Rain in Australia

Ivana Nestorović

May 2023

Sadržaj

1 Uvod

U ovom radu bavimo se proučavanjem skupa podataka "Rain in Australia" koji sadrži razne meteorološke atribute, i na osnovu njih želimo da predvidimo da li će padati kiša ili ne. Koristićemo algoritme klasifikacije i klasterovanja, kao i programski jezik Python i neke njegove biblioteke, kao što je sklearn, koja nam omogućava da koristimo ove algoritme.

2 Skup podataka "Rain in Australia"

Ovo je veliki skup podataka koji sadrži 145460 instanci i 23 atributa. Rezultat je svakodnevnog prikupljanja podataka o vremenskim uslovima u periodu od 10 godina.

Atributi skupa podataka "Rain in Australia":

1. Date (Datum) - datum merenja vremeskih uslova
2. Location (Lokacija) - naziv lokacije gde su prikupljani podaci
3. MinTemp - Minimalna dnevna temperatura u °C
4. MaxTemp - Maximalna dnevna temperatura u °C
5. Rainfall - Ukupna količina padavina u toku 24h, u milimetrima
6. Evaporation - Količina isparavanja u toku 24h, u milimetrima
7. Sunshine - broj sunčanih sati u toku dana
8. WindGustDir - Smer najjačeg udara vetra u toku 24h
9. WindGustSpeed - Jačina najjačeg udara vetra, u kilometrima na čas
10. WindDir9am - Smer vetra u 9 časova ujutru
11. WindDir3pm - Smer vetra u 3 časa popodne
12. WindSpeed9am - Brzina vetra zabeležena u kilometrima na čas, u 9 ujutru
13. WindSpeed3pm - Brzina vetra zabeležena u kilometrima na čas, u 3 popodne
14. Humidity9am - Vlažnost vazduha u 9 ujutru, izražena u procentima
15. Humidity3pm- Vlažnost vazduha u 3 popodne, izražena u procentima
16. Pressure9am - Atmosferski pritisak u 9 časova ujutru izražen u hektopaskalima
17. Pressure3pm - Atmosferski pritisak u 3 časa popodne izražen u hektopaskalima
18. Cloud9am - Količina oblačnosti u 9 ujutru, izražena u oktama
19. Cloud3pm - Količina oblačnosti u 3 popodne, izražena u oktama

20. Temp9am - Temperatura u 9 ujutru, izražena u °C
21. Temp3pm - Temperatura u 3 popodne, izražena u °C
22. RainToday - Da li je bilo padavina tog dana (Da/Ne)
23. RainTomorrow - Da li će sutra biti padavina (Da/Ne)

Kategorički atributi su: 'Date', 'Location', 'WindGustDir', 'WindDir9am', 'WindDir3pm', 'RainToday', 'RainTomorrow.

Numerički atributi su: 'MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation', 'Sunshine', 'WindGustSpeed', 'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am', 'Temp3pm'.

