pitanju su podaci o količini SMS poruka, poziva i internet saobraćaja, datih za grad Milano za period od 1.11.2013. do 7.11.2013. Površina grada je izdeljena na područja (jedinice) od približno 235x235 metara, te takvih područja ima 1000, svaki sa jedinstvenim identifikacionim brojem. Za svako područje su dati podaci o mobilnoj aktivnosti agregirani svakih sat vremena. Treba napraviti algoritam koji će otkriti karakteristične obrasce ponašanja za određena područja, tj. izvršiti klasterizaciju delova grada Milano na osnovu mobilne aktivnosti korisnika (koristiti samo *sms-call-internet-mi-2013-11-0X.csv* fajlove i to samo podatke sa *countrycode=39* i eventualno po potrebi *milano-grid.geojson* fajl).

7. **Klasterizacija gena**

   Baza sadrži genske ekspresije za 801 uzorak koji potiču od pacijenata sa dijagnozama 5 različitih tumora. Metodatama klasterizacije analizirati različite klastere, uvideti pravilnosti i specifičnosti baze. Da li se prema genskim ekspresijama mogu identifikovati klasteri koji odgovaraju vrstama tumora ili se izdvajaju podvrste ili se uočavaju sličnosti uzoraka koji potiču od različitih vrsta tumora.

8. **Prepoznavanje šaka**

   Baza podataka sadrži 1208 slika šaka, po 8 od 151 osobe. Slike šake jedne osobe razlikuju se po osvetljaju, prisustvu aksesoara, položaja, itd. U pitanju je podskup originalne baze koja je sadržala 11000 slika. Slike su smanjene rezolucije u odnosu na originalnu (150x200). Slike su vektorizovane (vektor je niz vrsta originalne slike) i smeštene po vrstama u csv fajl. Student ako želi može da koristi i celokupnu bazu sa datog sajta ili neki njen podskup, a cilj je kreirati algoritam za identifikaciju osobe.

9. **CIFAR10**

   Baza se sastoji od 60.000 slika dimenzije 32x32 koje potiču iz 10 različitih klasa: avion, automobil, ptica, mačka, jelen, pas, žaba, konj, brod i kamion. Iako je čoveku jako lako da razlikuje ove pojmove na slikama, za mašine to predstavlja veliki izazov. Baza je podeljena u 6 skupova po 10.000 slika, od kojih je jedan označen kao test skup. Svaki skup predstavlja matricu vektorizovanih slika, gde je vektor niz vrsta slike, i to prvo crveni, pa zeleni i na kraju plavi kanal (3x32x32=3072 piksela). Kreirati model koji će vršiti klasifikaciju slika.

10. **Fashion MNIST**

    Baza podataka sadrži 60.000 slika za obuku i 10.000 slika za testiranje, dimenzija 28x28 piksela. Slike predstavljaju odevne predmete iz 10 različith klasa: majica, pantalone, džemper, haljina, kaput, sandala, košulja, patika, torba i čizma. Baza je napravljena po ugledu na popularnu MNIST bazu za klasifikaciju rukom pisanih cifara, predstavljajući nešto kompleksniji problem. Treba kreirati model koji će na osnovu slike moći da identifikuje vrstu odevnog predmeta. Slike su vektorizovane (vektor predstavlja niz vrsta originalne slike) i smestene po vrstama u matricu.

11. **Analiza audio scene**

    Audio-vizuelna analiza scene u cilju detekcije anomalija (sudara, pucnjeva i sl.) kao i u cilju ostvarenja pametnih gradova, danas je jedna od izuzetno aktuelnih oblasti istraživanja. Baza za projekat se sastoji od audio fajlova iz kojih su izdvojena akustička obeležja i data u vidu *csv* fajla, a zadatak je da se primenom algoritama mašinskog učenja napravi model koji će vršiti klasifikaciju ovih zvukova. U pitanju su zvukovi iz urbane sredine, poput saobraćaja, sirene, ptica, govora ljudi i dr.

12. **Iznajmljivanje bicikla**

    Baza sadrži podatke o broju inajmljenih biciklova i meteorološke podatke. Potrebno je na osnovu meteoroloških podataka predvideti ukupan broj iznajmljenih biciklova ili broj iznajmljenih biciklova od strane registrovanih ili neregistrovanih korisnika.

13. **Nokia telekomunikacioni podaci**

    Baza sadrži izdvojena obeležja iz logova mobilnih telefona – ceo i redukovan skup. Ta obeležja podrazumevaju vreme pokretanja aplikacije, primanja poziva/poruka, promenu GSM ćelije, rastojanja koja korisnik pređe, senzorna ubrzanja i dr. Treba napraviti algoritam koji će na osnovu izdvojenih obeležja određivati da li je korisnih muškarac ili žena. Podaci su deo NOKIA izazova.

14. **Prepoznavanje lica**

    Baza se sastoji od mnoštva slika sa licima 200 osoba. Slike jedne osobe razlikuju se po osvetljaju, položaju lica, prisustvu naočara, brkova kod muškaraca, šminke kod žena, itd. Treba napraviti algoritam koji će na osnovu slike izvršiti identifikaciju osobe. Za trening se koriste 3 slike po osobi, a za test 5 slika po osobi. Redukcija dimenzionalnosti može se raditi lokalno (posebno na svakoj od podslika u okviru jedne slike) ili globalno (nad celom slikom).

15. **Prepoznavanje cifara**

    Baza se sastoji od 70000 slika cifara pisanih rukom. Podaci su razdvojeni na deo za obuku i deo za testiranje. Sve slike su predobrađene u smislu ujednačenja veličine slika i centriranja napisane cifre. Treba napraviti algoritam koji će na osnovu slike moći da prepozna koja cifra je napisana.

16. **Prepoznavanje zvukova beba**

    U pitanju je deo *Oxford Vocal Sounds* baze podataka sa AARHUS Univerziteta iz Danske koji se odnosi na zvuke koje proizvode bebe – plač, smeh i neutralni zvukovi. Iz wav signala izvučeno je 13 MFCC obeležja, osnovna frekvencija i prva dva formanta nad prozorima širine 25ms uz preklapanje od 10ms. Potom su nad signalima za ova obeležja izračunati srednja vrednost, minimum, maksimum i standardna devijacija (uz interpolaciju delova gde veličine nisu definisane) i to je dato kao konačna obeležja – 16x4, dakle 64 obeležja. Treba napraviti algoritam koji će na osnovu obeležja moći da prepozna da li beba plače ili se smeje ili proizvodi neutralne zvukove.

17. **Predviđanje vrednosti toplotne energije**

    Dati su podaci nastali u okviru ECOTECT programa za simulaciju koji prate okruženje definisano sa 8 parametara na 768 zgrada. Cilj je određivanje 2 odziva na sistem koji predstavljaju količinu toplotne energije koju je potrebnu dovesti ili odvesti da bi se održavala temperatura u prihvatljivom opsegu.

18. **Predviđanje produktivnosti radnika**

    Dato je 1197 podataka o radnicima u industriji odeće, koji imaju za cilj predviđanje potencijalne produktivnosti datog radnika u zavisnosti od seta poznatih informacija.

*Ako postoji problem sa nekim linkom i preuzimanjem neke od baza, javiti se asistentu.

*Ako ima mnogo podataka u bazi pa se algoritmi izvršavaju sporo, oni se mogu testirati na nekom manjem podskupu radi određivanja najbolje kombinacije parametara, a tek potom izvršiti konačna unakrsna validacija na svim podacima.

*Kada nešto zapne, prvo pogledati na internetu. Ukoliko ne možete da se snađete, **obratiti se asistentu.**