

Klasifikacija zvezda po tipu

Dragoslav Tamindžija SV47/2021

Način pretprocesiranja podataka

Potrebno je da se boja zvezde i spektralna klasa enkoduju u numeričke vrednosti kako bi se podaci mogli obrađivati klasifikacionim algoritmom. Vrednosti temperature, sjaja, radijusa i apsolutne magnitude se standardizuju radi osiguranja jednakih skala između različitih karakteristika.

Programsko rešenje

Za rešavanja datog problema korišćena su četiri algoritma:

- **KNN** – algoritam daje najbolje rezultate kada je broj komšija jedan
- **RandomForestClassifier** – algoritam daje najbolje rezultate sa sto estimatora i maksimalnom dubinom pet
- **DecisionTreeClassifier** – algoritam daje najbolje rezultate kada se koristi kriterijum entropije, nasumično deljenje i maksimalna dubina pet
- **Gaussian Naive Bayes**

Analiza rezultata

Nakon treniranja i evaluacije dobijene su sledeće tačnosti:

- **KNN** – 95.8%
- **RandomForestClassifier** – 100%
- **DecisionTreeClassifier** – 93%
- **Gaussian Naive Bayes** – 97.2%

Ovako visoki procenti tačnosti ukazuju na to da modeli uspešno prepoznaju i klasifikuju tip zvezde. Podaci korišćeni za testiranje se razlikuju od podataka korišćenih za obuku, što osigurava da rezultati reflektuju realne performanse modela.

Unapređenje projekta

Povećanje veličine skupa podataka je ključno za poboljšanje tačnosti i generalizacije modela. Veći skup podataka pruža reprezentativnije uzorke prostora problema, što pomaže modelu da nauči osnovne obrasce umesto da pamti šum ili specifične instance, čime se smanjuje rizik od overfitovanja. Sa više podataka model može bolje da primeti varijabilnost i složenost stvarnih scenarija, što će dovesti do pouzdanijih predikcija.