**Univerzitet u Beogradu**

Master akademske studije pri Univerzitetu

Računarstvo u društvenim nakama



**Analiza i vizualizacija podataka**

Primena algoritma klasifikacionog stabla odlučivanja (*classification tree*) radi predviđanja kiše na dnevnom nivou

|  |  |
| --- | --- |
| Student: Ana Fotev Nikolić 89/2019 | Profesor: dr Dragan Đurić |

Beograd, februar 2022.

Sadržaj

[Uvod 3](#_Toc95334316)

[Podaci 4](#_Toc95334317)

[Sređivanje podataka 5](#_Toc95334318)

[Eksploratorna analiza 7](#_Toc95334319)

[Izrada klasifikacion modela stabla odlučivanja (*classification tree*) 14](#_Toc95334320)

[Zaključak 24](#_Toc95334321)

[Literatura 26](#_Toc95334322)

[Paketi 26](#_Toc95334323)

## Uvod

Stablo odlučivanja (*decision tree*) je prediktivni model nadglednog mašinskog učenja (*supervised learning*) za klasifikaciju podataka. Sadrži koren koji nema ulaznih grana (*root node*), već iz koga se redom nižu unutrašnji čvorovi (internal nodes) koji imaju po jednu ulaznu i dve izlazne grane/ivice, kao i listove (leaf, terminal nodes) koji imaju po jednu ulaznu granu/ivicu i nijednu izlaznu. Svakom čvoru je dodeljen po jedan klasifikacioni kriterijum za određeni atribut observacija, na osnovu koga se odlučuje dalja putanja observacije: ukoliko observacija zadovoljava kriterijum, duž leve ivice “silazi” do narednog unutrašnjeg čvora gde se vrši dalja klasifikacija na osnovu novog kriterijuma, ili pak stigne do poslednjeg čvora odnosno lista. Ukoliko observacija ne zadovoljava kriterijum, njena putanja se od trenutnog čvora nastavlja duž desne ivice.[[1]](#footnote-1)

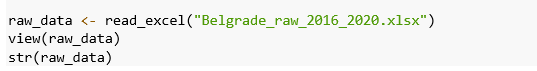
U zavisnosti od toga da li je cilj predviđanje diskretne ili kontinualne varijable, stablo odlučivanja može biti klasifikaciono (*classification tree*) ili regresivno (*regression tree*), respektivno. S obzirom da je u pitanju nadgledno učenje, model prvo uči na trening setu, a potom biva proveravan na test setu. Prilikom formiranja test i trenging setova, potrebno je sprovesti stratifikaciju po ispitivanom atributu, kako bi raspodela atributa bila ekvivalentna u oba seta. Prilikom treniranja modela potrebno je takođe definisati ograničenja kao što su minimalan broj obzervacija koje su potrebne da bi mogao da se “odvoji” zaseban čvor, parameter kompleksnosti, maksimalan broj čvorova, podela i nivoa. Faza provere, odnosno evaluacije treniranog seta podrazumeva proveru metrika izvedenih na osnovu matrije konfuzije (*confusion matrix*) poput *accuracy*, *precission*, *recall* i *F1*. O svemu će biti više reči u nastavku rada, prilikom prikazivanja procesa i rezultata.

Prednosti klasifikacionog stabla odlučivianja su to što je relativno lako za razumevanje i objašnjavanje, može da uoči nelinearne veze, ne zahteva normalizaciju ili standardizaciju numeričkih podataka i nije potrebno kreiranje *dummy* varijabli. Takođe, nedostajuće vrednosti ne predstavljaju problem i modeli su otporni na outlajere.[[2]](#footnote-2) Sa druge strane nedostaci ovakvih modela su “overfitting” podataka koji se može javiti kod kompletno razgranatih klasifikacionih stabala je, kao i slabija tačnost predviđanja u poređenju sa drugim pristupima mašinskog učenja.[[3]](#footnote-3)

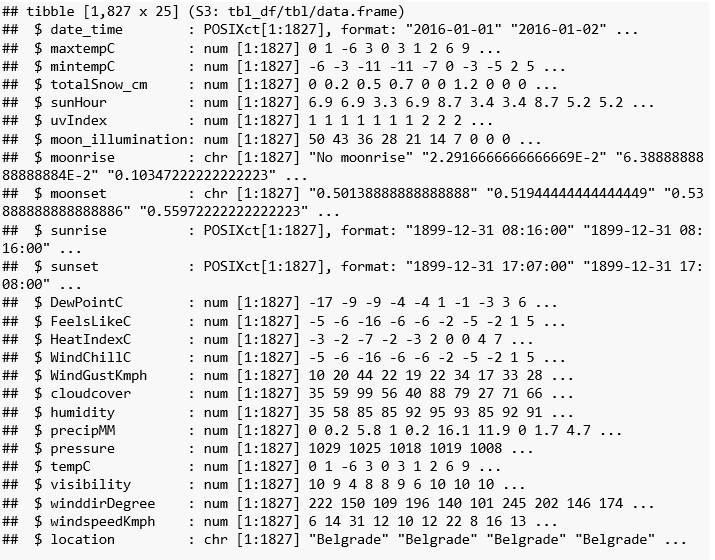
Klasifikaciona stabla odlučivanja se koriste u najrazličitijim poljima – medicine, bankarstvu, prepoznavanju, filtriranje spam e-mailova i tako dalje.[[4]](#footnote-4) Cilj ovog rada je naći najbolji algoritam stabla odlučivanja za predviđanje da li će određenog dana padati kiša spram ostalih vremenskih varijabli.

