Domaći zadatak

I DEO: ANALIZA PODATAKA

1. Broj svog indeksa (bez godine) podeliti po modulu 5 - dobijeni broj označava bazu na kojoj treba raditi. U pitanju su baze koje se tiču vremenskih uslova u 5 različitih gradova u Kini.

   Baza 0: Peking - BeijingPM20100101_20151231.csv

   Baza 1: Čengdu - ChengduPM20100101_20151231.csv

   Baza 2: Guangdžou - GuangzhouPM20100101_20151231.csv

   Baza 3: Šangaj - ShanghaiPM20100101_20151231.csv

   Baza 4: Šenjang - ShenyangPM20100101_20151231.csv

   Pojašnjenje podataka iz baze:

   - *No*: redni broj vrste
   - *year*: godina
   - *month*: mesec
   - *day*: dan u mesecu
   - *hour*: sat u danu
   - *season*: godišnje doba
   - *PM*: koncentracija PM2.5 čestica na nekoliko lokacija ($\mu g/m^3$)
   - *DEWP*: temperatura rose/kondenzacije (°C)
   - *TEMP*: temperatura (°C)
   - *HUMI*: vlažnost vazduha (%)
   - *PRES*: vazdušni pritisak (hPa)
   - *cbwd*: pravac vetra (N-sever, S-jug, E-istok, W-zapad, cv-calm/variable)
   - *Iws*: kumulativna brzina vetra (m/s)
   - *precipitation*: padavine na sat (mm)
   - *Iprec*: kumulativne padavine (mm)

2. Sa moodle platforme skinuti bazu podataka koja je dobijena na osnovu broja indeksa.
3. Učitati bazu u *DataFrame*. Proveriti kako izgleda prvih nekoliko vrsta u bazi.
4. Upoznati se sa bazom. Koliko ima obeležja? Koliko ima uzoraka? Šta predstavlja jedan uzorak baze? Kojim obeležjima raspolažemo? Koja obeležja su kategorička, a koja numerička? Postoje li nedostajući podaci? Gde se javljaju i koliko ih je? Postoje li nelogične/nevalidne vrednosti?
5. Izbaciti obeležja koja se odnose na sve lokacije merenja koncentracije PM čestica osim *US Post*.
6. Ukoliko postoje nedostajući podaci, rešiti taj problem na odgovarajući način. Objasniti zašto je rešeno na odabrani način.

7. Analizirati obeležja (statističke veličine, raspodela, …)
8. Analizirati detaljno vrednosti obeležja $PM_{2.5}$ (’*PM_US Post’*).
9. Vizuelizovati i iskomentarisati zavisnost promene $PM_{2.5}$ od preostalih obeležja u bazi.
10. Analizirati međusobne korelacije obeležja.
11. Uraditi još nešto po sopstvenom izboru (takođe **obavezna stavka**).

II DEO: LINEARNA REGRESIJA

Napraviti model linearne regresije koji predviđa koncengtraciju $PM_{2.5}$ čestica.

1. Potrebno je 15% nasumično izabranih uzoraka ostaviti kao test skup, 15% kao validacioni a preostalih 70% koristiti za obuku modela.
2. Isprobati različite hipoteze, regularizaciju modela, selekciju obeležja (unapred ili unazad).
3. Odabrati najbolji model linearne regresije i objasniti zašto je baš taj model odabran.

III DEO: KNN KLASIFIKATOR

Napraviti klasifikator koji koristi kNN metodu za klasifikaciju uzoraka u jednu od 3 grupe zagađenja.

1. Prvo je potrebno uzorcima iz date baze dodeliti labele: *bezbedno*, *nebezbedno* ili *opasno*. Uzorcima čija je vrednost koncentracije PM2.5 čestica do 55.4 µg/m3 dodeliti labelu *bezbedno*, onima čija je vrednost koncentracije PM2.5 čestica od 55.5 µg/m3 do 150.4 µg/m3 dodeliti labelu *nebezbedno*, dok onima sa vrednošću preko 150.5 µg/m3 dodeliti labelu *opasno*.
2. Koristiti 15% uzoraka za testiranje finalnog klasifikatora, a preostalih 85% uzoraka koristiti za metodu unakrsne validacije sa 10 podskupova. Ovom metodom odrediti optimalne parametre klasifikatora, oslanjajući se na željenu meru uspešnosti. Obratiti pažnju da u svakom od podskupova za unakrsnu validaciju, kao i u test skupu, bude dovoljan broj uzoraka svake klase.
3. Za konačno odabrane parametre prikazati i analizirati matricu konfuzije dobijenu akumulacijom matrica iz svake od 10 iteracija unakrsne validacije. Odrediti prosečnu tačnost klasifikatora, kao i tačnost za svaku klasu.
4. Klasifikator sa konačno odabranim parametrima obučiti na celokupnom trening skupu, pa testirati na izdvojenom test skupu. Na osnovu dobijene matrice konfuzije izračunati mere uspešnosti klasifikatora, kao i mere uspešnosti za svaku klasu (tačnost, osetljivost, specifičnost, preciznost, F-mera).
5. Rezultate prikazati i diskutovati u izveštaju.

Kod možete pisati u PyCharmu, Jupyteru, Colabu, Spyderu...

