Inteligentni sistemi

Prezime Ime: _____ Broj indeksa: _____

R uputstvo:

- U prilogu uz ovaj zadatak dat je materijal koji obuhvata dataset i cheatsheet. Na desktopu kreirati folder sa nazivom 22-ImePrezime. R projekat kreirati u pomenutom direktorijumu na desktopu i nazvati ga RProjekat.
- Zadatak se priznaje samo ako je kreiran PROJEKAT, inače se ne pregleda. Nije dovoljno kreirati samo R script fajl. R script fajl ***ne*** nazivati isto kao ceo projekat!
- Folder u kojem se nalazi kreirani projekat mora da sadrži SVE elemente koji su neophodni da se na nekom DRUGOM računaru (na kojem postoje instalirani R i RStudio) projekat otvori, da se u okviru otvorenog projekta otvori odgovarajući R script, kao i da se taj R script izvršava bez dodatnih intervencija

---

ZADATAK (LR): Predikcija količine proteina u proizvodima brze hrane

U fajlu ***fastfood.csv*** dat je sveobuhvatan pregled nutritivnog sadržaja različitih proizvoda brze hrane iz popularnih lanaca brze hrane. Varijable dataset-a su:

- **restaurant:** naziv restorana
- **item:** naziv proizvoda
- **calories:** ukupne kalorije
- **cal_fat:** masne kalorije
- **total_fat:** ukupno masti
- **sat_fat:** zasićene masti
- **trans_fat:** trans-zasićene masne kiseline
- **cholesterol:** holesterol
- **sodium:** natrijum
- **total_carb:** ukupni ugljeni hidrati
- **fiber:** vlakna
- **sugar:** šećer
- **calcium:** kalcijum
- **protein:** protein

Potrebno je uraditi sledeće:

1. Napraviti podskup podataka koji **ne** sadrži proizvode čije ukupne masti (total_fat) prelaze **125**. (**1 poena**).
2. U dataset-u dobijenom na osnovu prethodnog zahteva proveriti da li postoje nedostajuće vrednosti (NAs ili “-”, ” ”, ””) i ako je moguće, zameniti ih adekvatnim vrednostima. Prokomentarisati postupak zamene vrednosti, odnosno zašto je određeni metod zamene vrednosti odabran. (**4 poena**)
3. Proceniti koje atribute je potrebno uključiti u model **linearne regresije** za predviđanje vrednosti varijable **protein.** Obavezno navesti u komentaru zašto su baš ti atributi uključeni u model. Takođe, ukoliko se neki atribut izostavi, obrazložiti zašto je izostavljen. Tako dobijeni redukovani dataset eventualno dodatno obraditi da bi bio pogodan za predviđanje vrednosti izl. varijable primenom linearne regresije. (**7 poena**)
4. Primenom linearne regresije, kreirati model za predviđanje vrednosti varijable **protein**, i na osnovu njega kreirati predviđanja**.** (**2 poena**)
5. Na osnovu kreiranog modela (**6 poena**):
    - protumačiti koeficijente svake varijable, odnosno interpretirati relaciju nezavisne i zavisne varijable na osnovu vrednosti koeficijenta, **u kontekstu problema koji se rešava u zadatku**[1]
    - navesti koji atributi su značajni za predikciju i na osnovu čega je to zaključeno; interpretirati taj zakljucak u kontekstu problema koji se rešava u zadatku
    - objasniti šta je koeficijent determinacije (R-squared) i protumačiti njegovu vrednost, takođe **u kontekstu problema koji se rešava u zadatku**
6. Napisati šta predstavlja svaki od četiri grafikona za dijagnostiku modela linearne regresije. Objasniti šta se može uvideti na grafikonima i **protumačiti** svaki grafikon **u kontekstu problema koji se rešava u zadatku**[2]. (**12 poena**)
7. Proveriti postojanje multikolinearnosti na urađenom modelu. Prokomentarisati rezultate provere kao i mogućnosti za poboljšanje modela. (**3 poena**)

Napomena: Varijable koje **očigledno** nema smisla koristiti za predviđanje izlazne se **mogu izbaciti bez analiziranja** (npr. Izbacivanje ID-a pri predikciji plate neke osobe) - potrebno je samo napisati komentar zasto se ta kolona izbacuje.

Svaka varijabla čiji odnos sa **izlaznom nije očigledan na prvi pogled**, mora se na neki način uporediti sa izlaznom, i na osnovu toga doneti zaključak da li se koristi u modelu li ne, sa propratnim komentarima.

---

[1] Značenje tog izraza je: ***ne*** navoditi u komentarima samo brojke koje se dobiju, već protumačiti šta one ***znače*** u kontekstu problema koji se rešava u zadatku i dataset-a koji se koristi.

[2] Značenje tog izraza je: ***ne*** opisivaiti u komentarima samo oblik grafikona, ***ne*** navoditi samo kakav tj oblik *treba* da bude, već protumačiti šta tako dobijeni grafikoni i njihovi oblici ***znače*** u kontekstu problema koji se rešava u zadatku i dataset-a koji se koristi.