## Podaci

Podaci o vremenskim prilikama u Beogradu za period 01.01.2016-31.12.2020. su skrejpovani sa sajta <https://www.worldweatheronline.com> pomoću besplatne verzije API-ja i Phytong skripte (Jupyter Notebook). Sirova baza (raw\_data) sadrži 1827 observacija, gde je svaka observacija dan u godini, i 25 varjabli. Varijable obuhvataju datum, minimalnu i maksimalnu temperature izraženu u C°, snežne padavine izražene u cm, broj sunčanih sati, varijable o izlasku i zalasku meseca i sunca, subjektivni osećaj temperature, varijable vezane za vertovitost, oblačnost, vlažnost, padavinu kiše, pritisak i vetar.



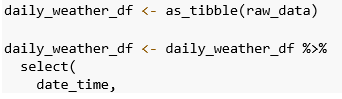
Slika : Učitavanje podataka



Slika : Struktura podataka

## Sređivanje podataka

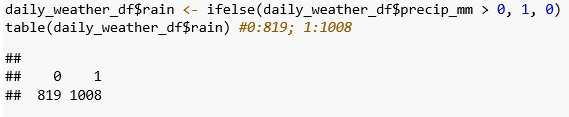
Prvi korak u sređivanju podataka je selekcija potrebnih varijabli i preimenovanje određenih varijabli radi preglednosti.



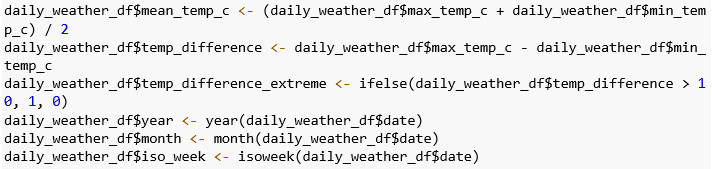


Slika : Selekcija i promena naziva varijabli

Kako ćemo pre izgradnje klasifikacionog modela uraditi i eksplorativnu deskriptivnu analizu radi boljeg upoznavanja sa podacima, sledeći korak u sređivanju podataka je konstruisanje zavisne varijable kiša (rain) kao binanrne (0/1), a koja će kasnije biti transformisana u kategorijsku, a za potrebe klasifikacionog modela. Pored ovoga, kreirano je i više nezavisnih varijabli: prosečna dnevna temperature (mean\_temp\_c), razlika između maksimalne i minimalne dnevne temperature (temp\_difference), ekstremna razlika u temperaturi (temp\_difference\_extreme), godina (year), mesec (month), nedelja prema ISO standard (iso\_week). U ovoj fazi izvršena je i selekcija varijabli koje će biti podobnije istražene u eksploratornoj analizi radi upoznavanja sa podacima, a pre treniranja modela.

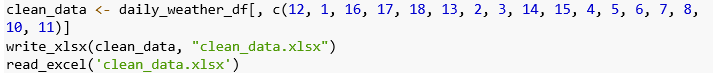


Slika : Konstruisanje zavisne varijable 'rain'



Slika : Konstruisanje novih nezavisnih varijabli

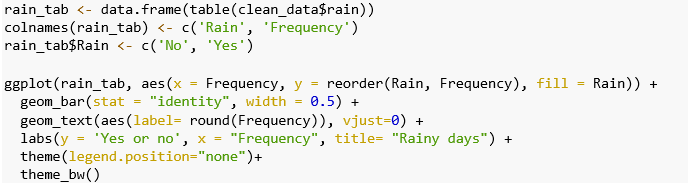
Nakon sređivanja podataka, novi set je sačuvan kao eksel fajl “clean\_data” koji je u prilogu ovog rada.



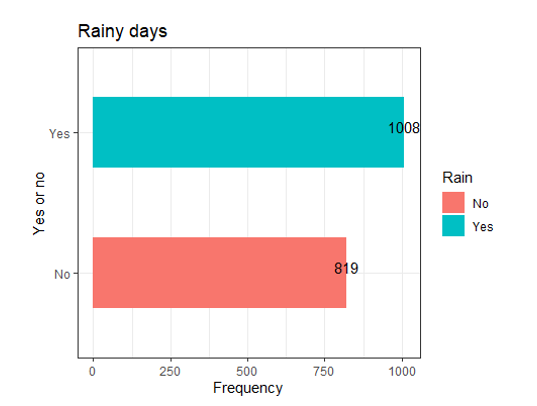
Slika : Eksportovanje sređene baze podataka

## Eksploratorna analiza

Pre samog treniranja modela, urađena je eksploratorna deskriptivna analiza radi boljeg upoznavanja sa podacima. Prvo je ispitan odnos kišnih i suvih dana za celokupan period što je pokazalo da Beograd ima više kišnih dana od suvih. Vizualizovan je štapićasti dijagram.

