Analiza

"Women's Tennis Association Matches"

WTA matches from 2000-2016

skupa podataka

Maja Gavrilović 489/2017

Septembar 2019

**Sažetak:**

U ovom radu predstavljamo rezultate dobijene analizom teniskih mečeva na turnirima u periodu od 2000 do 2016 godine. Prvo ćemo se upoznati sa opisom skupa podataka, dok kasnije stavljamo akcenat i bavimo se predikcijom mečeva koja omogućava korisniku da za dva tenisera dobije ishod meča. Jedan od izazova je bio postupak pretprocesiranja o kome ćemo detaljnije u nastavku...

**Sadržaj:**

Sažetak....................................................................................................2 Opis i vizualizacija skupa podataka.........................................................3 Opis i vizualizacija skupa podataka.........................................................4 Opis i vizualizacija skupa podataka.........................................................5 Opis i vizualizacija skupa podataka.........................................................6 Opis i vizualizacijaskupa podataka..........................................................7 Korišćeni alati .........................................................................................7 Pretprocesiranje......................................................................................7 Pretprocesiranje......................................................................................8Klasifikacija..............................................................................................8 Klasifikacija..............................................................................................9 Klasifikacija.............................................................................................10 Klasifikacija.............................................................................................11Zaključak................................................................................................11

Opis i vizualizacija skupa podataka

“Women’s tennis association matches” je skup podataka koji sadrži rezultate teniskih mečeva u periodu 2000 do 2016 godine. U istraživanje je uključeno i sakupljeno 2876 instanci. Ovaj skup koristićemo za razvoj prediktivnog modeliranja teniskih mečeva i za statističko istraživanje. Skup se sastoji od sledećih fajlova:

• wta\_matches\_2000.csv: schema koja uključuje podatke iz 2000 godine.

• wta\_matches\_2001.csv: schema koja uključuje podatke iz 2001 godine.

• wta\_matches\_2002.csv: schema koja uključuje podatke iz 2002 godine.

• wta\_matches\_2003.csv: schema koja uključuje podatke iz 2003 godine.

• wta\_matches\_2004.csv: schema koja uključuje podatke iz 2004 godine.

• wta\_matches\_2005.csv: schema koja uključuje podatke iz 2005 godine.

• wta\_matches\_2006.csv: schema koja uključuje podatke iz 2006 godine.

• wta\_matches\_2007.csv: schema koja uključuje podatke iz 2007 godine.

• wta\_matches\_2008.csv: schema koja uključuje podatke iz 2008 godine.

• wta\_matches\_2009.csv: schema koja uključuje podatke iz 2009 godine.

• wta\_matches\_2010.csv: schema koja uključuje podatke iz 2010 godine.

• wta\_matches\_2011.csv: schema koja uključuje podatke iz 2011 godine.

• wta\_matches\_2012.csv: schema koja uključuje podatke iz 2012 godine.

• wta\_matches\_2013.csv: schema koja uključuje podatke iz 2013 godine.

• wta\_matches\_2014.csv: schema koja uključuje podatke iz 2014 godine.

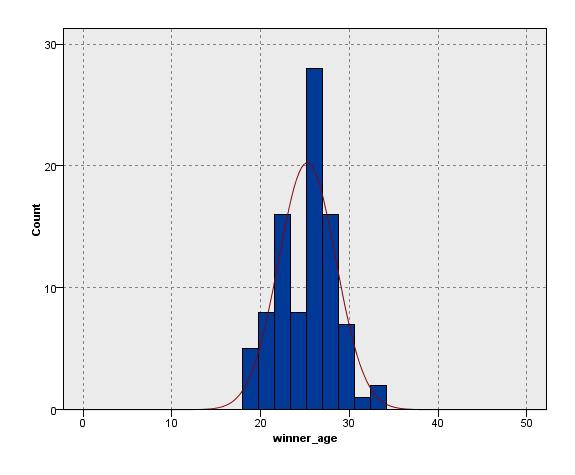
• wta\_matches\_2015.csv: schema koja uključuje podatke iz 2015 godine.

