PRIRODNO-MATEMATIČKI FAKULET
INFORMATIKA

TIM: DatAlex
PREDMET: Uvod u nauku o podacima
SEMINARSKI RAD NA TEMU: Patient survival prediction

Članovi tima
Aleksandra Stanić 82/2020
Aleksandra Janković 46/2020
Pavle Oprić 69/2020

Predmetni profesor
Branko Arsić

Sadržaj

Predstavljanje problema

Ishod preživljavanja pacijenta u bolnici može biti faktor slučajnosti ili možda greška bolničkog osoblja tokom tretiranja pacijenta. Naš cilj je da na osnovu podataka koji su dobijeni analizom pacijenta tokom prijema u bolnicu, i podataka anamneze pacijenta predktujemo da li će pacijent da preživi u bolnici. Skup podataka koji smo koristili se nalazi u folderu seminarski rad i naziva se dataset.csv.
Link koji vodi do sajta odakle je preuzet dataset: Patient Survival Prediction.

Za početak ćemo uraditi import dataseta:

```
library(readr)
dataset <- read_csv("dataset.csv")
View(dataset)
```

Zatim ćemo učitati potrebne biblioteke za rad:

```
library(tidyverse)
library(dplyr)
library(mice)
library(ggplot2)
library(plotly)
library(rio)
library(validate)
library(leaps)
library(MASS)
library(glmnet)
library(rpart)
library(randomForest)
library(caret)
library(ROCR)
library(pROC)
library(irr)
```

Predstavićemo dataset i objasniti svaku varijablu.

Funkcijom **str** proveravamo kakva je struktura datih kolona/obeležja. Možemo videti da postoji *7* obeležja znakovnog tipa(chr) i *78* obeležja numeričkog tipa, jedno obeležje je tipa *logic*.

```
str(dataset)
```

```
## spc_tbl_ [91,713 × 85] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ encounter_id                : num [1:91713] 66154 114252 119783 79267 92056
...
##  $ patient_id                  : num [1:91713] 25312 59342 50777 46918 34377 ..
.
##  $ hospital_id                 : num [1:91713] 118 81 118 118 33 83 83 33 118 1
18 ...
##  $ age                         : num [1:91713] 68 77 25 81 19 67 59 70 45 50 ..
.
##  $ bmi                         : num [1:91713] 22.7 27.4 31.9 22.6 NA ...
##  $ elective_surgery            : num [1:91713] 0 0 0 1 0 0 0 0 0 0 ...
```

```
##  $ ethnicity                   : chr [1:91713] "Caucasian" "Caucasian" "Caucasi
an" "Caucasian" ...
##  $ gender                      : chr [1:91713] "M" "F" "F" "F" ...
##  $ height                      : num [1:91713] 180 160 173 165 188 ...
##  $ icu_admit_source            : chr [1:91713] "Floor" "Floor" "Accident & Emer
gency" "Operating Room / Recovery" ...
##  $ icu_id                      : num [1:91713] 92 90 93 92 91 95 95 91 114 114
...
##  $ icu_stay_type               : chr [1:91713] "admit" "admit" "admit" "admit"
...
##  $ icu_type                    : chr [1:91713] "CTICU" "Med-Surg ICU" "Med-Surg
ICU" "CTICU" ...
##  $ pre_icu_los_days            : num [1:91713] 0.541667 0.927778 0.000694 0.000
694 0.073611 ...
##  $ weight                      : num [1:91713] 73.9 70.2 95.3 61.7 NA ...
##  $ apache_2_diagnosis          : num [1:91713] 113 108 122 203 119 301 108 113
116 112 ...
##  $ apache_3j_diagnosis         : num [1:91713] 502 203 703 1206 601 ...
##  $ apache_post_operative       : num [1:91713] 0 0 0 1 0 0 0 0 0 0 ...
##  $ arf_apache                  : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ gcs_eyes_apache             : num [1:91713] 3 1 3 4 NA 4 4 4 4 4 ...
##  $ gcs_motor_apache            : num [1:91713] 6 3 6 6 NA 6 6 6 6 6 ...
##  $ gcs_unable_apache           : num [1:91713] 0 0 0 0 NA 0 0 0 0 0 ...
##  $ gcs_verbal_apache           : num [1:91713] 4 1 5 5 NA 5 5 5 5 5 ...
##  $ heart_rate_apache           : num [1:91713] 118 120 102 114 60 113 133 120 8
2 94 ...
##  $ intubated_apache            : num [1:91713] 0 0 0 1 0 0 1 0 0 0 ...
##  $ map_apache                  : num [1:91713] 40 46 68 60 103 130 138 60 66 58
...
##  $ resprate_apache             : num [1:91713] 36 33 37 4 16 35 53 28 14 46 ...
##  $ temp_apache                 : num [1:91713] 39.3 35.1 36.7 34.8 36.7 36.6 35
36.6 36.9 36.3 ...
##  $ ventilated_apache           : num [1:91713] 0 1 0 1 0 0 1 1 1 0 ...
##  $ d1_diasbp_max               : num [1:91713] 68 95 88 48 99 100 76 84 65 83 .
..
##  $ d1_diasbp_min               : num [1:91713] 37 31 48 42 57 61 68 46 59 48 ..
.
##  $ d1_diasbp_noninvasive_max   : num [1:91713] 68 95 88 48 99 100 76 84 65 83 .
..
##  $ d1_diasbp_noninvasive_min   : num [1:91713] 37 31 48 42 57 61 68 46 59 48 ..
.
##  $ d1_heartrate_max            : num [1:91713] 119 118 96 116 89 113 112 118 82
96 ...
##  $ d1_heartrate_min            : num [1:91713] 72 72 68 92 60 83 70 86 82 57 ..
.
##  $ d1_mbp_max                  : num [1:91713] 89 120 102 84 104 127 117 114 93
101 ...
##  $ d1_mbp_min                  : num [1:91713] 46 38 68 84 90 80 97 60 71 59 ..
.
##  $ d1_mbp_noninvasive_max      : num [1:91713] 89 120 102 84 104 127 117 114 93
101 ...
##  $ d1_mbp_noninvasive_min      : num [1:91713] 46 38 68 84 90 80 97 60 71 59 ..
.
##  $ d1_resprate_max             : num [1:91713] 34 32 21 23 18 32 38 28 24 44 ..
.
##  $ d1_resprate_min             : num [1:91713] 10 12 8 7 16 10 16 12 19 14 ...
##  $ d1_spo2_max                 : num [1:91713] 100 100 98 100 100 97 100 100 97
100 ...
##  $ d1_spo2_min                 : num [1:91713] 74 70 91 95 96 91 87 92 97 96 ..
.
##  $ d1_sysbp_max                : num [1:91713] 131 159 148 158 147 173 151 147
104 135 ...
```

```
##  $ d1_sysbp_min                 : num [1:91713] 73 67 105 84 120 107 133 71 98 7
8 ...
##  $ d1_sysbp_noninvasive_max     : num [1:91713] 131 159 148 158 147 173 151 147
104 135 ...
##  $ d1_sysbp_noninvasive_min     : num [1:91713] 73 67 105 84 120 107 133 71 98 7
8 ...
##  $ d1_temp_max                  : num [1:91713] 39.9 36.3 37 38 37.2 36.8 37.2 3
8.5 36.9 37.1 ...
##  $ d1_temp_min                  : num [1:91713] 37.2 35.1 36.7 34.8 36.7 36.6 35
36.6 36.9 36.4 ...
##  $ h1_diasbp_max                : num [1:91713] 68 61 88 62 99 89 107 74 65 83 .
..
##  $ h1_diasbp_min                : num [1:91713] 63 48 58 44 68 89 79 55 59 61 ..
.
##  $ h1_diasbp_noninvasive_max    : num [1:91713] 68 61 88 NA 99 89 NA 74 65 83 ..
.
##  $ h1_diasbp_noninvasive_min    : num [1:91713] 63 48 58 NA 68 89 NA 55 59 61 ..
.
##  $ h1_heartrate_max             : num [1:91713] 119 114 96 100 89 83 79 118 82 9
6 ...
##  $ h1_heartrate_min             : num [1:91713] 108 100 78 96 76 83 72 114 82 60
...
##  $ h1_mbp_max                   : num [1:91713] 86 85 91 92 104 111 117 88 93 10
1 ...
##  $ h1_mbp_min                   : num [1:91713] 85 57 83 71 92 111 117 60 71 77
...
##  $ h1_mbp_noninvasive_max       : num [1:91713] 86 85 91 NA 104 111 117 88 93 10
1 ...
##  $ h1_mbp_noninvasive_min       : num [1:91713] 85 57 83 NA 92 111 117 60 71 77
...
##  $ h1_resprate_max              : num [1:91713] 26 31 20 12 NA 12 18 28 24 29 ..
.
##  $ h1_resprate_min              : num [1:91713] 18 28 16 11 NA 12 18 26 19 17 ..
.
##  $ h1_spo2_max                  : num [1:91713] 100 95 98 100 100 97 100 96 97 1
00 ...
##  $ h1_spo2_min                  : num [1:91713] 74 70 91 99 100 97 100 92 97 96
...
##  $ h1_sysbp_max                 : num [1:91713] 131 95 148 136 130 143 191 119 1
04 135 ...
##  $ h1_sysbp_min                 : num [1:91713] 115 71 124 106 120 143 163 106 9
8 103 ...
##  $ h1_sysbp_noninvasive_max     : num [1:91713] 131 95 148 NA 130 143 NA 119 104
135 ...
##  $ h1_sysbp_noninvasive_min     : num [1:91713] 115 71 124 NA 120 143 NA 106 98
103 ...
##  $ d1_glucose_max               : num [1:91713] 168 145 NA 185 NA 156 197 129 36
5 134 ...
##  $ d1_glucose_min               : num [1:91713] 109 128 NA 88 NA 125 129 129 288
134 ...
##  $ d1_potassium_max             : num [1:91713] 4 4.2 NA 5 NA 3.9 5 5.8 5.2 4.1
...
##  $ d1_potassium_min             : num [1:91713] 3.4 3.8 NA 3.5 NA 3.7 4.2 2.4 5.
2 3.3 ...
##  $ apache_4a_hospital_death_prob: num [1:91713] 0.1 0.47 0 0.04 NA 0.05 0.1 0.11
NA 0.02 ...
##  $ apache_4a_icu_death_prob     : num [1:91713] 0.05 0.29 0 0.03 NA 0.02 0.05 0.
06 NA 0.01 ...
##  $ aids                         : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ cirrhosis                    : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ diabetes_mellitus            : num [1:91713] 1 1 0 0 0 1 1 0 0 0 ...
##  $ hepatic_failure              : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ immunosuppression            : num [1:91713] 0 0 0 0 0 0 0 1 0 0 ...
```

```
##  $ leukemia                     : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ lymphoma                     : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ solid_tumor_with_metastasis  : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ apache_3j_bodysystem         : chr [1:91713] "Sepsis" "Respiratory" "Metaboli
c" "Cardiovascular" ...
##  $ apache_2_bodysystem          : chr [1:91713] "Cardiovascular" "Respiratory" "
Metabolic" "Cardiovascular" ...
##  $ ...84                        : logi [1:91713] NA NA NA NA NA NA ...
##  $ hospital_death               : num [1:91713] 0 0 0 0 0 0 0 0 1 0 ...
```

Nakon učitavanja podataka, funkcija *dim* daje informacije o dimenzijama okvira podataka. Vidi se da skup podataka sadrži 91713 redova i 85 kolona/obeležja.

```
dim(dataset)
```

```
## [1] 91713    85
```

Obeležja i njihov opis koje sadrži okvir podataka *Patient Survival Prediction*

1. encounter_id - jedinstveni identifikator povezan sa boravkom pacijenta na odeljenju
2. patient_id - jedinstveni identifikator povezan sa pacijentom
3. hospital_id - jedinstveni identifikator povezan sa bolnicom
4. age - starost pacijenta prilikom prijema na odeljenje
5. bmi - body mass index pacijenta prilikom prijema u bolnicu
6. elective_surgery - da li je pacijent primljen na neobaveznu hiruršku operaciju
7. ethnicity - nacionalnost ili kulturna tradicija kojoj osoba pripada
8. gender - pol pacijenta
9. height - visina pacijenta na prijemu na odeljenje
10. icu_admit_source - lokacija pacijenta pre prijema na odeljenje
11. icu_id - jedinstveni identifikator jedinice u koju je pacijent primljen
12. icu_stay_type - koje je stanje nakon javljanja pacijenta na odeljenje (da li je primljen, prebačen ili je ponovo primljen)
13. icu_type - klasifikacija koja ukazuje na vrstu nege koju jedinica može da pruži
14. pre_icu_los_days - dužina boravka između prijema u bolnicu i prijema na odeljenje
15. weight - težina (body mass) pacijenta prilikom prijema na odeljenje
16. apache_2_diagnosis - APACHE II dijagnoza za prijem na intezivnu negu
17. apache_3j_diagnosis - šifra poddijagnoze APACHE III-J koja najbolje opisuje razlog prijema na intezivnu negu
18. apache_post_operative - APACHE operativni status; 1 za postoperativno; 0 za neoperativno
19. arf_apache - da li je pacijent imao akutnu bubrežnu insuficijenciju tokom prva 24 sata boravka na odeljenju, definisano kao 24-časovno izlučivanje urina <410ml, kreatinin >=133mikromol/L i bez hronične dijalize
20. gcs_eyes_apache - komponenta otvaranja očiju prema Glasgow Coma Scale, merena tokom prva 24 sata, što rezultira najvišim APACHE III rezultatom
21. gcs_motor_apache - motorna komponenta prema Glasgow Coma Scale, merena tokom prva 24 sata , što rezultira najvišim APACHE III rezultatom

22. gcs_unable_apache - da li Glasgow Coma Scale nije mogla da se proceni zbog sedacije pacijenta
23. gcs_verbal_apache - verbalna komponenta prema Glasgow Coma Scale, merena tokom prva 24 sata, što rezultira najvišim APACHE III rezultatom
24. heart_rate_apache - broj otkucaja srca izmeren tokom prva 24 sata što rezultira najvišim APACHE III rezultatom
25. intubated_apache - da li je pacijent intubiran u trenutku kada je vrednost parcijalnog pritiska gasova u arterisjkoj krvi bio najviši
26. map_apache - srednji arterijski pritisak izmeren tokom prva 24 sata koji rezultira najvišim APACHE III rezultatom
27. resprate_apache - brzina disanja izmerena tokom prva 24 sata što rezultira najvišim APACHE III rezultatom
28. temp_apache - temperatura izmerena tokom prva 24 sata što rezultira najvišim APACHE III rezultatom
29. ventilated_apache - da li je pacijent bio invazivno ventiliran u vreme najvećeg nivoa gasa arterijske krvi koristeći algoritam za ocenjivanje oksigenacije, uključujući bilo koji način ventilacije sa pozitivnim pritiskom koji se isporučuje kroz kolo spojeno na endotrahealnu cev ili traheostomiju
30. d1_diasbp_max - najviši dijastolni krvni pritisak pacijenta tokom prva 24 sata boravka u odeljenju, bilo invazivno ili neinvazivno meren
31. d1_diasbp_min - najniži dijastolni krvni pritisak pacijenta tokom prva 24 sata boravka u odeljenju, bilo invazivno ili neinvazivno meren
32. d1_diasbp_noninvasive_max - najviši dijastolni krvni pritisak pacijenta tokom prva 24 sata boravka u odeljenju, neinvazivno meren
33. d1_diasbp_noninvasive_min - najniži dijastolni krvni pritisak pacijenta tokom prva 24 sata boravka u odeljenju, neinvazivno meren
34. d1_heartrate_max - najveći broj otkucaja srca tokom prva 24 sata boravka na odeljenju
35. d1_heartrate_min - najmanji broj otkucaja srca tokom prva 24 sata boravka na odeljenju
36. d1_mbp_max - najviši srednji krvi pritisak pacijenta tokom prva 24 sata na odeljenju, bilo neinvazivno ili invazivno meren
37. d1_mbp_min - najniži srednji krvi pritisak pacijenta tokom prva 24 sata na odeljenju, bilo neinvazivno ili invazivno meren
38. d1_mbp_noninvasive_max - najviši srednji krvi pritisak pacijenta tokom prva 24 sata na odeljenju, neinvazivno meren
39. d1_mbp_noninvasive_min - najniži srednji krvi pritisak pacijenta tokom prva 24 sata na odeljenju, neinvazivno meren
40. d1_resprate_max - najveća brzina disanja izmerena tokom prva 24 sata na odeljenju
41. d1_resprate_min - najmanja brzina disanja izmerena tokom prva 24 sata na odeljenju
42. d1_spo2_max - najveća saturacija pacijenta tokom prva 24 sata boravka na odeljenju
43. d1_spo2_min - najmanja saturacija pacijenta tokom prva 24 sata boravka na odeljenju
44. d1_sysbp_max - najviši sistolni krvni pritisak pacijenta tokom prva 24 sata, bilo neinvazivno ili invazivno meren
45. d1_sysbp_min - najniži sistolni krvni pritisak pacijenta tokom prva 24 sata, bilo neinvazivno ili invazivno meren

46. d1_sysbp_noninvasive_max - najviši sistolni krvni pritisak pacijenta tokom prva 24 sata, neinvazivno meren
47. d1_sysbp_noninvasive_min - najniži sistolni krvni pritisak pacijenta tokom prva 24 sata, neinvazivno meren
48. d1_temp_max - najviša temperatura tela pacijenta izmerena tokom prva 24 sata, invazivno merena
49. d1_temp_min - najniža temperatura tela pacijenta izmerena tokom prva 24 sata
50. h1_diasbp_max - najviši dijastolni krvni pritisak pacijenta tokom prva 24 sata, bilo neinvazivno ili invazivno meren
51. h1_diasbp_min - najniži dijastolni krvni pritisak pacijenta tokom prva 24 sata, bilo neinvazivno ili invazivno meren
52. h1_diasbp_noninvasive_max - najviši dijastolni krvni pritisak pacijenta tokom prva 24 sata, neinvazivno meren
53. h1_diasbp_noninvasive_min - najviši dijastolni krvni pritisak pacijenta tokom prva 24 sata, neinvazivno meren
54. h1_heartrate_max - najveći broj otkucaja srca pacijenta tokom prvog sata boravka na odeljenju
55. h1_heartrate_min - najmanji broj otkucaja srca pacijenta tokom prvog sata boravka na odeljenju
56. h1_mbp_max - najviši srednji krvni pritisak pacijenta tokom prvog sata boravka na odeljenju, bilo neinvazivno ili invazivno meren
57. h1_mbp_min - najniži srednji krvni pritisak pacijenta tokom prvog sata boravka na odeljenju, bilo neinvazivno ili invazivno meren
58. h1_mbp_noninvasive_max - najviši srednji krvni pritisak pacijenta tokom prvog sata boravka na odeljenju, neinvazivno meren
59. h1_mbp_noninvasive_min - najniži srednji krvni pritisak pacijenta tokom prvog sata boravka na odeljenju, neinvazivno meren
60. h1_resprate_max - najveća brzina disanja pacijenta tokom prvog sata boravka na odeljenju
61. h1_resprate_min - najniža brzina disanja pacijenta tokom prvog sata boravka na odeljenju
62. h1_spo2_max - najveća saturacija kiseonikom tokom prvog sata boravka u jedinici
63. h1_spo2_min - najmanja saturacija kiseonikom tokom prvog sata boravka u jedinici
64. h1_sysbp_max - najviši sistolni pritisak pacijenta tokom prvog sata boravka na odeljenju, bilo neinvazivno ili invazivno meren
65. h1_sysbp_min - najniži sistolni pritisak pacijenta tokom prvog sata boravka na odeljenju, bilo neinvazivno ili invazivno meren
66. h1_sysbp_noninvasive_max - najviši sistolni pritisak pacijenta tokom prvog sata borvaka na odeljenju, neinvazivno meren
67. h1_sysbp_noninvasive_min - najniži sistolni pritisak pacijenta tokom prvog sata borvaka na odeljenju, neinvazivno meren
68. d1_glucose_max - najveća koncentracija glukoze kod pacijenta u serumu ili plazmi tokom prva 24 sata boravka na odeljenju
69. d1_glucose_min - najmanja koncentracija glukoze kod pacijenta u serumu ili plazmi tokom prva 24 sata boravka na odeljenju
70. d1_potassium_max - najveća koncentracija kalijuma kod pacijenta u serumu ili plazmi tokom prva 24 sata boravka na odeljenju

71. d1_potassium_min - najmanja koncentracija kalijuma kod pacijenta u serumu ili plazmi tokom prva 24 sata boravka na odeljenju
72. apache_4a_hospital_death_prob - predikcija za bolnički mortalitet APACHE IVa, koristi APACHE III skor i druge kovarijente, uključujući dijagnozu
73. apache_4a_icu_death_prob - predikcija za mortalitet na intenzivnoj nezi APACHE IVa, koristi APACHE III skor i druge kovarijente, uključujući dijagnozu
74. aids - da li pacijent ima konačnu dijagnozu sindroma stečene imunodeficijencije(AIDS)(ne samo HIV pozitivan)
75. cirrhosis - bilo da pacijent ima istoriju teške upotrebe alkohola sa portonom hipertenzijom i varikozitetima, drugim uzorcima ciroze sa dokazima portne hipertenzije i varikoziteta ili cirozom dokazanom biopsijom. Ovaj komorbiditet se ne odnosi na pacijente sa funkcionalnom transplatacijom jetre
76. diabetes_mellitus - da li je pacijentu dijagnostikovan dijabetes, bilo juvenilni ili adultni, koji zahteva lekove
77. hepatic_failure - da li pacijent ima cirozu i dodatne komplikacije uključujući žuticu i ascites, krvarenje u gornjem delu gastroinfestilnog trakta, hepatičnu encefalopatiju ili komu
78. immunosuppression - da li je imuni sistem pacijenta ugrožen u periodu od 6 meseci pre prijema na intezivnu negu iz bilo kog od sledećih razloga: terapija zračenjem, hemoterapija, upotreba necitotoksičnih imunosupresivnih lekova, visoke doze steroida (najmanje 0,3 mg/kg/dan metilprednizolona ili ekvivalent najmanje 6 meseci)
79. leukemia - da li je pacijentu dijagnostikovana akutna ili hronična mijelogena leukemija, akutna ili hronična limfocitna leukemija ili multipli mijelom
80. lymphoma - da li je pacijentu dijagnostikovan ne-Hodgkin limfom
81. solid_tumor_with_metastasis - da li je pacijentu dijagnostikovan bilo koji karcinom solidnog tumora (uključujući maligni melanom) koji ima dokaze o metastazama
82. apache_3j_bodysystem - grupa za prijmnu dijagnostiku APACHE III
83. apache_2_bodysystem - grupa za prijmnu dijagnostiku APACHE II
84. hospital_death - da li je pacijent preminuo tokom ove hospitalizacije

- APACHE (Acute Physiology and Chronic Health Evaluation) skor je sistem za procenu ozbiljnosti bolesti i predviđanje ishoda pacijenata smeštenih u intenzivnu negu ili intenzivno odeljenje. APACHE skor je razvijen kako bi se pružila kvantitativna ocena težine pacijentovog stanja i kako bi se podržalo medicinsko osoblje u donošenju odluka o tretmanu i brizi za pacijenta. Osnovna ideja APACHE skora je prikupljanje kliničkih podataka o pacijentu, uključujući vitalne znake, laboratorijske rezultate i druge parametre koji ukazuju na fiziološko stanje pacijenta. Ovi podaci se koriste kako bi se izračunao numerički skor koji reflektuje ozbiljnost bolesti. APACHE skor može uključivati parametre kao što su krvni pritisak, puls, temperatura, nivo kiseonika u krvi, pH vrednost, nivo natrijuma i drugi vitalni znaci. Na osnovu ovih podataka, APACHE skor generiše ukupan broj bodova. Ovaj broj se zatim koristi kako bi se predvideli različiti ishodi, kao što su smrtnost, dužina boravka u intenzivnoj nezi, potreba za ventilacijom i drugi parametri. Različite verzije APACHE skora su razvijane tokom vremena kako bi se poboljšala tačnost i pouzdanost sistema za procenu. APACHE skor je često deo

protokola u intenzivnoj nezi i pomaže medicinskom osoblju da prioritetizuje pacijente i pruži optimalnu negu.

