



Mašinsko učenje 2024

Zadatak 2

Sadržaj



Zadatak 1 - Rekapitulacija



Zadatak 2

Zadatak 1 - Rekapitulacija

Zadatak 1 - Rekapitulacija

- Procenat uspešnosti: **82%** (28/34).
- Najveće preklapanje izvornih kodova prema alatu za detekciju plagijata: **18%**.
- Najbolji rezultati po terminima:

Termin	Tim	RMSE
Ponedeljak - G4	tim7_24	77.66
Utorak - G5	placeholder	93.54
Utorak - G3	T&M	81.90
Četvrtak - G2	@	81.30
Petak - G1	Tehno trube	81.30

Zadatak 1 - Rekapitulacija

- Dobre stvari (na nivou generacije):
 - Vizuelizacija podataka
 - Rad sa outlier-ima
 - Rad sa trening skupom podataka
 - Implementacija algoritama
 - Računanje metrike
 - Prpratni izveštaji.
- Stvari koje mogu biti bolje (na nivou generacije):
 - Normalizacija podataka.

Zadatok 2

Zadatak 2

- Višestruka regresija:
 - Prediktovati cenu (kolona **Cena** u **evrima**) automobila u Srbiji na osnovu više atributa.
 - Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije **RMSE (Root Mean Square Error) manji od 4100**.
 - Algoritmi mašinskog učenja se samostalno implementiraju - **zabranjena upotreba algoritama iz biblioteka**.
 - Rok za izradu zadatka je **16.04.2024. u 23:59h**.
 - Instalirane biblioteke za Zadatak 2:
 - NumPy
 - Pandas
 - SciPy.

Zadatak 2

- Sledeći termin vežbi (odbrana Zadatka 2 i predstavljanje Zadatka 3):

Termin	Datum
Ponedeljak - G4	22.04.2024.
Utorak - G5	23.04.2023.
Utorak - G3	23.04.2023.
Četvrtak - G2	25.04.2023.
Petak - G1	26.04.2023.

Zadatak 2

- Atributi (kolone) na osnovu kojih se prediktuje cena:
 - **Marka** - marka/proizvođač
 - **Grad** - mesto gde se automobil nalazi
 - **Godina proizvodnje** - godina proizvodnje automobila
 - **Karoserija** - vrsta karoserije automobila:
 - Hečbek, Limuzina, Karavan, Džip/SUV, Monovolumen (MiniVan), Kupe, Kabriolet/Roadster, Pickup
 - **Gorivo** - pogonsko gorivo:
 - Dizel, Benzin, Benzin + Gas (TNG), Benzin + Metan (CNG), Hibridni pogon, Hibridni pogon (Dizel), Hibridni pogon (Benzin)
 - **Zapremina motora** - zapremina motora u cm^3

Zadatak 2

- Atributi (kolone) na osnovu kojih se prediktuje cena:
 - **Kilometraza** - broj pređenih kilometara
 - **Konjske snage** - broj konjskih snaga
 - **Menjac** - vrsta menjača:
 - **Manuelni, Automatski**

Zadatak 2

- Gradivo za Zadatak 2 obuhvata kompletno gradivo od početka semestra zaključno sa **Metodom maksimalne verodostojnosti** (Predavanje 5 planirano za 27.03.2024.).
- Koncepti za Zadatak 2:
 - Rad sa kategoričkim podacima
 - Matrica korelacije
 - Višestruka linearna regresija $y=h(x_1, x_2, \dots, x_d)$.
 - Regularizacija
 - Neparametarski pristup

Zadatak 2

- Rad sa kategoričkim podacima:
 - Atributi (kolone) **Marka**, **Grad**, **Karoserija**, **Gorivo** i **Menjac** sadrže kategoričke podatke.
 - Neke od tehnika za rad sa kategoričkim podacima su:
 - **Label Encoding** - konvertovanje kategoričkih podataka u broj iz opsega [0, broj-klasa-1], npr.: za kolonu **Menjac** vrednosti [**Manuelni**, **Automatski**] će se konvertovati u vrednosti [0, 1].
 - **One Hot Encoding** - konvertovanje svake klase u novu kolonu i pridruživanje vrednosti 1 ili 0 (*True* ili *False*), npr.: za kolonu **Menjac** ćemo dobiti dve binarne kolone **Manuelni** i **Automatski**.
 - **Custom Binary Encoding** - kombinacija **Label Encoding**-a i **One Hot Encoding**-a kako bi se kreirala dodatna kolona od značaja.