• wta\_matches\_2016.csv: schema koja uključuje podatke iz 2016 godine.

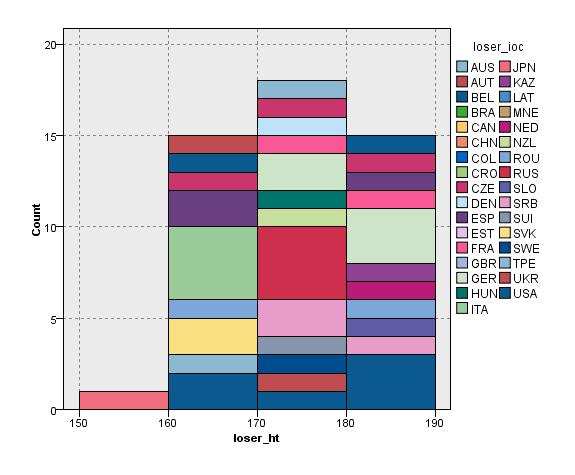
Neki od glavnih korišćenih atributa i njihove vizualizacije će biti prikazane u nastavku:

Shema 2016:

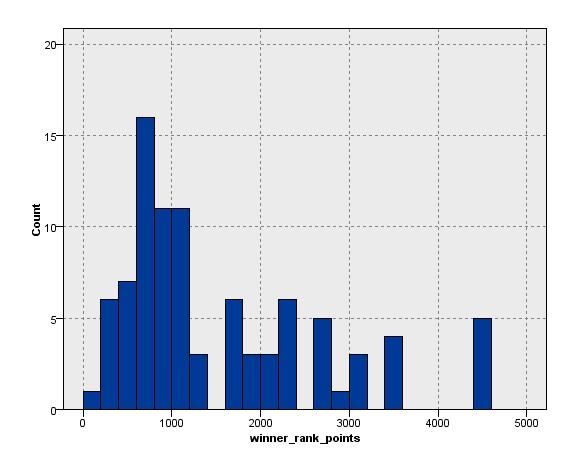
• winner\_age – starost teniserki koje su odnele pobedu, u odnosu na normalnu raspodelu



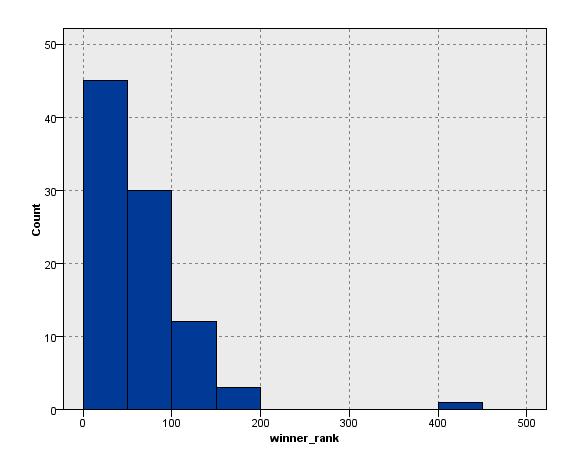
• looser\_ht-visina teniserki koje su dozivele poraz prikazane po zemlji porekla



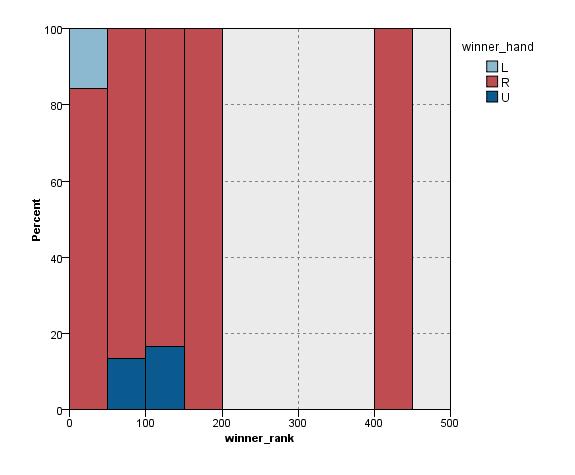
• winner\_rank\_points-rangiranje na osnovu broja poena u osvojenim mečevima, na skali do 5000



• winner\_rank-rangiranje teniserki koje su odnele pobedu



• Isti histogram prikazan u procentima kojom rukom teniserke igraju



Korišćeni alati

Za obradu tj. klasifikovanje i vizualizaciju korišćen je IBM SPSS, dok je za pretprocesiranje korišćen jezik Python i biblioteke numpy I pandas.

