Predviđanje srčanog napada u R programskom jeziku

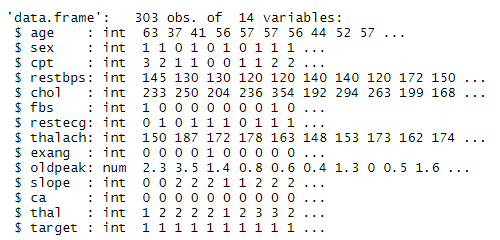
1. Analiza skupa podataka

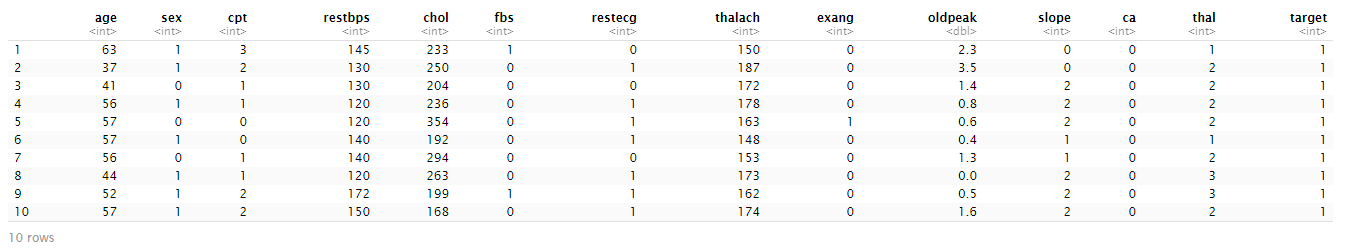
Skup podataka je vezan za dijagnostičke podatke i analiza krvi i srca. Na osnovu ovih podataka se može izvršiti predikcija ili verovatnoća da će osoba imati srčani napad. Svi atributi su već bili brojčane vrednosti tako da nije bilo potrebe za numeričkom kategorizacijom atributa ili dodeljivanjem nekih numeričkih vrednosti, kao niti bilo kakvog skaliranja vrednosti.

Srčani napad je uzrokovan otkazom dela srčanog mišića usled prestanka dovoda krvi. Dovod krvi je obično prekinut usled stvaranja ugruška u krvi u arteriji koja snabdeva srčani mišić. Ako neki od delova srčanog mišića otkaže, osoba oseća jak bol u grudima i električnu nestabilnost tkiva srčanog mišića.

**Atributi:**  
1) age – **godine starosti,**  
2) sex – **pol (1 = muško, 0 = žensko),**  
3) cpt = chest pain type (4 vrednosti) – **tip bola u grudima,**  
4) restbps = resting blood pressure – **krvni pritisak u mirovanju,**  
5) chol = serum cholestoral in mg/dl – **holesteralni serum,**6) fbs = fasting blood sugar > 120 mg/dl – **visok šećer u krvi,**  
7) restecg = resting electrocardiographic results (vrednosti 0,1,2) – **ostalih elektrokardiografskih rezultata,**8) thalach = maximum heart rate achieved – **maksimalni broj otkucaja srca,**  
9) exang = exercise induced angina – **angina izazvana vežbanjem,**  
10) oldpeak = ST depression induced by exercise relative to rest – **ST depresija izazvana vežbanjem u odnosu na odmor,**  
11) slope = the slope of the peak exercise ST segment – **nagib otkucaja u toku fizičkog napora,**  
12) ca = number of major vessels (0-3) colored by flourosopy – **broj glavnih žila obojen flourosopijom**,  
13) thal: 0 = **normalno**; 1 = **fiksni defekt**; 2 = **reverzibilni defekt**  
14) target: 0= **manja šansa za srčani napad**; 1= **veća šansa za srčani napad**

Ispod je prikaz strukture skupa podataka kao i prvog head skupa.