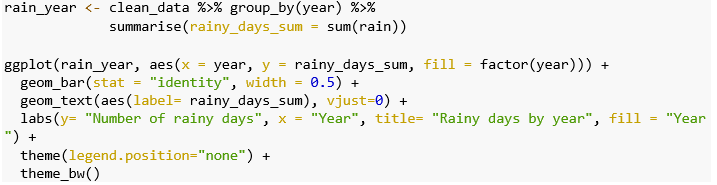


Slika : Kod za kreiranje štapićastog dijagrama

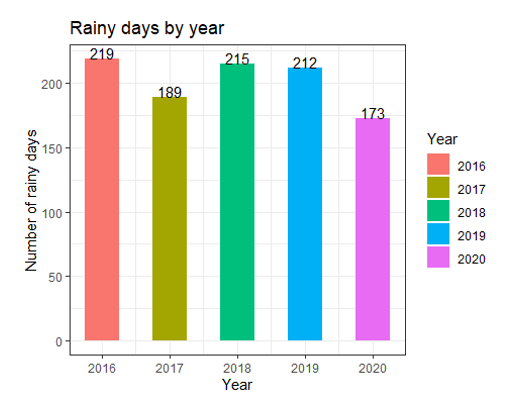


Slika : Štapićasti dijagram: kišni i suvi dani za period 01.01.2016-31.12.2020.

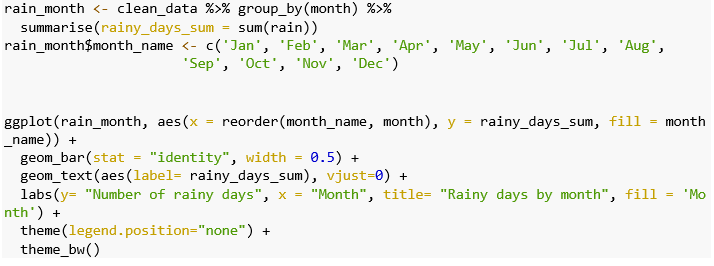
Sledeći korak je bio ispitivanja raspodele kišnih dana po godinama i po meseicma, respektivno, a potom takođe vizualizovana pomoću štapićastih dijagrama.



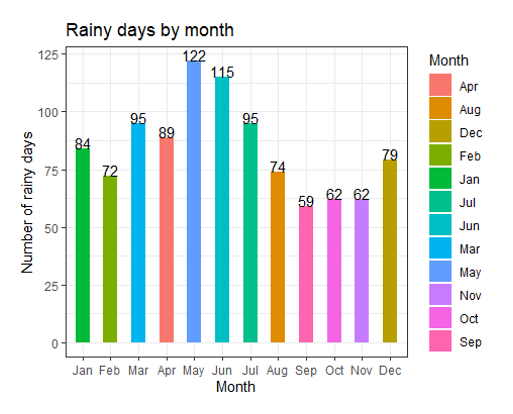
Slika : Kod za izradu štapićastog dijagrama - kišni dani po godinama



Slika : Štapićasti dijagram kišnih dana po godinama

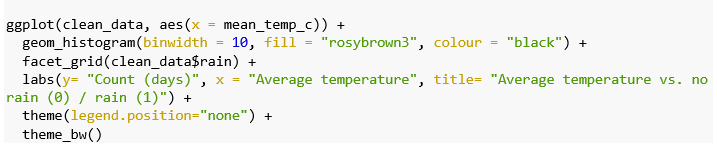


Slika : Kod za izradu štapišastog dijagrama kišnih dana po mesecima

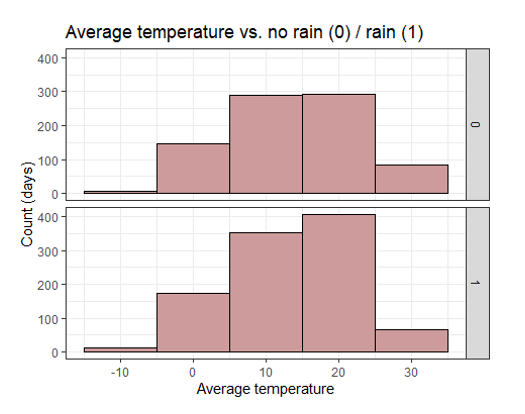


Slika : Štapićasti dijagram kišnih dana po mesecima

Prilikom ispitivanja odnosa kontinualnih varijabli (prosečne temperature, oblačnosti, vlažnosti i pritiska) korišćen je histogram.