Proverićemo koliko NA vrednosti varijable imaju procentualno.

```
(colMeans(is.na(dataset)))*100
```

```
##                   encounter_id                    patient_id
##                     0.00000000                    0.00000000
##                    hospital_id                           age
##                     0.00000000                    4.61003347
##                             bmi               elective_surgery
##                     3.73883746                    0.00000000
##                      ethnicity                        gender
##                     1.52104936                    0.02725895
##                          height               icu_admit_source
##                     1.45453752                    0.12212009
##                          icu_id                  icu_stay_type
##                     0.00000000                    0.00000000
##                        icu_type               pre_icu_los_days
##                     0.00000000                    0.00000000
##                          weight             apache_2_diagnosis
##                     2.96577366                    1.81217494
##            apache_3j_diagnosis         apache_post_operative
##                     1.20048412                    0.00000000
##                     arf_apache              gcs_eyes_apache
##                     0.77960594                    2.07277049
##               gcs_motor_apache             gcs_unable_apache
##                     2.07277049                    1.13070121
##              gcs_verbal_apache             heart_rate_apache
##                     2.07277049                    0.95733429
##                intubated_apache                   map_apache
##                     0.77960594                    1.08381582
##                 resprate_apache                  temp_apache
##                     1.34550173                    4.47919052
##              ventilated_apache                d1_diasbp_max
##                     0.77960594                    0.17990906
##                  d1_diasbp_min     d1_diasbp_noninvasive_max
##                     0.17990906                    1.13397228
##     d1_diasbp_noninvasive_min              d1_heartrate_max
##                     1.13397228                    0.15810190
##               d1_heartrate_min                   d1_mbp_max
##                     0.15810190                    0.23987875
##                    d1_mbp_min        d1_mbp_noninvasive_max
##                     0.23987875                    1.61263943
##        d1_mbp_noninvasive_min               d1_resprate_max
##                     1.61263943                    0.41978782
##                d1_resprate_min                  d1_spo2_max
##                     0.41978782                    0.36308920
##                   d1_spo2_min                 d1_sysbp_max
##                     0.36308920                    0.17336692
##                  d1_sysbp_min      d1_sysbp_noninvasive_max
##                     0.17336692                    1.11979763
##      d1_sysbp_noninvasive_min                   d1_temp_max
##                     1.11979763                    2.53399191
##                   d1_temp_min                 h1_diasbp_max
##                     2.53399191                    3.94600547
##                  h1_diasbp_min     h1_diasbp_noninvasive_max
##                     3.94600547                    8.01413104
##     h1_diasbp_noninvasive_min              h1_heartrate_max
```

```
##                    8.01413104                    3.04209872
##              h1_heartrate_min                    h1_mbp_max
##                    3.04209872                    5.05817060
##                    h1_mbp_min        h1_mbp_noninvasive_max
##                    5.05817060                    9.90481175
##        h1_mbp_noninvasive_min                h1_resprate_max
##                    9.90481175                    4.75068965
##               h1_resprate_min                   h1_spo2_max
##                    4.75068965                    4.56314808
##                   h1_spo2_min                  h1_sysbp_max
##                    4.56314808                    3.93728261
##                  h1_sysbp_min      h1_sysbp_noninvasive_max
##                    3.93728261                    8.00431782
##      h1_sysbp_noninvasive_min                 d1_glucose_max
##                    8.00431782                    6.33170870
##                d1_glucose_min                d1_potassium_max
##                    6.33170870                   10.45108109
##               d1_potassium_min apache_4a_hospital_death_prob
##                   10.45108109                    8.66507474
##      apache_4a_icu_death_prob                          aids
##                    8.66507474                    0.77960594
##                     cirrhosis             diabetes_mellitus
##                    0.77960594                    0.77960594
##               hepatic_failure             immunosuppression
##                    0.77960594                    0.77960594
##                      leukemia                      lymphoma
##                    0.77960594                    0.77960594
##   solid_tumor_with_metastasis          apache_3j_bodysystem
##                    0.77960594                    1.81217494
##           apache_2_bodysystem                          ...84
##                    1.81217494                  100.00000000
##                hospital_death
##                         0.00000000
```

Priprema Podataka

- Pre nego što krenemo sa obradom podataka, prvo ćemo odraditi pripremu podataka tako što ćemo uraditi transformaciju podataka i srediti NA vrednosti. Funkcija summary daje detaljnu statistiku o svakoj koloni/obeležju to jest: maksimum, minimum, medijanu, broj nedostajućih vrednosti, prvi kvartil, treći kvartil.

```
summary(dataset)
```

```
##   encounter_id      patient_id      hospital_id         age
##  Min.   :     1   Min.   :     1   Min.   :  2.0   Min.   :16.00
##  1st Qu.: 32852   1st Qu.: 32830   1st Qu.: 47.0   1st Qu.:52.00
##  Median : 65665   Median : 65413   Median :109.0   Median :65.00
##  Mean   : 65606   Mean   : 65537   Mean   :105.7   Mean   :62.31
##  3rd Qu.: 98342   3rd Qu.: 98298   3rd Qu.:161.0   3rd Qu.:75.00
##  Max.   :131051   Max.   :131051   Max.   :204.0   Max.   :89.00
##                                                    NA's   :4228
##       bmi         elective_surgery  ethnicity           gender
##  Min.   :14.85   Min.   :0.0000   Length:91713       Length:91713
##  1st Qu.:23.64   1st Qu.:0.0000   Class :character   Class :character
##  Median :27.66   Median :0.0000   Mode  :character   Mode  :character
##  Mean   :29.19   Mean   :0.1837
```

```
##  3rd Qu.:32.93   3rd Qu.:0.0000
##  Max.   :67.81   Max.   :1.0000
##  NA's   :3429
##      height      icu_admit_source      icu_id      icu_stay_type
##  Min.   :137.2   Length:91713       Min.   : 82.0   Length:91713
##  1st Qu.:162.5   Class :character   1st Qu.:369.0   Class :character
##  Median :170.1   Mode  :character   Median :504.0   Mode  :character
##  Mean   :169.6                      Mean   :508.4
##  3rd Qu.:177.8                      3rd Qu.:679.0
##  Max.   :195.6                      Max.   :927.0
##  NA's   :1334
##    icu_type          pre_icu_los_days        weight       apache_2_diagnosis
##  Length:91713       Min.   :-24.94722   Min.   : 38.60   Min.   :101.0
##  Class :character   1st Qu.:  0.03542   1st Qu.: 66.80   1st Qu.:113.0
##  Mode  :character   Median :  0.13889   Median : 80.30   Median :122.0
##                     Mean   :  0.83577   Mean   : 84.03   Mean   :185.4
##                     3rd Qu.:  0.40903   3rd Qu.: 97.10   3rd Qu.:301.0
##                     Max.   :159.09097   Max.   :186.00   Max.   :308.0
##                                         NA's   :2720     NA's   :1662
##  apache_3j_diagnosis apache_post_operative   arf_apache    gcs_eyes_apache
##  Min.   :   0.01     Min.   :0.0000        Min.   :0.000   Min.   :1.000
##  1st Qu.: 203.01     1st Qu.:0.0000        1st Qu.:0.000   1st Qu.:3.000
##  Median : 409.02     Median :0.0000        Median :0.000   Median :4.000
##  Mean   : 558.22     Mean   :0.2011        Mean   :0.028   Mean   :3.465
##  3rd Qu.: 703.03     3rd Qu.:0.0000        3rd Qu.:0.000   3rd Qu.:4.000
##  Max.   :2201.05     Max.   :1.0000        Max.   :1.000   Max.   :4.000
##  NA's   :1101                              NA's   :715     NA's   :1901
##  gcs_motor_apache gcs_unable_apache gcs_verbal_apache heart_rate_apache
##  Min.   :1.000    Min.   :0.0000    Min.   :1.000     Min.   : 30.00
##  1st Qu.:6.000    1st Qu.:0.0000    1st Qu.:4.000     1st Qu.: 86.00
##  Median :6.000    Median :0.0000    Median :5.000     Median :104.00
##  Mean   :5.471    Mean   :0.0095    Mean   :3.995     Mean   : 99.71
##  3rd Qu.:6.000    3rd Qu.:0.0000    3rd Qu.:5.000     3rd Qu.:120.00
##  Max.   :6.000    Max.   :1.0000    Max.   :5.000     Max.   :178.00
##  NA's   :1901     NA's   :1037      NA's   :1901      NA's   :878
##  intubated_apache   map_apache     resprate_apache  temp_apache
##  Min.   :0.0000   Min.   : 40.00   Min.   : 4.00   Min.   :32.10
##  1st Qu.:0.0000   1st Qu.: 54.00   1st Qu.:11.00   1st Qu.:36.20
##  Median :0.0000   Median : 67.00   Median :28.00   Median :36.50
##  Mean   :0.1512   Mean   : 88.02   Mean   :25.81   Mean   :36.41
##  3rd Qu.:0.0000   3rd Qu.:125.00   3rd Qu.:36.00   3rd Qu.:36.70
##  Max.   :1.0000   Max.   :200.00   Max.   :60.00   Max.   :39.70
##  NA's   :715      NA's   :994      NA's   :1234    NA's   :4108
##  ventilated_apache d1_diasbp_max    d1_diasbp_min   d1_diasbp_noninvasive_max
##  Min.   :0.0000    Min.   : 46.00   Min.   :13.00   Min.   : 46.00
##  1st Qu.:0.0000    1st Qu.: 75.00   1st Qu.:42.00   1st Qu.: 75.00
##  Median :0.0000    Median : 86.00   Median :50.00   Median : 87.00
##  Mean   :0.3257    Mean   : 88.49   Mean   :50.16   Mean   : 88.61
##  3rd Qu.:1.0000    3rd Qu.: 99.00   3rd Qu.:58.00   3rd Qu.: 99.00
##  Max.   :1.0000    Max.   :165.00   Max.   :90.00   Max.   :165.00
##  NA's   :715       NA's   :165      NA's   :165     NA's   :1040
##  d1_diasbp_noninvasive_min d1_heartrate_max d1_heartrate_min   d1_mbp_max
##  Min.   :13.00             Min.   : 58      Min.   :  0.00   Min.   : 60.0
##  1st Qu.:42.00             1st Qu.: 87      1st Qu.: 60.00   1st Qu.: 90.0
##  Median :50.00             Median :101      Median : 69.00   Median :102.0
##  Mean   :50.24             Mean   :103      Mean   : 70.32   Mean   :104.7
##  3rd Qu.:58.00             3rd Qu.:116      3rd Qu.: 81.00   3rd Qu.:116.0
##  Max.   :90.00             Max.   :177      Max.   :175.00   Max.   :184.0
##  NA's   :1040              NA's   :145      NA's   :145      NA's   :220
##    d1_mbp_min     d1_mbp_noninvasive_max d1_mbp_noninvasive_min d1_resprate_max
##  Min.   : 22.00   Min.   : 60.0          Min.   : 22.00         Min.   :14.00
```

```
##  1st Qu.: 55.00   1st Qu.: 90.0          1st Qu.: 55.00         1st Qu.:22.00
##  Median : 64.00   Median :102.0          Median : 64.00         Median :26.00
##  Mean   : 64.87   Mean   :104.6          Mean   : 64.94         Mean   :28.88
##  3rd Qu.: 75.00   3rd Qu.:116.0          3rd Qu.: 75.00         3rd Qu.:32.00
##  Max.   :112.00   Max.   :181.0          Max.   :112.00         Max.   :92.00
##  NA's   :220      NA's   :1479           NA's   :1479           NA's   :385
##  d1_resprate_min  d1_spo2_max      d1_spo2_min      d1_sysbp_max
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 90.0
##  1st Qu.: 10.00   1st Qu.: 99.00   1st Qu.: 89.00   1st Qu.:130.0
##  Median : 13.00   Median :100.00   Median : 92.00   Median :146.0
##  Mean   : 12.85   Mean   : 99.24   Mean   : 90.45   Mean   :148.3
##  3rd Qu.: 16.00   3rd Qu.:100.00   3rd Qu.: 95.00   3rd Qu.:164.0
##  Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :232.0
##  NA's   :385      NA's   :333      NA's   :333      NA's   :159
##   d1_sysbp_min    d1_sysbp_noninvasive_max d1_sysbp_noninvasive_min
##  Min.   : 41.00   Min.   : 90.0            Min.   : 41.03
##  1st Qu.: 83.00   1st Qu.:130.0            1st Qu.: 84.00
##  Median : 96.00   Median :146.0            Median : 96.00
##  Mean   : 96.92   Mean   :148.2            Mean   : 96.99
##  3rd Qu.:110.00   3rd Qu.:164.0            3rd Qu.:110.00
##  Max.   :160.00   Max.   :232.0            Max.   :160.00
##  NA's   :159      NA's   :1027             NA's   :1027
##   d1_temp_max     d1_temp_min    h1_diasbp_max    h1_diasbp_min
##  Min.   :35.10   Min.   :31.89   Min.   : 37.00   Min.   : 22.00
##  1st Qu.:36.90   1st Qu.:36.10   1st Qu.: 62.00   1st Qu.: 52.00
##  Median :37.11   Median :36.40   Median : 74.00   Median : 62.00
##  Mean   :37.28   Mean   :36.27   Mean   : 75.36   Mean   : 62.84
##  3rd Qu.:37.60   3rd Qu.:36.66   3rd Qu.: 86.00   3rd Qu.: 73.00
##  Max.   :39.90   Max.   :37.80   Max.   :143.00   Max.   :113.00
##  NA's   :2324    NA's   :2324    NA's   :3619     NA's   :3619
##  h1_diasbp_noninvasive_max h1_diasbp_noninvasive_min h1_heartrate_max
##  Min.   : 37.00            Min.   : 22.00            Min.   : 46.00
##  1st Qu.: 63.00            1st Qu.: 52.00            1st Qu.: 77.00
##  Median : 74.00            Median : 62.00            Median : 90.00
##  Mean   : 75.81            Mean   : 63.27            Mean   : 92.23
##  3rd Qu.: 87.00            3rd Qu.: 74.00            3rd Qu.:106.00
##  Max.   :144.00            Max.   :114.00            Max.   :164.00
##  NA's   :7350              NA's   :7350              NA's   :2790
##  h1_heartrate_min  h1_mbp_max      h1_mbp_min     h1_mbp_noninvasive_max
##  Min.   : 36.00   Min.   : 49.00   Min.   : 32.0   Min.   : 49.00
##  1st Qu.: 69.00   1st Qu.: 77.00   1st Qu.: 66.0   1st Qu.: 77.00
##  Median : 82.00   Median : 90.00   Median : 78.0   Median : 90.00
##  Mean   : 83.66   Mean   : 91.61   Mean   : 79.4   Mean   : 91.59
##  3rd Qu.: 97.00   3rd Qu.:104.00   3rd Qu.: 92.0   3rd Qu.:104.00
##  Max.   :144.00   Max.   :165.00   Max.   :138.0   Max.   :163.00
##  NA's   :2790     NA's   :4639     NA's   :4639    NA's   :9084
##  h1_mbp_noninvasive_min h1_resprate_max h1_resprate_min  h1_spo2_max
##  Min.   : 32.00         Min.   :10.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 66.00         1st Qu.:18.00   1st Qu.: 14.00   1st Qu.: 97.00
##  Median : 79.00         Median :21.00   Median : 16.00   Median : 99.00
##  Mean   : 79.71         Mean   :22.63   Mean   : 17.21   Mean   : 98.05
##  3rd Qu.: 92.00         3rd Qu.:26.00   3rd Qu.: 20.00   3rd Qu.:100.00
##  Max.   :138.00         Max.   :59.00   Max.   :189.00   Max.   :100.00
##  NA's   :9084           NA's   :4357    NA's   :4357     NA's   :4185
##   h1_spo2_min      h1_sysbp_max    h1_sysbp_min   h1_sysbp_noninvasive_max
##  Min.   :  0.00   Min.   : 75.0   Min.   : 53.0   Min.   : 75.0
##  1st Qu.: 94.00   1st Qu.:113.0   1st Qu.: 98.0   1st Qu.:113.0
##  Median : 96.00   Median :131.0   Median :115.0   Median :130.0
##  Mean   : 95.17   Mean   :133.2   Mean   :116.4   Mean   :133.1
##  3rd Qu.: 99.00   3rd Qu.:150.0   3rd Qu.:134.0   3rd Qu.:150.0
##  Max.   :100.00   Max.   :223.0   Max.   :194.0   Max.   :223.0
```

```
##  NA's   :4185     NA's   :3611    NA's   :3611    NA's   :7341
##  h1_sysbp_noninvasive_min d1_glucose_max  d1_glucose_min  d1_potassium_max
##  Min.   : 53.0            Min.   : 73.0   Min.   : 33.0   Min.   :2.800
##  1st Qu.: 98.0            1st Qu.:117.0   1st Qu.: 91.0   1st Qu.:3.800
##  Median :115.0            Median :150.0   Median :107.0   Median :4.200
##  Mean   :116.5            Mean   :174.6   Mean   :114.4   Mean   :4.252
##  3rd Qu.:134.0            3rd Qu.:201.0   3rd Qu.:131.0   3rd Qu.:4.600
##  Max.   :195.0            Max.   :611.0   Max.   :288.0   Max.   :7.000
##  NA's   :7341             NA's   :5807    NA's   :5807    NA's   :9585
##  d1_potassium_min apache_4a_hospital_death_prob apache_4a_icu_death_prob
##  Min.   :2.400    Min.   :-1.000                Min.   :-1.000
##  1st Qu.:3.600    1st Qu.: 0.020                1st Qu.: 0.010
##  Median :3.900    Median : 0.050                Median : 0.020
##  Mean   :3.935    Mean   : 0.087                Mean   : 0.044
##  3rd Qu.:4.300    3rd Qu.: 0.130                3rd Qu.: 0.060
##  Max.   :5.800    Max.   : 0.990                Max.   : 0.970
##  NA's   :9585     NA's   :7947                  NA's   :7947
##       aids          cirrhosis      diabetes_mellitus hepatic_failure
##  Min.   :0e+00   Min.   :0.0000   Min.   :0.0000    Min.   :0.000
##  1st Qu.:0e+00   1st Qu.:0.0000   1st Qu.:0.0000    1st Qu.:0.000
##  Median :0e+00   Median :0.0000   Median :0.0000    Median :0.000
##  Mean   :9e-04   Mean   :0.0157   Mean   :0.2252    Mean   :0.013
##  3rd Qu.:0e+00   3rd Qu.:0.0000   3rd Qu.:0.0000    3rd Qu.:0.000
##  Max.   :1e+00   Max.   :1.0000   Max.   :1.0000    Max.   :1.000
##  NA's   :715     NA's   :715      NA's   :715       NA's   :715
##  immunosuppression    leukemia         lymphoma
##  Min.   :0.0000    Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000    1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000    Median :0.0000   Median :0.0000
##  Mean   :0.0262    Mean   :0.0071   Mean   :0.0041
##  3rd Qu.:0.0000    3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000    Max.   :1.0000   Max.   :1.0000
##  NA's   :715       NA's   :715      NA's   :715
##  solid_tumor_with_metastasis apache_3j_bodysystem apache_2_bodysystem
##  Min.   :0.0000              Length:91713         Length:91713
##  1st Qu.:0.0000              Class :character     Class :character
##  Median :0.0000              Mode  :character     Mode  :character
##  Mean   :0.0206
##  3rd Qu.:0.0000
##  Max.   :1.0000
##  NA's   :715
##    ...84          hospital_death
##  Mode:logical    Min.   :0.0000
##  NA's:91713      1st Qu.:0.0000
##                  Median :0.0000
##                  Mean   :0.0863
##                  3rd Qu.:0.0000
##                  Max.   :1.0000
##
```

Možemo da primetimo da imamo dosta NA vrednosti. Više od 2/3 feature-a ima NA vrednosti i treba da nađemo način da ih pripremimo za dalji rad.

```
has_all_na_row <- apply(dataset, 1, function(row) all(is.na(row)))
#has_all_na_row
```

Ovim smo proverili i zaključili da ne postoji red kojim ima sve NA vrednosti. Moramo proveriti koliko svaka kolona ima nedostajućih vrednosti:

- Ukoliko kolona ima više od 99% nedostajućih vrednosti, nemoguće je popuniti te vrednosti zato ćemo ih obrisati.
- Ukoliko kolona ima više od 10% nedostajućih vrednosti, tada ćemo primeniti proces prediktovanja nedostajućih vrednosti.

```
(colMeans(is.na(dataset)))*100 >= 99
```

```
##                   encounter_id                    patient_id
##                          FALSE                         FALSE
##                    hospital_id                           age
##                          FALSE                         FALSE
##                            bmi                elective_surgery
##                          FALSE                         FALSE
##                      ethnicity                        gender
##                          FALSE                         FALSE
##                         height                icu_admit_source
##                          FALSE                         FALSE
##                         icu_id                   icu_stay_type
##                          FALSE                         FALSE
##                       icu_type                pre_icu_los_days
##                          FALSE                         FALSE
##                         weight              apache_2_diagnosis
##                          FALSE                         FALSE
##            apache_3j_diagnosis          apache_post_operative
##                          FALSE                         FALSE
##                     arf_apache                 gcs_eyes_apache
##                          FALSE                         FALSE
##                gcs_motor_apache               gcs_unable_apache
##                          FALSE                         FALSE
##                gcs_verbal_apache               heart_rate_apache
##                          FALSE                         FALSE
##                intubated_apache                     map_apache
##                          FALSE                         FALSE
##                  resprate_apache                    temp_apache
##                          FALSE                         FALSE
##                ventilated_apache                 d1_diasbp_max
##                          FALSE                         FALSE
##                   d1_diasbp_min     d1_diasbp_noninvasive_max
##                          FALSE                         FALSE
##     d1_diasbp_noninvasive_min                d1_heartrate_max
##                          FALSE                         FALSE
##                d1_heartrate_min                     d1_mbp_max
##                          FALSE                         FALSE
##                     d1_mbp_min         d1_mbp_noninvasive_max
##                          FALSE                         FALSE
##         d1_mbp_noninvasive_min                d1_resprate_max
##                          FALSE                         FALSE
##                 d1_resprate_min                    d1_spo2_max
##                          FALSE                         FALSE
##                     d1_spo2_min                   d1_sysbp_max
##                          FALSE                         FALSE
##                    d1_sysbp_min      d1_sysbp_noninvasive_max
##                          FALSE                         FALSE
##       d1_sysbp_noninvasive_min                    d1_temp_max
##                          FALSE                         FALSE
```

```
##                   d1_temp_min                   h1_diasbp_max
##                         FALSE                           FALSE
##                h1_diasbp_min     h1_diasbp_noninvasive_max
##                         FALSE                           FALSE
##     h1_diasbp_noninvasive_min              h1_heartrate_max
##                         FALSE                           FALSE
##              h1_heartrate_min                    h1_mbp_max
##                         FALSE                           FALSE
##                    h1_mbp_min        h1_mbp_noninvasive_max
##                         FALSE                           FALSE
##        h1_mbp_noninvasive_min               h1_resprate_max
##                         FALSE                           FALSE
##               h1_resprate_min                   h1_spo2_max
##                         FALSE                           FALSE
##                   h1_spo2_min                   h1_sysbp_max
##                         FALSE                           FALSE
##                   h1_sysbp_min     h1_sysbp_noninvasive_max
##                         FALSE                           FALSE
##      h1_sysbp_noninvasive_min                d1_glucose_max
##                         FALSE                           FALSE
##                d1_glucose_min              d1_potassium_max
##                         FALSE                           FALSE
##              d1_potassium_min apache_4a_hospital_death_prob
##                         FALSE                           FALSE
##      apache_4a_icu_death_prob                           aids
##                         FALSE                           FALSE
##                     cirrhosis            diabetes_mellitus
##                         FALSE                           FALSE
##               hepatic_failure            immunosuppression
##                         FALSE                           FALSE
##                      leukemia                     lymphoma
##                         FALSE                           FALSE
##   solid_tumor_with_metastasis          apache_3j_bodysystem
##                         FALSE                           FALSE
##           apache_2_bodysystem                          ...84
##                         FALSE                           TRUE
##                hospital_death
##                         FALSE
```

Feature koji ima preko 99% NA vrednosti je logic feture …84. U nastavku ćemo ga rešiti.

```
(colMeans(is.na(dataset)))*100 >= 10
```

```
##                  encounter_id                    patient_id
##                         FALSE                           FALSE
##                   hospital_id                           age
##                         FALSE                           FALSE
##                           bmi              elective_surgery
##                         FALSE                           FALSE
##                     ethnicity                        gender
##                         FALSE                           FALSE
##                        height              icu_admit_source
##                         FALSE                           FALSE
##                        icu_id                 icu_stay_type
##                         FALSE                           FALSE
##                     icu_type              pre_icu_los_days
##                         FALSE                           FALSE
##                        weight            apache_2_diagnosis
##                         FALSE                           FALSE
```

```
##           apache_3j_diagnosis         apache_post_operative
##                         FALSE                         FALSE
##                   arf_apache               gcs_eyes_apache
##                         FALSE                         FALSE
##              gcs_motor_apache             gcs_unable_apache
##                         FALSE                         FALSE
##             gcs_verbal_apache             heart_rate_apache
##                         FALSE                         FALSE
##              intubated_apache                   map_apache
##                         FALSE                         FALSE
##               resprate_apache                  temp_apache
##                         FALSE                         FALSE
##             ventilated_apache                d1_diasbp_max
##                         FALSE                         FALSE
##                d1_diasbp_min     d1_diasbp_noninvasive_max
##                         FALSE                         FALSE
##     d1_diasbp_noninvasive_min              d1_heartrate_max
##                         FALSE                         FALSE
##             d1_heartrate_min                   d1_mbp_max
##                         FALSE                         FALSE
##                   d1_mbp_min        d1_mbp_noninvasive_max
##                         FALSE                         FALSE
##        d1_mbp_noninvasive_min               d1_resprate_max
##                         FALSE                         FALSE
##              d1_resprate_min                  d1_spo2_max
##                         FALSE                         FALSE
##                  d1_spo2_min                 d1_sysbp_max
##                         FALSE                         FALSE
##                 d1_sysbp_min      d1_sysbp_noninvasive_max
##                         FALSE                         FALSE
##      d1_sysbp_noninvasive_min                  d1_temp_max
##                         FALSE                         FALSE
##                  d1_temp_min                h1_diasbp_max
##                         FALSE                         FALSE
##                h1_diasbp_min     h1_diasbp_noninvasive_max
##                         FALSE                         FALSE
##     h1_diasbp_noninvasive_min              h1_heartrate_max
##                         FALSE                         FALSE
##             h1_heartrate_min                   h1_mbp_max
##                         FALSE                         FALSE
##                   h1_mbp_min        h1_mbp_noninvasive_max
##                         FALSE                         FALSE
##        h1_mbp_noninvasive_min               h1_resprate_max
##                         FALSE                         FALSE
##              h1_resprate_min                  h1_spo2_max
##                         FALSE                         FALSE
##                  h1_spo2_min                 h1_sysbp_max
##                         FALSE                         FALSE
##                 h1_sysbp_min      h1_sysbp_noninvasive_max
##                         FALSE                         FALSE
##      h1_sysbp_noninvasive_min                d1_glucose_max
##                         FALSE                         FALSE
##                d1_glucose_min              d1_potassium_max
##                         FALSE                          TRUE
##             d1_potassium_min apache_4a_hospital_death_prob
##                          TRUE                         FALSE
##      apache_4a_icu_death_prob                          aids
##                         FALSE                         FALSE
##                    cirrhosis             diabetes_mellitus
##                         FALSE                         FALSE
##               hepatic_failure             immunosuppression
```

```
##                          FALSE                          FALSE
##                       leukemia                       lymphoma
##                          FALSE                          FALSE
##   solid_tumor_with_metastasis          apache_3j_bodysystem
##                          FALSE                          FALSE
##           apache_2_bodysystem                          ...84
##                          FALSE                           TRUE
##                hospital_death
##                          FALSE
```

Primećujemo da imamo dva feature-a koji imaju preko 10% NA vrednosti i to su:

1. d1_potassium_min - najveća koncentracija kalijuma kod pacijenta u serumu ili plazmi tokom prva 24 sata boravka na odeljenju
2. d1_potassium_max - najmanja koncentracija glukoze kod pacijenta u serumu ili plazmi tokom prva 24 sata boravka na odeljenju

Pored njih je tu *logic* feature sa 100% *NA* vrednosti.

