Klasifikacija zvezda po tipu

Dragoslav Tamindžija SV47/2021

Definicija problema

Projekat se sastoji od implementacije klasifikatora za određivanje tipa zvezde na osnovu njenih spektralnih karakteristika.

Motivacija

Rešavanje problema kategorizacije zvezda na osnovu njihovih spektralnih karakteristika je ključno za različite astronomske studije, kao što su razumevanje zvezdanih populacija, formiranje galaksija i evolucija. U praksi, klasifikacija tipova zvezda se može koristiti u astronomskim istraživanjima, posebno prilikom analize obimnih podataka o nebeskim telima.

Skup podataka

Skup podataka se sastoji od 240 zvezda. Svaka zvezda ima atribute:

1. Temperatura – temperatura zvezde izražena u K
2. Sjaj – sjaj zvezde izražen u L/Lo
3. Radijus – radijus zvezde izražen u R/Ro
4. Apsolutna magnituda – mera koliko bi zvezda bila svetla ako bi se videla sa standardne udaljenosti izraženo u Mv
5. Tip zvezde
6. Boja zvezde
7. Spektralna klasa – jedna od mogućih klasa: O, B, A, F, G, K, M

Ciljno obeležje je tip zvezde i vrednost obeležja je broj 0-5 koji predstavlja jednu od mogućih tipova: crveni patuljak, braon patuljak, beli patuljak, glavni niz, superdžinovi, hiperdžinovi. Za svaku od navedenih tipova ima 40 zvezda u skupu podataka. Najznačajniji atributi, koji pokazuju najveću korelaciju sa ciljnim atributom, su sjaj, radijus i temperatura.

Skup podataka: https://www.kaggle.com/datasets/deepu1109/star-dataset/data

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.

Metodologija

Koraci u rešavanju problema:

1. Učitavanje podataka
2. Pretprocesiranje podataka
3. Fitovanje svih modela
4. Poređenje rezultata modela

Za rešavanja datog problema biće korišćena četiri algoritma: KNN, RandomForestClassifier, DecisionTreeClassifier i Gaussian Naive Bayes. Model koristi pretprocesirane podatke kao ulaz, pri čemu je ciljno obeležje uklonjeno. Izlaz modela predstavlja predviđenu klasifikaciju, na osnovu koje se može doneti zaključak o tačnosti.

Način evaluacije

Kao metrika evaluacije koristi se tačnost. Podaci se podele na dva skupa: skup za obuku i skup za testiranje. Za obuku se koristi 70% podataka, dok se preostalih 30% koristi za testiranje.

Tehnologije

Za implementaciju projekta je korišćen programski jezik Python i biblioteke pandas i scikit-learn.

Relevantna literatura

Primeri gotovih rešenja:

https://www.kaggle.com/code/shantanulekule/star-type-prediction

https://www.kaggle.com/code/cchen002/star-classification-boosting-accuracy-to-93-75