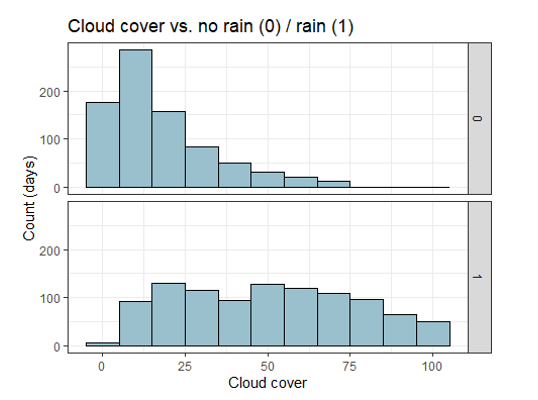
Slika : Kod za histogram prosecne temperature tokom kisnih i suvih dana



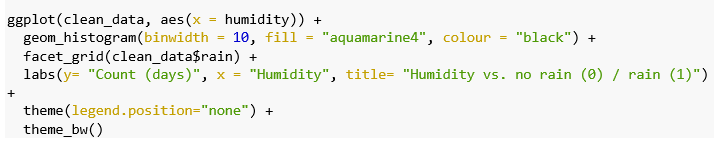
Slika : Histogram prosečne temperature tokom kišnih i suvih dana



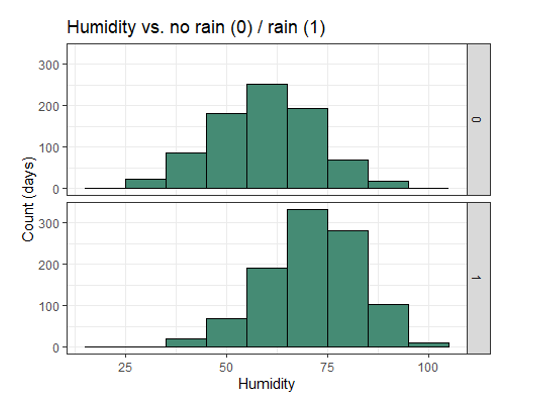
Slika : Kod za histogram oblačnosti tokom kišnih i suvih dana



Slika : Histogram oblačnosti tokom kišnih i suvih dana



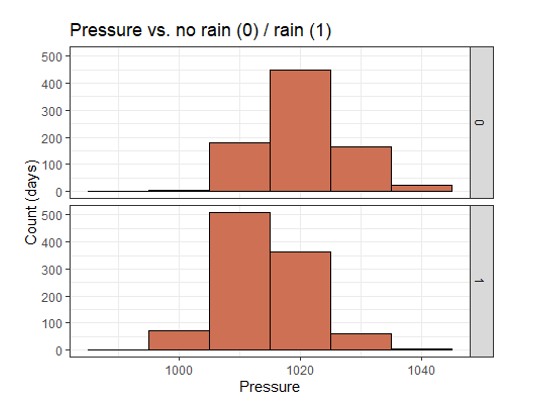
Slika : Kod za histogram vlažnosti tokom kišnih i suvih dana



Slika : Histogram vlažnosti tokom kišnih i suvih dana



Slika : Kod za histogram vazdušnog pritiska tokom kišnih i suvih dana



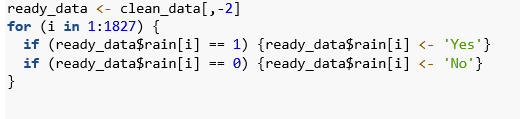
Slika : Histogram vazdušnog pritiska tokom kišnih i suvih dana

Eksploratornom analizom ustanovljeno je da ne postoji posebna razlika u raspodeli prosečne temperature zavisno od toga da li je u toku dana bilo kiše ili ne, dok kod nivoa vlažnosti i pritiska postoji izvensa korelacija, a najveća veza je svakako uočena između oblačnosti i padavine kiše.

## 

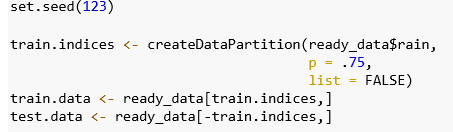
## Izrada klasifikacion modela stabla odlučivanja (*classification tree*)

Pre treniranja modela, izbačena je varijabla datuma kao varijable koja je suvišna za model. Pored toga bilo je neophodno izvršiti transformaciju binarne “rain” u kategorijsku varijablu sa dva nivoa (Yes/No) s obzirom na to da ćemo trenirato klasifikacioni model stabla odlučivanja (classification tree). Istovremeno, konstruisana je i nova dejtafrejm struktura “ready\_data”.

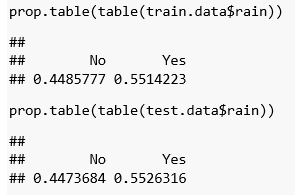


Slika : Izbacivanje varijable datum i transformisanje varijable 'rain' iz binarne u kategorijsku

Pre samog konsturisanja i treniranja modela, ostalo je još podeliti bazu podataka na trening i test set stratifikacijom podataka kako bi odnos kišnih dana i dana bez kiše bude približno jednak u oba seta. Određeno je da od ukupe baze (ready\_data) 75% baze čini trening set, 25% podataka test set.