Irelevantne/nepotrebne vrednosti

```
head(dataset)
```

Uklonićemo kolonu koja nema smisla i ima 100% NA vrednosti.

```
dataset <- subset(dataset, select = -c(...84))
```

Kolone koje možemo odmah da obrišemo: icu_admit_source, icu_id, icu_stay_type, patient_id, hospital_id.

```
dataset <- subset(dataset, select = -c(icu_admit_source, icu_id, icu_stay_t
ype, patient_id, hospital_id))
```

Primećujemo irelevantne vrednosti u sledećoj koloni:
pre_icu_los_days - dužina boravka između prijema u bolnicu i prijema na odeljenje

```
summary(dataset$pre_icu_los_days)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -24.94722   0.03542   0.13889   0.83577   0.40903 159.09097
```

Ovde imamo vrednosti koje su negativne i vrednosti koje nam govore u prilog tome da je neko primljen na odeljenje pola godine nakon što se prijavio u bolnicu. Ovo je kolona koju ćemo obrisati. Takođe nam ovaj feture ne daje značajne podatke tako da ga možemo obrisati.

```
dataset <- subset(dataset, select = -c(pre_icu_los_days))
```

Sada posmatramo kolone kao što su:

1. d1_diasbp_noninvasive_max - najviši dijastolni krvni pritisak pacijenta tokom prva 24 sata boravka u odeljenju, neinvazivno meren
2. d1_diasbp_noninvasive_min - najniži dijastolni krvni pritisak pacijenta tokom prva 24 sata boravka u odeljenju, neinvazivno meren
3. d1_mbp_noninvasive_max - najviši srednji krvi pritisak pacijenta tokom prva 24 sata na odeljenju, neinvazivno meren
4. d1_mbp_noninvasive_min - najniži srednji krvi pritisak pacijenta tokom prva 24 sata na odeljenju, neinvazivno meren
5. d1_sysbp_noninvasive_max - najviši sistolni krvni pritisak pacijenta tokom prva 24 sata, neinvazivno meren
6. d1_sysbp_noninvasive_min - najniži sistolni krvni pritisak pacijenta tokom prva 24 sata, neinvazivno meren
7. h1_diasbp_noninvasive_max - najviši dijastolni krvni pritisak pacijenta tokom prva 24 sata, neinvazivno meren
8. h1_diasbp_noninvasive_min - najviši dijastolni krvni pritisak pacijenta tokom prva 24 sata, neinvazivno meren
9. h1_mbp_noninvasive_max - najviši srednji krvni pritisak pacijenta tokom prvog sata boravka na odeljenju, neinvazivno meren
10. h1_mbp_noninvasive_min - najniži srednji krvni pritisak pacijenta tokom prvog sata boravka na odeljenju, neinvazivno meren
11. h1_sysbp_noninvasive_max - najviši sistolni pritisak pacijenta tokom prvog sata borvaka na odeljenju, neinvazivno meren
12. h1_sysbp_noninvasive_min - najniži sistolni pritisak pacijenta tokom prvog sata borvaka na odeljenju, neinvazivno meren

Ovim kolonama je zajedničko to da su njihove vrednosti dobijene neinvazivnim merenjem. Neinvazivno merenje je neprecizno (npr. kod pritiska to je merenje aparatom za pritisak). Takođe za sve ove kolone imamo vrednosti koje su merene invazivno/neinvazivno (invazivno merenje pritiska je direktno ubadanje iglom u arteriju). Zbog toga što nam je invazivno merenje relevantnije, kolone koje sadrže vrednosi neinvazivnog merenja ćemo obrisati.

Prvo ćemo da proverimo da li nam ove kolone mogu pomoći u popunjavanju NA vrednosti kod invazivno merenih vrednosti.

1. d1_diasbp_noninvasive_max

```
brojac <- 0
for (i in 1:nrow(dataset)) {
  if (is.na(dataset$d1_diasbp_noninvasive_max[i]) && is.na(dataset$d1_diasb
p_max[i])) {
    brojac <- brojac+1 }}
cat("Broj redova sa NA vrednostima u obema kolonama:", brojac, "\n")
## Broj redova sa NA vrednostima u obema kolonama: 165
cat("Broj redova sa NA vrednostima u d1_diasbp_noninvasive_max koloni:", su
m(is.na(dataset$d1_diasbp_noninvasive_max)), "\n")
```

```
## Broj redova sa NA vrednostima u d1_diasbp_noninvasive_max koloni: 1040
cat("Broj redova sa NA vrednostima u d1_diasbp_max koloni:", sum(is.na(data
set$d1_diasbp_max)), "\n")
## Broj redova sa NA vrednostima u d1_diasbp_max koloni: 165
```

2. d1_diasbp_noninvasive_min

```
brojac <- 0
for (i in 1:nrow(dataset)) {
  if (is.na(dataset$d1_diasbp_noninvasive_min[i]) && is.na(dataset$d1_diasb
p_min[i])) {
    brojac <- brojac+1}}
cat("Broj redova sa NA vrednostima u obema kolonama:", brojac, "\n")
## Broj redova sa NA vrednostima u obema kolonama: 165
cat("Broj redova sa NA vrednostima u d1_diasbp_noninvasive_min koloni:", su
m(is.na(dataset$d1_diasbp_noninvasive_min)), "\n")
## Broj redova sa NA vrednostima u d1_diasbp_noninvasive_min koloni: 1040
cat("Broj redova sa NA vrednostima u d1_diasbp_min koloni:", sum(is.na(data
set$d1_diasbp_min)), "\n")
## Broj redova sa NA vrednostima u d1_diasbp_min koloni: 165
```

3. d1_mbp_noninvasive_max

```
brojac <- 0
for (i in 1:nrow(dataset)) {
  if (is.na(dataset$d1_mbp_noninvasive_max[i]) && is.na(dataset$d1_mbp_max[
i])) {
    brojac <- brojac+1}}
cat("Broj redova sa NA vrednostima u obema kolonama:", brojac, "\n")
## Broj redova sa NA vrednostima u obema kolonama: 220
cat("Broj redova sa NA vrednostima u d1_mbp_noninvasive_max koloni:", sum(i
s.na(dataset$d1_mbp_noninvasive_max)), "\n")
## Broj redova sa NA vrednostima u d1_mbp_noninvasive_max koloni: 1479
cat("Broj redova sa NA vrednostima u d1_mbp_max koloni:", sum(is.na(dataset
$d1_mbp_max)), "\n")
## Broj redova sa NA vrednostima u d1_mbp_max koloni: 220
```

4. d1_mbp_noninvasive_min

```
brojac <- 0
for (i in 1:nrow(dataset)) {
  if (is.na(dataset$d1_mbp_noninvasive_min[i]) && is.na(dataset$d1_mbp_min[
i])) {
    brojac <- brojac+1}}
cat("Broj redova sa NA vrednostima u obema kolonama:", brojac, "\n")
## Broj redova sa NA vrednostima u obema kolonama: 220
cat("Broj redova sa NA vrednostima u d1_mbp_noninvasive_min koloni:", sum(i
s.na(dataset$d1_mbp_noninvasive_min)), "\n")
## Broj redova sa NA vrednostima u d1_mbp_noninvasive_min koloni: 1479
```

```
cat("Broj redova sa NA vrednostima u d1_mbp_min koloni:", sum(is.na(dataset
$d1_mbp_min)), "\n")
## Broj redova sa NA vrednostima u d1_mbp_min koloni: 220
```

5. d1_sysbp_noninvasive_max

```
brojac <- 0
for (i in 1:nrow(dataset)) {
  if (is.na(dataset$d1_sysbp_noninvasive_max[i]) && is.na(dataset$d1_sysbp_
max[i])) {
    brojac <- brojac+1}}
cat("Broj redova sa NA vrednostima u obema kolonama:", brojac, "\n")
## Broj redova sa NA vrednostima u obema kolonama: 159
cat("Broj redova sa NA vrednostima u d1_sysbp_noninvasive_max koloni:", sum
(is.na(dataset$d1_sysbp_noninvasive_max)), "\n")
## Broj redova sa NA vrednostima u d1_sysbp_noninvasive_max koloni: 1027
cat("Broj redova sa NA vrednostima u d1_sysbp_max koloni:", sum(is.na(datas
et$d1_sysbp_max)), "\n")
## Broj redova sa NA vrednostima u d1_sysbp_max koloni: 159
```

6. d1_sysbp_noninvasive_min

```
brojac <- 0
for (i in 1:nrow(dataset)) {
  if (is.na(dataset$d1_sysbp_noninvasive_min[i]) && is.na(dataset$d1_sysbp_
min[i])) {
    brojac <- brojac+1}}
cat("Broj redova sa NA vrednostima u obema kolonama:", brojac, "\n")
## Broj redova sa NA vrednostima u obema kolonama: 159
cat("Broj redova sa NA vrednostima u d1_sysbp_noninvasive_min koloni:", sum
(is.na(dataset$d1_sysbp_noninvasive_min)), "\n")
## Broj redova sa NA vrednostima u d1_sysbp_noninvasive_min koloni: 1027
cat("Broj redova sa NA vrednostima u d1_sysbp_min koloni:", sum(is.na(datas
et$d1_sysbp_min)), "\n")
## Broj redova sa NA vrednostima u d1_sysbp_min koloni: 159
```

7. h1_diasbp_noninvasive_max

```
brojac <- 0
for (i in 1:nrow(dataset)) {
  if (is.na(dataset$h1_diasbp_noninvasive_max[i]) && is.na(dataset$h1_diasb
p_max[i])) {
    brojac <- brojac+1}}
cat("Broj redova sa NA vrednostima u obema kolonama:", brojac, "\n")
## Broj redova sa NA vrednostima u obema kolonama: 3619
cat("Broj redova sa NA vrednostima u h1_diasbp_noninvasive_max koloni:", su
m(is.na(dataset$h1_diasbp_noninvasive_max)), "\n")
## Broj redova sa NA vrednostima u h1_diasbp_noninvasive_max koloni: 7350
```

```
cat("Broj redova sa NA vrednostima u h1_diasbp_max koloni:", sum(is.na(data
set$h1_diasbp_max)), "\n")
## Broj redova sa NA vrednostima u h1_diasbp_max koloni: 3619
```

8. h1_diasbp_noninvasive_min

```
brojac <- 0
for (i in 1:nrow(dataset)) {
  if (is.na(dataset$h1_diasbp_noninvasive_min[i]) && is.na(dataset$h1_diasb
p_min[i])) {
    brojac <- brojac+1}}
cat("Broj redova sa NA vrednostima u obema kolonama:", brojac, "\n")
## Broj redova sa NA vrednostima u obema kolonama: 3619
cat("Broj redova sa NA vrednostima u h1_diasbp_noninvasive_min koloni:", su
m(is.na(dataset$h1_diasbp_noninvasive_min)), "\n")
## Broj redova sa NA vrednostima u h1_diasbp_noninvasive_min koloni: 7350
cat("Broj redova sa NA vrednostima u h1_diasbp_min koloni:", sum(is.na(data
set$h1_diasbp_min)), "\n")
## Broj redova sa NA vrednostima u h1_diasbp_min koloni: 3619
```

9. h1_mbp_noninvasive_max

```
brojac <- 0
for (i in 1:nrow(dataset)) {
  if (is.na(dataset$h1_mbp_noninvasive_max[i]) && is.na(dataset$h1_mbp_max[
i])) {
    brojac <- brojac+1}}
cat("Broj redova sa NA vrednostima u obema kolonama:", brojac, "\n")
## Broj redova sa NA vrednostima u obema kolonama: 4639
cat("Broj redova sa NA vrednostima u h1_mbp_noninvasive_max koloni:", sum(i
s.na(dataset$h1_mbp_noninvasive_max)), "\n")
## Broj redova sa NA vrednostima u h1_mbp_noninvasive_max koloni: 9084
cat("Broj redova sa NA vrednostima u h1_mbp_max koloni:", sum(is.na(dataset
$h1_mbp_max)), "\n")
## Broj redova sa NA vrednostima u h1_mbp_max koloni: 4639
```

10. h1_mbp_noninvasive_min

```
brojac <- 0
for (i in 1:nrow(dataset)) {
  if (is.na(dataset$h1_mbp_noninvasive_min[i]) && is.na(dataset$h1_mbp_min[
i])) {
    brojac <- brojac+1}}
cat("Broj redova sa NA vrednostima u obema kolonama:", brojac, "\n")
## Broj redova sa NA vrednostima u obema kolonama: 4639
cat("Broj redova sa NA vrednostima u h1_mbp_noninvasive_min koloni:", sum(i
s.na(dataset$h1_mbp_noninvasive_min)), "\n")
## Broj redova sa NA vrednostima u h1_mbp_noninvasive_min koloni: 9084
```

```
cat("Broj redova sa NA vrednostima u h1_mbp_min koloni:", sum(is.na(dataset
$h1_mbp_min)), "\n")
## Broj redova sa NA vrednostima u h1_mbp_min koloni: 4639
```

11. h1_sysbp_noninvasive_max

```
brojac <- 0
for (i in 1:nrow(dataset)) {
  if (is.na(dataset$h1_sysbp_noninvasive_max[i]) && is.na(dataset$h1_sysbp_
max[i])) {
    brojac <- brojac+1}}
cat("Broj redova sa NA vrednostima u obema kolonama:", brojac, "\n")
## Broj redova sa NA vrednostima u obema kolonama: 3611
cat("Broj redova sa NA vrednostima u h1_sysbp_noninvasive_max koloni:", sum
(is.na(dataset$h1_sysbp_noninvasive_max)), "\n")
## Broj redova sa NA vrednostima u h1_sysbp_noninvasive_max koloni: 7341
cat("Broj redova sa NA vrednostima u h1_sysbp_max koloni:", sum(is.na(datas
et$h1_sysbp_max)), "\n")
## Broj redova sa NA vrednostima u h1_sysbp_max koloni: 3611
```

12. h1_sysbp_noninvasive_min

```
brojac <- 0
for (i in 1:nrow(dataset)) {
  if (is.na(dataset$h1_sysbp_noninvasive_min[i]) && is.na(dataset$h1_sysbp_
min[i])) {
    brojac <- brojac+1}}
cat("Broj redova sa NA vrednostima u obema kolonama:", brojac, "\n")
## Broj redova sa NA vrednostima u obema kolonama: 3611
cat("Broj redova sa NA vrednostima u h1_sysbp_noninvasive_min koloni:", sum
(is.na(dataset$h1_sysbp_noninvasive_min)), "\n")
## Broj redova sa NA vrednostima u h1_sysbp_noninvasive_min koloni: 7341
cat("Broj redova sa NA vrednostima u h1_sysbp_min koloni:", sum(is.na(datas
et$h1_sysbp_min)), "\n")
## Broj redova sa NA vrednostima u h1_sysbp_min koloni: 3611
```

Sada smo potvrdili da možemo obrisati ove kolone.

```
dataset <- subset(dataset, select = -c(d1_diasbp_noninvasive_max, d1_diasbp
_noninvasive_min, d1_mbp_noninvasive_max, d1_mbp_noninvasive_min, d1_sysbp_
noninvasive_max, d1_sysbp_noninvasive_min, h1_diasbp_noninvasive_max, h1_di
asbp_noninvasive_min, h1_mbp_noninvasive_max, h1_mbp_noninvasive_min, h1_sy
sbp_noninvasive_max, h1_sysbp_noninvasive_min))

str(dataset)
## tibble [91,713 × 66] (S3: tbl_df/tbl/data.frame)
##  $ encounter_id                : num [1:91713] 66154 114252 119783 79267 92056
...
##  $ age                         : num [1:91713] 68 77 25 81 19 67 59 70 45 50 ..
.
```

```
##  $ bmi                          : num [1:91713] 22.7 27.4 31.9 22.6 NA ...
##  $ elective_surgery             : num [1:91713] 0 0 0 1 0 0 0 0 0 0 ...
##  $ ethnicity                    : chr [1:91713] "Caucasian" "Caucasian" "Caucasi
an" "Caucasian" ...
##  $ gender                       : chr [1:91713] "M" "F" "F" "F" ...
##  $ height                       : num [1:91713] 180 160 173 165 188 ...
##  $ icu_type                     : chr [1:91713] "CTICU" "Med-Surg ICU" "Med-Surg
ICU" "CTICU" ...
##  $ weight                       : num [1:91713] 73.9 70.2 95.3 61.7 NA ...
##  $ apache_2_diagnosis           : num [1:91713] 113 108 122 203 119 301 108 113
116 112 ...
##  $ apache_3j_diagnosis          : num [1:91713] 502 203 703 1206 601 ...
##  $ apache_post_operative        : num [1:91713] 0 0 0 1 0 0 0 0 0 0 ...
##  $ arf_apache                   : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ gcs_eyes_apache              : num [1:91713] 3 1 3 4 NA 4 4 4 4 4 ...
##  $ gcs_motor_apache             : num [1:91713] 6 3 6 6 NA 6 6 6 6 6 ...
##  $ gcs_unable_apache            : num [1:91713] 0 0 0 0 NA 0 0 0 0 0 ...
##  $ gcs_verbal_apache            : num [1:91713] 4 1 5 5 NA 5 5 5 5 5 ...
##  $ heart_rate_apache            : num [1:91713] 118 120 102 114 60 113 133 120 8
2 94 ...
##  $ intubated_apache             : num [1:91713] 0 0 0 1 0 0 1 0 0 0 ...
##  $ map_apache                   : num [1:91713] 40 46 68 60 103 130 138 60 66 58
...
##  $ resprate_apache              : num [1:91713] 36 33 37 4 16 35 53 28 14 46 ...
##  $ temp_apache                  : num [1:91713] 39.3 35.1 36.7 34.8 36.7 36.6 35
36.6 36.9 36.3 ...
##  $ ventilated_apache            : num [1:91713] 0 1 0 1 0 0 1 1 1 0 ...
##  $ d1_diasbp_max                : num [1:91713] 68 95 88 48 99 100 76 84 65 83 .
..
##  $ d1_diasbp_min                : num [1:91713] 37 31 48 42 57 61 68 46 59 48 ..
.
##  $ d1_heartrate_max             : num [1:91713] 119 118 96 116 89 113 112 118 82
96 ...
##  $ d1_heartrate_min             : num [1:91713] 72 72 68 92 60 83 70 86 82 57 ..
.
##  $ d1_mbp_max                   : num [1:91713] 89 120 102 84 104 127 117 114 93
101 ...
##  $ d1_mbp_min                   : num [1:91713] 46 38 68 84 90 80 97 60 71 59 ..
.
##  $ d1_resprate_max              : num [1:91713] 34 32 21 23 18 32 38 28 24 44 ..
.
##  $ d1_resprate_min              : num [1:91713] 10 12 8 7 16 10 16 12 19 14 ...
##  $ d1_spo2_max                  : num [1:91713] 100 100 98 100 100 97 100 100 97
100 ...
##  $ d1_spo2_min                  : num [1:91713] 74 70 91 95 96 91 87 92 97 96 ..
.
##  $ d1_sysbp_max                 : num [1:91713] 131 159 148 158 147 173 151 147
104 135 ...
##  $ d1_sysbp_min                 : num [1:91713] 73 67 105 84 120 107 133 71 98 7
8 ...
##  $ d1_temp_max                  : num [1:91713] 39.9 36.3 37 38 37.2 36.8 37.2 3
8.5 36.9 37.1 ...
##  $ d1_temp_min                  : num [1:91713] 37.2 35.1 36.7 34.8 36.7 36.6 35
36.6 36.9 36.4 ...
##  $ h1_diasbp_max                : num [1:91713] 68 61 88 62 99 89 107 74 65 83 .
..
##  $ h1_diasbp_min                : num [1:91713] 63 48 58 44 68 89 79 55 59 61 ..
.
##  $ h1_heartrate_max             : num [1:91713] 119 114 96 100 89 83 79 118 82 9
6 ...
##  $ h1_heartrate_min             : num [1:91713] 108 100 78 96 76 83 72 114 82 60
...
```

```
##  $ h1_mbp_max                   : num [1:91713] 86 85 91 92 104 111 117 88 93 10
1 ...
##  $ h1_mbp_min                   : num [1:91713] 85 57 83 71 92 111 117 60 71 77
...
##  $ h1_resprate_max              : num [1:91713] 26 31 20 12 NA 12 18 28 24 29 ..
.
##  $ h1_resprate_min              : num [1:91713] 18 28 16 11 NA 12 18 26 19 17 ..
.
##  $ h1_spo2_max                  : num [1:91713] 100 95 98 100 100 97 100 96 97 1
00 ...
##  $ h1_spo2_min                  : num [1:91713] 74 70 91 99 100 97 100 92 97 96
...
##  $ h1_sysbp_max                 : num [1:91713] 131 95 148 136 130 143 191 119 1
04 135 ...
##  $ h1_sysbp_min                 : num [1:91713] 115 71 124 106 120 143 163 106 9
8 103 ...
##  $ d1_glucose_max               : num [1:91713] 168 145 NA 185 NA 156 197 129 36
5 134 ...
##  $ d1_glucose_min               : num [1:91713] 109 128 NA 88 NA 125 129 129 288
134 ...
##  $ d1_potassium_max             : num [1:91713] 4 4.2 NA 5 NA 3.9 5 5.8 5.2 4.1
...
##  $ d1_potassium_min             : num [1:91713] 3.4 3.8 NA 3.5 NA 3.7 4.2 2.4 5.
2 3.3 ...
##  $ apache_4a_hospital_death_prob: num [1:91713] 0.1 0.47 0 0.04 NA 0.05 0.1 0.11
NA 0.02 ...
##  $ apache_4a_icu_death_prob     : num [1:91713] 0.05 0.29 0 0.03 NA 0.02 0.05 0.
06 NA 0.01 ...
##  $ aids                         : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ cirrhosis                    : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ diabetes_mellitus            : num [1:91713] 1 1 0 0 0 1 1 0 0 0 ...
##  $ hepatic_failure              : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ immunosuppression            : num [1:91713] 0 0 0 0 0 0 0 1 0 0 ...
##  $ leukemia                     : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ lymphoma                     : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ solid_tumor_with_metastasis  : num [1:91713] 0 0 0 0 0 0 0 0 0 0 ...
##  $ apache_3j_bodysystem         : chr [1:91713] "Sepsis" "Respiratory" "Metaboli
c" "Cardiovascular" ...
##  $ apache_2_bodysystem          : chr [1:91713] "Cardiovascular" "Respiratory" "
Metabolic" "Cardiovascular" ...
##  $ hospital_death               : num [1:91713] 0 0 0 0 0 0 0 0 1 0 ...
```

Validacija

U nastavku ćemo rešiti sve “nelogične” vrednosti (sve zabeležene vrednosti koje je nemoguće dostići, u zavisnosti od metrike). Takve vrednosti ćemo tretirati kao greške, i pretvorićemo ih u NA vrednosti koje ćemo u nastavku popuniti. Želimo da sačuvamo što veći broj vrednosti u datasetu. Granice i uslovi za svaku od metrika dobijene su domenskim i ekspertnim znanjem, u skladu sa medicinskom dokumentacijom.

Proverićemo vrednosti svake od relevantnih kolona, pamtiti indekse onih redova koji krše zadato pravilo, zatim ćemo proći kroz čitav dataset i za svaku kolonu i odgovarajuće indekse pretvoriti “nelogične” vrednosti u NA.

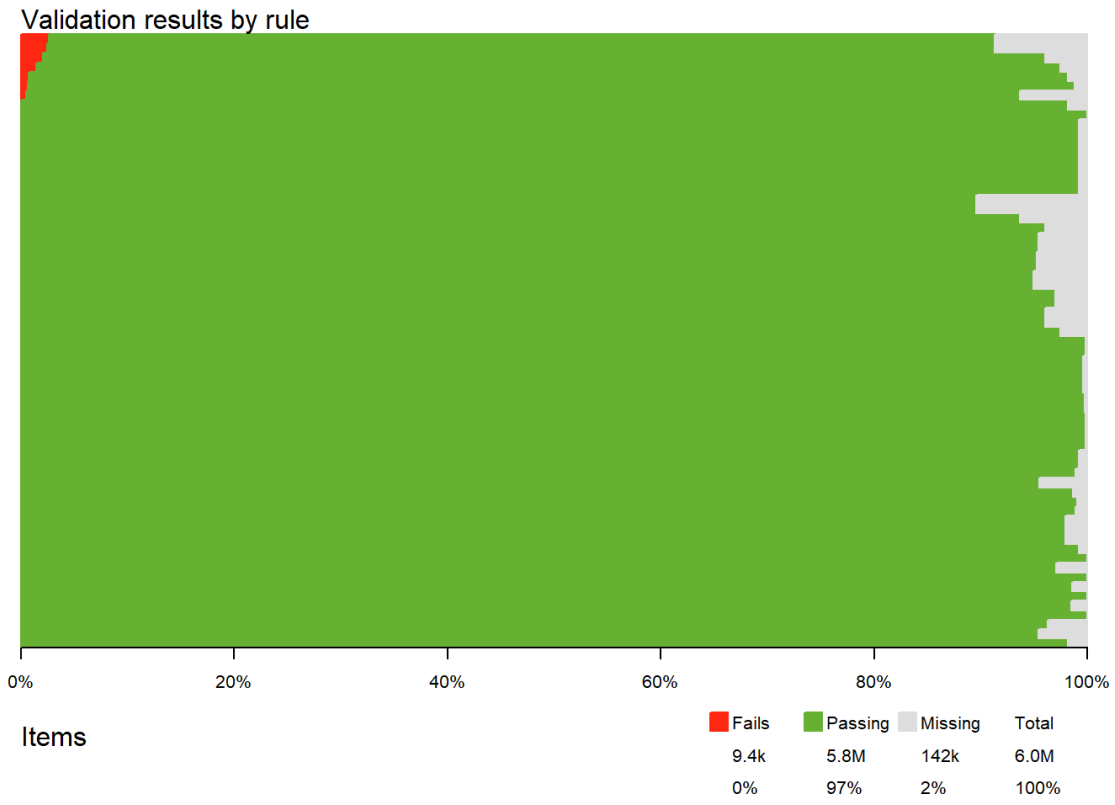
```
any(is.na(dataset$encounter_id))
## [1] FALSE
```

```
all(!duplicated(dataset$encounter_id))
## [1] TRUE
```

