1. Domaći zadatak
Analiza podataka

1. Broj svog indeksa 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

Link na originalni rad je https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2016JD024877

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 (ug/m^3)
- DEWP: temperatura rose/kondenzacije (stepeni Celzijusa)
- TEMP: temperatura (stepeni Celzijusa)
- HUMI: vlažnost vazduha (%)
- PRES: vazdušni pritisak (hPa)
- cbwd: pravac vetra (N-sever, S-jug, E-istok, W-zapad, cv-calm/variable)
- Iws: brzina vetra (m/s)
- precipitation: padavine na sat (mm)
- Iprec: ukupne padavine (mm)

2. Sa moodle platforme skinuti bazu podataka koja je dobijena na osnovu broja indeksa (ostatak pri deljenju sa 5).
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 proizvoljan način (neke od mogućnosti rađene su na vežbama). Objasniti zašto je rešeno na odabrani način.
7. Analizirati obeležja (vrsta obeležja, osnovne statistike, raspodela, …)
8. Analizirati detaljno vrednosti obeležja koje će biti postavljeno kao izlaz linearne regresije, a to je $PM_{2.5}$ (*'PM_US Post'*).
9. Vizuelizovati i iskomentarisati zavisnost promene promenljive koja se predviđa linearnom regresijom od preostalih obeležja u bazi.
10. Analizirati međukorelaciju obeležja.
11. Po sopstvenom izboru uraditi još neku vrstu analize (takođe obavezna stavka).

Nakon sprovedene analize, napraviti model linearne regresije koji predviđa koncengtraciju $PM_{2.5}$ čestica.

0. **Pročitajte lekciju o linearnoj regresiji dostupnu u PDF formatu sa predavanja.**
1. Potrebno je 10% nasumično izabranih uzoraka ostaviti kao test skup, a preostalih 90% koristiti za obuku modela.
2. Isprobati različite hipoteze, primeniti selekciju obeležja, kao i regularizaciju.
3. Odabrati konačni model linearne regresije koji po vama predstavlja najbolji model od svih ispitanih modela i objasniti zašto je baš taj model odabran.

O temi i obeležjima treba zadravorazumski promisliti, nemojte reda radi raditi analize koje nemaju nikakvog smisla. Isto tako, razmislite šta može da utiče na izlaznu varijablu i koje kombinacije obeležja je interesatno posmatrati.

- 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 interpretaciji rezultata analize i vizuelizaciji.** U izveštaju se ne koriste parametri funkcija da bi se objasnilo šta je urađeno.
- Ako se u zadatku traži da napravite i analizirate više slika, to treba i da uradite u kodu, ali nije neophodno sve slike da stavite u izveštaj, izaberite neke reprezentativne i interesantne slike jer ćete na te slike i njihov sadržaj ukazivati i u tekstu izveštaja.
- Izveštaj se ne može napisati za sat-dva (za potreban broj bodova), tako da ostavite sebi dovoljno vremena da ga uradite kvalitetno
- **Domaći se radi samostalno**
- **Dva ista koda ili dva ista izveštaja dobijaju 0 bodova bez daljeg istraživanja kako su nastali**.

Izveštaj predati putem moodla najkasnije do 23:59 27.11.2021. Potrebno je postaviti 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...