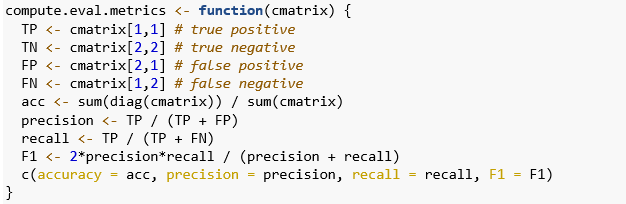


Slika : Podela podataka na trening i test setove



Slika : Provera da li je zastupljenost klase jednaka u trening i test setu

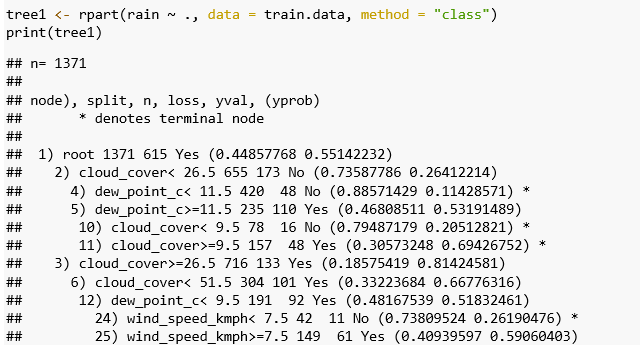
Poslednji korak pre trenitranja modela je pravljenjje matrice konfuzije kao alata za evaluaciju performansa binarnih klasifikacionih modela.

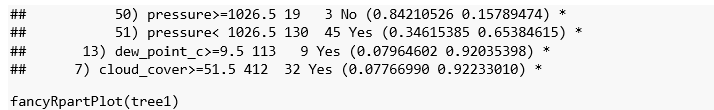


Slika : Kod za funkciju 'comput.eval.matrix' koja obračunava metrike za evaluaciju modela

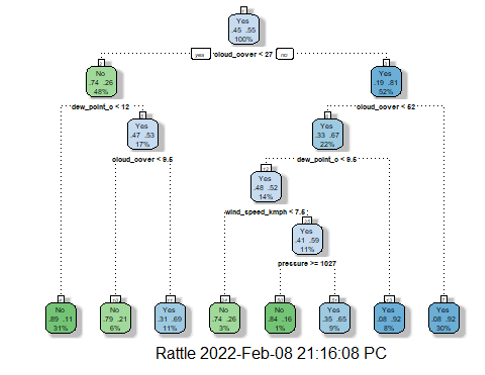
Pored same matrice, funkcija obračunava i sledeće metrike: *accuracy*, odnosno odnos tačnih predviđanja i svih predviđanja, potom *precision* – odnos tačno predviđenih pozitivnih ishoda i svih predviđenih pozitivnih ishoda, *recall* – odnos tačno predviđenih pozitivnih ishoda i stvarno pozitivnih ishoda, kao i F1 odnosno ponderisana prosečna vrednost metrika *precision* i *recall*.

Prvo je izrađeno Stablo 1 (tree1) sa standardnim parametrima kontrole (minsplit = 20, cp = 0.01)



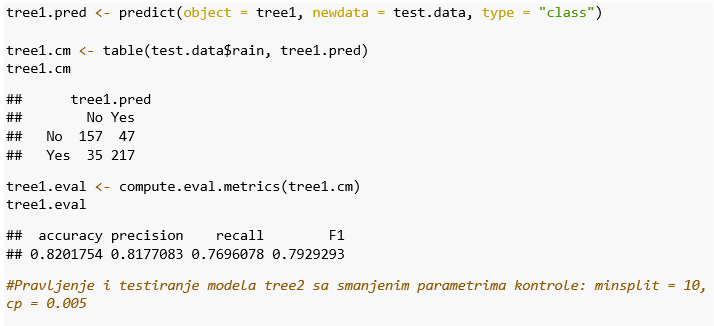


Slika : Kod za izradu modela stabla odlučivanja tree1



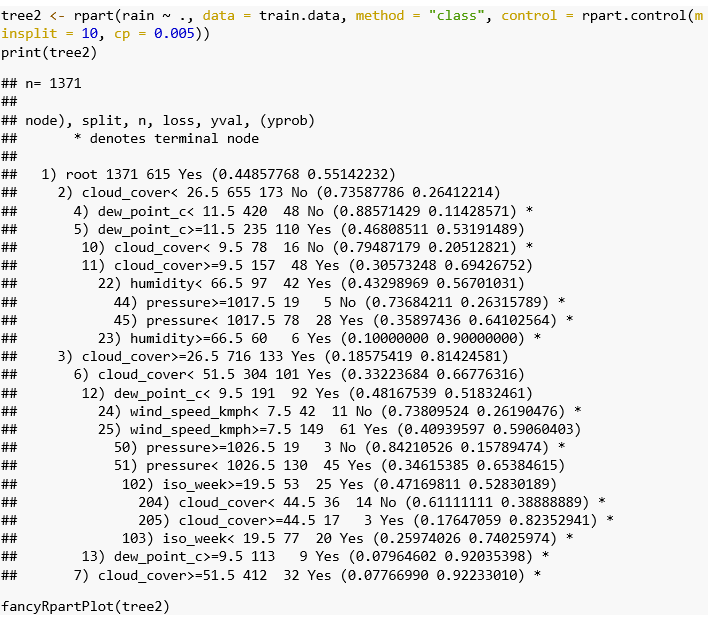
Slika : Stablo odlučivanja Tree1

Potom je stablo odlučivanja testirano na test setu i evalurano korišćenjem matrije konfuzije i izvedenih metrika.