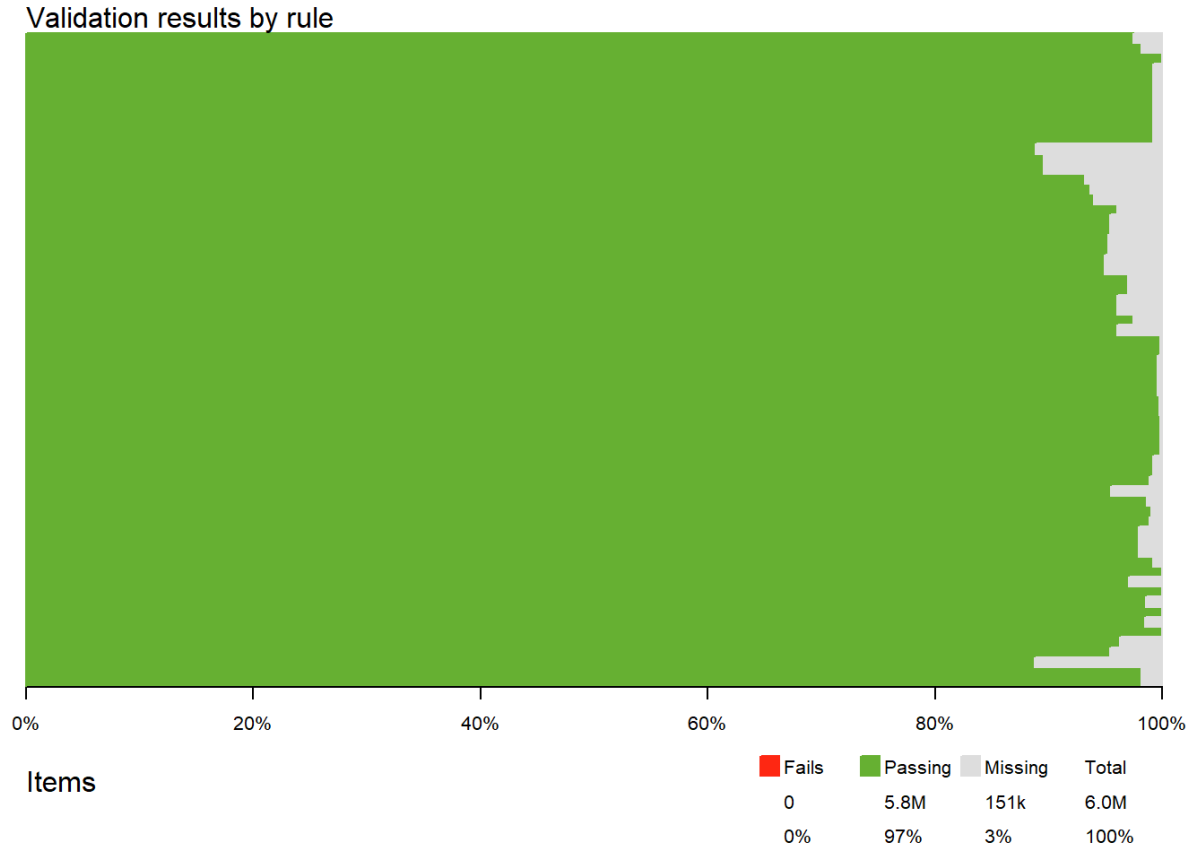
```
#ne postoji nijedna nedostajuća vrednost za encounter_id i svaka vrednost j
e jedinstvena, pa je mozemo u nastavku koristiti

rules <- validator(  "apache_3j_diagnosis" = dataset$apache_3j_diagnosis  >
= 100 & dataset$apache_3j_diagnosis  <= 3000
                   ,"apache_2_diagnosis"= dataset$apache_2_diagnosis >= 10
0 & dataset$apache_2_diagnosis <= 3000
                   ,"apache_4a_hospital_death_prob"=dataset$apache_4a_hosp
ital_death_prob >= 0 & dataset$apache_4a_hospital_death_prob <= 1
                   ,"age"=dataset$age >= 0 & dataset$age < 130
                   ,"bmi"=dataset$bmi >= 0 & dataset$bmi <= 200
                   ,"elective_surgery" = dataset$elective_surgery == 0 | d
ataset$elective_surgery == 1
                   ,"ethnicity"= dataset$ethnicity == "Caucasian" | datase
t$ethnicity == "Hispanic" | dataset$ethnicity == "African American" | datas
et$ethnicity == "Asian" |dataset$ethnicity == "Native American" | dataset$e
thnicity == "Other/Unknown" #moramo ovako za stringove inace petlja ne radi
:(
                   ,"gender" = dataset$gender == 'F' | dataset$gender == '
M'
                   ,"height" = dataset$height >= 0 & dataset$height <= 280
                   ,"icu_type" = dataset$icu_type == "CTICU" | dataset$icu
_type == "Med-Surg ICU" | dataset$icu_type == "CCU-CTICU" | dataset$icu_typ
e == "Neuro ICU" | dataset$icu_type == "MICU" | dataset$icu_type == "SICU"
| dataset$icu_type == "Cardiac ICU" | dataset$icu_type == "CSICU"
                   , "weight" = dataset$weight >= 0 & dataset$weight < 640
                   ,"apache_post_operative" = dataset$apache_post_operativ
e == 0 | dataset$apache_post_operative == 1
                   ,"arf_apache" =  dataset$arf_apache == 0 | dataset$arf_
apache == 1
                   ,"gcs_eyes_apache" = dataset$gcs_eyes_apache >= 1 & dat
aset$gcs_eyes_apache <= 4
                   ,"gcs_verbal_apache" = dataset$gcs_verbal_apache >= 1 &
dataset$gcs_verbal_apache <= 5
                   ,"gcs_motor_apache" = dataset$gcs_motor_apache >= 1 & d
ataset$gcs_motor_apache <= 6
                   , "gcs_unable_apache" =  dataset$gcs_unable_apache == 0
| dataset$gcs_unable_apache == 1
                   ,"heart_rate_apache" = dataset$heart_rate_apache >= 0 &
dataset$heart_rate_apache <= 350
                   ,"resprate_apache" = dataset$resprate_apache >= 0 & dat
aset$resprate_apache <= 200
                   ,"temp_apache" = dataset$temp_apache >= 0 & dataset$tem
p_apache <= 47
                   ,"map_apache" = dataset$map_apache >= 0 & dataset$map_a
pache <= 370
                   ,"intubated_apache" = dataset$intubated_apache == 0 | d
ataset$intubated_apache == 1
```

```
                    ,"ventilated_apache" =  dataset$ventilated_apache == 0
| dataset$ventilated_apache == 1
                    ,"d1_diasbp_max" = dataset$d1_diasbp_max >= 0 & dataset
$d1_diasbp_max <= 370
                    ,"d1_diasbp_min" = dataset$d1_diasbp_min >= 0 & dataset
$d1_diasbp_min <= 370
                    ,"d1_heartrate_max" = dataset$d1_heartrate_max >= 0 & d
ataset$d1_heartrate_max <= 350
                    ,"d1_heartrate_min" = dataset$d1_heartrate_min >= 0 & d
ataset$d1_heartrate_min <= 350
                    ,"d1_mbp_max" = dataset$d1_mbp_max >= 0 & dataset$d1_mb
p_max <= 370
                    ,"d1_mbp_min" = dataset$d1_mbp_min >= 0 & dataset$d1_mb
p_min <= 370
                    ,"d1_resprate_max" = dataset$d1_resprate_max >= 0 & dat
aset$d1_resprate_max <= 200
                    ,"d1_resprate_min" = dataset$d1_resprate_min >= 0 & dat
aset$d1_resprate_min <= 200
                    ,"d1_spo2_max" = dataset$d1_spo2_max >= 0 & dataset$d1_
spo2_max <= 100
                    ,"d1_spo2_min" = dataset$d1_spo2_min >= 0 & dataset$d1_
spo2_min <= 100
                    ,"d1_sysbp_max" = dataset$d1_sysbp_max >= 0 & dataset$d
1_sysbp_max <= 300
                    ,"d1_sysbp_min" = dataset$d1_sysbp_min >=40 & dataset$d
1_sysbp_min <= 160
                    , "d1_temp_max" = dataset$d1_temp_max >=36 & dataset$d1
_temp_max <= 41
                    , "d1_temp_min" = dataset$d1_temp_min >=31 & dataset$d1
_temp_min <= 38
                    , "h1_diasbp_max" = dataset$h1_diasbp_max >= 37 & datas
et$h1_diasbp_max < 150
                    , "h1_diasbp_min" = dataset$h1_diasbp_min >= 22 & datas
et$h1_diasbp_min <= 115
                    , "h1_heartrate_max" = dataset$h1_heartrate_max >= 46 &
dataset$h1_heartrate_max <= 164
                    , "h1_heartrate_min" = dataset$h1_heartrate_min >= 36 &
dataset$h1_heartrate_min <= 144
                    , "h1_mbp_max" = dataset$h1_mbp_max >= 49 & dataset$h1_
mbp_max <= 165
                    , "h1_mbp_min" = dataset$h1_mbp_min >= 32 & dataset$h1_
mbp_min <= 138
                    , "h1_resprate_max" = dataset$h1_resprate_max >= 10 & d
ataset$h1_resprate_max < 100
                    , "h1_resprate_min" = dataset$h1_resprate_min >= 0 & da
taset$h1_resprate_min < 200
                    , "h1_spo2_max" = dataset$h1_spo2_max >= 0 & dataset$h1
_spo2_max <= 100
                    , "h1_spo2_min" = dataset$h1_spo2_min >= 0 & dataset$h1
_spo2_min <= 100
                    , "h1_sysbp_max" = dataset$h1_sysbp_max >= 75 & dataset
$h1_sysbp_max <= 223
```

```
                    , "h1_sysbp_min" = dataset$h1_sysbp_min >= 53 & dataset
$h1_sysbp_min <= 175
                    , "d1_glucose_max" = dataset$d1_glucose_max >= 73 & dat
aset$d1_glucose_max < 620
                    , "d1_glucose_min" = dataset$d1_glucose_min >= 33 & dat
aset$d1_glucose_min < 288
                    , "d1_potassium_max" = dataset$d1_potassium_max >= 2.5
& dataset$d1_potassium_max <= 7
                    , "d1_potassium_min" = dataset$d1_potassium_min >= 2.3
& dataset$d1_potassium_min <= 6
                    , "apache_4a_icu_death_prob"=dataset$apache_4a_icu_deat
h_prob >= 0 & dataset$apache_4a_icu_death_prob <= 1
                    , "aids" = dataset$aids == 0 | dataset$aids == 1
                    , "cirrhosis" = dataset$cirrhosis == 0 | dataset$cirrho
sis == 1
                    , "diabetes_mellitus" = dataset$diabetes_mellitus == 0
| dataset$diabetes_mellitus == 1
                    , "hepatic_failure " = dataset$hepatic_failure == 0 | d
ataset$hepatic_failure == 1
                    , "immunosuppression" = dataset$immunosuppression == 0
| dataset$immunosuppression == 1
                    , "leukemia" = dataset$leukemia == 0 | dataset$leukemia
== 1
                    , "lymphoma" = dataset$lymphoma == 0 | dataset$lymphoma
== 1
                    , "solid_tumor_with_metastasis" = dataset$solid_tumor_w
ith_metastasis == 0 | dataset$solid_tumor_with_metastasis == 1
                    , "hospital_death" = dataset$hospital_death == 0 | data
set$hospital_death == 1
                    , "apache_2_bodysystem"= dataset$apache_2_bodysystem ==
"Cardiovascular" | dataset$apache_2_bodysystem == "Respiratory" | dataset$a
pache_2_bodysystem == "Metabolic" | dataset$apache_2_bodysystem == "Trauma"
| dataset$apache_2_bodysystem == "Neurologic" | dataset$apache_2_bodysystem
== "Gastrointestinal" | dataset$apache_2_bodysystem == "Renal/Genitourinary
" | dataset$apache_2_bodysystem == "Undefined diagnoses" | dataset$apache_2
_bodysystem == "Haematologic" | dataset$apache_2_bodysystem == "Undefined D
iagnoses"
                    ,  "apache_3j_bodysystem"= dataset$apache_3j_bodysystem
== "Cardiovascular" | dataset$apache_3j_bodysystem == "Respiratory" | datas
et$apache_3j_bodysystem == "Metabolic" | dataset$apache_3j_bodysystem == "T
rauma" | dataset$apache_3j_bodysystem == "Neurological" | dataset$apache_3j
_bodysystem == "Gastrointestinal" | dataset$apache_3j_bodysystem == "Genito
urinary" | dataset$apache_3j_bodysystem == "Musculoskeletal/Skin" | dataset
$apache_3j_bodysystem == "Haematological" | dataset$apache_3j_bodysystem ==
"Sepsis" | dataset$apache_3j_bodysystem == "Gynecological")


output<- confront(dataset, rules)
plot(output)
```




```
set_values_to_na <- function(dataset, column_name, ids) {
  rows_to_update <- dataset$encounter_id %in% ids
  dataset[rows_to_update, column_name] <- NA
  return(dataset)}
rules_len<- length(rules)
for(i in 1:rules_len)
{
  column_name <- names(rules)[i]
  column_name
  string_without_dot <- sub("\\.$", "", column_name)
  string_without_dot
  violating_rows<-violating(dataset, rules[i])
  violating_ids<-violating_rows$encounter_id
  dataset <- set_values_to_na(dataset, string_without_dot, violating_ids)}

output <- confront(dataset, rules)
plot(output)
```




```
dataset <- subset(dataset, select = -c(encounter_id))
```

Nedostajuće vrednosti

Posvetićemo se NA vrednostima kojih ima uglavnom ispod 9%.

```
jedinstvene_vrednosti <- sapply(dataset, n_distinct)
jedinstvene_vrednosti
```

```
##                          age                          bmi
##                           75                        34889
##              elective_surgery                    ethnicity
##                            2                            7
##                       gender                       height
##                            3                          402
##                     icu_type                       weight
##                            8                         3410
##           apache_2_diagnosis          apache_3j_diagnosis
##                           45                          382
##         apache_post_operative                  arf_apache
##                            2                            3
##               gcs_eyes_apache              gcs_motor_apache
##                            5                            7
##             gcs_unable_apache             gcs_verbal_apache
##                            3                            6
##             heart_rate_apache              intubated_apache
##                          150                            3
##                   map_apache              resprate_apache
##                          162                           75
##                  temp_apache             ventilated_apache
```

```
##                            192                              3
##                d1_diasbp_max                 d1_diasbp_min
##                            121                             79
##             d1_heartrate_max              d1_heartrate_min
##                            121                            155
##                  d1_mbp_max                    d1_mbp_min
##                            126                             92
##              d1_resprate_max               d1_resprate_min
##                             80                             56
##                  d1_spo2_max                   d1_spo2_min
##                             44                            102
##                 d1_sysbp_max                  d1_sysbp_min
##                            144                            121
##                  d1_temp_max                   d1_temp_min
##                            171                            210
##                h1_diasbp_max                 h1_diasbp_min
##                            108                             93
##             h1_heartrate_max              h1_heartrate_min
##                            120                            110
##                  h1_mbp_max                    h1_mbp_min
##                            118                            108
##              h1_resprate_max               h1_resprate_min
##                             51                             92
##                  h1_spo2_max                   h1_spo2_min
##                             73                            101
##                 h1_sysbp_max                  h1_sysbp_min
##                            150                            124
##               d1_glucose_max                d1_glucose_min
##                            539                            256
##             d1_potassium_max              d1_potassium_min
##                            101                            117
## apache_4a_hospital_death_prob     apache_4a_icu_death_prob
##                            101                             99
##                           aids                     cirrhosis
##                              3                              3
##             diabetes_mellitus               hepatic_failure
##                              3                              3
##             immunosuppression                      leukemia
##                              3                              3
##                       lymphoma   solid_tumor_with_metastasis
##                              3                              3
##          apache_3j_bodysystem           apache_2_bodysystem
##                             11                             11
##                hospital_death
##                              2
```

```
xtabs(~ ethnicity, data = dataset)
## ethnicity
## African American            Asian        Caucasian         Hispanic
##             9547             1129            70684             3796
##  Native American    Other/Unknown
##              788             4374
```

Možemo da primetimo da je *Caucasian* etička pripadnost koja je zastupljena kod skoro 80% pacijenata, tako da ćemo NA vrednosti zameniti tim podatkom.

```
dataset <- dataset %>%
```

```
  mutate(ethnicity = ifelse(is.na(ethnicity), "Caucasian", ethnicity))
```

Na ovaj način ćemo da nadomestimo ostale NA vrednosti s obzirom na to da uglavnom nema kolona koje imaju NA vrednosti preko 5%.

```
xtabs(~ gender, data = dataset)
```

```
## gender
##     F     M
## 42219 49469
```

Ima 54% procenata muškaraca i samo 25 nedostajućih vrednosti za gender.

```
dataset <- dataset %>%
  mutate(gender = ifelse(is.na(gender), "M", gender))
```

```
xtabs(~ apache_2_bodysystem, data = dataset)
```

```
## apache_2_bodysystem
##      Cardiovascular    Gastrointestinal        Haematologic           Metabolic
##               38816                9026                 638                7650
##          Neurologic Renal/Genitourinary         Respiratory             Trauma
##               11896                2460               11609                3842
## Undefined diagnoses Undefined Diagnoses
##                3768                 346
```

Možemo da primetimo da je *Cardiovascular* grupa za prijemnu dijagnostiku APACHE II koja je zastupljena kod skoro 40% pacijenata.

```
dataset <- dataset %>%
  mutate(apache_2_bodysystem = ifelse(is.na(apache_2_bodysystem), "Cardiova
scular", apache_2_bodysystem))
```

```
xtabs(~ apache_3j_bodysystem, data = dataset)
```

```
## apache_3j_bodysystem
##       Cardiovascular     Gastrointestinal        Genitourinary
##                29999                9026                2172
##        Gynecological           Metabolic Musculoskeletal/Skin
##                  313                7650                1166
##         Neurological         Respiratory              Sepsis
##                11896               11609               11740
##               Trauma
##                 3842
```

Možemo da primetimo da je *Cardiovascular* grupa za prijemnu dijagnostiku APACHE III koja je zastupljena kod skoro 30% pacijenata.

```
dataset <- dataset %>%
  mutate(apache_3j_bodysystem = ifelse(is.na(apache_3j_bodysystem), "Cardio
vascular", apache_3j_bodysystem))
```

Kada je u pitanje feature *age* pacijente ćemo podeliti u grupe po životnom dobu kako bismo nadomestili NA vrednosti.

```
minimum <- min(dataset$age, na.rm = TRUE)
maximum <- max(dataset$age, na.rm = TRUE)
```

Vidimo da nam se godine pacijenata kreću između 16 i 89 godina. Što znači da pacijente možemo podeliti na sledeće kategorije *puberty*, *adolescent*, *adult*, *middle-age*, *pensioner*.

```
puberty <- seq(16,18,1)
adolescent <- seq(19,20,1)
adult <- seq(21,40,1)
middle_age <- seq(41,60,1)
pensioner <- seq(61,90,1)
```

```
dataset$age[ dataset$age %in% puberty ] <- "puberty"
dataset$age[ dataset$age %in% adolescent ] <- "adolescent"
dataset$age[ dataset$age %in% adult ] <- "adult"
dataset$age[ dataset$age %in% middle_age ] <- "middle_age"
dataset$age[ dataset$age %in% pensioner ] <- "pensioner"
```

```
xtabs(~ age, data = dataset)
## age
## adolescent      adult middle_age  pensioner    puberty
##        681       9314      25364      51697        429
```

Na ovaj način smo podelili feture godine na životne dobi pacijenta i primećujemo da najveći broj pacijenata su penzioneri tačnije između 60 i 90 godina. Tako da ćemo NA vrednosti popuniti tim podatkom.

```
dataset <- dataset %>%
  mutate(age = ifelse(is.na(age), "pensioner", age))
```

Na osnovu životne dobi i na osnovu područja sa kog dolazi (rase) možemo da odredimo prosečnu visinu i težinu pacijenta. Za stare osobe važi da izgube otprilike 2.5 cm visine, nezavisno od područja sa kog dolaze.

Prosečna visina

1.1. African American dečaka u pubertetu: oko 150 cm, u odraslom periodu: 180 cm, stare osobe: 177.5 cm
1.2. African American devojčice u pubertetu: oko 145 cm, u odraslom periodu: 170 cm, stare osobe: 167.5 cm
2.1. Asian dečaka u pubertetu: oko 150 cm, u odraslom periodu: 170 cm, stare osobe: 167.5 cm
2.2. Asian devojčice u pubertetu: oko 145 cm, u odraslom periodu: 157 cm, stare osobe: 154.5 cm

3.1. Caucasian dečaka u pubertetu: oko 150 cm, u odraslom periodu: 180 cm, stare osobe: 177.5 cm
3.2. Caucasian devojčice u pubertetu: oko 145 cm, u odraslom periodu: 167 cm, stare osobe: 164.5 cm
4.1. Hispanic dečaka u pubertetu: oko 150 cm, u odraslom periodu: 173 cm, stare osobe: 170.5 cm
4.2. Hispanic devojčice u pubertetu: oko 145 cm, u odraslom periodu: 160 cm, stare osobe: 157.5 cm
5.1. Native American dečaka u pubertetu: oko 150 cm, u odraslom periodu: 177 cm, stare osobe: 174.5 cm
5.2. Native American devojčice u pubertetu: oko 145 cm, u odraslom periodu: 164 cm, stare osobe: 161.5 cm

Prosečna težina

1.1 African American dečaka u pubertetu: oko 40 kg, u odraslom periodu: 75 kg, stare osobe: 77 kg
1.2. African American devojčice u pubertetu: oko 37 kg, u odraslom periodu: 66 kg, stare osobe: 57 kg
2.1. Asian dečaka u pubertetu: oko 40 kg, u odraslom periodu: 67 kg, stare osobe: 70 kg
2.2. Asian devojčice u pubertetu: oko 37 kg, u odraslom periodu: 57 kg, stare osobe: 62 kg
3.1. Caucasian dečaka u pubertetu: oko 40 kg, u odraslom periodu: 77 kg, stare osobe: 80 kg
3.2. Caucasian devojčice u pubertetu: oko 37 kg, u odraslom periodu: 62 kg, stare osobe: 70 kg
4.1. Hispanic dečaka u pubertetu: oko 40 kg, u odraslom periodu: 77 kg, stare osobe: 80 kg
4.2. Hispanic devojčice u pubertetu: oko 37 kg, u odraslom periodu: 57 kg, stare osobe: 68 kg
5.1. Native American dečaka u pubertetu: oko 40 kg, u odraslom periodu: 77 kg, stare osobe: 70 kg
5.1. Native American devojčice u pubertetu: oko 37 kg, u odraslom periodu: 57 kg, stare osobe: 60 kg

! S obzirom na to da imamo Other/Unknown poreklo pacijenta, globalna prosečna visina i težina je najprihližnija *Caucasian* poreklu tako da ćemo iskoristiti te podatke.

Za početak se bavimo *height* feature-om:

```
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "African American") & (dataset$age == "puberty")] <- 150
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "African American") & (dataset$age == "pensioner")] <- 177.5
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "African American") & (dataset$age == "adolescent" | dataset$
age == "adult" | dataset$age == "middle_age")] <- 180


dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "African American") & (dataset$age == "puberty")] <- 145
```

```
dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "African American") & (dataset$age == "pensioner")] <- 167.5
dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "African American") & (dataset$age == "adolescent"| dataset$a
ge == "adult" | dataset$age == "middle_age")] <- 170
```

```
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "Asian") & (dataset$age == "puberty")] <- 150
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "Asian") & (dataset$age == "pensioner")] <- 167.5
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "Asian") & (dataset$age == "adolescent" | dataset$age == "adu
lt" | dataset$age == "middle_age")] <- 170


dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "Asian") & (dataset$age == "puberty")] <- 145
dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "Asian") & (dataset$age == "pensioner")] <- 154.5
dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "Asian") & (dataset$age == "adolescent" | dataset$age == "adu
lt" | dataset$age == "middle_age")] <- 157
```

```
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "Caucasian" | dataset$ethnicity == "Other/Unknown") & (datase
t$age == "puberty")] <- 150
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "Caucasian" | dataset$ethnicity == "Other/Unknown") & (datase
t$age == "pensioner")] <- 177.5
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "Caucasian" | dataset$ethnicity == "Other/Unknown") & (datase
t$age == "adolescent" | dataset$age == "adult" | dataset$age == "middle_age
")] <- 180


dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "Caucasian" | dataset$ethnicity == "Other/Unknown") & (datase
t$age == "puberty")] <- 145
dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "Caucasian" | dataset$ethnicity == "Other/Unknown") & (datase
t$age == "pensioner")] <- 164.5
dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "Caucasian" | dataset$ethnicity == "Other/Unknown") & (datase
t$age == "adolescent" | dataset$age == "adult" | dataset$age == "middle_age
")] <- 167
```

```
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "Hispanic") & (dataset$age == "puberty")] <- 150
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "Hispanic") & (dataset$age == "pensioner")] <- 170.5
```

```
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "Hispanic") & (dataset$age == "adolescent" | dataset$age == "
adult" | dataset$age == "middle_age")] <- 173


dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "Hispanic") & (dataset$age == "puberty")] <- 145
dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "Hispanic") & (dataset$age == "pensioner")] <- 157.5
dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "Hispanic") & (dataset$age == "adolescent" | dataset$age == "
adult" | dataset$age == "middle_age")] <- 160
```

```
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "Native American") & (dataset$age == "puberty")] <- 150
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "Native American") & (dataset$age == "pensioner")] <- 174.5
dataset$height[(is.na(dataset$height)) & (dataset$gender == "M") & (dataset
$ethnicity == "Native American") & (dataset$age == "adolescent" | dataset$a
ge == "adult" | dataset$age == "middle_age")] <- 177


dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "Native American") & (dataset$age == "puberty")] <- 145
dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "Native American") & (dataset$age == "pensioner")] <- 161.5
dataset$height[(is.na(dataset$height)) & (dataset$gender == "F") & (dataset
$ethnicity == "Native American") & (dataset$age == "adolescent" | dataset$a
ge == "adult" | dataset$age == "middle_age")] <- 164
```

Kada smo završili sa visinom, na isti način ćemo da rešimo problem NA vrednosti kod feature-a *weight*:

```
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "African American") & (dataset$age == "puberty")] <- 40.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "African American") & (dataset$age == "pensioner")] <- 77.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "African American") & (dataset$age == "adolescent" | dataset$
age == "adult" | dataset$age == "middle_age")] <- 75.00


dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "African American") & (dataset$age == "puberty")] <- 37.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "African American") & (dataset$age == "pensioner")] <- 57.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "African American") & (dataset$age == "adolescent" | dataset$
age == "adult" | dataset$age == "middle_age")] <- 66.00
```

```
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "Asian") & (dataset$age == "puberty")] <- 40.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "Asian") & (dataset$age == "pensioner")] <- 70.00
```

```
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "Asian") & (dataset$age == "adolescent" | dataset$age == "adu
lt" | dataset$age == "middle_age")] <- 67


dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "Asian") & (dataset$age == "puberty")] <- 37.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "Asian") & (dataset$age == "pensioner")] <- 62.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "Asian") & (dataset$age == "adolescent" | dataset$age == "adu
lt" | dataset$age == "middle_age")] <- 57.00
```

```
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "Caucasian" | dataset$ethnicity == "Other/Unknown") & (datase
t$age == "puberty")] <- 40.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "Caucasian" | dataset$ethnicity == "Other/Unknown") & (datase
t$age == "pensioner")] <- 80.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "Caucasian" | dataset$ethnicity == "Other/Unknown") & (datase
t$age == "adolescent" | dataset$age == "adult" | dataset$age == "middle_age
")] <- 77.00


dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "Caucasian" | dataset$ethnicity == "Other/Unknown") & (datase
t$age == "puberty")] <- 37.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "Caucasian" | dataset$ethnicity == "Other/Unknown") & (datase
t$age == "pensioner")] <- 70.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "Caucasian" | dataset$ethnicity == "Other/Unknown") & (datase
t$age == "adolescent" | dataset$age == "adult" | dataset$age == "middle_age
")] <- 62.00
```

```
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "Hispanic") & (dataset$age == "puberty")] <- 40.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "Hispanic") & (dataset$age == "pensioner")] <- 80.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "Hispanic") & (dataset$age == "adolescent" | dataset$age == "
adult" | dataset$age == "middle_age")] <- 77.00


dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "Hispanic") & (dataset$age == "puberty")] <- 37.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "Hispanic") & (dataset$age == "pensioner")] <- 68.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "Hispanic") & (dataset$age == "adolescent" | dataset$age == "
adult" | dataset$age == "middle_age")] <- 57.00
```

```
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "Native American") & (dataset$age == "puberty")] <- 40.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "Native American") & (dataset$age == "pensioner")] <- 77.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "M") & (dataset
$ethnicity == "Native American") & (dataset$age == "adolescent" | dataset$a
ge == "adult" | dataset$age == "middle_age")] <- 70.00


dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "Native American") & (dataset$age == "puberty")] <- 37.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "Native American") & (dataset$age == "pensioner")] <- 60.00
dataset$weight[(is.na(dataset$weight)) & (dataset$gender == "F") & (dataset
$ethnicity == "Native American") & (dataset$age == "adolescent" | dataset$a
ge == "adult" | dataset$age == "middle_age")] <- 57.00
```

Na osnovu *height* i *weight* feature-a možemo da izračunamo BMI(Body mass index) na sledeći način: telesna masa(kg) / visina(m)^2.

```
dataset$bmi[is.na(dataset$bmi)] = dataset$weight / (dataset$height/100)^2
```

```
summary(dataset$bmi)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.92   23.66   27.63   29.17   32.90   68.24
```

1. d1_diasbp_max - najviši dijastolni krvni pritisak pacijenta tokom prva 24 sata boravka u odeljenju, bilo invazivno ili neinvazivno meren
2. d1_diasbp_min - najniži dijastolni krvni pritisak pacijenta tokom prva 24 sata boravka u odeljenju, bilo invazivno ili neinvazivno meren

Što se tiče podataka za pritisak, imamo podatke za izmeren najviši donji (dijastolni) i najviši gornji (sistolni) pritisak, meren u toku 1h (h1) i u toku 24h (d1) boravka na odeljenju.

```
sum(is.na(dataset$d1_diasbp_max))
## [1] 165
sum(is.na(dataset$d1_diasbp_min))
## [1] 165
```

Vidimo da ima 165 nedostajućih vrednosti za d1_diasbp_max i d1_diasbp_min. Na krvni pritisak najviše utiče starost pacijenta i bmi. Prvo što ćemo odraditi jeste da ćemo dodati novu kolonu koja će predstavljati bmi kao kategorijsku promenljivu.