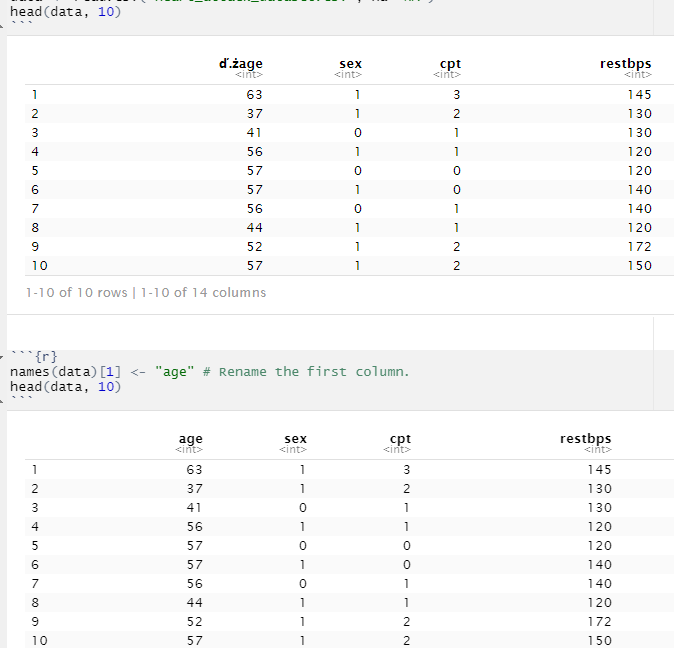




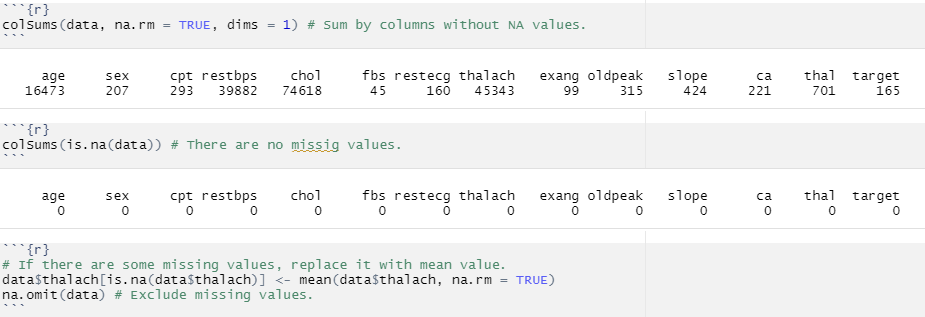
Na osnovu predstavljenih podataka, može se izvršiti predikcija ili verovatnoća za šansu dobijanja srčanog napada upotrebom stabala odlučivanja (Random Forest) koja će izdvojiti najbitnije atribute prema informativnosti ansamblom stabala koji je jedan od osnovnih načina za binarnu klasifikaciju.

1. Čišćenje podataka

Najpre je bilo potrebno pravilno izmeniti prvu kolonu „age“.



Zatim je bilo potrebno proveriti da li postoje nedostajuće vrednosti (NA vrednosti). U koršćenom skupu podataka ih nije bilo od početka, ali je prikazano da bi rešili problem tako što bi za nedostajuće vrednosti uneli srednje vrednosti atributa.