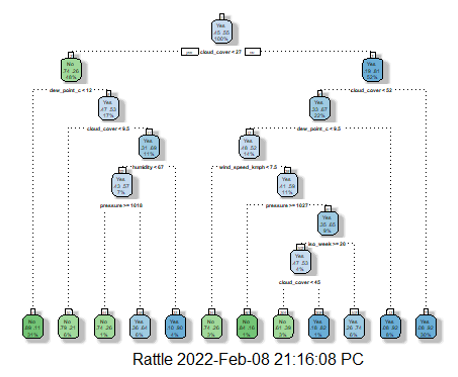


Slika : Testiranje i evaluiranje tree1

Nakon izrade Tree1 sa defoltnim parametrima kontrole, izrađeno je i Staloblo 2 (tree2) sa smanjenim parametrima kontrole (minsplit = 10, cp = 0.005).

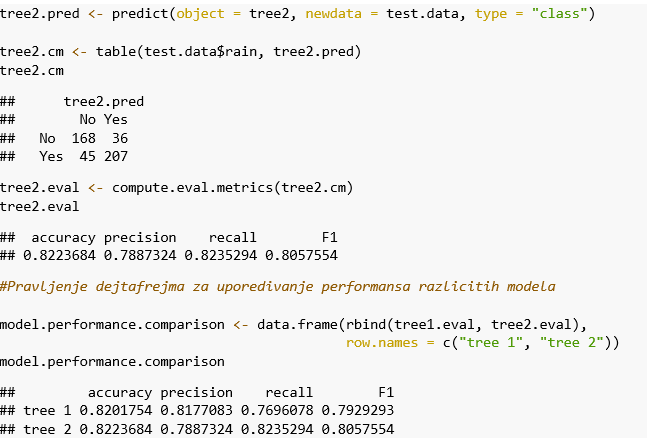


Slika : Kod za izradu modela stabla odlučivanja tree2



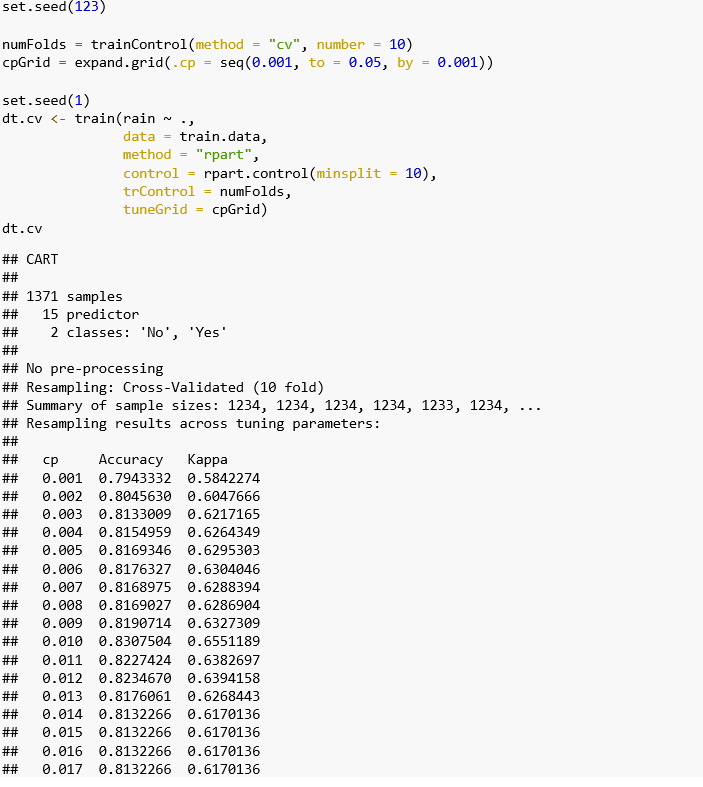
Slika : Stablo odlučivanja Tree2

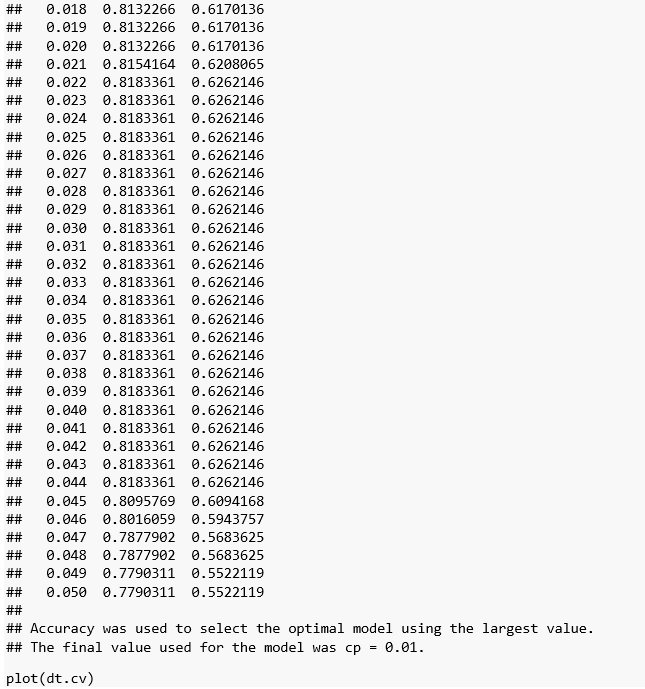
Potpm je i Stablo 2 (tree2) testirano i evaluirano korišćenjem matrice konfuzije i izvedenih metrika.