1. BMI manje od 18.5: Nedovoljna težina
2. BMI 18.5 - 24.9: Normalna težina
3. BMI 25.0 - 29.9: Prekomerna težina
4. BMI 30.0 i više: Gojaznost

```
dataset$BMI_category <- ifelse(dataset$bmi < 18.5, "underweight",
                           ifelse(dataset$bmi < 25.0, "normal weight",
                           ifelse(dataset$bmi < 30.0, "overweight","obesi
ty")))
```

Sada možemo iskoristiti grupe po godinama i bmi po kategorijama kako bismo odredili srednje vrednosti d1_diasbp_max za svaku od kombinacija kategorija.

```
group_bmi_age_diasbp_max <- aggregate(d1_diasbp_max ~ BMI_category + age, d
ata = dataset, FUN = mean, na.rm = TRUE)
```

1. d1_diasbp_max - za penzioneri, odrasli, adolescenti, osobe u srednjim godinama, osobe u pubertetu

```
#penzioneri
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "pe
nsioner" & dataset$BMI_category == "underweight")]<-86
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "pe
nsioner" & dataset$BMI_category == "normal weight")] <-87
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "pe
nsioner" & dataset$BMI_category == "overweight")] <-87
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "pe
nsioner" & dataset$BMI_category == "obesity")]<-88

#odrasli
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "ad
ult" & dataset$BMI_category == "underweight")]<-88
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "ad
ult" & dataset$BMI_category == "normal weight")] <-88
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "ad
ult" & dataset$BMI_category == "overweight")] <-90
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "ad
ult" & dataset$BMI_category == "obesity")]<-92

#adolescenti.
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "ad
olescent" & dataset$BMI_category == "underweight")]<-82
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "ad
olescent" & dataset$BMI_category == "normal weight")] <-83
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "ad
olescent" & dataset$BMI_category == "overweight")] <-85
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "ad
olescent" & dataset$BMI_category == "obesity")]<-87

#osobe u srednjim godinama
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "mi
ddle_age" & dataset$BMI_category == "underweight")]<-90
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "mi
ddle_age" & dataset$BMI_category == "normal weight")] <-90
```

```
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "mi
ddle_age" & dataset$BMI_category == "overweight")] <-91
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "mi
ddle_age" & dataset$BMI_category == "obesity")]<-92


#osobe u pubertetu
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "pu
berty" & dataset$BMI_category == "underweight")]<-80
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "pu
berty" & dataset$BMI_category == "normal weight")] <-82
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "pu
berty" & dataset$BMI_category == "overweight")] <-86
dataset$d1_diasbp_max[(is.na(dataset$d1_diasbp_max) & dataset$age == "pu
berty" & dataset$BMI_category == "obesity")]<-85
```

2. d1_diasbp_min - za penzioneri, odrasli, adolescenti, osobe u srednjim godinama, osobe u pubertetu

```
group_bmi_age_diasbp_min <- aggregate(d1_diasbp_min ~ BMI_category + age
, data = dataset, FUN = mean, na.rm = TRUE)
```

```
#penzioneri
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "pensi
oner" & dataset$BMI_category == "underweight")]<-47
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "pensi
oner" & dataset$BMI_category == "normal weight")]<-48
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "pensi
oner" & dataset$BMI_category == "overweight")] <-48
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "pensi
oner" & dataset$BMI_category == "obesity")]<-48


#osdrasli
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "adult
" & dataset$BMI_category == "underweight")]<-53
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "adult
" & dataset$BMI_category == "normal weight")] <-54
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "adult
" & dataset$BMI_category == "overweight")] <-55
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "adult
" & dataset$BMI_category == "obesity")]<-55


#adolescenti
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "adole
scent" & dataset$BMI_category == "underweight")]<-49
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "adole
scent" & dataset$BMI_category == "normal weight")] <-50
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "adole
scent" & dataset$BMI_category == "overweight")] <-49
```

```
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "adole
scent" & dataset$BMI_category == "obesity")]<-52

#osobe u srednjim godinama
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "middl
e_age" & dataset$BMI_category == "underweight")]<-54
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "middl
e_age" & dataset$BMI_category == "normal weight")] <-54
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "middl
e_age" & dataset$BMI_category == "overweight")] <-55
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "middl
e_age" & dataset$BMI_category == "obesity")]<-53

#osobe u pubertetu
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "puber
ty" & dataset$BMI_category == "underweight")]<-48
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "puber
ty" & dataset$BMI_category == "normal weight")] <-48
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "puber
ty" & dataset$BMI_category == "overweight")] <-47
dataset$d1_diasbp_min[(is.na(dataset$d1_diasbp_min) & dataset$age == "puber
ty" & dataset$BMI_category == "obesity")]<-52
```

1. d1_heartrate_max - najveći broj otkucaja srca tokom prva 24 sata boravka na odeljenju
2. d1_heartrate_min - najmanji broj otkucaja srca tokom prva 24 sata boravka na odeljenju

Nakon istraživanja saznali smo da na broj otkucaja srca najviše utiče startost zato ćemo za nedostajuće vrednosti koristi srednju vrednost najvećeg broja otkucaja srca za svaku starosnu grupu.

```
sum(is.na(dataset$d1_heartrate_max))
## [1] 145
sum(is.na(dataset$d1_heartrate_min))
## [1] 145
```

Imamo 145 nedostajućih vrednosti za d1_heartrate_min i d1_heartrate_max.

```
group_age_heartrate_max <- aggregate(d1_heartrate_max ~ age, data = dataset
, FUN = mean,na.rm = TRUE)
group_age_heartrate_min <- aggregate(d1_heartrate_min ~ age, data = dataset
, FUN = mean,na.rm = TRUE)
```

```
#d1_heartrate_max
dataset$d1_heartrate_max[(is.na(dataset$d1_heartrate_max) & dataset$age ==
"puberty")]<-113
dataset$d1_heartrate_max[(is.na(dataset$d1_heartrate_max) & dataset$age ==
"middle_age")] <-104
```

```
dataset$d1_heartrate_max[(is.na(dataset$d1_heartrate_max) & dataset$age ==
"adolescent")] <-113
dataset$d1_heartrate_max[(is.na(dataset$d1_heartrate_max) & dataset$age ==
"adult")]<-110
dataset$d1_heartrate_max[(is.na(dataset$d1_heartrate_max) & dataset$age ==
"pensioner")]<-101


#d1_heartrate_min
group_age_heartrate_min <- aggregate(d1_heartrate_min ~ age, data = dataset
, FUN = mean,na.rm = TRUE)
dataset$d1_heartrate_min[(is.na(dataset$d1_heartrate_min) & dataset$age ==
"puberty")]<-75
dataset$d1_heartrate_min[(is.na(dataset$d1_heartrate_min) & dataset$age ==
"middle_age")]<-72
dataset$d1_heartrate_min[(is.na(dataset$d1_heartrate_min) & dataset$age ==
"adolescent")] <-75
dataset$d1_heartrate_min[(is.na(dataset$d1_heartrate_min) & dataset$age ==
"adult")]<-75
dataset$d1_heartrate_min[(is.na(dataset$d1_heartrate_min) & dataset$age ==
"pensioner")]<-69
```

1. d1_resprate_max - najveća brzina disanja izmerena tokom prva 24 sata na odeljenju
2. d1_resprate_min - najmanja brzina disanja izmerena tokom prva 24 sata na odeljenju

```
#d1_resprate_max
group_age_respirate_max <- aggregate(d1_resprate_max ~ age, data = datas
et, FUN = mean,na.rm = TRUE)
dataset$d1_resprate_max[(is.na(dataset$d1_resprate_max) & dataset$age ==
"puberty")]<-27
dataset$d1_resprate_max[(is.na(dataset$d1_resprate_max) & dataset$age ==
"middle_age")]<-29
dataset$d1_resprate_max[(is.na(dataset$d1_resprate_max) & dataset$age ==
"adolescent")] <-28
dataset$d1_resprate_max[(is.na(dataset$d1_resprate_max) & dataset$age ==
"adult")]<-29
dataset$d1_resprate_max[(is.na(dataset$d1_resprate_max) & dataset$age ==
"pensioner")]<-29


#d1_resprate_min
group_age_respirate_min <- aggregate(d1_resprate_min ~ age, data = datas
et, FUN = mean,na.rm = TRUE)
dataset$d1_resprate_min[(is.na(dataset$d1_resprate_min) & dataset$age ==
"puberty")]<-13
dataset$d1_resprate_min[(is.na(dataset$d1_resprate_min) & dataset$age ==
"middle_age")]<-12
dataset$d1_resprate_min[(is.na(dataset$d1_resprate_min) & dataset$age ==
"adolescent")] <-13
dataset$d1_resprate_min[(is.na(dataset$d1_resprate_min) & dataset$age ==
"adult")]<-13
```

```
dataset$d1_resprate_min[(is.na(dataset$d1_resprate_min) & dataset$age ==
"pensioner")]<-13
```

1. d1_spo2_max - najveća saturacija pacijenta tokom prva 24 sata boravka na odeljenju
2. d1_spo2_min - najmanja saturacija pacijenta tokom prva 24 sata boravka na odeljenju

```
sum(is.na(dataset$ d1_spo2_max))
## [1] 333
sum(is.na(dataset$ d1_spo2_min))
## [1] 333
```

Imamo 333 nedostajućih vrednosti za d1_spo2_max i d1_spo2_min (približno 0.0036 od ukupnog broja podataka). I za ova dva feature-a možemo potražiti prosek po starosnim grupama.

```
#d1_spo2_max
group_age_spo2_max <- aggregate(d1_spo2_max ~ age, data = dataset, FUN = me
an,na.rm = TRUE)
dataset$d1_spo2_max[(is.na(dataset$d1_spo2_max) & dataset$age == "puberty")
]<-100
dataset$d1_spo2_max[(is.na(dataset$d1_spo2_max) & dataset$age == "middle_ag
e")]<-99
dataset$d1_spo2_max[(is.na(dataset$d1_spo2_max) & dataset$age == "adolescen
t")] <-100
dataset$d1_spo2_max[(is.na(dataset$d1_spo2_max) & dataset$age == "adult")]<
-99
dataset$d1_spo2_max[(is.na(dataset$d1_spo2_max) & dataset$age == "pensioner
")]<-99


#d1_spo2_min
group_age_spo2_min <- aggregate(d1_spo2_min ~ age, data = dataset, FUN = me
an,na.rm = TRUE)
dataset$d1_spo2_min[(is.na(dataset$d1_spo2_min) & dataset$age == "puberty")
]<-93
dataset$d1_spo2_min[(is.na(dataset$d1_spo2_min) & dataset$age == "middle_ag
e")]<-91
dataset$d1_spo2_min[(is.na(dataset$d1_spo2_min) & dataset$age == "adolescen
t")] <-93
dataset$d1_spo2_min[(is.na(dataset$d1_spo2_min) & dataset$age == "adult")]<
-92
dataset$d1_spo2_min[(is.na(dataset$d1_spo2_min) & dataset$age == "pensioner
")]<-90
```

1. d1_temp_max - najviša temperatura tela pacijenta izmerena tokom prva 24 sata, invazivno merena
2. d1_temp_min - najniža temperatura tela pacijenta izmerena tokom prva 24 sata

```
sum(is.na(dataset$d1_temp_min))
## [1] 2324
sum(is.na(dataset$d1_temp_max))
## [1] 3611
```

Za ove dve kolone nam fali oko 0.025% vrednosti. Popunićemo ih prosečnim vrednostima po godinama.

```
group_age_temp_max <- aggregate(d1_temp_max ~ age, data = dataset, FUN = mean,na.rm = TRUE)
dataset$d1_temp_max[(is.na(dataset$d1_temp_max) & dataset$age == "puberty")]<-37
dataset$d1_temp_max[(is.na(dataset$d1_temp_max) & dataset$age == "middle_age")]<-37
dataset$d1_temp_max[(is.na(dataset$d1_temp_max) & dataset$age == "adolescent")] <-37
dataset$d1_temp_max[(is.na(dataset$d1_temp_max) & dataset$age == "adult")]<-37
dataset$d1_temp_max[(is.na(dataset$d1_temp_max) & dataset$age == "pensioner")]<-37

group_age_temp_min <- aggregate(d1_temp_min ~ age, data = dataset, FUN = mean,na.rm = TRUE)
dataset$d1_temp_min[(is.na(dataset$d1_temp_min) & dataset$age == "puberty")]<-36
dataset$d1_temp_min[(is.na(dataset$d1_temp_min) & dataset$age == "middle_age")]<-36
dataset$d1_temp_min[(is.na(dataset$d1_temp_min) & dataset$age == "adolescent")] <-36
dataset$d1_temp_min[(is.na(dataset$d1_temp_min) & dataset$age == "adult")]<-36
dataset$d1_temp_min[(is.na(dataset$d1_temp_min) & dataset$age == "pensioner")]<-36
```

1. d1_sysbp_max - najviši sistolni krvni pritisak pacijenta tokom prva 24 sata, bilo neinvazivno ili invazivno meren
2. d1_sysbp_min - najniži sistolni krvni pritisak pacijenta tokom prva 24 sata, bilo neinvazivno ili invazivno meren

```
sum(is.na(dataset$d1_sysbp_max))
## [1] 159
sum(is.na(dataset$d1_sysbp_min))
## [1] 159
```

Vidimo da nam fali 159 za d1_sysbp_min i d1_sysbp_max,što je oko 0.0017% podataka, tako da ćemo primeniti isti princip kao i kod dijastolnog pritiska.

```
group_bmi_age_sysbp_max <- aggregate(d1_sysbp_max ~ BMI_category + age, dat
a = dataset, FUN = mean, na.rm = TRUE)
group_bmi_age_sysbp_min <- aggregate(d1_sysbp_min ~ BMI_category + age, dat
a = dataset, FUN = mean, na.rm = TRUE)
```

1. d1_sysbp_max - penzioneri, odrasli, adolescenti, osobe u srednjim godinama, osobe u pubertetu

```
#penzioneri
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "pens
ioner" & dataset$BMI_category == "underweight")]<-148
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "pens
ioner" & dataset$BMI_category == "normal weight")] <-149
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "pens
ioner" & dataset$BMI_category == "overweight")] <-150
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "pens
ioner" & dataset$BMI_category == "obesity")]<-151

#odrasli
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "adul
t" & dataset$BMI_category == "underweight")]<-134
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "adul
t" & dataset$BMI_category == "normal weight")] <-138
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "adul
t" & dataset$BMI_category == "overweight")] <-142
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "adul
t" & dataset$BMI_category == "obesity")]<-147

#adolescenti
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "adol
escent" & dataset$BMI_category == "underweight")]<-130
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "adol
escent" & dataset$BMI_category == "normal weight")] <-132
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "adol
escent" & dataset$BMI_category == "overweight")] <-137
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "adol
escent" & dataset$BMI_category == "obesity")]<-141

#osobe u srednjim godinama
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "midd
le_age" & dataset$BMI_category == "underweight")]<-143
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "midd
le_age" & dataset$BMI_category == "normal weight")] <-144
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "midd
le_age" & dataset$BMI_category == "overweight")] <-147
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "midd
le_age" & dataset$BMI_category == "obesity")]<-150

#osobe u pubertetu
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "pube
rty" & dataset$BMI_category == "underweight")]<-130
```

```
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "pube
rty" & dataset$BMI_category == "normal weight")] <-130
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "pube
rty" & dataset$BMI_category == "overweight")] <-140
dataset$d1_sysbp_max[(is.na(dataset$d1_sysbp_max) & dataset$age == "pube
rty" & dataset$BMI_category == "obesity")]<-143
```

2. d1_sysbp_min - penzioneri, odrasli, adolescenti, osobe u srednjim godinama, osobe u pubertetu

```
#penzioneri
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "pens
ioner" & dataset$BMI_category == "underweight")]<-92
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "pens
ioner" & dataset$BMI_category == "normal weight")] <-95
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "pens
ioner" & dataset$BMI_category == "overweight")] <-97
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "pens
ioner" & dataset$BMI_category == "obesity")]<-97

#odrasli
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "adul
t" & dataset$BMI_category == "underweight")]<-94
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "adul
t" & dataset$BMI_category == "normal weight")] <-98
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "adul
t" & dataset$BMI_category == "overweight")] <-100
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "adul
t" & dataset$BMI_category == "obesity")]<-102

#adolescenti
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "adol
escent" & dataset$BMI_category == "underweight")]<-93
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "adol
escent" & dataset$BMI_category == "normal weight")] <-97
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "adol
escent" & dataset$BMI_category == "overweight")] <-99
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "adol
escent" & dataset$BMI_category == "obesity")]<-104

#osobe u srednjim godinama
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "midd
le_age" & dataset$BMI_category == "underweight")]<-92
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "midd
le_age" & dataset$BMI_category == "normal weight")] <-96
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "midd
le_age" & dataset$BMI_category == "overweight")] <-98
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "midd
le_age" & dataset$BMI_category == "obesity")]<-99

#osobe u pubertetu
```

```
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "pube
rty" & dataset$BMI_category == "underweight")]<-95
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "pube
rty" & dataset$BMI_category == "normal weight")] <-96
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "pube
rty" & dataset$BMI_category == "overweight")] <-97
dataset$d1_sysbp_min[(is.na(dataset$d1_sysbp_min) & dataset$age == "pube
rty" & dataset$BMI_category == "obesity")]<-103
```

1. d1_mbp_max - najviši srednji krvi pritisak pacijenta tokom prva 24 sata na odeljenju, bilo neinvazivno ili invazivno meren
2. d1_mbp_min - najniži srednji krvi pritisak pacijenta tokom prva 24 sata na odeljenju, bilo neinvazivno ili invazivno meren

Konačno možemo odrediti i srednji i krvi pritisak pacijenta tokom prva 24 sata na odeljenju. Dobija se po sledećoj formuli: map = dbp + (sbp-dbp)/3. Ovu formulu ćemo primeniti za sve vrednosti, za slučaj da neke vrednosti nisu bile dobro izračunate.

```
dataset$d1_mbp_max <- dataset$d1_diasbp_max+(dataset$d1_sysbp_max-dataset$d
1_diasbp_max)/3
dataset$d1_mbp_min <- dataset$d1_diasbp_min+(dataset$d1_sysbp_min-dataset$d
1_diasbp_min)/3
```

Sada je vreme da sredimo ove feture za period od 1h. Način na koji ćemo to da uradimo se neće razlikovati od načina na koji smo to uradili za period od 24h.

1. h1_diasbp_max - najviši dijastolni krvni pritisak pacijenta tokom prva 24 sata, bilo neinvazivno ili invazivno meren
2. h1_diasbp_min - najniži dijastolni krvni pritisak pacijenta tokom prva 24 sata, bilo neinvazivno ili invazivno meren

```
sum(is.na(dataset$h1_diasbp_max))
## [1] 3619
sum(is.na(dataset$h1_diasbp_min))
## [1] 3619
```

Vidimo da ima 3619 nedostajućih vrednosti za h1_diasbp_max i h1_diasbp_min. Na krvni pritisak, kao što smo ranije zaključili, najviše utiče starost pacijenta i bmi. Sada možemo da koristimo varijablu koju smo već kreirali - BMI_category. Nema potrebe ponovo komentarisati sve korake, jer će proces sređivanja biti identičan kao prethodni za podatke merene toku 24h.

```
group_bmi_age_diasbp_max_h1 <- aggregate(h1_diasbp_max ~ BMI_category + age
, data = dataset, FUN = mean, na.rm = TRUE)
```

1. h1_diasbp_max - penzioneri, adolescenti, odrasle osobe, osobe u srednjim godinama i osobe u pubertetu.

```
#penzioneri
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "pe
nsioner" & dataset$BMI_category == "underweight")]<-86
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "pe
nsioner" & dataset$BMI_category == "normal weight")] <-87
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "pe
nsioner" & dataset$BMI_category == "overweight")] <-87
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "pe
nsioner" & dataset$BMI_category == "obesity")]<-88

#odrasli
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "ad
ult" & dataset$BMI_category == "underweight")]<-88
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "ad
ult" & dataset$BMI_category == "normal weight")] <-88
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "ad
ult" & dataset$BMI_category == "overweight")] <-90
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "ad
ult" & dataset$BMI_category == "obesity")]<-92

#adolescenti
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "ad
olescent" & dataset$BMI_category == "underweight")]<-82
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "ad
olescent" & dataset$BMI_category == "normal weight")] <-83
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "ad
olescent" & dataset$BMI_category == "overweight")] <-85
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "ad
olescent" & dataset$BMI_category == "obesity")]<-87

#osobe u srednjim godinama
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "mi
ddle_age" & dataset$BMI_category == "underweight")]<-90
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "mi
ddle_age" & dataset$BMI_category == "normal weight")] <-90
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "mi
ddle_age" & dataset$BMI_category == "overweight")] <-91
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "mi
ddle_age" & dataset$BMI_category == "obesity")]<-92

#osobe u pubertetu
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "pu
berty" & dataset$BMI_category == "underweight")]<-80
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "pu
berty" & dataset$BMI_category == "normal weight")] <-82
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "pu
berty" & dataset$BMI_category == "overweight")] <-86
```

```
dataset$h1_diasbp_max[(is.na(dataset$h1_diasbp_max) & dataset$age == "pu
berty" & dataset$BMI_category == "obesity")]<-85
```

2. h1_diasbp_min - penzioneri, adolescenti, odrasle osobe, osobe u srednjim godinama i osobe u pubertetu.

```
#penzioneri
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "pe
nsioner" & dataset$BMI_category == "underweight")]<-47
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "pe
nsioner" & dataset$BMI_category == "normal weight")] <-48
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "pe
nsioner" & dataset$BMI_category == "overweight")] <-48
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "pe
nsioner" & dataset$BMI_category == "obesity")]<-48

#odrasli
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "ad
ult" & dataset$BMI_category == "underweight")]<-53
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "ad
ult" & dataset$BMI_category == "normal weight")] <-54
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "ad
ult" & dataset$BMI_category == "overweight")] <-55
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "ad
ult" & dataset$BMI_category == "obesity")]<-55

#adolescenti
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "ad
olescent" & dataset$BMI_category == "underweight")]<-49
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "ad
olescent" & dataset$BMI_category == "normal weight")] <-50
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "ad
olescent" & dataset$BMI_category == "overweight")] <-49
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "ad
olescent" & dataset$BMI_category == "obesity")]<-52

#osobe u srednjim godinama
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "mi
ddle_age" & dataset$BMI_category == "underweight")]<-54
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "mi
ddle_age" & dataset$BMI_category == "normal weight")] <-54
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "mi
ddle_age" & dataset$BMI_category == "overweight")] <-55
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "mi
ddle_age" & dataset$BMI_category == "obesity")]<-53

#osobe u pubertetu
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "pu
berty" & dataset$BMI_category == "underweight")]<-48
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "pu
berty" & dataset$BMI_category == "normal weight")] <-48
```

```
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "puberty" & dataset$BMI_category == "overweight")] <-47
dataset$h1_diasbp_min[(is.na(dataset$h1_diasbp_min) & dataset$age == "puberty" & dataset$BMI_category == "obesity")]<-52
```

1. h1_heartrate_max - najveći broj otkucaja srca pacijenta tokom prvo sata boravka na odeljenju
2. h1_heartrate_min - najmanji broj otkucaja srca pacijenta tokom prvo sata boravka na odeljenju

```
sum(is.na(dataset$h1_heartrate_max))
## [1] 2790
sum(is.na(dataset$h1_heartrate_min))
## [1] 2790
```