Pretprocesiranje

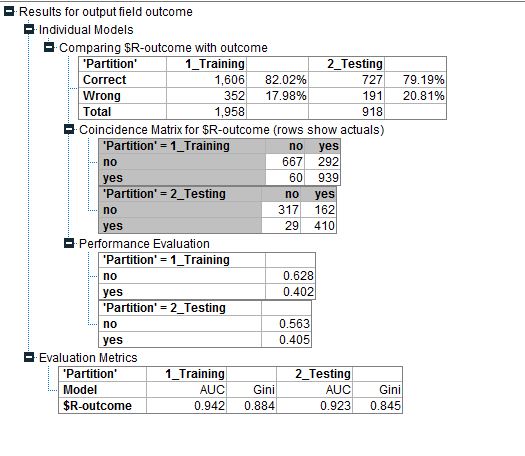
Jedan od težih delova za pretprocesiranje bilo je nalaženje statistika. Iz rada kog navodim[[1]](#footnote-8359) sam izvukla statistike za računanje. Sve sto sam dobila iz kolona je dato kroz apsolutne brojeve. Nek od kolona koje sam izračunala su procenat osvojenih poena na prvom servisu itd.. Ovde nailazimo do problema. U ovom data setu mi nemamo output kolonu, zapravo nemamo target.. Vec znamo unapred podatke o pobedniku I gubitniku meča... Morala sam da se dosetim 2 stvari. Ideja jedan je bila da napravim nove kolone po referenci na rad koji sam malopre navela.

Polovinu mečeva proglasićemo klasom YES a polovinu sa NO. Samim tim za svaku kolonu računamo razlike pobednik-gubitnik. Npr razlike u broju asova i tome dodelimo klasu YES. Ako računamo razlike gubitnik-pobednik(dobijamo npr pozitivan broj duplih grešaka, negativan broj asova) dodelićemo klasi NO. Medjutim ono sto nije dobro kod ove ideje jeste da neko zaista moze da predvidi ishod meča na osnovu statistika datih kroz razlike ali mečevi su vec odigrani pa naš model nema bas smisla... Tu dolazmo na ideju broj dva koja vec govori o tome ko će pobediti u meču koji jos uvek nije odigran. Uz svakog igrača grupišemo njegove statistike. Pri susretu dva igrača njihove statistike koristimo za model. Sada imamo statistike iz prethodnih mečeva i opet pravimo razlike. Sve to pod pretpostavkom da igrači vec godinama igraju sličnim stilom. Kroz ovu ideju računali smo procente ubačenog prvog servisa,wta rangove, medjusobne susrete itd... Cilj je uneti dva igrača u model. Prelazimo na izracunavanje modela.

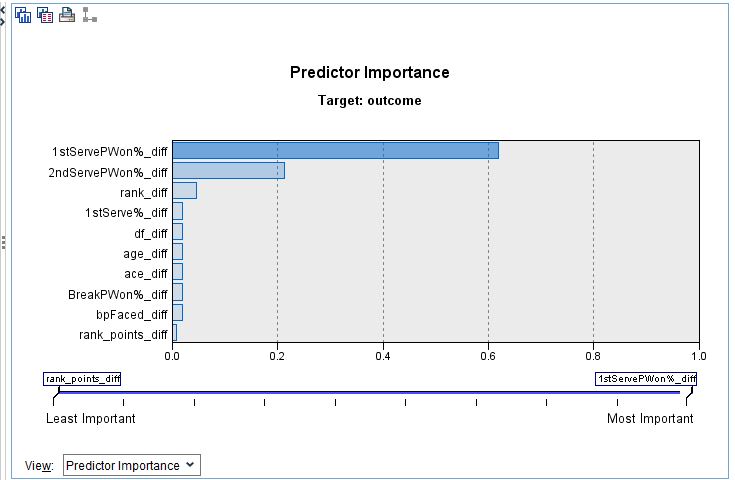
Klasifikacija

Koristila sam različite modele. Pušten je KNN u python-u(dobijena je tacnost od 0.69 na test skupu) kao i u spss-u.