Za sva eventualna pitanja, nejasnoće ili ako smatrate da je traženo nešto što se ne može uraditi ili deluje preterano zahtevno, obratiti se mailom asistentu. Pri pisanju izveštaja pratiti uputstva koja su data. Postavljeni su izveštaji koji su uspešno urađeni, ali treba imati na umu da se od vas ne traže iste stvari, te se ne treba slepo držati formata i informacija koje postoje u tim izveštajima. **U izveštajima ne treba objašnjavati kod niti ga prepisivati, akcenat je na objašnjenu korišćene metodologije, prikazu i interpretaciji rezultata analize i vizuelizaciji.** Ako se u zadatku traži 15 slika da napravite i analizirate, to treba i da uradite u kodu, ali nije neophodno svih 15 slika staviti u izveštaj. Ako je ostalo bilo šta nejasno povodom pisanja izveštaja, stojimo na raspolaganju. Izdvojite dovoljno vremena za pisanje izveštaja kako biste ga uradili kvalitetno.

Izveštaj u PDF formatu i skriptu koja sadrži kod (.py ili .ipynb, ako imate više skripti, smestite sve u jednu) **potrebno je predati putem moodla najkasnije do 23:59 30.12.2022.**

Domaći radite samostalno – **dva ista koda ili dva suštinski ista izveštaja dobijaju 0 bodova bez daljeg istraživanja kako su nastali.**

UPUTSTVO ZA IZRADU DOMAĆEG ZADATKA

1. **Pročitati lekcije na koje se domaći oslanja**, sa razumevanjem. Uočiti potencijalne slabosti i dobre strane metode koja se primenjuje, biti svestan problema koji mogu da se jave.
2. **Pročitati tekst domaćeg zadatka u celosti** i ako je bilo šta nejasno, odmah pitati asistenta na konsultacijama ili mejlom.
3. **Pogledati kodove u vežbama relevantne za ovaj domaći**, jer je sigurno bar većina potrebnih funkcija korišćena na vežbama.
4. **Osmisliti kod**. Ako je moguće, razmisliti kako se neke stvari mogu uraditi kompaktnije (npr. u petlji uraditi nešto za sve klase).
5. Za sve nedoumice konsultovati sav materijal sa predavanja i vežbi. Ako je i dalje nešto nejasno, pokušati pronaći odgovor na internetu. Ako je idalje nejasno, obratiti se asistentu.
6. **Napisati izveštaj da pokriva sve traženo**, paziti na formatiranje i na slike. Pomoću slika, tabela i teksta predstaviti i protumačiti rezultate tako da rezultati budu pregledni i jasni (npr. budućem klijentu).
7. Pročitati izveštaj još jednom i nazvati ga **domaci_ime_prezime.pdf**
8. **Pred odbranu se još jednom podsetite šta ste pisali u izveštaju, šta ste radili u kodu i zašto** (čemu služe prosleđeni parametri, šta vraća neka funkcija i sl.).

UPUTSTVO ZA PISANJE IZVEŠTAJA

Izveštaj sadrži uvod, opis korišćene baze, teorijske osnove korišćenih metoda, prikaz i interpretaciju rezultata, zaključak i spisak korišćene literature.

Uvod treba da sadrži nekoliko rečenica o temi koja se obrađuje. Ako je tema recimo Parkinsonova bolest, onda nešto kratko o toj bolesti, kao i značaju analize podataka i istraživanja teme i značaju kreiranja modela regresije/klasifikacije. Ako je tema npr. potrošnja električne energije, onda o problemu potrošnje, razlozima za uštedu i značaju analize i istraživanja ove oblasti kao i pravljenju modela zasnovanog na algoritmu mašinskog učenja.

Poglavlje o bazi podataka treba da sadrži kratak opis baze. Objasniti šta baza predstavlja. Zatim, navesti koliko uzoraka se nalazi u bazi; koliko ima obeležja; da li su obeležja numerička ili kategorička, itd.

Student sam određuje broj i nazive poglavlja u kojima će opisati šta je uradio i prokomentarisati dobijene rezultate.

Neke od analiza mogu biti ilustrovane korišćenjem slike ili tabele. Svaka slika i tabela moraju imati i odgovarajući naziv koji jasno objašnjava šta se može videti na slici i/ili u tabeli, ali moraju biti i referencirane u samom tekstu.

Zaključak nije neophodan. Iako zaključak može da sadrži pregled ključnih rezultata i objasni njihov značaj, nemojte da ponavljate deo opisan u uvodu. Preporučljivo je navesti i korišćenu literaturu.

Stil pisanja treba da bude naučni, što se odlikuje jasnim i konciznim rečenicama i korišćenjem stručnih termina. Obratiti pažnju na pravopis i gramatiku, ali i formatiranje teksta (detalji o formatiranju dati su u šablonu za pisanje izveštaja). Bodovi će se smanjivati, kako za nepotpuno urađen zadatak, tako i za nepismeno napisane i neformatirane izveštaje. Ukoliko asistent posumnja na prepisivanje, student će biti sankcionisan. Predajom izveštaja i propratnog koda, student obezbeđuje sebi pravo na usmenu odbranu, kada će i dobiti do 25 bodova, koliko maksimalno nosi domaći zadatak. Tih 25 bodova raspoređeno je na sledeći način: 10 bodova nosi sam izveštaj – preciznost, jasnoća, formatiranje i potpunost urađenog zadatka; 5 bodova nosi sam kod – da li je u kodu urađeno sve što je u zadatku traženo i da li je tačno urađeno (kod može biti pitan i na odbrani – ispitivanje funkcija, parametara ili da se traži da se nešto izmeni u kodu na licu mesta); preostalih 10 bodova nosi usmena odbrana – asistentova i profesorova procena studentovog razumevanja zadatka koje će biti provereno kroz usmena pitanja na samoj odbrani.

Iako su bodovi raspoređeni kako je u prethodnom pasusu napisano, bez predaje izveštaja kao ni bez usmene odbrane, student ne može dobiti bodove za domaći zadatak (dobija 0 bodova bez mogućnosti naknadne predaje).