LV5 –VIŠESTRUKA LINEARNA REGRESIJA – ZADACI

Cilj vježbe

Studenti će analizirati dani dataset koristeći višestruku linearnu regresiju. Naučit će postaviti model, interpretirati koeficijente i procijeniti prikladnost modela korištenjem statističkih mjera

Zadaci i koraci koje studenti moraju napraviti

1. Učitavanje podataka

- **Korak 1.1:** Učitajte dataset pomoću Pythona koristeći biblioteku `pandas`.
- **Korak 1.2:** Pregledajte osnovne informacije o datasetu:
    - Broj redaka i stupaca (što redovi i stupci predstavljaju).
    - Provjerite tipove podataka u datasetu (numerički, kategorijski).
    - Identificirajte ima li dataset nedostajuće vrijednosti koristeći `isnull()`s

2. Opisna analiza podataka

- **Korak 2.1:** Prikazati osnovne statistike za sve varijable:
    - Srednju vrijednost, medijan, standardnu devijaciju, minimalnu i maksimalnu vrijednost.
- **Korak 2.2:**Vizualizirati odnose među značajkama i ciljnom varijablom (scatter plotovi, korelacijski grafovi).
- **Korak 2.3:** Napraviti grafički prikaz i interpretirati razdiobe pojedinačnih varijabli i njihove međusobne zavisnosti. To nam daje dobar uvid u sadržaj skupa podataka i međusobni odnos varijabli. Možemo koristiti funkciju pairplot() iz biblioteke **seaborn**.

3. Priprema podataka

- **Korak 3.1:** Pretvaranje kategorijskih varijabli u numerički format:
    - Koristite **One-Hot Encoding** za značajke s više kategorija.
    - Koristite **Label Encoding** za binarne kategorijske značajke.
- **Korak 3.2:** Identificirajte i uklonite značajke s velikom multikolinearnošću (ako postoje):
    - Izračunajte korelacijske koeficijente među varijablama i interpretirati koji bi bile najbolje koristiti u modelu
    - Koristite `Variance Inflation Factor (VIF)` kako biste detektirali redundantne varijable.
- **Korak 3.3:** Normalizacija ili standardizacija numeričkih značajki (ako je potrebno):

- Primijenite tehniku standardizacije (npr., `StandardScaler` iz `sklearn`) na numeričke varijable.

4. Postavljanje višestruke linearne regresije

- **Korak 4.1:** Definirajte ciljne i ulazne varijable (npr. `X` za značajke i `y` za ciljnu varijablu).
- **Korak 4.2:** Koristite biblioteku `statsmodels` za kreiranje modela višestruke linearne regresije:
  - Importirajte funkciju `OLS` iz `statsmodels.api`.
  - Postavite model i prikazujte rezultate korištenjem metode `summary()`.

5. Analiza rezultata

- Interpretirajte koeficijente modela:
  - Koje značajke imaju najveći utjecaj na ciljnu varijablu?
  - Jesu li koeficijenti statistički značajni (`p-vrijednost < 0.05`)?
- Provjerite kvalitetu modela:
  - Što znači R^2 (koeficijent determinacije)?
  - Kako se model ponaša na temelju preostalih (rezidualnih) vrijednosti?

6. Vizualizacija rezultata

- Prikazati stvarne vrijednosti ciljne varijable naspram predviđenih vrijednosti (koristite scatter plot).
- Analizirati i interpretirati rezidualne vrijednosti i prikazati ih grafički.

Zaključak

- Koji su ključni faktori koji utječu na ciljnu varijablu?
- Kako se višestruka linearna regresija može koristiti za interpretaciju međusobnih odnosa među varijablama?