Analiza podataka i obrada informacija

Primjer kolokvija #2

Kolokvij nosi ukupno **30 bodova** i piše se **120 minuta**.

Na pitanja odgovarajte tako da pitanja kopirate u kôd i zakomentirate, **ispod svakog napišete odgovarajuće naredbe**, a **ispod naredbi kao komentar upišete odgovore**. Kôd treba pokazati kako ste došli do rezultata.

- **Obavezno navedite redni broj pitanja i zadatka!**

Zadatak 1. (12 bodova)

Standardizirana mjera fertiliteta i socioekonomski pokazatelji za svaku od 47 frankofonskih pokrajina Švicarske oko 1888. godine.

Varijabla	**Opis**
Fertility	Standardizirana mjera fertiliteta (*uobičajena mjera fertiliteta*)
Agriculture	Postotak muškaraca zaposlenih u poljoprivredi kao zanimanju
Examination	Postotak regruta koji su dobili najvišu ocjenu na vojnom ispitu
Education	Postotak regruta s obrazovanjem iznad osnovne škole
Catholic	Postotak katolika (*naspram protestanata*) u populaciji

1. Učitajte skup podataka `swiss`. Izdvojite `80%` podataka koristeći svoj JMBAG.
2. **Vizualizirajte** odnos između varijabli `Fertility` i `Education`. Opišite **smjer** i **snagu** korelacije?
3. Izračunajte **Pearsonov** i **Spearmanov** koeficijent korelacije između `Agriculture` i `Education`. Koji koeficijent je jači?
4. Napravite **korelacijsku matricu** svih varijabli? Koje varijable imaju najveću **pozitivnu** a koje **negativnu** korelaciju?
5. Napravite **linearni regresijski model** za predviđanje `Fertility` na temelju svih ostalih varijabli. Koje varijable su **značajne** i u kojem **smjeru**? Postoji li problem **multikolearnosti**?

Zadatak 2. (8 bodova)

Prosječne temperature zraka na dvorcu Nottingham, izražene u stupnjevima Fahrenheita, u razdoblju od 20 godina.

- **Vremenski period:** od siječnja 1920. do prosinca 1939
- **Frekvencija**: 12 (*mjesečna opažanja*)
- **Broj mjerenja**: 240 (*20 godina × 12 mjeseci*)

1. Učitajte vremenski niz `nottem`.
2. Provedite **STL dekompoziciju** niza i prikažite rezultate. Kakav je **trend**?
3. Provedite **ADF test**. Je li niz stacionaran? Objasnite.
4. Izradite **ARIMA model**, koliko iznosi **veličina pogreške** (*RMSE*) u modelu? Napravite predikciju za sljedeće 2 godine i prikažite grafički.

Zadatak 3. (10 bodova)

1. Učitajte skup podataka **adult.data**. Naznačite nazive varijabli.
   - `"age", "workclass", "fnlwgt", "education", "education.num", "sex", "capital.gain", "capital.loss", "hours.per.week", "native.country", "income"`
2. Pretvorite ciljnu varijablu `income` u binarnu (`1` ako osoba zarađuje >50K, `0` inače).
   - izvucite `70%` podataka u varijablu `train` koristeći svoj **JMBAG**
   - ostale u varijablu `test`
3. Izradite **klasifikacijsko stablo** koristeći `train` podatke i prikažite ga grafički.
   - hoće li osoba koja **nije oženjenja** i ima **capital gain 6500** imati income `>50k`
   - hoće li osoba koja **nema dijete**, ima **zanimanje ribolov**, **capital gain 5200** imati income `>50k`
4. Izračunajte i interpretirajte **točnost** klasifikacijskog stabla koristeći `test` podatke.

Predajte sljedeće datoteke:

- `R` ili `Rmd` datoteku
- **Opcionalno**:
  - **JPG datoteku** s grafičkim prikazom
  - **Rezultate izvođenja** u `PDF` formatu