Imamo 2790 nedostajućih vrednosti za *h1_heartrate_min* i *h1_heartrate_max*.

```
group_age_heartrate_max_h1 <- aggregate(h1_heartrate_max ~ age, data = dataset, FUN = mean,na.rm = TRUE)
group_age_heartrate_min_h1 <- aggregate(h1_heartrate_min ~ age, data = dataset, FUN = mean,na.rm = TRUE)
```

```
#h1_heartrate_max
dataset$h1_heartrate_max[(is.na(dataset$h1_heartrate_max) & dataset$age == "puberty")]<-113
dataset$h1_heartrate_max[(is.na(dataset$h1_heartrate_max) & dataset$age == "middle_age")] <-104
dataset$h1_heartrate_max[(is.na(dataset$h1_heartrate_max) & dataset$age == "adolescent")] <-113
dataset$h1_heartrate_max[(is.na(dataset$h1_heartrate_max) & dataset$age == "adult")]<-110
dataset$h1_heartrate_max[(is.na(dataset$h1_heartrate_max) & dataset$age == "pensioner")]<-101

#h1_heartrate_min
dataset$h1_heartrate_min[(is.na(dataset$h1_heartrate_min) & dataset$age == "puberty")]<-75
dataset$h1_heartrate_min[(is.na(dataset$h1_heartrate_min) & dataset$age == "middle_age")]<-72
dataset$h1_heartrate_min[(is.na(dataset$h1_heartrate_min) & dataset$age == "adolescent")] <-75
dataset$h1_heartrate_min[(is.na(dataset$h1_heartrate_min) & dataset$age == "adult")]<-75
dataset$h1_heartrate_min[(is.na(dataset$h1_heartrate_min) & dataset$age == "pensioner")]<-69
```

1. h1_resprate_max - najveća brzina disanja pacijenta tokom prvog sata boravka na odeljenju
2. h1_resprate_min - najniža brzina disanja pacijenta tokom prvog sata boravka na odeljenju

```
sum(is.na(dataset$h1_resprate_max))
## [1] 4357
sum(is.na(dataset$h1_resprate_min))
## [1] 4357
```

Imamo 4357 nedostajućih vrednosti za h1_resprate_min i h1_resprate_max.

```
group_age_respirate_max_h1 <- aggregate(h1_resprate_max ~ age, data = datas
et, FUN = mean,na.rm = TRUE)
group_age_respirate_min_h1 <- aggregate(h1_resprate_min ~ age, data = datas
et, FUN = mean,na.rm = TRUE)
```

```
#h1_resprate_max
dataset$h1_resprate_max[(is.na(dataset$h1_resprate_max) & dataset$age == "p
uberty")]<-27
dataset$h1_resprate_max[(is.na(dataset$h1_resprate_max) & dataset$age == "m
iddle_age")]<-29
dataset$h1_resprate_max[(is.na(dataset$h1_resprate_max) & dataset$age == "a
dolescent")] <-28
dataset$h1_resprate_max[(is.na(dataset$h1_resprate_max) & dataset$age == "a
dult")]<-29
dataset$h1_resprate_max[(is.na(dataset$h1_resprate_max) & dataset$age == "p
ensioner")]<-29


#h1_resprate_min
dataset$h1_resprate_min[(is.na(dataset$h1_resprate_min) & dataset$age == "p
uberty")]<-13
dataset$h1_resprate_min[(is.na(dataset$h1_resprate_min) & dataset$age == "m
iddle_age")]<-12
dataset$h1_resprate_min[(is.na(dataset$h1_resprate_min) & dataset$age == "a
dolescent")] <-13
dataset$h1_resprate_min[(is.na(dataset$h1_resprate_min) & dataset$age == "a
dult")]<-13
dataset$h1_resprate_min[(is.na(dataset$h1_resprate_min) & dataset$age == "p
ensioner")]<-13
```

1. h1_spo2_max - najveća saturacija kiseonikom tokom prvog sata boravka u jedinici
2. h1_spo2_min - najmanja saturacija kiseonikom tokom prvog sata boravka u jedinici

```
sum(is.na(dataset$h1_spo2_max))
## [1] 4185
sum(is.na(dataset$h1_spo2_min))
```

```
## [1] 4185
```

Imamo 4185 nedostajućih vrednosti za h1_spo2_min i h1_spo2_max.

```
group_age_spo2_max_h1 <- aggregate(h1_spo2_max ~ age, data = dataset, FUN = mean,na.rm = TRUE)
group_age_spo2_min_h1 <- aggregate(h1_spo2_min ~ age, data = dataset, FUN = mean,na.rm = TRUE)
```

```
#h1_spo2_max
dataset$h1_spo2_max[(is.na(dataset$h1_spo2_max) & dataset$age == "puberty")]<-100
dataset$h1_spo2_max[(is.na(dataset$h1_spo2_max) & dataset$age == "middle_age")]<-99
dataset$h1_spo2_max[(is.na(dataset$h1_spo2_max) & dataset$age == "adolescent")] <-100
dataset$h1_spo2_max[(is.na(dataset$h1_spo2_max) & dataset$age == "adult")]<-99
dataset$h1_spo2_max[(is.na(dataset$h1_spo2_max) & dataset$age == "pensioner")]<-99

#h1_spo2_min
dataset$h1_spo2_min[(is.na(dataset$h1_spo2_min) & dataset$age == "puberty")]<-93
dataset$h1_spo2_min[(is.na(dataset$h1_spo2_min) & dataset$age == "middle_age")]<-91
dataset$h1_spo2_min[(is.na(dataset$h1_spo2_min) & dataset$age == "adolescent")] <-93
dataset$h1_spo2_min[(is.na(dataset$h1_spo2_min) & dataset$age == "adult")]<-92
dataset$h1_spo2_min[(is.na(dataset$h1_spo2_min) & dataset$age == "pensioner")]<-90
```

1. h1_sysbp_max - najviši sistolni pritisak pacijenta tokom prvog sata borvaka na odeljenju, bilo neinvazivno ili invazivno meren
2. h1_sysbp_min - najniži sistolni pritisak pacijenta tokom prvog sata borvaka na odeljenju, bilo neinvazivno ili invazivno meren

```
sum(is.na(dataset$h1_sysbp_max))
## [1] 3611
sum(is.na(dataset$h1_sysbp_min))
## [1] 5477
```

Imamo 3611 nedostajućih vrednosti za h1_sysbp_min i 5477 h1_sysbp_max.

```
group_bmi_age_sysbp_max_h1 <- aggregate(h1_sysbp_max ~ BMI_category + age, data = dataset, FUN = mean, na.rm = TRUE)
```

1. h1_sysbp_max - penzioneri, odrasli, adolescenti, osobe u srednjim godinama, osobe u pubertetu.

```
#penzioneri
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "pens
ioner" & dataset$BMI_category == "underweight")]<-148
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "pens
ioner" & dataset$BMI_category == "normal weight")] <-149
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "pens
ioner" & dataset$BMI_category == "overweight")] <-150
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "pens
ioner" & dataset$BMI_category == "obesity")]<-151

#odrasli
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "adul
t" & dataset$BMI_category == "underweight")]<-134
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "adul
t" & dataset$BMI_category == "normal weight")] <-138
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "adul
t" & dataset$BMI_category == "overweight")] <-142
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "adul
t" & dataset$BMI_category == "obesity")]<-147

#adolescenti
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "adol
escent" & dataset$BMI_category == "underweight")]<-130
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "adol
escent" & dataset$BMI_category == "normal weight")] <-132
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "adol
escent" & dataset$BMI_category == "overweight")] <-137
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "adol
escent" & dataset$BMI_category == "obesity")]<-141

#osobe u srednjim godinama.
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "midd
le_age" & dataset$BMI_category == "underweight")]<-143
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "midd
le_age" & dataset$BMI_category == "normal weight")] <-144
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "midd
le_age" & dataset$BMI_category == "overweight")] <-147
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "midd
le_age" & dataset$BMI_category == "obesity")]<-150

#osobe u pubertetu.
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "pube
rty" & dataset$BMI_category == "underweight")]<-130
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "pube
rty" & dataset$BMI_category == "normal weight")] <-130
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "pube
rty" & dataset$BMI_category == "overweight")] <-140
```

```
dataset$h1_sysbp_max[(is.na(dataset$h1_sysbp_max) & dataset$age == "pube
rty" & dataset$BMI_category == "obesity")]<-143
```

2. h1_sysbp_min - penzioneri, odrasli, adolescenti, osobe u srednjim godinama, osobe u pubertetu.

```
group_bmi_age_sysbp_min_h1 <- aggregate(h1_sysbp_min ~ BMI_category + ag
e, data = dataset, FUN = mean, na.rm = TRUE)
```

```
#penzioneri
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "pension
er" & dataset$BMI_category == "underweight")]<-92
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "pension
er" & dataset$BMI_category == "normal weight")] <-95
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "pension
er" & dataset$BMI_category == "overweight")] <-97
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "pension
er" & dataset$BMI_category == "obesity")]<-97

#odrasli
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "adult"
& dataset$BMI_category == "underweight")]<-94
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "adult"
& dataset$BMI_category == "normal weight")] <-98
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "adult"
& dataset$BMI_category == "overweight")] <-100
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "adult"
& dataset$BMI_category == "obesity")]<-102

#adolescenti
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "adolesc
ent" & dataset$BMI_category == "underweight")]<-93
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "adolesc
ent" & dataset$BMI_category == "normal weight")] <-97
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "adolesc
ent" & dataset$BMI_category == "overweight")] <-99
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "adolesc
ent" & dataset$BMI_category == "obesity")]<-104

#osobe u srednjim godinama
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "middle_
age" & dataset$BMI_category == "underweight")]<-92
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "middle_
age" & dataset$BMI_category == "normal weight")] <-96
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "middle_
age" & dataset$BMI_category == "overweight")] <-98
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "middle_
age" & dataset$BMI_category == "obesity")]<-99
```

```
#osobe u pubertetu
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "puberty
" & dataset$BMI_category == "underweight")]<-95
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "puberty
" & dataset$BMI_category == "normal weight")] <-96
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "puberty
" & dataset$BMI_category == "overweight")] <-97
dataset$h1_sysbp_min[(is.na(dataset$h1_sysbp_min) & dataset$age == "puberty
" & dataset$BMI_category == "obesity")]<-103
```

1. h1_mbp_max - najviši srednji krvni pritisak pacijenta tokom prvog sata boravka na odeljenju, bilo neinvazivno ili invazivno meren
2. h1_mbp_min - najniži srednji krvni pritisak pacijenta tokom prvog sata boravka na odeljenju, bilo neinvazivno ili invazivno meren

Takođe ćemo formulu primeniti na sve vrednosti, u slučaju da ima grešaka prilikom računa.

```
dataset$h1_mbp_max <- dataset$h1_diasbp_max + (dataset$h1_sysbp_max - datas
et$h1_diasbp_max) / 3
dataset$h1_mbp_min <- dataset$h1_diasbp_min + (dataset$h1_sysbp_min - datas
et$h1_diasbp_min) / 3
```

1. aids - da li pacijent ima konačnu dijagnozu sindroma stečene imunodeficijencije(AIDS)(ne samo HIV pozitivan)
2. cirrhosis - bilo da pacijent ima istoriju teške upotrebe alkohola sa portonom hipertenzijom i varikozitetima, drugim uzorcima ciroze sa dokazima portne hipertenzije i varikoziteta ili cirozom dokazanom biopsijom. Ovaj komorbiditet se ne osnosi na pacijente sa funkcionalnom transplatacijom jetre
3. hepatic_failure - da li pacjient ima cirozu i dodatne komplikacije uključujući žuticu i ascites, krvarenje u gornjem delu gastroinfestilnog trakta, hepatičnu encefalopatiju ili komu
4. immunosuppression - da li je imuni sistem pacijenta ugrožen u preiodu od 6 meseci pre prijema na intezivnu negu iz bilo kog od sledećih razloga: terapija zračenjem, hemoterapija, upotreba necitotoksičnih imunosupresivnih lekova, visoke doze steroida (najmanje 0,3 mg/kg/dan metilprednizolona ili ekvivalent najmanje 6 meseci)
5. leukemia - da li je pacijentu dijagnostikovana akutna ili hronična mijelogena leukemija, akutna ili hronična limfocitna leukemija ili multipli mijelom
6. lymphoma - da li je pacijentu dijagnostikovan ne-Hodgkin limfom
7. solid_tumor_with_metastasis - da li je pacijentu dijagnostikovan bilo koji karcinom solidnog tumora (uključujući maligni melanom) koji ima dokaze o metastazama
8. diabetes_mellitus - da li je pacijentu dijagnostikovan dijabetes, bilo juvenilni ili adultni, koji zahteva lekove

Kada su u pitanju ovi fature-i, njihovi podaci se dobijaju određenim testovima, i našim istraživanjem smo zaključili da oni ne mogu da se prediktuju. Tako da ćemo pretpostaviti da podaci nedostaju u slučaju da nije bilo sumnje za testiranjem (tačnije

pretpostavlja se da nema ovih oboljenja) i da su rezultati testa negativni (tačnije da se nisu uneli negativni rezultati u tabelu).

```
#aids
dataset$aids[is.na(dataset$aids)] <- 0
#cirrhosis
dataset$cirrhosis[is.na(dataset$cirrhosis)] <- 0
#hepatic_failure
dataset$hepatic_failure[is.na(dataset$hepatic_failure)] <- 0
#immunosuppression
dataset$immunosuppression[is.na(dataset$immunosuppression)] <- 0
#leukemia
dataset$leukemia[is.na(dataset$leukemia)] <- 0
#lymphoma
dataset$lymphoma[is.na(dataset$lymphoma)] <- 0
#solid_tumor_with_metastasis
dataset$solid_tumor_with_metastasis[is.na(dataset$solid_tumor_with_metastas
is)] <- 0
#diabetes_mellitus
dataset$diabetes_mellitus[is.na(dataset$diabetes_mellitus)] <- 0
```

1. apache_2_diagnosis - APACHE II dijagnoza za prijem na intezivnu negu

Jasno nam je da vrednosti ove promenljive predstavlja kod koji se koristi za kategorizaciju bolesti sa kojom je pacijent primljen. Međutim, pošto dataset nije dosao ni sa kakvom dokumentacijom, ostaje na nama da zaključimo šta koji kod predstavlja.

Hajde da prvo vidimo sve različite vrednosti sa njom povezane apache_2_bodysystem promenljive.

```
na_counts <- dataset %>%
  summarise_all(~ sum(is.na(.)))
unique_body_systems <- unique(dataset$apache_2_bodysystem)
print(unique_body_systems)
```

```
##  [1] "Cardiovascular"      "Respiratory"         "Metabolic"
##  [4] "Trauma"              "Neurologic"          "Gastrointestinal"
##  [7] "Renal/Genitourinary" "Undefined diagnoses" "Haematologic"
## [10] "Undefined Diagnoses"
```

Vidimo da postoji 10 jedinstvenih vrednosti, međutim jedna se ponavlja (razlika je u velikom i malom slovu, značenje je isto). Hajde da to ispravimo:

```
dataset$apache_2_bodysystem[dataset$apache_2_bodysystem == "Undefined diagn
oses"] <- "Undefined Diagnoses"
unique_bodysystem_values <- unique(dataset$apache_2_bodysystem)
unique_bodysystem_values
```

```
## [1] "Cardiovascular"      "Respiratory"         "Metabolic"
## [4] "Trauma"              "Neurologic"          "Gastrointestinal"
## [7] "Renal/Genitourinary" "Undefined Diagnoses" "Haematologic"
```

Vidimo da je 113 najčešća vrednost kategorijske promenljive *apache_2_diagnosis*, ali se javlja u svega 13% slučajeva. Iz tog razloga je nećemo iskoriti za popunjavanje NA vrednosti ove kolone (spoiler alert-jos uvek). Umesto toga, probaćemo da vrednosti ove kolone tačnije odredimo pomoću *apache_2_bodysystem* promenljive. Pronaćićemo najzastupljeniju vrstu bolesti. U ovom slucaju to je Cardiovascular, sa 44.14%.

```
mode_value_diagnosis <- as.numeric(names(sort(table(dataset$apache_2_diagnosis), decreasing = TRUE)[1]))
mode_value_diagnosis
```

```
## [1] 113
```

```
percentage_table <- prop.table(table(dataset$apache_2_diagnosis)) * 100
percentage_table
```

```
##
##         101         102         103         104         105         106
##  0.41754117  2.09103730  0.32315021  0.40310491  1.15156967  2.74955303
##         107         108         109         110         112         113
##  0.21876492  1.30703712  1.18932605  3.56020477  4.83059600 13.03705678
##         114         115         116         117         118         119
##  2.50635751  0.27873094  0.22653829  4.35642025  1.50470289  1.91113924
##         120         121         122         123         124         202
##  1.52913349  2.45971727  4.18429557  2.29203451  4.33754206  2.75510544
##         203         207         208         209         212         213
##  2.92167772  0.71403982  0.13658927  0.53303128  0.10993770  0.83286138
##         214         215         216         217         218         219
##  1.29593231  0.17656661  0.02887253  0.53636273  0.61409646  0.51193213
##         301         302         303         304         305         306
##  7.55904987  7.44244928  3.69679404  3.44693563  2.52634618  0.70848741
##         307         308
##  2.01885598  4.56852228
```

```
any(is.na(dataset$apache_2_bodysystem))
```

```
## [1] FALSE
```

```
percentage_table <- prop.table(table(dataset$apache_2_bodysystem)) * 100
percentage_table
```

```
##
##      Cardiovascular    Gastrointestinal        Haematologic           Metabolic
##           44.1355097           9.8415710           0.6956484           8.3412384
##          Neurologic Renal/Genitourinary         Respiratory              Trauma
##          12.9708983           2.6822806          12.6579656           4.1891553
## Undefined Diagnoses
##           4.4857327
```

Sada prelazimo na enkodiranje. Mapiramo svaku od različitih vrednosti ove kolone radi olakšanog daljnjeg rada. Enkodirane vrednosti pamtimo u novoj koloni encoded_apache_2_bodysystem.

```
encoding_map <- c(
  "Cardiovascular" = 1,
  "Respiratory" = 2,
  "Metabolic" = 3,
```

```
  "Trauma" = 4,
  "Neurologic" = 5,
  "Gastrointestinal" = 6,
  "Renal/Genitourinary" = 7,
  "Haematologic" = 8,
  "Undefined Diagnoses" = 9
)
dataset$encoded_apache_2_bodysystem <- encoding_map[dataset$apache_2_bodysystem]
#dataset$encoded_apache_2_bodysystem
```

Hajde da pokušamo da nađemo povezanost između ove 2 promenljive. Nema smisla raditi korelaciju, pa ćemo probati malo drugačiji pristup. Uporedićemo svaku vrednost prve sa svakom vrednošću druge promenljive, i da izračunamo procenat slučajeva u kojima se 2 ista para pojavljuju. Ignorisaćemo NA vrednosti kolone *apache_2_diagnosis*, jer njih nema smisla uporedjivati.

```
result_list <- list()
for (diagnosis_value in unique(dataset$apache_2_diagnosis)) {
  non_na_diagnosis_df <- dataset[!is.na(dataset$apache_2_diagnosis), ]
  subset_df <- non_na_diagnosis_df[non_na_diagnosis_df$apache_2_diagnosis == diagnosis_value, ]
  percentage_results <- list()

  for (target_bodysystem in unique(dataset$encoded_apache_2_bodysystem)) {
    matching_rows <- subset_df$encoded_apache_2_bodysystem == target_bodysystem
    percentage_matching <- sum(matching_rows) / nrow(subset_df) * 100
    percentage_results[[as.character(target_bodysystem)]] <- percentage_matching}
  result_list[[as.character(diagnosis_value)]] <- percentage_results}
#print(result_list)
```

Dolazimo do zanimljivog zaključka:

- Vrednost 113 odgovara vrednosti 1 (100%).
- Vrednost 114 odgovara vrednosti 1 (100%).
- Vrednost 108 odgovara vrednosti 2 (100%).
- Vrednost 122 odgovara vrednosti 3 (100%).
- Vrednost 203 odgovara vrednosti 1 (100%).
- Vrednost 119 odgovara vrednosti 4 (100%).
- Vrednost 301 odgovara vrednosti 5 (100%).
- Vrednost 116 odgovara vrednosti 1 (100%).
- Vrednost 112 odgovara vrednosti 1 (100%).
- Vrednost 303 odgovara vrednosti 2 (100%).
- Vrednost 102 odgovara vrednosti 2 (100%).
- Vrednost 217 odgovara vrednosti 5 (100%).

- Vrednost 218 odgovara vrednosti 5 (100%).
- Vrednost 304 odgovara vrednosti 6 (100%).
- Vrednost 302 odgovara vrednosti 1 (100%).
- Vrednost 305 odgovara vrednosti 7 (100%).
- Vrednost 124 odgovara vrednosti 6 (100%).
- Vrednost 202 odgovara vrednosti 1 (100%).
- Vrednost 207 odgovara vrednosti 4 (100%).
- Vrednost 110 odgovara vrednosti 1 (100%).
- Vrednost 209 odgovara vrednosti 2 (100%).
- Vrednost 109 odgovara vrednosti 1 (100%).
- Vrednost 106 odgovara vrednosti 2 (100%).
- Vrednost 117 odgovara vrednosti 1 (100%).
- Vrednost 120 odgovara vrednosti 5 (100%).
- Vrednost 308 odgovara vrednosti 9 (100%).
- Vrednost 105 odgovara vrednosti 2 (100%).
- Vrednost 212 odgovara vrednosti 6 (100%).
- Vrednost 219 odgovara vrednosti 5 (100%).
- Vrednost 306 odgovara vrednosti 3 (100%).
- Vrednost 121 odgovara vrednosti 5 (100%).
- Vrednost 214 odgovara vrednosti 6 (100%).
- Vrednost 123 odgovara vrednosti 3 (100%).
- Vrednost 213 odgovara vrednosti 6 (100%).
- Vrednost 208 odgovara vrednosti 4 (100%).
- Vrednost 101 odgovara vrednosti 2 (100%).
- Vrednost 118 odgovara vrednosti 4 (100%).
- Vrednost 307 odgovara vrednosti 3 (100%).
- Vrednost 215 odgovara vrednosti 7 (100%).
- Vrednost 103 odgovara vrednosti 2 (100%).
- Vrednost 115 odgovara vrednosti 1 (100%).
- Vrednost 104 odgovara vrednosti 2 (100%).
- Vrednost 216 odgovara vrednosti 7 (100%).
- Vrednost 107 odgovara vrednosti 2 (100%).
- Vrednosti 113, 114, 203, 116, 112, 302, 202, 110, 109, 117, 115 uvek odgovaraju vrednosti 1.
- Vrednosti 108, 303, 102, 209, 106, 105, 101, 103, 104, 107 uvek odgovaraju vrednosti 2.
- Vrednosti 122, 306, 123, 307 uvek odgovaraju vrednosti 3.
- Vrednosti 119, 207, 208, 118 uvek odgovaraju vrednosti 4.
- Vrednosti 301, 217, 218, 120, 219, 121 uvek odgovaraju vrednosti 5.
- Vrednosti 304, 124, 212, 214, 213 uvek odgovaraju vrednosti 6.
- Vrednosti 305, 215, 216 uvek odgovaraju vrednosti 7.
- Vrednosti 306 uvek odgovara vrednosti 8.
- Vrednosti 308 uvek odgovara vrednosti 9.

Vidimo da smo pokrili sve vrednosti.

Zaključak?
Svaki kod x uvek odgovara tipu bolesti y.

Šta možemo uraditi sa tom činjenicom?
Možemo odrediti koja grupa kodove je najčešće zastupljena u datasetu, a zatim odrediti koji kod iz te grupe se pojavljuje najviše puta. To će biti kod kojim ćemo popuniti NA vrednosti promenljive *apache_2_diagnosis*.

