01 Domaci Linearna Regresija

Predmet: Tehnike i metode analize podataka

Student 1636 Milica Jovanovic

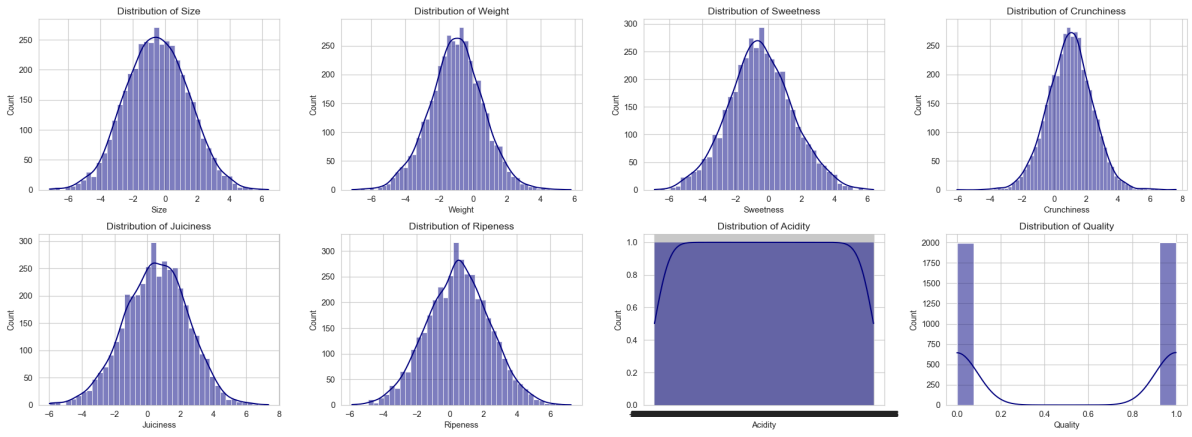
Za potrebe ovog domaceg zadatka koristila sam sledeci dataset: https://www.kaggle.com/datasets/nelgiriyewithana/apple-quality . Dataset sadrži informacije o kvalitetu jabuka, a ključni podaci uključuju različite parametre koji pomažu u klasifikaciji jabuka kao zrelih ili nezrelih, odnosno dobrih ili loših. Dataset sadrzi sledece kolone:

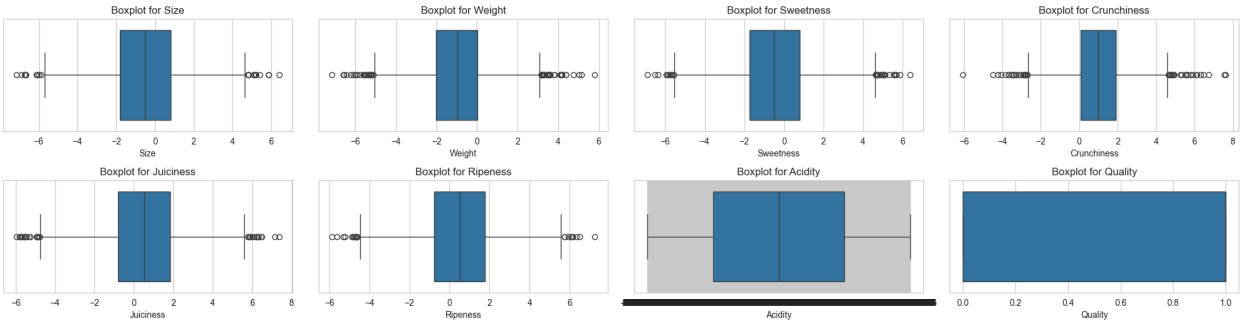
- **A_id:** Jedinstveni identifikator za svaki plod
- **Size:** Veličina ploda
- **Weight:** Težina ploda
- **Sweetness:** Nivo slatkoće ploda
- **Crunchiness:** Tekstura koja ukazuje na hrskavost ploda
- **Juiciness:** Nivo sočnosti ploda
- **Ripeness:** Stadijum zrelosti ploda
- **Acidity:** Nivo kiselosti ploda
- **Quality:** Ukupni kvalitet ploda

Nakon učitavanja podataka, sprovedena je početna analiza kako bi se identifikovale nedostajuće vrednosti, koje su potom uklonjene iz dataset-a.

