1. Domaći zadatak
Analiza podataka

1. Broj svog indeksa podeliti po modulu 5 - dobijeni broj označava bazu na kojoj treba raditi.

Baza 0: PRSA_Data_Changping_20130301-20170228.csv

- Izlazna promenljiva za linearnu regresiju: $PM_{2.5}$

Baza 1: PRSA_Data_Guanyuan_20130301-20170228.csv

- Izlazna promenljiva za linearnu regresiju: PM_{10}

Baza 2: PRSA_Data_Tiantan_20130301-20170228.csv

- Izlazna promenljiva za linearnu regresiju: NO_2

Baza 3: bikesharing_modified.csv

- Izlazna promenljiva za linearnu regresiju: *cnt* (broj iznajmljenih bicikala)
- Obeležje *dtday* ne treba koristiti

Baza 4: SeoulBikeData.csv

- Izlazna promenljiva za linearnu regresiju: *Rented Bike Count* (broj iznajmljenih bicikala)
- Izbaciti obeležje *Date,* a umesto njega napraviti obeležja *Mesec* i *DanUNedelji* (hint: https://medium.com/@swethalakshmanan14/simple-ways-to-extract-features-from-date-variable-using-python-60c33e3b0501)
- Pri učitavanju potrebno je podesiti parametar *encoding='latin1'*

2. Sa moodle platforme skinuti bazu podataka koja je dobijena na osnovu broja indeksa.
3. Pročitati tekstualni fajl dat uz svaku bazu sa objašnjenjem obeležja i same baze.
4. Potom učitati bazu u DataFrame. Proveriti kako izgleda prvih nekoliko vrsta u bazi.
5. 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?
6. Izbaciti obeležja koja neće biti korišćena. Objasniti zašto su izbačena (nekad je po uslovu zadatka).
7. Ukoliko postoje nedostajući podaci, rešiti taj problem na proizvoljan način (neke od mogućnosti rađene su na vežbama). Objasniti zašto je rešeno na odabrani način.
8. Izanalizirati obeležja (statističke veličine, raspodela, …)
9. Izanalizirati detaljno vrednosti obeležja koje će biti postavljeno kao izlaz linearne regresije (dato za svaku od baza koje je to obeležje).

10. Vizuelizovati i iskomentarisati zavisnost promene promenljive koja se predviđa linearnom regresijom od preostalih obeležja u bazi.
11. Analizirati korelaciju svih obeležja međusobno.
12. Uraditi još nešto po sopstvenom izboru (takođe obavezna stavka).

Nakon sprovedene analize, napraviti model linearne regresije koji predviđa promenljivu zadatu uz bazu.

1. Potrebno je 10% nasumično izabranih uzoraka ostaviti kao test skup, a preostalih 90% koristiti za pravljenje modela.
2. Isprobati različite hipoteze, primeniti selekciju obeležja, kao i regularizaciju modela.
3. Odabrati najbolji model linearne regresije i objasniti zašto je baš taj model odabran.

Za sva eventualna pitanja, nejasnoće ili ako smatrate da je traženo nešto što se ne može uraditi ili deluje preterano zahtevno, obratiti se mailom na tijana.delic92@gmail.com. Pri pisanju izveštaja pratiti uputstva koja su data. Postavljeni su izveštaji za koje smatram da 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 interpretaciji rezultata analize i vizuelizaciji.** Ako se u zadatku traži 15 slika da napravite i izanalizirate, 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, stojim na raspolaganju. Izveštaj se ne može napisati za sat-dva, tako da ostavite sebi dovoljno vremena da ga uradite kvalitetno. **Putem moodla najkasnije do 23:59 6.12.2020. treba predati 2 fajla: skriptu** koja sadrži kod (.py ili .ipynb, ako imate više skripti, smestite sve u jednu) **i izveštaj** (u .pdf formatu). Možete pisati kod u PyCharmu, Jupyteru, Colabu, Spyderu... Domaći radite samostalno – **dva ista koda ili dva ista izveštaja dobijaju 0 bodova bez daljeg istraživanja kako su nastali.**