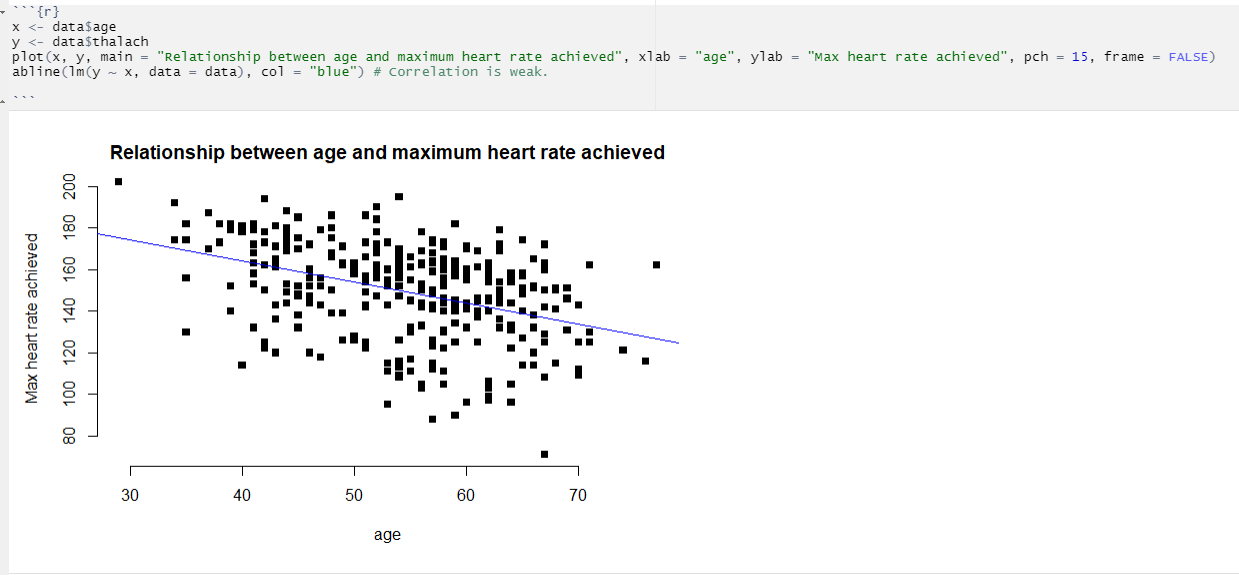


1. Vizualizacija podataka

Prikaz histograma sa raspodelama za godine, najveći otkucaj srca, holesteralnog seruma i krvnog pritiska u stanju mirovanja. Možemo primetiti da raspodela godina i maksimalnog pulsa imaju slične raspodele, kao i holerestarlni serum sa krvnim pritiskom u toku mirovanja.



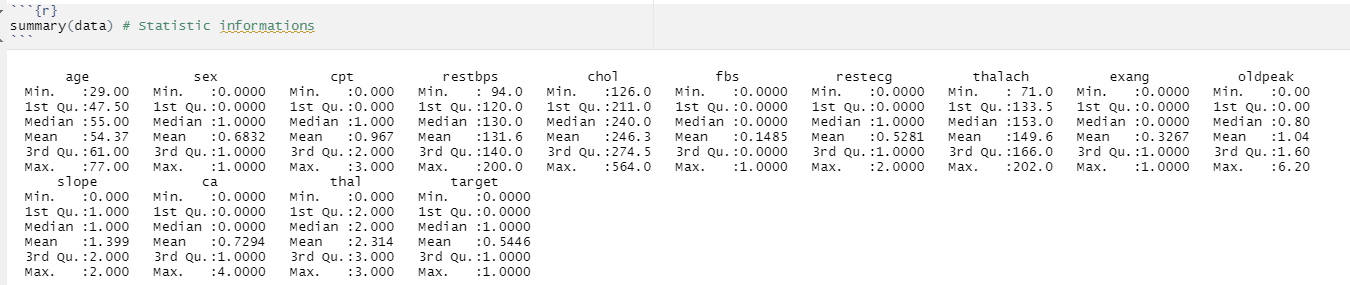
Ispod vidimo da je mala zavisnost između atrbita godina i maksimalnog pulsa ali da opet postoji zavisnost, s obzirom da je retka velika zavisnost između 2 originalna atributa od svih ostalih atributa u skupu podataka.