3 Priprema podataka

3.1 Uklanjanje nedostajućih vrednosti

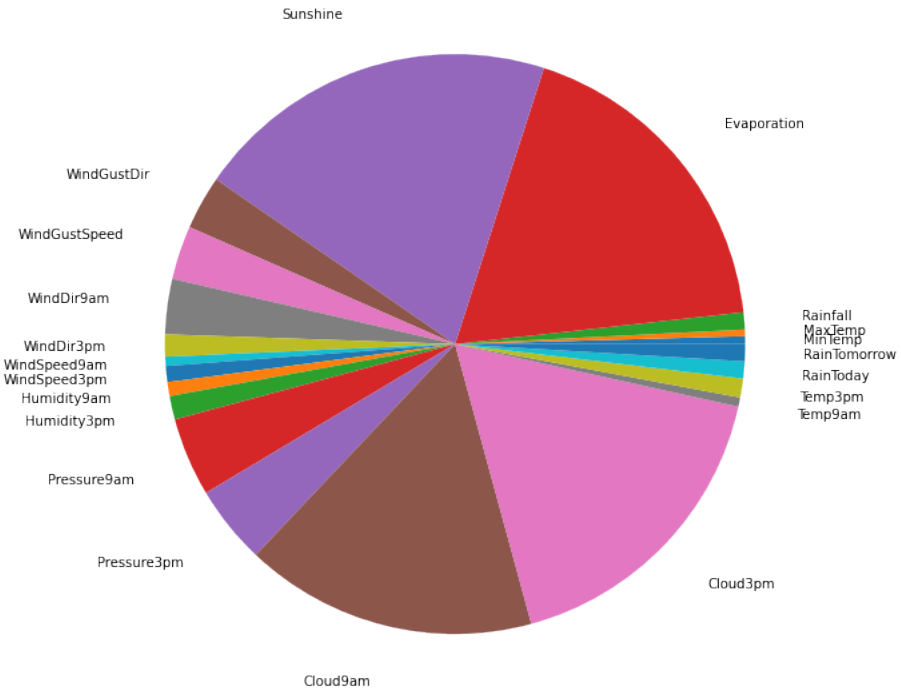


Figure 1: Nedostajuće vrednosti po atributima

Za ciljni atribut RainTomorrow imamo 3267 nedostajucih vrednosti, pa izbacujemo instance sa ovim vrednostima.
Ispod se nalazi heatmapa na kojoj su prikazane nedostajuće vrednosti po atributima, pre rada sa ostalim atributima.

Vidimo da najviše nedostajućih vrednosti ima u atributima Evaporation, Sunshine, Cloud9am and Cloud3pm.
Ove nedostajuće vrednosti zamenjujemo sa vrednostima koje ima instanca sa jednakim atributima sa najvećom korelacijom u odnosu na traženi atribut.
Na primer, atribut 'Sunhine' ima najveću korelaciju sa atributima 'Cloud3pm', 'Cloud9am', 'Humidity3pm'. Odnosno broj sunčanih sati će u velikoj meri zavisiti od odblačnosti i vlažnosti vazduha.
Nedostajuće vrednosti preostalih numeričkih atributa, dopunjavamo njihovim prosekom.
Za kategoričke atribute, njihove nedostajuće vrednosti zamenjene su najčešćim pojavljivanjima.
Npr. za lokaciju, mesto koje se najčešće pojavljuje je Canberra.

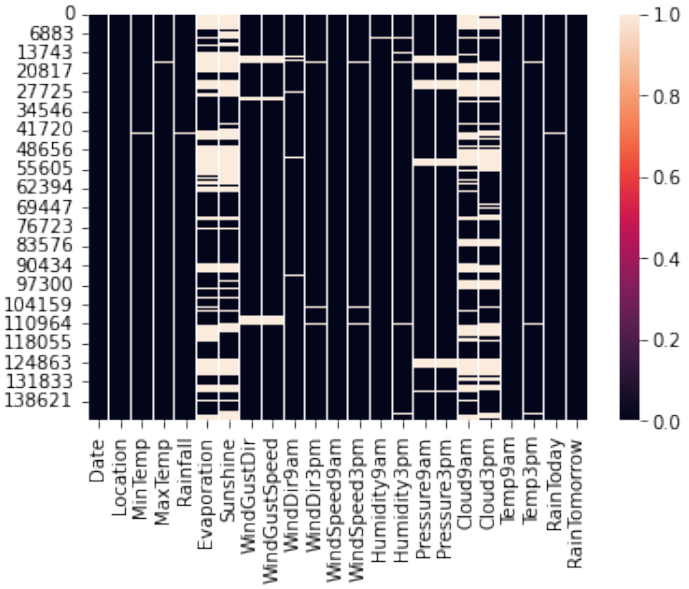


Figure 2: Heatmap

3.2 Enkodiranje

Kao što smo videli u analizi skupa podataka, imamo 7 atributa sa kategoričkim vrednostima.
Vrednosti Da i Ne, možemo zameniti sa 1 odnosno 0. Jedinica označava da je bilo padavina, a 0 da nije.
Atribut datum možemo razdvojiti na dan, mesec i godinu.
Strane sveta, su zamenjene rednim brojevima u smeru kazaljke na satu.
Mesta iz atributa 'Location' smo enkodirali u odnosu na to koja je verovatnoća da je atribut 'RainToday' za tu lokaciju jednak 1.