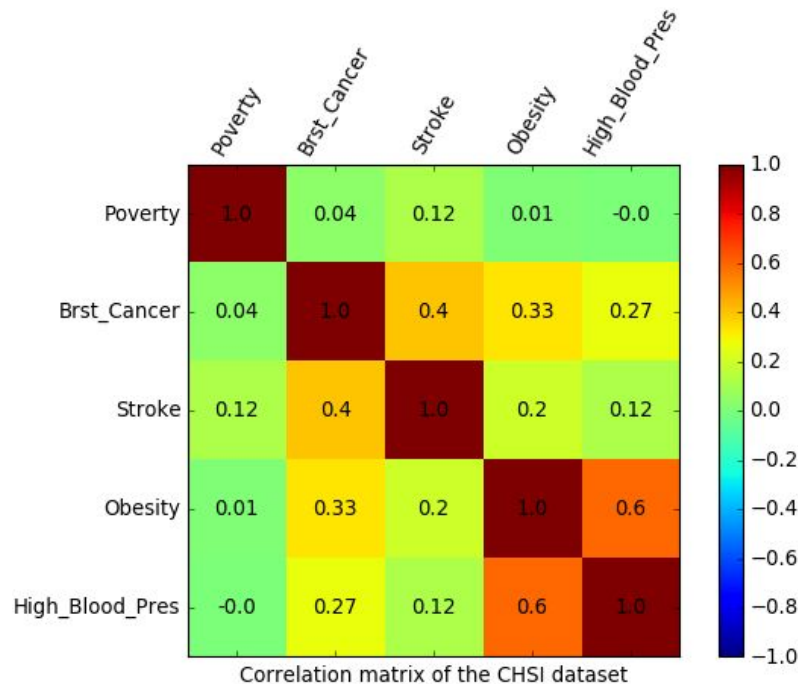
Zadatak 2

- Matrica korelacije:
 - Svaki element u matrici pokazuje koeficijent korelacije između 2 promenljive iz opsega $[-1, 1]$ gde:
 - 1 označava savršenu pozitivnu korelaciju (kako jedna promenljiva raste, tako raste i druga)
 - 0 označava da nema korelacije
 - -1 označava savršenu negativnu korelaciju (kako jedna promenljiva raste, druga opada).
 - Razumevanje matrice korelacije može pomoći u:
 - Izboru promenljivih za modele
 - Otkrivanju mogućih uzročno-posledičnih veza
 - Izbegavanju problema multikolinearnosti.

Zadatak 2

- Matrica korelacije:

- *NumPy*:
 - `correlation_matrix = np.corrcoef(data)`
- *Pandas*:
 - `correlation_matrix = df.corr()`



Zadatak 2

- Regularizacija:
 - Modifikacija optimizacionog problema koje ograničava prilagodljivost modela i čini ga manje podložnim preprilagođavanju
 - *Lasso (L1):*
 - *Gradient Descent*
 - *Coordinate Descent*
 - *Ridge (L2):*
 - *Gradient Descent*
 - *Closed Form Solution*
 - *Elastic Net:*
 - Linearna kombinacija L1 i L2

Zadatak 2

- Neparametarski pristup:
 - Jednostavnost i fleksibilnost
 - *Nearest Neighbors*
 - *Kernel Regression*

Zadatak 2

- Saveti za rešavanje zadatka:
 - Podsetiti se gradiva sa predavanja
 - Vizuelizovati i analizirati podatke
 - Pokušati smanjiti dimenzionalnost problema
 - Isprobati i parametarske i neparametarske pristupe
 - Ako se radi normalizacija podataka, obratiti pažnju kako se radi i kako će se računati RMSE metrika

Zadatak 2

- Dodatno istraživanje:
 - Gradivo obrađeno na predavanjima je dovoljno kako bi se zadatak uspešno uradio
 - **Dodatnim istraživanjem (pod)oblasti i problema moguće je ostvariti bolje rezultate.**
 - **Ohrabruje se dodatno istraživanje i primena istraženog, uz (jedino) ograničenje da (novi, istraženi) algoritmi moraju biti algoritmi višestruke regresije**
 - Ne postoje ograničenja što se tiče tehnika za obradu podataka i rad sa trening skupom
 - Pošaljite asistentu e-mail ukoliko niste sigurni da li nešto sme ili ne sme da se iskoristi za izradu zadatka.