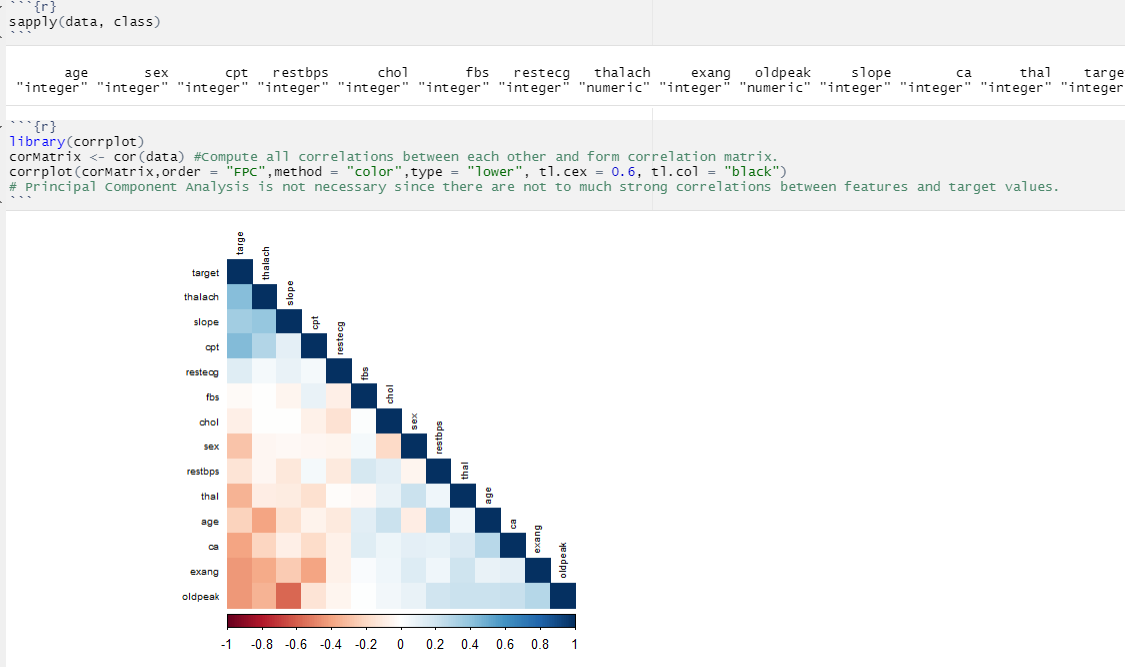
1. Statistički parametri i korelacija

Ispod je naredbom **summary()** prikazana lista parametara za svaki atribut redom:

* Minimalna vrednost,
* Prvi kvartil (25%),
* Drugi Kvartil - Medijana (50%),
* Srednja vrednost,
* Treći kvartil (75%),
* Maksimalna vrednost.

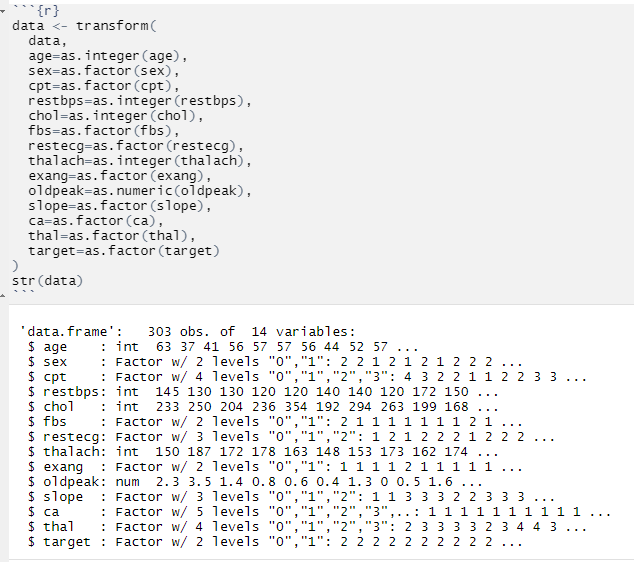


Proveravamo da li su tipovi numerički kako bi uspešno kreirali korelacionu matricu koja je prikazana kao Heatmap ispod. Korelacija se meri prosečnom udaljenošću vrednosti od prave koja je podešena tako da je prosečno rastojanje tačaka minimalno podešavajući **b0** i **b1** parametre.