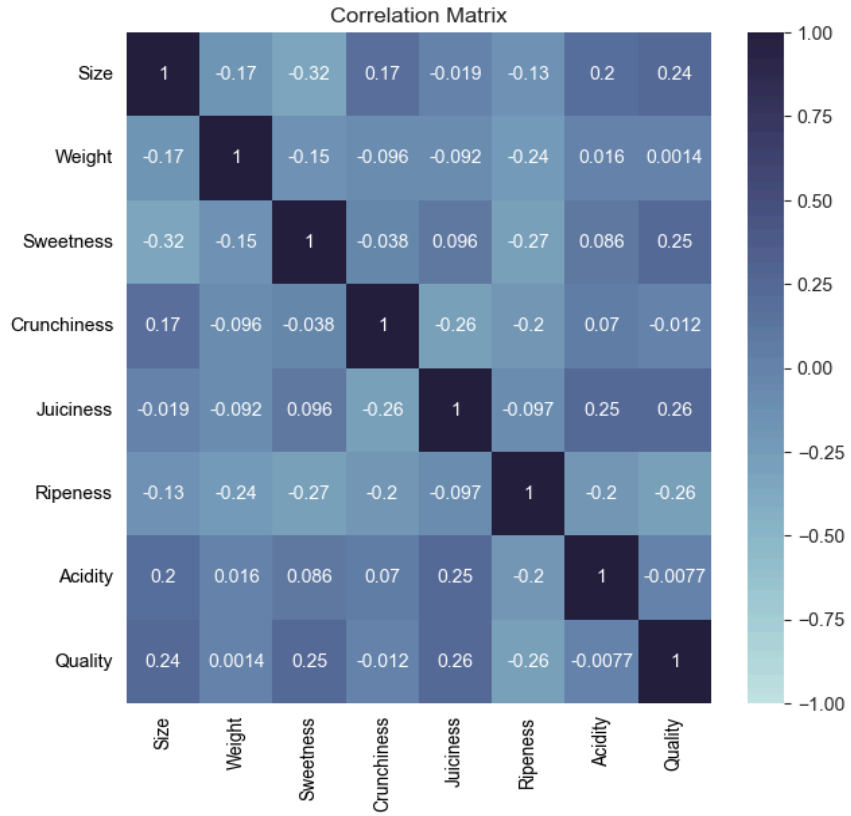
Kolona sa jedinstvenim identifikatorom svakog uzorka (*A_id*) nije bila relevantna za analizu, pa je izbačena. Kategorijske vrednosti ciljne promenljive su mapirane u numeričke vrednosti gde je "good" zamenjen brojem 1, a "bad" brojem 0. Pre treninga modela, svi podaci su normalizovani korišćenjem standardizacije kako bi se osigurala konzistentnost u obučavanju modela.

Vizuelna analiza dataset-a uključivala je prikaz distribucija svih atributa kroz histogram grafike, kao i analizu outliera pomoću boxplot prikaza.





Prikazana je i matrica korelacije koja otkriva međusobnu povezanost između različitih karakteristika voća. Na osnovu korelacije, atributi poput veličine i slatkoće pokazali su najveću povezanost sa ciljnim kvalitetom.



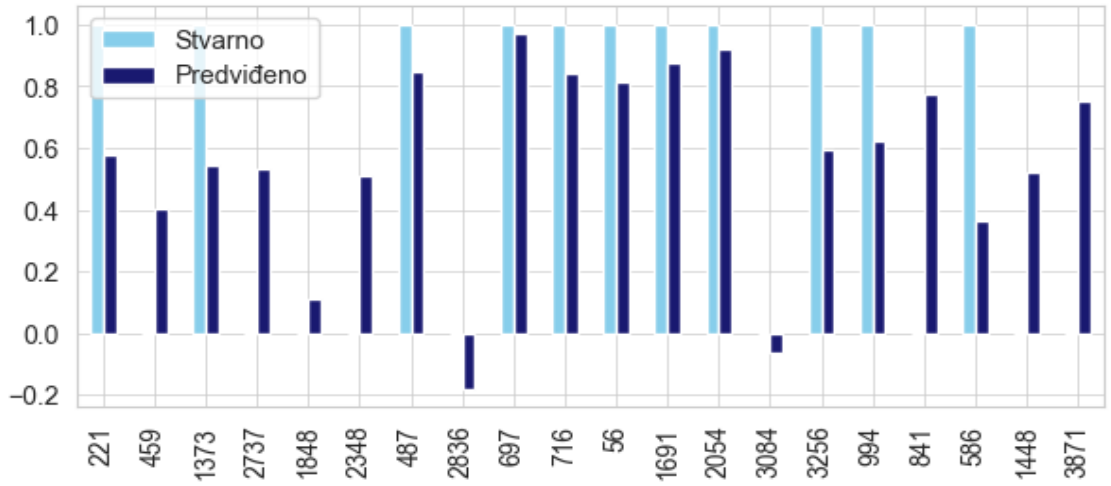
Nakon preprocesiranja, podaci su normalizovano koriscenjem `StandardScaler()`, i zatim podeljeni na trening i test skupove u odnosu 80:20.