* CRT model

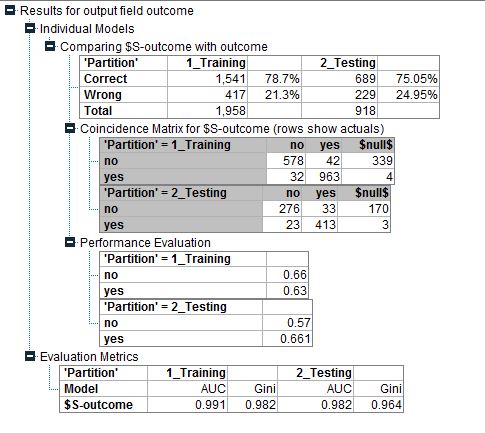


Ovim modelom dobijamo tačnost na trening skupu od 82.02% i na test skupu od 79.19%. Sa slike mozemo pročitati i matricu konfuzije.

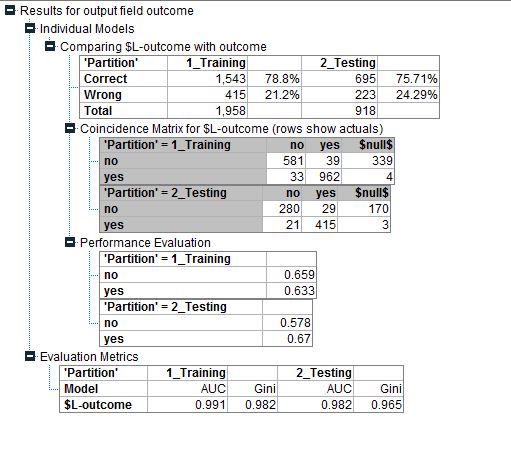


Uočavamo da je procenat prvog servisa jako bitan.

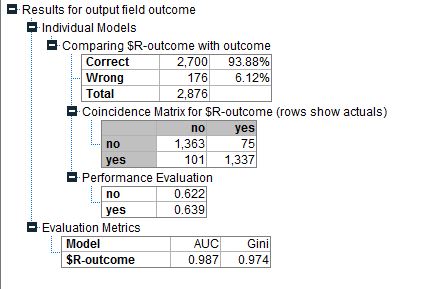
* SVM model uz kernel RBF



* Logistička regresija daje



* Kod C50 mozemo videti da je procenat prvog servisa jako bitan i da od njega zavisi ko odnosi pobedu u meču.
* CHAID



Iz kolone AUC zaključujemo da je model jako dobar posto je priblizan jedinici.

Zaključak

Zaključujemo da model sa velikom tačnošću radi. Na Kaggle mozemo naći da će ovaj data set biti obogaćen još nekim atributima poput povreda, trenutna forma itd. Samim tim to će nam zakomplikovati ali poboljšati model.

[1](https://www.researchgate.net/publication/330883807_Predicting_the_Outcome_of_a_Tennis_Tournament_Based_on_Both_Data_and_Judgments?fbclid=IwAR1ImVd6uXKdbYvQMoEfQ2y60bSYea-BNf-fHi5mwH0IVH7UiygDjaP1VFc)

1. https://www.researchgate.net/publication/330883807\_Predicting\_the\_Outcome\_of\_a\_Tennis\_Tournament\_Based\_on\_Both\_Data\_and\_Judgments?fbclid=IwAR1ImVd6uXKdbYvQMoEfQ2y60bSYea-BNf-fHi5mwH0IVH7UiygDjaP1VFc [↑](#footnote-ref-8359)