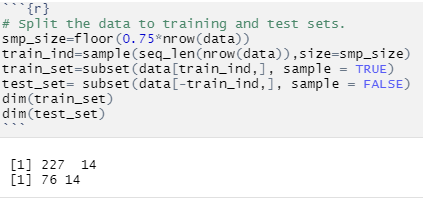


1. Priprema podataka za kreiranje modela

Izvršili smo transforamciju skupa podataka tasko što smo kategoričke vrednosti **as.factor()** funkcijom konvertovali u kategoričke vrednosti kako bi se pravilno upotrebile u obučavanju modela kod Random Forest algoritma.

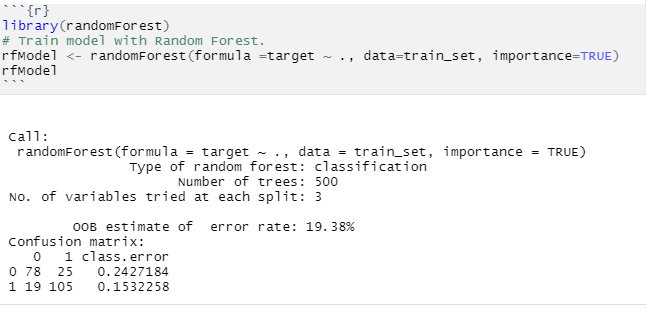


Skup podatak smo podelili na trening (75%) i validacioni test skup (25%)

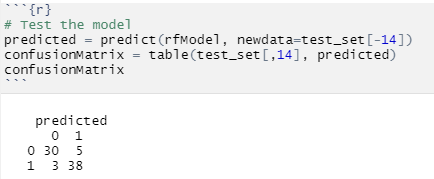


1. Obučavanje i testiranje modela

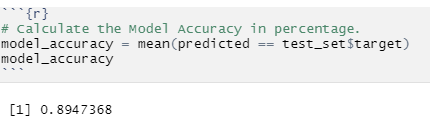
Model smo obučili koristeći Random Forest klasifikator za target kolonu i trening podatke. Koristili smo podrazumevani broj stabala = 500. Na kraju vidimo prikaz konfuzione matrice sa tačnim i pogrešnim klasifikacijama i OOB procenom greške od 19.38%.



Zatim je obučeni model testiran da bi se utvrdila greška klasifikacije nad neviđenim podacima i prikazali smo tačne i pogrešno klasifikovane konfuzionom matricom.



Zatim je na kraju tačno prikazana tačnost modela što iznosi približno 90% tačnosti, što je odličan rezultat za Random Forest algoritam, mada pošto je ovo donekle randomiziran algoritam, neće se uvek dobiti ova tačnost ali će varirati između 80-90% u zavinsoti i od broja stabala, maksimalne dubine stabla i ostlaih parametara.



1. Zaključak

Za ovakve podatke gde se za ciljni atribut traži klasifikacija je idealno za početak koristiti Random Forest koji je ansambl više stabala odluke gde se biraju favoriti i pronalazi značajnost atributa. Iako je Random Forest ograničen nivoom nasleđivanja kategorijalnih varijabli, ipak je jedan od najboljih modela za klasifikaciju. Lak je i jednostavan za razumevanje i korišćenje stoga što nam govori kako je izvršena neka klasifikacija tako što možemo videti putem od korena (glavnog atributa) do lista (rezultata klasifikacije) i opet nam pruža dobru klasifikaciju u poslovnom scenariju. Takođe se može uporediti Random Forest sa ostalim kompleksnijim modelima poput Logističke Regresije, Metode Nosećih Vektora (SVM) ili Neuronske mreže.