Budući da su bolesti koje pripadaju grupi kardiovaskularnih najčešće zastupljene, a dijagnoza sa kodom 113 (najčešće zastupljena dijagnoza u datasetu) njoj i pripada, na kraju ćemo ipak iskoristiti njen kod kako bi popunili nedostajuće vrednosti ove kolone.

```
unique_num<- unique(dataset$apache_2_diagnosis)
print(unique_num)
##  [1] 113 108 122 203 119 301 116 112 303 218 304 302 305 124 202 207 110 209 109
## [20] 106 117 120  NA 217 114 102 308 105 212 219 306 121 214 123 213 208 101 118
## [39] 307 215 103 115 104 216 107
```

```
dataset$apache_2_diagnosis[is.na(dataset$apache_2_diagnosis)] <- mode_value
_diagnosis
#dataset$apache_2_diagnosis
any(is.na(dataset$apache_2_diagnosis))
## [1] FALSE
any(is.na(dataset$apache_3j_bodysystem))
## [1] FALSE
```

1. apache_3j_diagnosis - šifra poddijagnoze APACHE III-J koja najbolje opisuje razlog prijema na intezivnu negu

Vidimo da je sepsa grupa bolesti koja odgovara najčešćem kodu, i to u 100% slučajeva (ne postoji drugi kod iz dataseta koji odgovara “Sepsis” vrsti). Budući da nam treba kod koji se najčešće pojavljuje iz grupe kardiovaskularnih bolesti, pronaći cemo ga i njime popuniti NA vrednosti *apache_3j_diagnosis* kolone.

```
most_common_value<-names(which.max(table(dataset$apache_3j_diagnosis[datase
t$apache_3j_bodysystem == "Cardiovascular"])))
most_common_value
## [1] "107.01"
```

```
dataset$apache_3j_diagnosis[is.na(dataset$apache_3j_diagnosis)] <- most_com
mon_value
#dataset$apache_3j_diagnosis
any(is.na(dataset$apache_3j_diagnosis))
## [1] FALSE
```

1. arf_apache - da li je pacijent imao akutnu bubrežnu insuficijenciju tokom prva 24 sata boravka u odeljenju, definisano kao 24-časovno izlučivanje urina <410ml, kreatinin >=133mikromol/L i bez hronične dijalize

Budući da je čak 97.2% vrednosti 0, ostale NA vrednosti ćemo popuniti nulom.

```
any(is.na(dataset$arf_apache))
## [1] TRUE
```

```
prop.table(table(dataset$arf_apache)) * 100
##
##         0         1
## 97.202136  2.797864
```

```
dataset$arf_apache[is.na(dataset$arf_apache)]<-0
any(is.na(dataset$arf_apache))
## [1] FALSE
```

1. gcs_unable_apache - da li Glasgow Coma Scale nije mogla da se proceni zbog sedacije pacijenta

Budući da je čak 99% vrednosti 0 (gotovo svi pacijenti su po proceni lekara mogli da odrade GCS test), ostale NA vrednosti ćemo popuniti nulom.

```
any(is.na(dataset$gcs_unable_apache))
## [1] TRUE
prop.table(table(dataset$gcs_unable_apache)) * 100
##
##          0          1
## 99.0471569  0.9528431
```

```
dataset$gcs_unable_apache[is.na(dataset$gcs_unable_apache)]<-0
any(is.na(dataset$gcs_unable_apache))
## [1] FALSE
```

1. gcs_eyes_apache - komponenta otvaranja očiju prema Glasgow Coma Scale, merena tokom prva 24 sata, što rezultira najvišim APACHE III rezultatom
2. gcs_motor_apache - motorna komponenta prema Glasgow Coma Scale, merena tokom prva 24 sata , što rezultira najvišim APACHE III rezultatom
3. gcs_verbal_apache - verbalna komponenta prema Glasgow Coma Scale, merena tokom prva 24 sata, što rezultira najvišim APACHE III rezultatom

Promenljive koje čine GCS test. U nastavku vidimo da svaki pacijent koji nije mogao da radi test zbog odluke lekara nema tačno određene vrednosti ovih parametara. Za ostale pacijente za koje smo naknadno “odobrili” rađenje testa (postavili *gcs_unable_apache* na 0) potrebno je odrediti ove parametre. Priroda testa

je takva da ne zavisi od ostalih parametara ovog dataseta, pa ćemo za njihovo popunjavanje koristiti biblioteku *mice*.

```
any(!is.na(dataset$gcs_eyes_apache[dataset$gcs_unable_apache == 1]) | !is.na(dataset$gcs_motor_apache[dataset$gcs_unable_apache == 1]) | !is.na(dataset$gcs_verbal_apache[dataset$gcs_unable_apache == 1]))
## [1] FALSE
```

```
pred_matrix <- make.predictorMatrix(dataset)
pred_matrix[c("gcs_eyes_apache", "gcs_motor_apache", "gcs_verbal_apache"), ] <- 0

imp <- mice(dataset, m = 5, maxit = 5, method = 'pmm', seed = 500, predictorMatrix = pred_matrix)
dataset_imp <- complete(imp, 1)

dataset$gcs_eyes_apache <- dataset_imp$gcs_eyes_apache
dataset$gcs_motor_apache <- dataset_imp$gcs_motor_apache
dataset$gcs_verbal_apache <- dataset_imp$gcs_verbal_apache
sum(is.na(dataset$gcs_eyes_apache))
## [1] 0
sum(is.na(dataset$gcs_motor_apache))
## [1] 0
sum(is.na(dataset$gcs_verbal_apache))
## [1] 0
```

1. heart_rate_apache - broj otkucaja srca izmeren tokom prva 24 sata što rezultira najvišim APACHE III rezultatom

Predstavlja “najgore” (najviše iznad ili ispod određenih granica) izmeren puls u prvih 24 sata nakon što je pacijent primljen na odeljenje. Ova vrednost bi trebalo da ima visoku korelaciju sa promenljivom *d1_heartrate_max* ili *d1_heartrate_min*, pa hajde to da proverimo. Izračunaćemo najbliže vrednost iz ove 2 kolone, i zapamtiti ih u novim kolonama.

```
dataset$min_diff <- with(dataset, ifelse(is.na(heart_rate_apache), NA, ifelse(abs(d1_heartrate_max - heart_rate_apache) < abs(d1_heartrate_min - heart_rate_apache), d1_heartrate_max, d1_heartrate_min)))

cor(dataset$heart_rate_apache, dataset$min_diff, use = "complete.obs")
## [1] 0.9379919
```

```
sum(is.na(dataset$d1_heartrate_max))
## [1] 0
sum(is.na(dataset$d1_heartrate_min))
## [1] 0
```

Kao što smo i pretpostavili, korelacija je vrlo visoka. To znači da ćemo moći popuniti NA vrednosti na osnovu vrednosti iz naše nove kolone. Izračunaćemo prosečnu razliku između *heart_rate_apache* i *min_diff* kolone i tom vrednošću ćemo popuniti NA vrednosti. Za vrednosti < 80 (procenjujemo da je puls niži od normalnog u tim slučajevima), koristićemo *d1_heartrate_min*, za vrednosti >= 80, koristićemo *d1_heartrate_max*.

```
avg_diff<-mean(dataset$heart_rate_apache - dataset$min_diff, na.rm = TRUE)
dataset$heart_rate_apache <- ifelse(is.na(dataset$heart_rate_apache) & data
set$d1_heartrate_min < 80,
                                    dataset$d1_heartrate_min + avg_diff,
                                    ifelse(is.na(dataset$heart_rate_apache)
,
                                           dataset$d1_heartrate_max + avg_d
iff,
                                           dataset$heart_rate_apache))
sum(is.na(dataset$heart_rate_apache))
## [1] 0
```

1. map_apache - srednji arterijski pritisak izmeren tokom prva 24 sata koji rezultira najvišim APACHE III rezultatom
2. resprate_apache - brzina disanja izmerena tokom prva 24 sata što rezultira najvišim APACHE III rezultatom
3. temp_apache - temperatura izmerena tokom prva 24 sata što rezultira najvišim APACHE III rezultatom

Možemo se voditi istom logikom i na ovaj način popuniti NA vrednosti kolone resprate_apache, temp_apache i map_apache.

resprate_apache:

```
dataset$min_diff_resp <- with(dataset, ifelse(is.na(resprate_apache), NA, i
felse(abs(d1_resprate_max - resprate_apache) < abs(d1_resprate_min - respra
te_apache), d1_resprate_max, d1_resprate_min)))


cor(dataset$resprate_apache, dataset$min_diff_resp, use = "complete.obs") #
jos uvek prilicno visoka korelacija
## [1] 0.8490546
```

```
sum(is.na(dataset$d1_resprate_max))
## [1] 0
sum(is.na(dataset$d1_resprate_min))
## [1] 0
```

```
#min_diff_resp vrednost je u proseku veca u odnosu na resprate_apache, pa c
emo je dodati kada budemo računali
avg_diff_resp<-mean(dataset$resprate_apache - dataset$min_diff_resp, na.rm
= TRUE)
```

```
avg_diff_resp
## [1] 2.235912
```

```
dataset$resprate_apache <- ifelse(is.na(dataset$resprate_apache) & dataset$
d1_resprate_min < 25, ##po istoj logici biramo vrednost od 25
                                  dataset$d1_resprate_min + avg_diff_resp
,
                                  ifelse(is.na(dataset$resprate_apache),
                                         dataset$d1_resprate_max + avg_di
ff_resp,
                                         dataset$resprate_apache))
sum(is.na(dataset$resprate_apache))
## [1] 0
```

temp_apache:

```
dataset$min_diff_temp <- with(dataset, ifelse(is.na(temp_apache), NA, ifels
e(abs(d1_temp_max - temp_apache) < abs(d1_temp_min - temp_apache), d1_temp_
max, d1_temp_min)))

cor(dataset$temp_apache, dataset$min_diff_temp, use = "complete.obs") #jos
uvek prilicno visoka korelacija
## [1] 0.9554557
```

```
sum(is.na(dataset$d1_temp_max))
## [1] 0
sum(is.na(dataset$d1_temp_min))
## [1] 0
```

```
#min_diff_temp vrednost je u proseku veca u odnosu na temp_apache, pa cemo
je dodati
avg_diff_temp<-mean(dataset$temp_apache - dataset$min_diff_temp, na.rm = TR
UE)
dataset$temp_apache <- ifelse(is.na(dataset$temp_apache) & dataset$d1_temp_
min < 37, ##po istoj logici biramo vrednost od 37
                                  dataset$d1_temp_min + avg_diff_temp,
                                  ifelse(is.na(dataset$temp_apache),
                                         dataset$d1_temp_max + avg_diff_t
emp,
                                         dataset$temp_apache))

sum(is.na(dataset$temp_apache))
## [1] 0
```

map_apache:

```
dataset$min_diff_map <- with(dataset, ifelse(is.na(map_apache), NA, ifelse(
abs(d1_mbp_max - map_apache) < abs(d1_mbp_min - map_apache), d1_mbp_max, d1
_mbp_min)))

cor(dataset$map_apache, dataset$min_diff_map, use = "complete.obs") #jos uv
ek prilicno visoka korelacija
## [1] 0.8767509
```

```
sum(is.na(dataset$d1_mbp_max))
## [1] 0
sum(is.na(dataset$d1_mbp_min))
## [1] 0
```

```
#min_diff_map vrednost je u proseku veca u odnosu na map_apache, pa cemo je
dodati
avg_diff_map<-mean(dataset$map_apache - dataset$min_diff_map, na.rm = TRUE)
dataset$map_apache <- ifelse(is.na(dataset$map_apache) & dataset$d1_mbp_min
< 85, ##po istoj logici biramo vrednost od 85
                             dataset$d1_mbp_min + avg_diff_map,
                             ifelse(is.na(dataset$map_apache),
                                    dataset$d1_mbp_max + avg_diff_ma
p,
                                    dataset$map_apache))

sum(is.na(dataset$map_apache))
## [1] 0
```

1. intubated_apache - da li je pacijent intubiran u trenutku kada je vrednost parcijalni pritiska gasova u arterisjkoj krvi bio najviši

Budući da je čak 85% vrednosti 0 (vecina pacijenata nije bilo intubirano), ostale NA vrednosti ćemo popuniti nulom.

```
any(is.na(dataset$intubated_apache))
## [1] TRUE
```

```
prop.table(table(dataset$intubated_apache)) * 100
##
##        0        1
## 84.87769 15.12231
```

```
dataset$intubated_apache[is.na(dataset$intubated_apache)]<-0
any(is.na(dataset$intubated_apache))
```

```
## [1] FALSE
```

1. ventilated_apache - da li je pacijent bio invazivno ventiliran u vreme najvećeg nivoa gasa arterijske krvi koristeći algoritam za ocenjivanje oksigenacije, uključujući bilo koji način ventilacije sa pozitivnim pritiskom koji se isporučuje kroz kolo spojeno na endotrahealnu cev ili traheostomiju

Budući da je čak 67% vrednosti 0 (većina pacijenata nije bilo ventilirano), ostale NA vrednosti ćemo popuniti nulom.

```
any(is.na(dataset$ventilated_apache))
## [1] TRUE
```

```
prop.table(table(dataset$ventilated_apache)) * 100
##
##        0        1
## 67.42786 32.57214
```

```
dataset$ventilated_apache[is.na(dataset$ventilated_apache)]<-0
any(is.na(dataset$ventilated_apache))
## [1] FALSE
```

1. d1_glucose_max - najveća koncentracija glukoze kod pacijenta u serumu ili plazmi tokom prva 24 sata boravka na odeljenju
2. d1_glucose_min - najmanja koncentracija glukoze kod pacijenta u serumu ili plazmi tokom prva 24 sata boravka na odeljenju

Nakon istraživanja i konsultovanjem sa stručnom osobom, zaključili smo da za predikciju glukoze nemamo određene fetures koji su nam potrebni, kao što su: drhtavica, pojačano znojenje, nekontrolisana glad. Visok nivo glukoze nam je pokazatelj da osoba ima dijabetes - ukoliko je koncentracija glukoze preko 11.1 mmol/L.

```
sum(is.na(dataset$d1_glucose_max))
## [1] 5807
sum(is.na(dataset$d1_glucose_min))
## [1] 6230
```

```
summary(dataset$d1_glucose_max)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    73.0   117.0   150.0   174.6   201.0   611.0    5807
summary(dataset$d1_glucose_min)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    33.0    90.0   107.0   113.5   130.0   287.0    6230
```

S obzirom na to da za glukozi imamo minimalnu i maksimalnu koncentraciju, ne znamo kako je je da predvidimo na koji način je glukoza skočila.

1. d1_potassium_max - najveća koncentracija kalijuma kod pacijenta u serumu ili plazmu tokom prva 24 sata boravka na odeljenju
2. d1_potassium_min - najmanja koncentracija kalijuma kod pacijenta u serumu ili plazmu tokom prva 24 sata boravka na odeljenju

Nakon istraživanja došli smo do sledećih zaključaka: Ukoliko je pacijent imao akutnu bubrežnu insuficijenciju tokom prva 24 sata boravka u odeljenju - izmerena koncentracija kalijuma je veća od 5 - to znači da je *d1_potassium_min* u gornjoj granici, shodno tome feture *d1_potassium_max* će takođe biti predstavljen vrednostima gornje granice. Ukoliko osoba nema bubrežnu insuficijenciju (a kako je to jedina bolest u našem datasetu koja može biti uzrokovana koncentracijom kalijuma) pretpostavićemo da osoba ima koncentraciju kalijuma zdrave osobe koja bi maksimalno trebalo da bude između 3.5 i 5.0 milimola po litri (mmol/L). Starije osobe mogu imati manju sposobnost bubrega da reguliše kalijum. Takođe koncentracija kalijuma zavisi i od gojaznosti pacijenta, ali je veza previše složena i zahteva dublje medicinske analize, kao što su test za insulinsku rezistenciju, funkcija bubrega…

```
sum(is.na(dataset$d1_potassium_max))
## [1] 9585
sum(is.na(dataset$d1_potassium_min))
## [1] 9585
```

Jako je čudan detalj koji nam predstavlja koncentraciju kalijuma kao min i max, s obzirom na to da je koncentracija kalijuma u organizmu konstantna i može da se poveća sa unosom hrane i lekova.

```
summary(dataset$d1_potassium_max)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.800   3.800   4.200   4.252   4.600   7.000    9585
summary(dataset$d1_potassium_min)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.400   3.600   3.900   3.935   4.300   5.800    9585
```

Ova dva parametra *glukoza* i *kalijum* ćemo da nadomestimo koristeći biblioteku mice, s obzirom na to da u dataset-u nemamo potrebne podatke da bismo ih odredili na drugi način.

```
dataset$d1_glucose_max <- dataset_imp$d1_glucose_max
dataset$d1_glucose_min <- dataset_imp$d1_glucose_min
dataset$d1_potassium_min <- dataset_imp$d1_potassium_min
dataset$d1_potassium_max <- dataset_imp$d1_potassium_max

sum(is.na(dataset$d1_glucose_max))
## [1] 0
sum(is.na(dataset$d1_glucose_min))
## [1] 0
```

```
sum(is.na(dataset$d1_potassium_max))
## [1] 0
sum(is.na(dataset$d1_potassium_min))
## [1] 0
```

Konačno, sada imamo vrednosti svih promenljivih sem *apache_4a_hospital_death_prob* i *apache_4a_icu_death_prob*. Smatramo da su ove dve kolone važne za predviđanje naše ciljne promenljive (hospital death), pa ćemo njene NA vrednosti popuniti što temeljnije moguće.
apache 4a verovatnoća smrti se zasniva na *APACHE III* skoru, nakon istraživanja utvrdili smo da nema šanse da ovo uradimo na osnovu podataka dostupnih u datasetu, možemo (kad bi imali sredstava) odrediti APACHE III skor i njime utvrditi šansu smrtnosti.

Za kraj, samo ćemo da iskoristimo mice biblioteku (ponovo :D).

```
dataset$apache_4a_hospital_death_prob <- dataset_imp$apache_4a_hospital_dea
th_prob
dataset$apache_4a_icu_death_prob <- dataset_imp$apache_4a_icu_death_prob
sum(is.na(dataset$apache_4a_icu_death_prob))
## [1] 0
sum(is.na(dataset$apache_4a_hospital_death_prob))
## [1] 0
```

Prilikom sređivanja dataseta kreirali smo kolone koje su nam bile od koristi samo prilikom ovog segmenta rada na seminarskom radu, tako da ih možemo obrisati jer su one neupotrebljive više.

```
dataset <- subset(dataset, select = -c(encoded_apache_2_bodysystem, min_dif
f, min_diff_resp, min_diff_temp, min_diff_map))
```

Proveravamo da li smo uspešno očistili naš dataset od nedostajućih vrednosti.

```
sum(is.na(dataset))
## [1] 0
```

Naši podaci su uspešno očišćeni od nedostajućih vrednosti.

Zbog dalje upotrebe sačuvaćemo naš sređeni dataset kao *cleaned_dataset*.

```
cleaned_dataset <- dataset
export(cleaned_dataset, "C:/Users/astan/Desktop/seminarski rad/cleaned_data
set.csv")
```

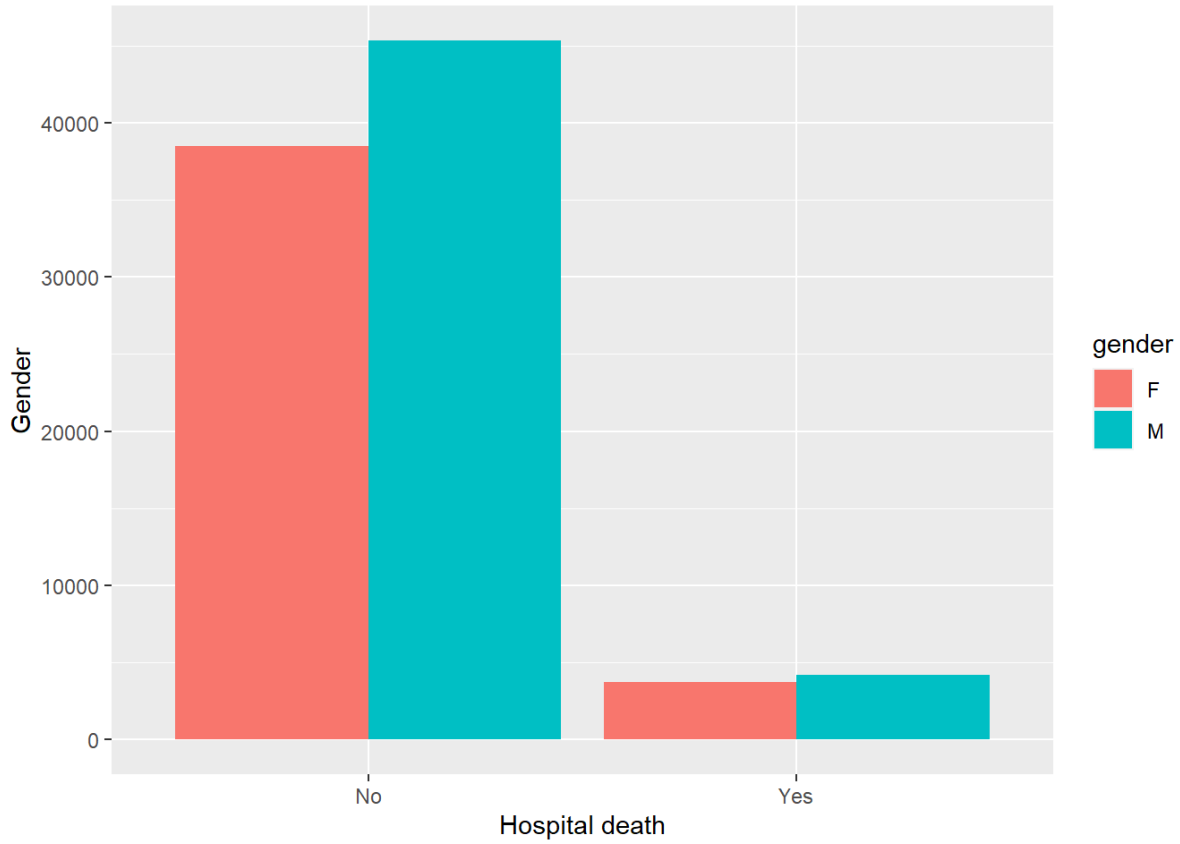
```
cleaned_dataset <- read_csv("cleaned_dataset.csv")
```

Analiza

U nastavku ćemo prikazati stopu smrtnosti pacijenata u zavisnosti od parametara koji opisuju jednog pacijenta.

hospital_death - da li je pacijent preminuo tokom ove hospitalizacije 0 - No 1 - Yes

```
gg_gender <- ggplot(cleaned_dataset, aes(x = factor(hospital_death), fill =
gender)) +
  geom_bar(position = "dodge") +
  ylab("Gender") +
  xlab("Hospital death") +
  scale_x_discrete(labels = c("No", "Yes"))
gg_gender
```




```
gender.survived <- xtabs(~ gender + hospital_death, data = cleaned_dataset)
gender.survived
```

```
##        hospital_death
## gender     0     1
##      F 38488  3731
##      M 45310  4184
```

```
gender.survived.prop <- prop.table(gender.survived, margin = 1)
```

```
gender.survived.prop
```

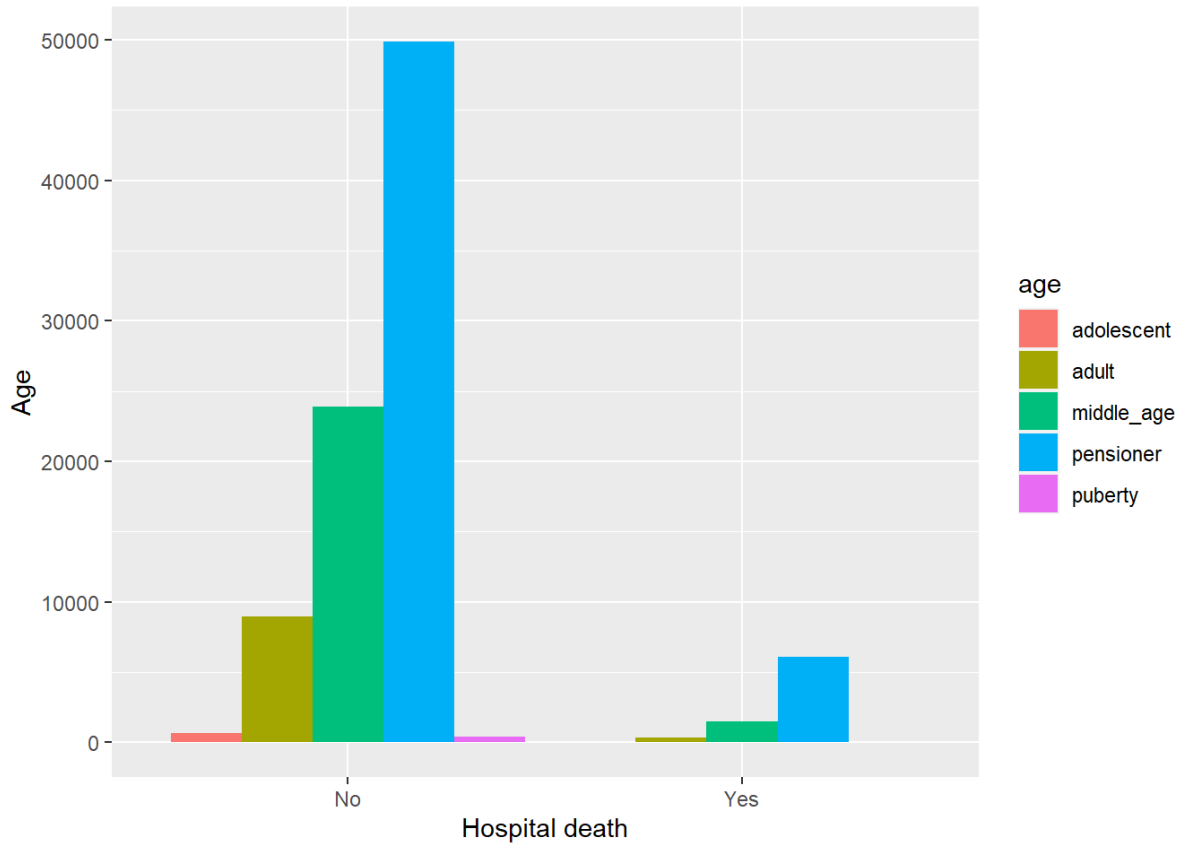
```
##       hospital_death
## gender          0          1
##      F 0.91162747 0.08837253
##      M 0.91546450 0.08453550
```

Nije toliko značajna razlika u broju preživelih pacijenata u odnosu na pol. Među preživelim pacijentima prednjače muškarci sa 54%. Također među svim pacijentima koji su preminuli također prednjače muškarci sa 53%. :(Što nije neočekivano s obzirom na to da muškaraca ima više nego žena.

```
xtabs(~gender, data = cleaned_dataset)
```

```
## gender
##     F     M
## 42219 49494
```

```
gg_age <- ggplot(cleaned_dataset, aes(x = factor(hospital_death), fill = age)) +
  geom_bar(position = "dodge") +
  ylab("Age") +
  xlab("Hospital death") +
  scale_x_discrete(labels = c("No", "Yes"))
gg_age
```



```
age.survived <- xtabs(~ age + hospital_death, data = cleaned_dataset)
age.survived
```

```
##             hospital_death
## age              0     1
##   adolescent   667    14
##   adult       8971   343
##   middle_age 23882  1482
##   pensioner  49861  6064
##   puberty      417    12
```

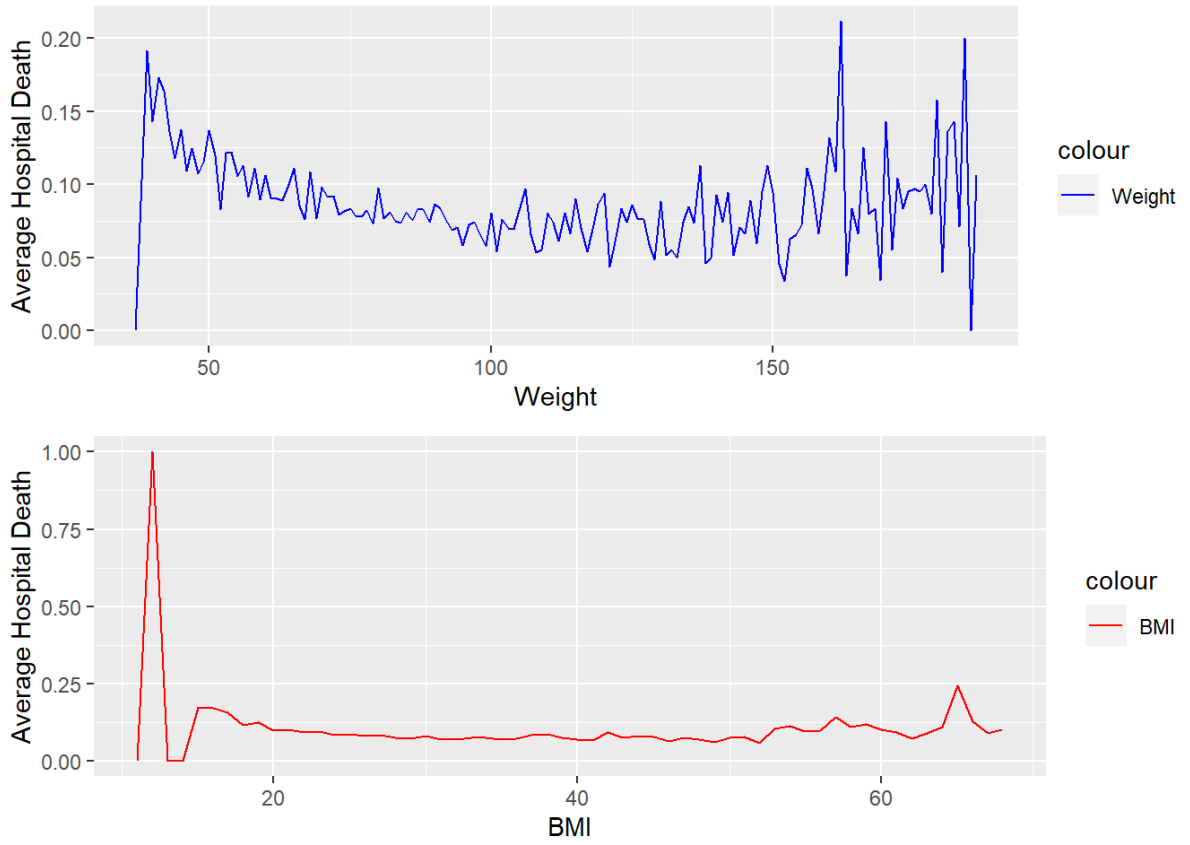
```
age.survived.prop <- prop.table(age.survived, margin = 1)
age.survived.prop
```

```
##             hospital_death
## age                    0          1
##   adolescent 0.97944200 0.02055800
##   adult      0.96317372 0.03682628
##   middle_age 0.94157073 0.05842927
##   pensioner  0.89156907 0.10843093
##   puberty    0.97202797 0.02797203
```