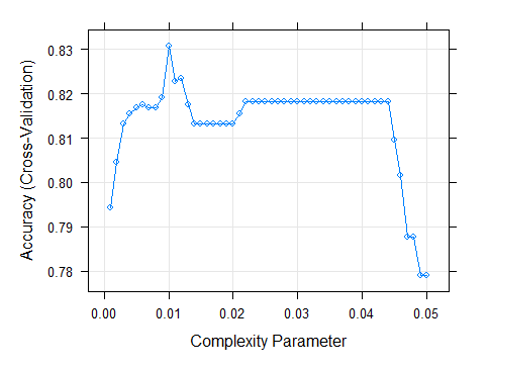
Slika : Testiranje i evaluacija Tree2

Nakon izrade dva modela stabla odlučivanja – sa default parametrima kontrole i sa sniženim parametrima kontrole, kako ne bismo nagađali u nedogled u potrazi za najoptimalnijim parametrom složenosti (*complexity parameter* – cp), sproveli smo proces sistemskog traženja parametra cp najoptimalnijeg za preciznost predviđanja modela (*cross-validation*).



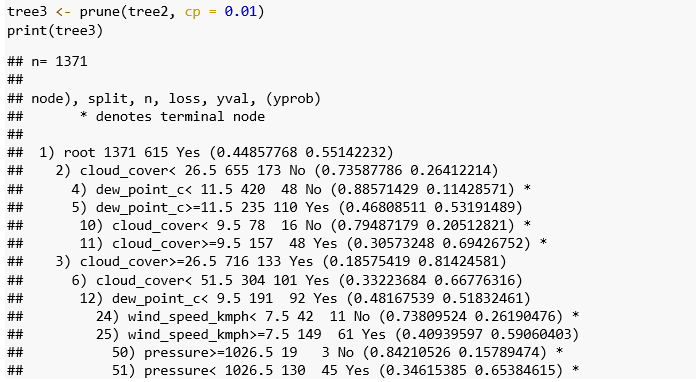


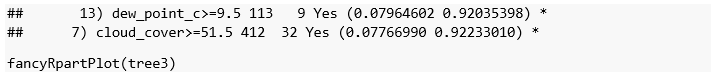
Slika : Sistemsko traženje optimalnog complexity parametra



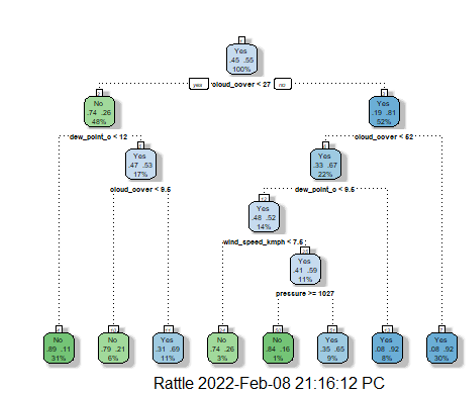
Slika : Rezultat sistemskog pretraživanja optimalnog complexity parametra

Potom je izvrešno i pravljenje i testiranje modela Stabla 3 (tree3) sa optimalnom vrednoću cp=0.01, ali primenom orezivanja modela stabla 2 (tree2) pozivanjem funkcije “prune”.



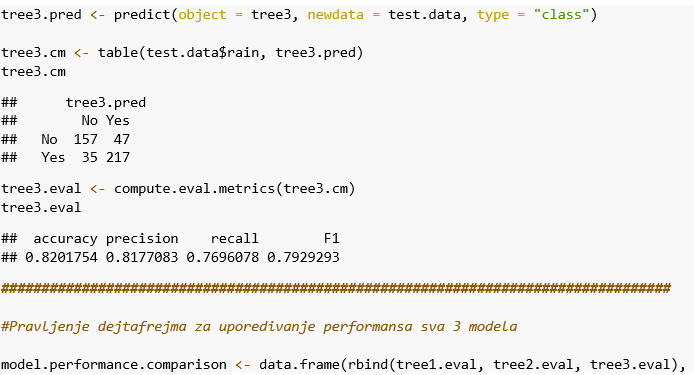


Slika : Kod za kostruisanje Tree3



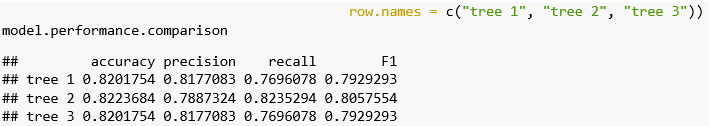
Slika : Vizualizacija Tree3

Nakon ovoga je i tree3 testirano i evaluirano korišćenjem matrice konfuzije i izvedenih metrika.