3.3 Feature Selection

Radi ubrzavanja i veće tačnosti algoritama, potrebno je da imamo manji broj atributa, ali moramo biti pažljivi pri odabiru istih, da ne bismo izgubili važne informacije.
U našem skupu podataka vremenske prilike se ponavljaju periodično, u periodu od jedne godine, što znači da nam atributi dan i godina nisu od velikog značaja.
Iz matrice korelacija, primećujemo da maksimalna temperatura ima veliku korelaciju sa temperaturom u 3 sata popodne, a da minimalna temperatura odgovara temperaturi u jutarnjim satima, pa i ove atribute izbacujemo.
Atributi koji označavaju atmosferski pritisak u 3 popodne i u 9 ujutru imaju slične vrednosti, zbog male dnevne promene pritiska, pa je potrebno zadržati samo jedan od njih.

4 Klasifikacija

U ovom poglavlju radićemo binarnu klasifikaciju, koja za cilj ima da predvidi da li će padati kiša ili ne. Koristićemo Decision Tree algoritam i K-nearest neighbors (KNN) algoritam.

4.1 Decision Tree

4.1.1 Priprema

Pre samog algoritma, potrebno je podeliti skup podataka na trening i test skup. Koristili smo stratifikovanu podelu, u odnosu 80:20.
Tek nakon podele vršimo izbacivanje autlajera. Detaljnom analzom podataka, možemo primetiti da nemamo šumove, odnosno greške koje izlaze iz skupa mogućih vrednosti.
Iscrtavanjem boxplot dijagrama, primećujemo da u atributima Evaporation i WindSpeed9am imamo ekstremne vrednosti. Instance koje imaju ekstremne vrednosti čine mali deo skupa podataka, pa možemo da ih izbacimo.
Za ciljni atribut RainTomorrow imamo mnogo više instanci iz klase No, nego iz klase Yes, pa moramo da koristimo balansiranje.
Ugradjena funkcija DecisionTreeClassifier podržava parametar Za ciljni atribut RainTomorrow imamo mnogo više instanci iz klase No, nego iz klase Yes, pa moramo da koristimo balansiranje. Ugradjena funkcija DecisionTreeClassifier podržava parametar class_weight = 'balanced' koji povećava težinu manje zastupljene klase.

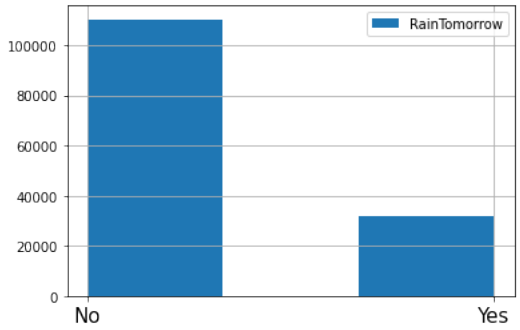


Figure 3: Histogram RainTomorrow

4.1.2 Bez podešavanja parametara

Kao početni model koristićemo stablo odlučivanja bez podešavanja parametara. Formiraće se stablo dubine 34. Vidimo da na trening skupu ima tačnost i preciznost 1, što nam govori da je možda došlo do preprilagodjavanja.

Na test skupu, za klasu koja je manje zastupljena, ima mnogo manju preciznost. Tačnost na trening skupu je 0.71.

4.1.3 Pošavanje parametara

Prosledjivanjem parametara za kriterijum Gini i Entropiju , niz = 3,7,11 za maksimalne dubine stabla, minimalan broj instanci za deljenje 2, 5, 10 i 1, 3, 5 za minimalan broj instanci za formiranje lista, GridSearh-u dobijamo da je najbolji klasifikator stablo sa parametrima Gini , maksimalnom dubinom 7,'min_samples_leaf': 5 i 'min_samples_split': 2.
I ovaj model ima malu preciznost za klasu Yes, ali ima dosta bolju tačnost u odnosu na prethodni model.