Nakon toga se model trenira, ovo su rezultati jednacine:

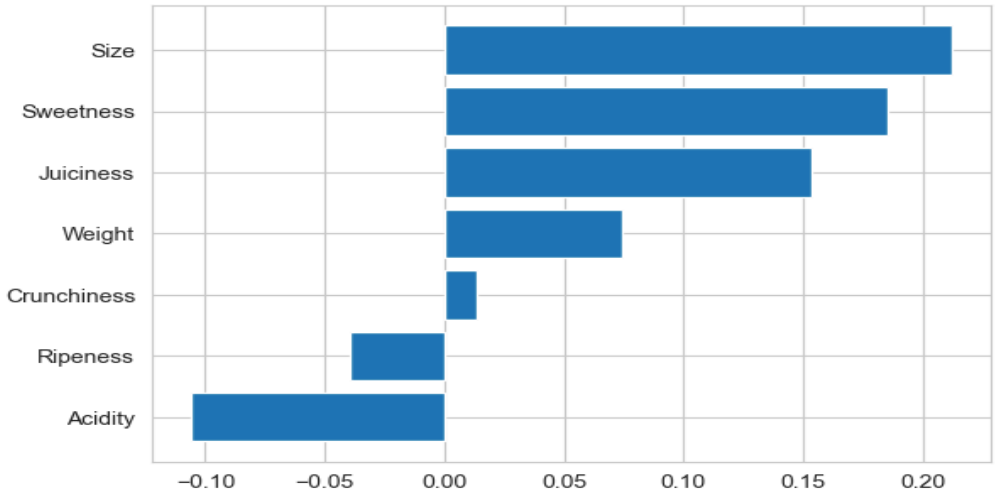
```
Slobodni član:  0.49736209021025507

Koeficijenti hiperravni:  [ 0.21254206  0.07408643  0.1854689   0.01372644
0.15348102 -0.03952068 -0.10634199]
```

	Coefficient
Size	0.212542
Weight	0.074086
Sweetness	0.185469
Crunchiness	0.013726
Juiciness	0.153481
Ripeness	-0.039521
Acidity	-0.106342

Rezultat modela nad prvih 20 redova:



Uticaj atributa na ciljnu promenljivu:



Evaluacija modela sprovedena je pomoću metričkih vrednosti kao što su srednja apsolutna greška, srednja kvadratna greška i koeficijent determinisanosti. Rezultat evaluacije:

```
Srednja apsolutna greška: 0.36041297703131364

Srednja kvadratna greška: 0.1725202578531845

Koren iz srednje kvadratne greške: 0.4153555800193185

Koeficijent determinisanosti: 0.309914655553859
```

Rezultati evaluacije modela linearne regresije ukazuju na sledeće:

- **Srednja apsolutna greška (MAE):** Vrednost od 0.36 znači da prosečna apsolutna razlika između stvarnih i predviđenih vrednosti iznosi 0.36. Ovo ukazuje na prosečan nivo greške bez obzira na njen smer.
- **Srednja kvadratna greška (MSE):** Vrednost od 0.17 predstavlja prosečnu kvadratnu razliku između stvarnih i predviđenih vrednosti. Veće kazne se dodeljuju većim greškama zbog kvadriranja.
- **Koren iz srednje kvadratne greške (RMSE):** Vrednost od 0.41 je u istim jedinicama kao originalni podaci i daje uvid u tipičnu veličinu greške. Manja vrednost ukazuje na bolju preciznost modela.
- **Koeficijent determinisanosti (R2R^2R2):** Vrednost od 0.31 znači da model objašnjava 31% varijanse u kvalitetu jabuka na osnovu korišćenih atributa. To sugeriše da je model donekle koristan, ali da postoji značajan prostor za poboljšanje, možda kroz dodavanje novih atributa ili korišćenje kompleksnijih algoritama.