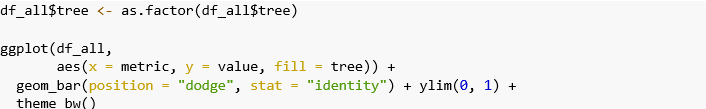


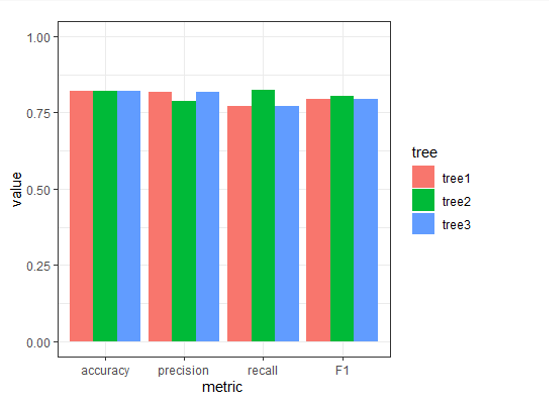
## Zaključak

Da bismo zaključili koji je model najuspešniji u predviđanju kišnih dana, performanse sva tri modela je upoređena i urađena je vizualizacija. Kako je sistemskim izračunavanjem optimalnog complexity parametra dobijena uobičajena vrednost cp=0.01, iako je drugačijim pristupom traniran model tree3, tree1 i tree3 su zapravo isti modeli – odnoso optimalni modeli za predviđanje kiše.









Slika : Poređenje performansa 3 modela

Poređenje performanska 3 modela govori sva tri modela predviđaju sa približno istom tačnošću (accuracy), mada je accuracy modela tree2 malo veća (oko 82,24% spram 82,01 posto modela 1 i 3).

Sa druge strance, ukoliko gledamo samo predviđaje pozitivnih ishoda (odnosno uspešno predviđanje dana kada pada kiša) model tree1 i tree3 su ostvarili bolji rezultat spram modela dva (precission od tree1 i tree3 je oko 81,77% dok je od tree 2 oko 78,87%). Metrika *recall* koja pokazuje odnos tačno predviđenih pozitivnih ishoda i stvarno pozitivnih ishoda ukazuje da je model tree2 u ovom aspektu najuspešniji, kao i prlikom poređenja metrike F1.

## Literatura

1. Man, Prem S. (2009) Uvod u statistiku, Centar za izdavačku delatnost Ekonomskog fakulteta u Beogradu, Beograd
2. Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016
3. <https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96> pristupljeno 30.01.2022.
4. <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052> pristupljeno 28.01.2022.
5. <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052> pristupljeno 24.02.2022.
6. [https://app.datacamp.com/learn/courses/machine-learning-with-tree-based-models-in-r](https://app.datacamp.com/learn/courses/machine-learning-with-tree-based-models-in-r%20%20)  pristupljeno 22. i 24.01.2022.
7. [https://youtu.be/ZVR2Way4nwQ](https://youtu.be/ZVR2Way4nwQ%20) pristupljeno 21.01.2022.
8. <https://youtu.be/tU3Adlru1Ng> 21.01.2022.
9. <https://youtu.be/uXIIk7suD6c> 21.01.2022.
10. <https://rti.etf.bg.ac.rs/rti/ms1psz/predavanja/StablaOdlucivanja.pdf> pristupljeno 1.02.2022.

## Paketi

1. Tidyverse: <https://CRAN.R-project.org/package=tidyverse>
2. Caret: <https://cran.r-project.org/package=caret>
3. ggplot2: <https://cran.r-project.org/package=ggplot2>
4. broom: <https://CRAN.R-project.org/package=broom>
5. dplyr: <https://cran.r-project.org/package=dplyr>
6. rpart: <https://cran.r-project.org/web/packages/rpart/index.html>
7. rpart.plot: <https://cran.r-project.org/package=rpart.plot>
8. rattle: <https://cran.r-project.org/package=rattle>
9. RColorBrewer: <https://cran.r-project.org/package=RColorBrewer>
10. ISLR: <https://cran.r-project.org/package=ISLR>
11. readr: <https://cran.r-project.org/package=readr>
12. e1071: <https://cran.r-project.org/package=e1071>
13. ISOweek: <https://cran.r-project.org/package=ISOweek>
14. tibble: <https://cran.r-project.org/package=tibble>
15. readxl: <https://cran.r-project.org/package=readxl>
16. writexl: <https://cran.r-project.org/package=writexl>
17. lubridate: <https://cran.r-project.org/package=lubridate>
18. corrplot: <https://cran.r-project.org/package=corrplot>

1. Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016. str. 163. [↑](#footnote-ref-1)
2. [https://app.datacamp.com/learn/courses/machine-learning-with-tree-based-models-in-r](https://app.datacamp.com/learn/courses/machine-learning-with-tree-based-models-in-r%20%20)  pristupljeno 22. i 24.01.2022. [↑](#footnote-ref-2)
3. [https://rti.etf.bg.ac.rs/rti/ms1psz/predavanja/StablaOdlucivanja.pdf](https://rti.etf.bg.ac.rs/rti/ms1psz/predavanja/StablaOdlucivanja.pdf%20) pristupljeno 01.02.2022. [↑](#footnote-ref-3)
4. [https://rti.etf.bg.ac.rs/rti/ms1psz/predavanja/StablaOdlucivanja.pdf](https://rti.etf.bg.ac.rs/rti/ms1psz/predavanja/StablaOdlucivanja.pdf%20) pristupljeno 01.02.2022. [↑](#footnote-ref-4)