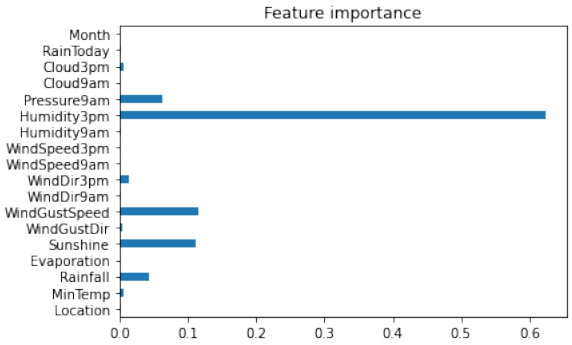


Figure 4: Feature importance

Sa grafika vidimo da je najznačajniji atribut za predvidjanje Humidity3pm, od ostalih značajnih atributa tu su: Pressure9am, Sunshine, WindGustSpeed, Sunshine i Rainfall.

4.1.4 Random Forrest model

Ensemble metoda koja kombinuje više stabala odlučivanja kako bi donela konačne predikcije.
Ako koristimo RandomForestClassifier iz biblioteke scikit-learn sa sledećim parametrima:
n_estimators=100 - što znači da će Random Forest model sadržati 100 stabala odlučivanja random_state = 42 - što znači da će model uvek davati iste rezultate za iste ulazne podatke
i class_weight='balanced', dobijamo model sa većom tačnošću i preciznošću nego za jedno pojedinačno stablo odlučivanja.

4.2 KNN algoritam

4.2.1 Priprema

Kao i za stablo odlučivanja, pre samog izvršavanja algoritma, potrebno je podatke podeliti na skup za trening i test skup. Nakon toga izbaciti nedostajuće vrednosti.
Atributi u našem skupu imaju različite opsege vrednosti, pa je potrebno izvršiti standardizaciju ili skaliranje radi postizanja boljih rezultata. Pomoću StandarScalera transformisali smo podatke, tako da svi atributi budu iz opsega od 0 do 1.

4.2.2 Bez podešavanja parametara

Treniranjem modela sa podrazumevanim parametrima i bez balansiranja podataka, na test skupu dobijamo tačnost 0.83, ali preciznost za podatke iz klase 1 je 0.67.
Balansiranje vršimo SMOTE i RandomUnderSampler tehnikom.

4.2.3 Podešavanje parametara

Pronalažanje najboljih paramatera prepuštamo unakrsnoj validaciji i GridSearhCV. Za najbolje parametre pronalazi metriku Menhetn, 20 susednih instanci i težine se dodeljuju na osnovu udaljenosti od suseda.

Na trening skupu imamo sa perfektnim performansama, što nam govori da je verovatno došlo do preprilagodjavanja.

4.3 Poredjenje rezultata

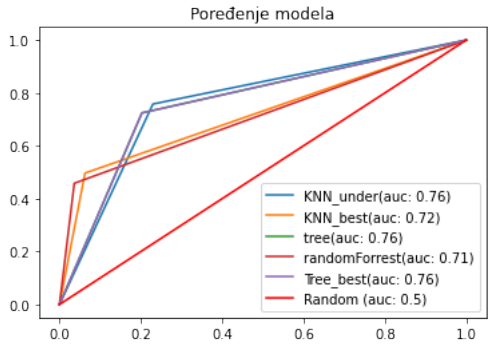


Figure 5: AUC

Posmatranjem površine ispod ROC krive, možemo da zaključimo da je najbolji klasifikator KNN algoritam sa balansiranjem RandomUnderSampler tehnikom kao i stablo sa podešenim parametrima.
Uporedićemo i ostale metrike:

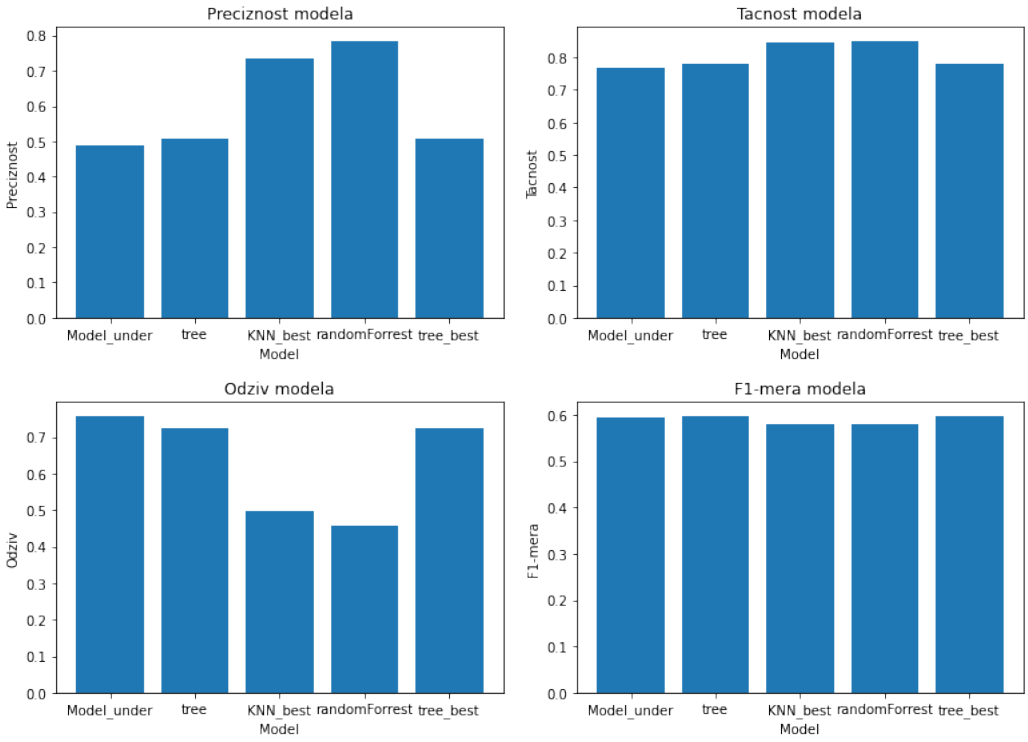


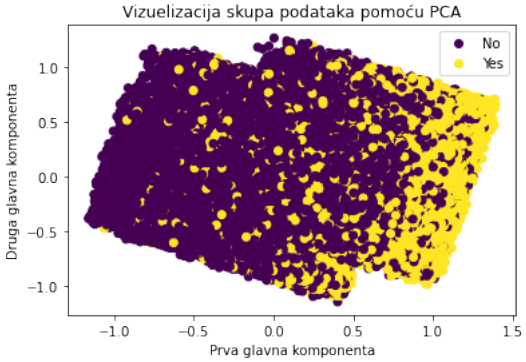
Figure 6: Metrike

Svi modeli imaju veliku tačnost, ali modeli RandomForrest i KNN sa GridSearh-om imaju veliku preciznost, a manji odziv, što znači da imamo minimizovan broj lažno pozitivnih rezultata, a veći broj lažno negativnih. Za naš konkretan problem to znači da ako model predvidi da će padati kiša, velika je verovatnoća da je to istina, dok ako predvidi da neće padati kiša, postoji mogućnost padavina. Model KNN sa balasiranjem i običan algoritam stabla odlučivanja, imaju suprotan problem.

5 Klasterovanje

Cilj klasterovanja je identifikovati prirodne strukture ili oblike u podacima, gde slični objekti pripadaju istom klasteru, dok se različiti objekti nalaze u različitim klasterima.

U našem skupu podataka imamo binarnu klasifikaciju, pa bi prirodno trebalo da postoje dva klastera.



Sa grafika možemo primetiti preklapanje klastera.

5.1 Kmeans

Radi na principu particionisanja podataka na K klastera, gde se K predstavlja unapred odredjenim brojem klastera koje želimo da formiramo. Algoritam nastoji da minimizuje sumu kvadratnih udaljenosti izmedju tačaka i centroida klastera.

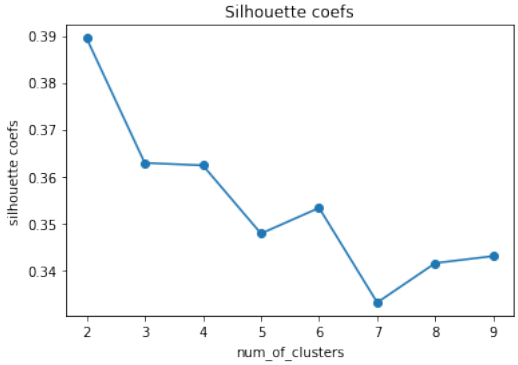


Figure 7: Caption

Poredjenjem koeficijenata siluete za broj klastera iz opsega od 2 do 9, vidimo da najveću vrednost ima za k = 2, pa nam je pretpostavka o prirodnim klaster-

ima tačna. Medjutim, koeficijent nije blizu jedinice, što nam govori da je došlo do preklapanja klastera.

5.2 Hijerarhijsko klasterovanje

Aglomerativno klasterovanje je jedan od popularnih algoritama za hijerarhijsko klasterovanje podataka. Radi na principu spajanja ili aglomeracije pojedinačnih instanci podataka kako bi se formirali klasteri.
Zbog velikog broja instanci, uzeli smo 20-procentni reprezentativan uzorak na kome ćemo uraditi klasterovanje.
AgglomerativeClustering kao parametre uzima broj klastera i način merenja udaljenosti izmedju klastera(linkages).
Za različite vrednosti broja klastera i drugačije načine merenja udaljenosti dobijamo različite rezultate.

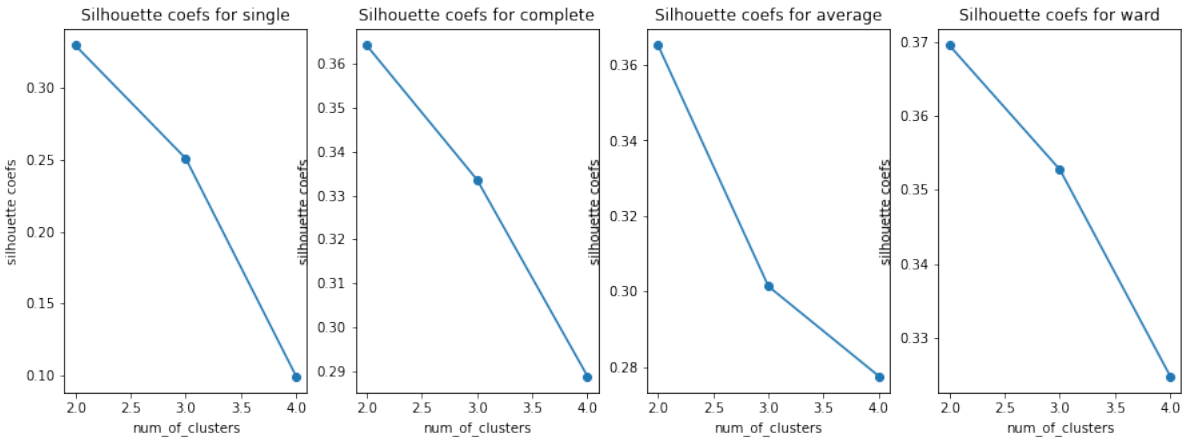


Figure 8: Caption

Za svaki način merenja udaljenosti dobijamo da je najbolja podela klastera na 2, kao što je i prirodno. Najmanji koeficijent siluete dobijamo za linkage = 'single', koji meri najmanje rastojanje izmedju tačaka klastera.