Primećujemo da je najviše penzionera preminulo u bolnici, zatim osobe u srednjim godinama i odrasli. Osobe u pubertetu i adolescenti su svi pacijenti preživeli. Vidimo da je starost dosta povezano sa smrtnoscu.

```
weight_df <- cleaned_dataset %>%
  dplyr::select(weight, hospital_death, bmi) %>%
  mutate(weight = round(weight),
         bmi = round(bmi))

weight_death <- weight_df %>%
  group_by(weight) %>%
  summarize(avg_hospital_death = mean(hospital_death)) %>%
  ungroup()

bmi_death <- weight_df %>%
  group_by(bmi) %>%
  summarize(avg_hospital_death = mean(hospital_death)) %>%
  ungroup()

gg_weight <- ggplot(weight_death, aes(x = weight, y = avg_hospital_death, c
olor = "Weight")) +
  geom_line() +
  labs(x = "Weight", y = "Average Hospital Death") +
  scale_color_manual(values = c("Weight" = "blue"))


gg_bmi <- ggplot(bmi_death, aes(x = bmi, y = avg_hospital_death, color = "B
MI")) +
  geom_line() +
```

```
  labs(x = "BMI", y = "Average Hospital Death") +
  scale_color_manual(values = c("BMI" = "red"))

library(gridExtra)
grid.arrange(gg_weight, gg_bmi)
```




```
#ovaj grafik rađen u R programskom jeziku ima dodatni deo koji nam omogućava hoover preko dobijenog grafika da bi za svaku osobu mogli da vidimo verovatnoću da će da premine u zavisnosti od težine i bmi.

gg_weight_hover <- plot_ly(data = weight_death, x = ~weight, y = ~avg_hospital_death, type = "scatter", mode = "lines",

        line = list(color = "blue"), name = "Weight", hoverinfo = "x+y") %>%

  layout(xaxis = list(title = "Weight"), yaxis = list(title = "Average Hospital Death"))

gg_bmi_hover <- plot_ly(data = bmi_death, x = ~bmi, y = ~avg_hospital_death, type = "scatter", mode = "lines",

          line = list(color = "red"), name = "BMI", hoverinfo = "x+y") %>%

  layout(xaxis = list(title = "BMI"), yaxis = list(title = "Average Hospital Death"))
```

```
subplot(gg_weight_hover, gg_bmi_hover, nrows = 1)
```

Zaključujemo da gojazni i neuhranjeni ljudi imaju najveću stopu smrtnosti.

```
unique_icu_type <- unique(cleaned_dataset$icu_type)
print(unique_icu_type)
```

```
## [1] "CTICU"        "Med-Surg ICU" "CCU-CTICU"    "Neuro ICU"    "MICU"
## [6] "SICU"         "Cardiac ICU"  "CSICU"
```

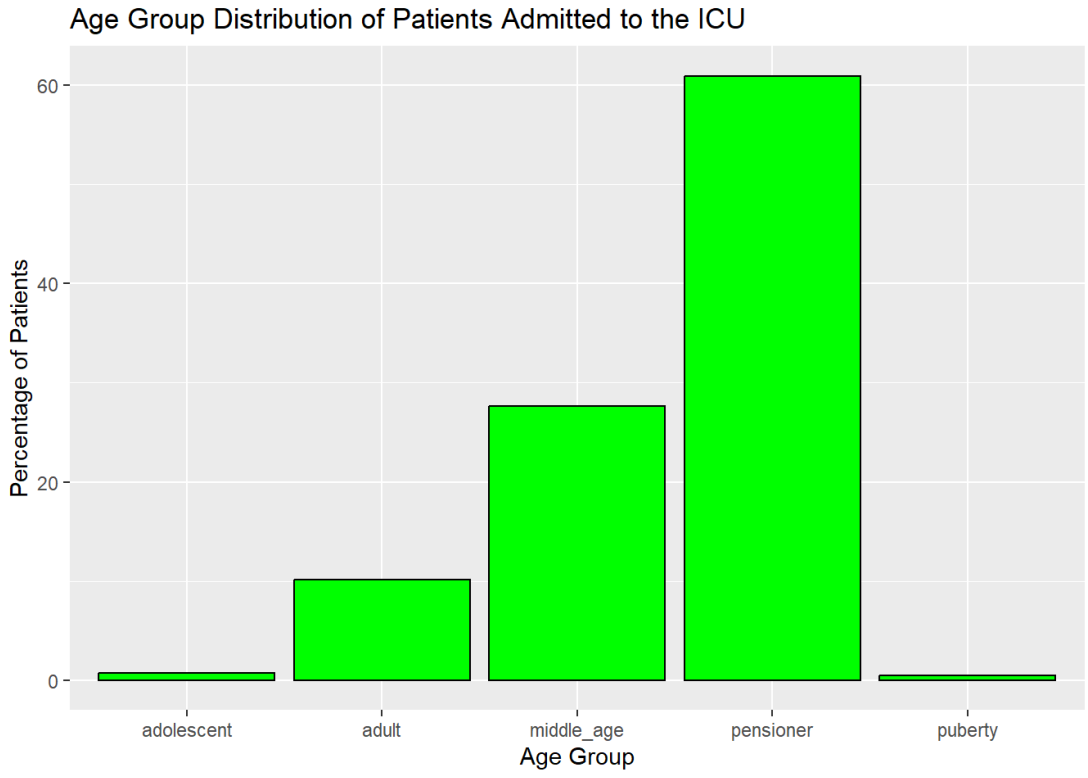
- CTICU - Cardiac Thoracic Intensive Care Unit (o je odeljenje intenzivne nege koje se specijalizuje za negu pacijenata koji su prošli kardiohirurške zahvate na srcu i toraksu ili imaju ozbiljne srčane ili plućne probleme. Ovo odeljenje je posebno opremljeno i ima stručno medicinsko osoblje koje se bavi pacijentima koji zahtevaju visok nivo monitoringa i medicinske intervencije nakon složenih kardiohirurških procedura ili u slučaju teških kardiovaskularnih bolesti).
- Med-Surg ICU - Medical-Surgical Intensive Care Unit (Ovo odeljenje pruža visoko stručno medicinsko osoblje i opremu za monitoring i podršku životnim funkcijama. Pacijenti ovde mogu biti različitih dijagnoza i potreba, uključujući pacijente koji su podvrgnuti hirurškim intervencijama, imaju teške medicinske bolesti ili zahtevaju posebne postupke i pažljivu kontrolu).
- CCU-CTICU - Cardiac Care Unit/Cardiac Thoracic Intensive Care Unit (Odeljenje intenzivne nege koje može pružati specijalizovanu negu za pacijente sa srčanim problemima i kardiovaskularnim operacijama, uključujući i pacijente koji su prošli hirurške zahvate na srcu i grudnom košu).
- Neuro ICU - Neurological Intensive Care Unit (Na ovom odeljenju medicinsko osoblje je stručno u upravljanju neurološkim hitnim slučajevima i komplikacijama. Odeljenje je opremljeno odgovarajućom medicinskom opremom za praćenje moždane aktivnosti, intrakranijalni pritisak, cerebralnu cirkulaciju i druge neurološke parametre. Cilj je pružiti optimalnu negu pacijentima sa oštećenjem nervnog sistema i smanjiti rizik od dodatnih komplikacija).
- MICU - Medical Intensive Care Unit (Na ovom odeljenju se pacijentima pruža visok nivo monitoringa i medicinske podrške, posebno onima koji imaju ozbiljne bolesti kao što su sepsa, plućne bolesti, zatajenje srca, komplikacije dijabetesa, i druga akutna ili hronična medicinska stanja. Odeljenje je opremljeno posebnom opremom za praćenje vitalnih znakova, funkcije organa i sastava krvi, kako bi medicinsko osoblje moglo brzo intervenisati u slučaju komplikacija).
- SICU - Surgical Intensive Care Unit (Na ovom odeljenju pacijenti koji su podvrgnuti različitim vrstama hirurških zahvata dobijaju visok nivo monitoringa, medicinske intervencije i podrške za oporavak nakon operacije.

Ovde se brinu o pacijentima sa različitim tipovima hirurških procedura, uključujući ortopedsku, abdominalnu, kardiohiruršku, plastičnu hirurgiju i druge).

- Cardiac ICU - Cardiac Intensive Care Unit (Na ovom odeljenju pacijentima sa stanjima kao što su srčani udar, aritmije, zatajenje srca, akutna insuficijencija srca i drugi hitni kardiovaskularni slučajevi pruža se intenzivna medicinska nega. Odeljenje je opremljeno posebnom opremom za praćenje srčane aktivnosti, elektrokardiografijom (EKG), monitoringom krvnog pritiska i drugim parametrima srčane funkcije).
- CSICU - Cardiothoracic Surgical Intensive Care Unit (Na ovom odeljenju pacijentima koji su podvrgnuti složenim operacijama srca, pluća ili toraksa pruža se visok nivo monitoringa, medicinske podrške i pažljive postoperativne nege. Odeljenje je opremljeno posebnom opremom za praćenje vitalnih znakova, funkcija organa i komplikacija nakon kardiohirurških procedura)

Sada ćemo proveriti raspodelu starosnih grupa pacijenata primljenih na odeljenje.

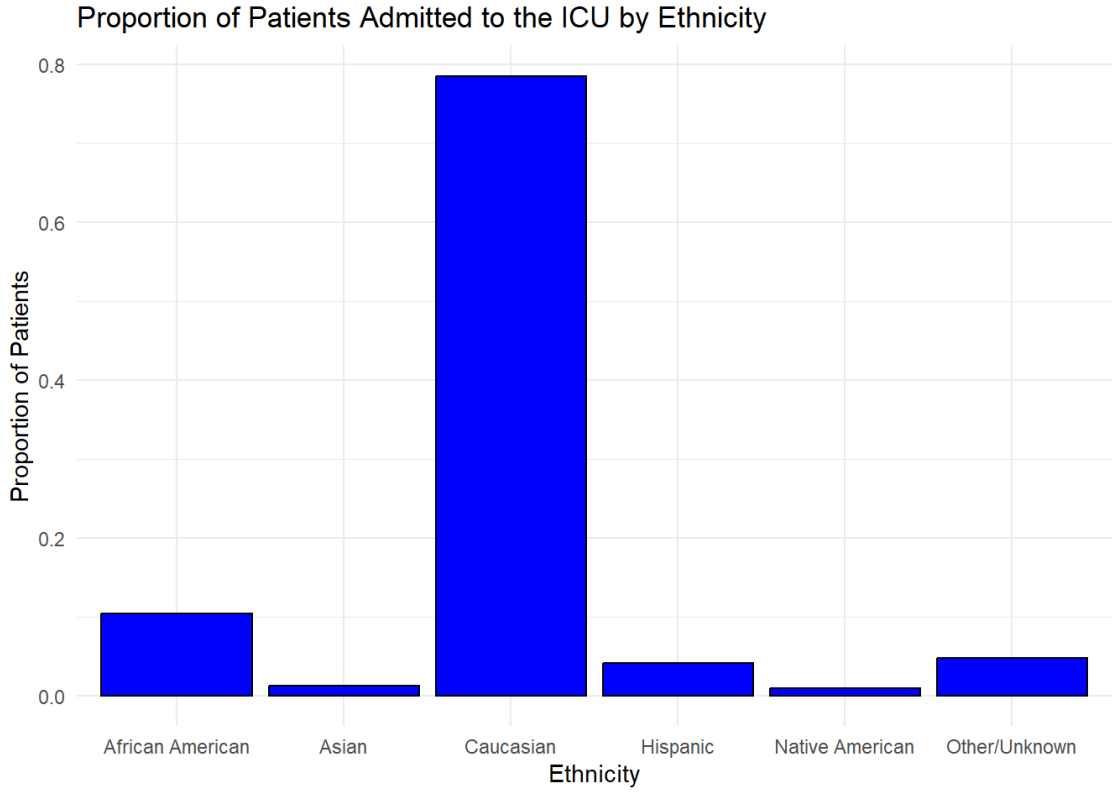
```
age_freq <- table(cleaned_dataset$age)
age_freq_df <- as.data.frame(age_freq)
age_freq_df$percentage <- (age_freq_df$Freq / sum(age_freq_df$Freq)) * 100

ggplot(age_freq_df, aes(x = Var1, y = percentage)) +
  geom_bar(stat = "identity", fill = "green", color = "black") +
  xlab("Age Group") +
  ylab("Percentage of Patients") +
  ggtitle("Age Group Distribution of Patients Admitted to the ICU")
```



Zaključujemo da, što je osoba starija, veća je i verovatnoća da će se naći na odeljenju.

Sada ćemo videti pacijenti koje rase su najčešće primljeni na odeljenje.

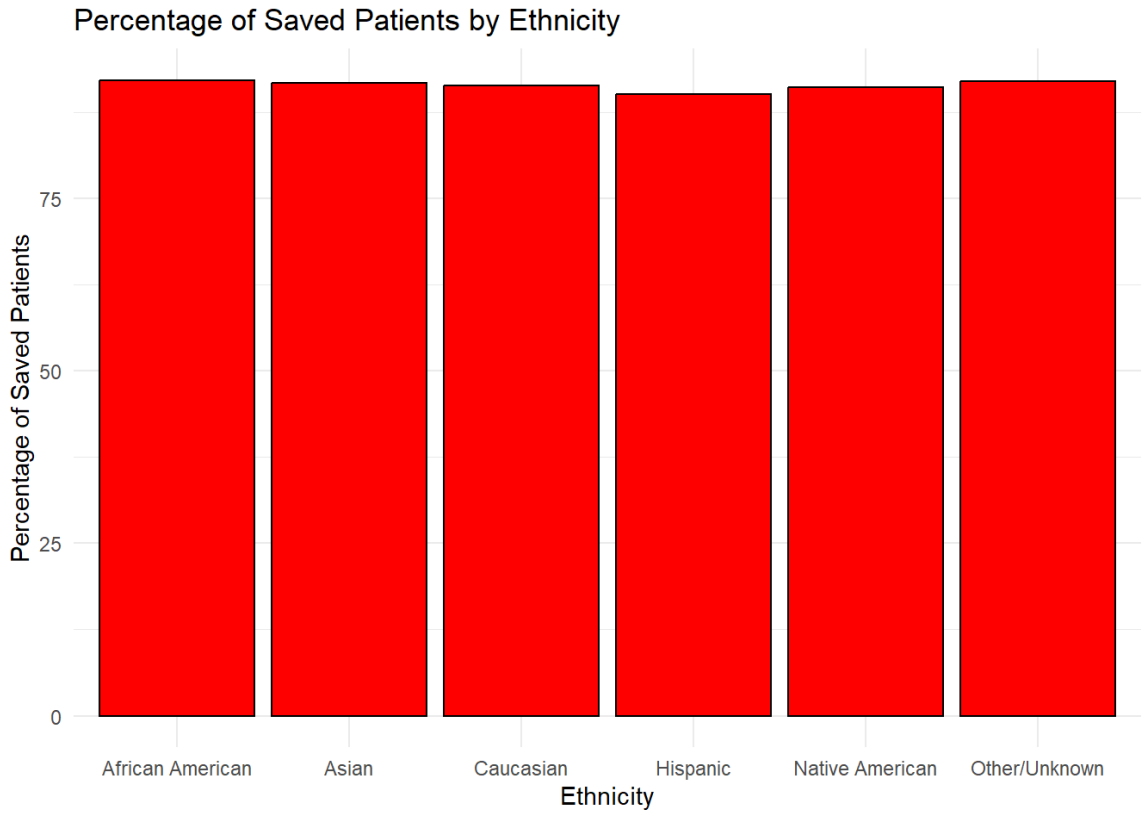
```
ethnicity_counts <- cleaned_dataset %>%
  group_by(ethnicity) %>%
  summarize(count = n()) %>%
  mutate(proportion = count / sum(count))

ggplot(ethnicity_counts, aes(x = ethnicity, y = proportion)) +
  geom_bar(stat = "identity", fill = "blue", color = "black") +
  labs(title = "Proportion of Patients Admitted to the ICU by Ethnicity",
       x = "Ethnicity",
       y = "Proportion of Patients") +
  theme_minimal()
```



Pacijenti bele rase su ubedljivo najčešći posetioci odeljenja. Amerika ima problema sa rasizmom, ali da li su doktori rasisti?

```
ethnicity_saved <- cleaned_dataset %>%
  group_by(ethnicity) %>%
  summarize(saved = sum(hospital_death == 0),
            total = n()) %>%
  mutate(percentage_saved = (saved / total) * 100)
ggplot(ethnicity_saved, aes(x = ethnicity, y = percentage_saved)) +
```

```
  geom_bar(stat = "identity", fill = "red", color = "black") +
  labs(title = "Percentage of Saved Patients by Ethnicity",
       x = "Ethnicity",
       y = "Percentage of Saved Patients")
```

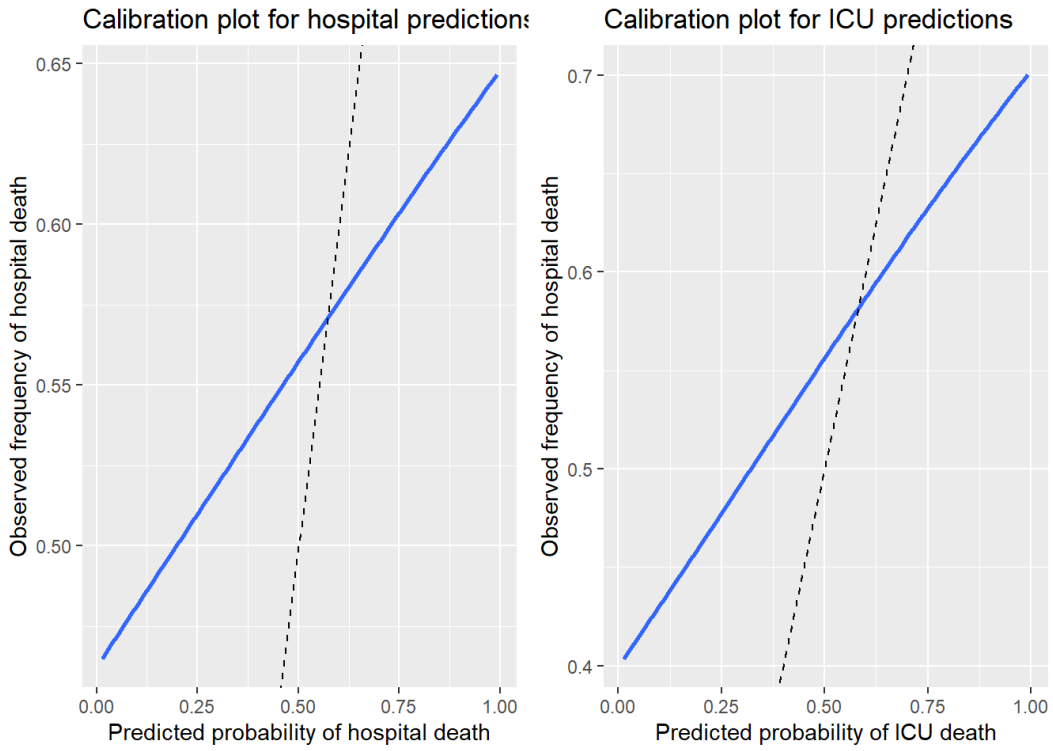


Srećom, nisu. Procenat preživelih pacijenata svih rasa je gotovo jednak.

Koliko je pouzdana apache 4a verovatnća smrti pri predviđanju realne stope smrtnosti? Sada ćemo da proverimo.

```
set.seed(1)
dataset <- data.frame(
  apache_4a_hospital_death_prob = runif(100),
  apache_4a_icu_death_prob = runif(100),
  hospital_death = sample(c(0, 1), 100, replace = TRUE)
)

p1 <- ggplot(dataset, aes(x = apache_4a_hospital_death_prob, y = hospital_d
eath)) +
  stat_smooth(method = "glm", method.args = list(family = "binomial"), se =
FALSE) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed") +
  xlab("Predicted probability of hospital death") +
  ylab("Observed frequency of hospital death") +
  ggtitle("Calibration plot for hospital predictions")
```

```
p2 <- ggplot(dataset, aes(x = apache_4a_icu_death_prob, y = hospital_death)
) +
  stat_smooth(method = "glm", method.args = list(family = "binomial"), se =
FALSE) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed") +
  xlab("Predicted probability of ICU death") +
  ylab("Observed frequency of hospital death") +
  ggtitle("Calibration plot for ICU predictions")

library(gridExtra)
grid.arrange(p1, p2, ncol=2)
```



Kao što vidimo, postoji solidno odstupanje od idealne krive. APACHE 4A predviđa smrt češće nego što se ona u stvarnosti dešava. Međutim, istraživanja su pokazala da "naštelovana" APACHE 4A metrika zapravo može biti solidan pokazatelj smrti. Moguće je da je naš dataset zasnovan na starijim podacima, ili da je model koji je služio za računanje ove metrike zastareo. *Zašto to mislimo?* Istraživanja su takođe pokazala da se predviđena stopa smrtnosti povećavala kako se starost modela povećavala."*Aggregate mortality was systematically overestimated as model age increased.*" Ovo ukazuje na napredak moderne medicine, metode lečenja za određene dijagnoze su poboljšane, što bi u realnom slučaju umanjilo izmereni APACHE skor i samim tim umanjilo i predviđenu verovatnoću smrtnosti. Kako bi uzeli ove stvari u obzir, važno je aktivno ažurirati APACHE model.

Citat:" Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's

critically ill patients. Crit Care Med. 2006 May;34(5):1297-310. doi: 10.1097/01.CCM.0000215112.84523.F0. PMID: 16540951."

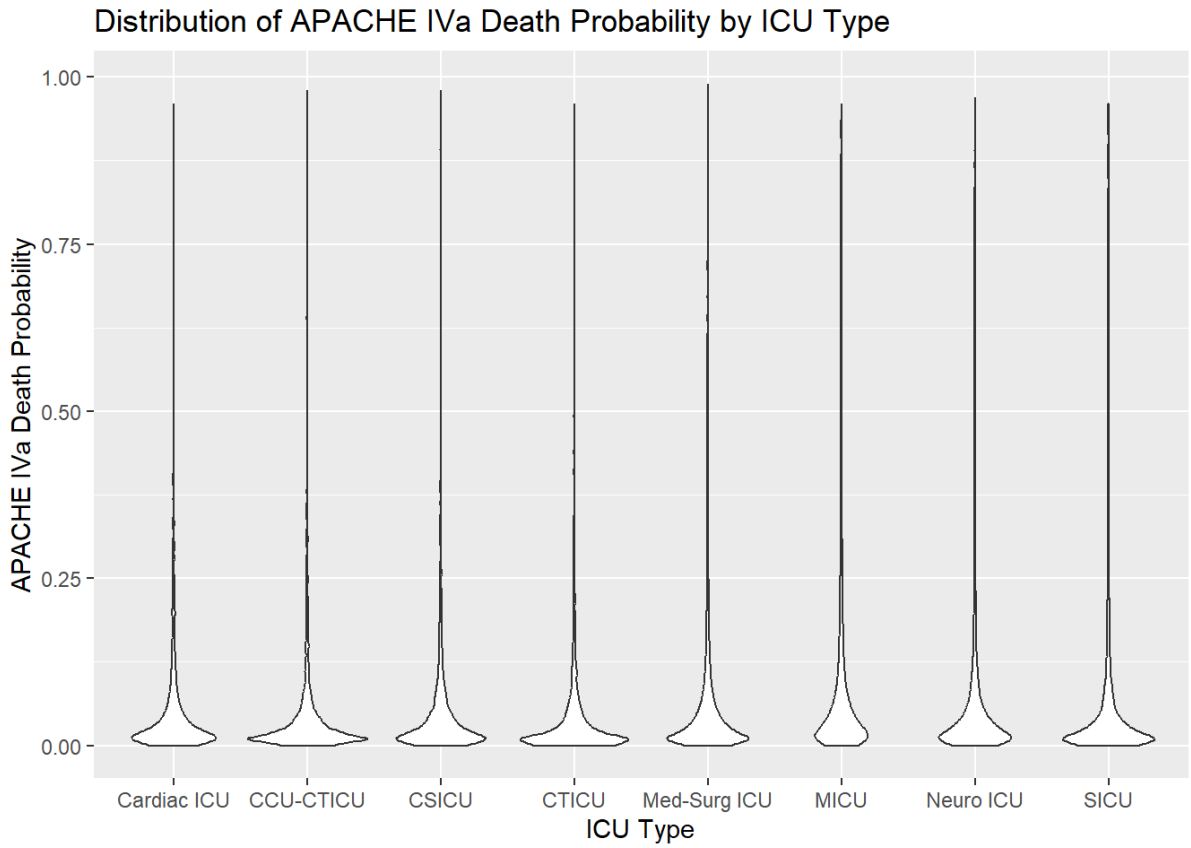
Nivo kalijuma i glukoze u krvi prilično je dobar indikator zdravlja osobe.

```
glucose_df <- cleaned_dataset %>%
  dplyr::select(d1_glucose_max, hospital_death) %>%
  mutate(d1_glucose_max = round(d1_glucose_max))

potassium_df <- cleaned_dataset %>%
  dplyr::select(d1_potassium_max, hospital_death) %>%
  mutate(d1_potassium_max = round(d1_potassium_max))

glucose_death <- glucose_df %>%
  group_by(d1_glucose_max) %>%
  summarize(avg_hospital_death = mean(hospital_death)) %>%
  ungroup()

potassium_death <- potassium_df %>%
  group_by(d1_potassium_max) %>%
  summarize(avg_hospital_death = mean(hospital_death)) %>%
  ungroup()

gg_glucose <- ggplot(glucose_death, aes(x = d1_glucose_max, y = avg_hospita
l_death)) +
  geom_line() +
  labs(title = "Relationship Between Glucose Levels and Hospital Death",
       x = "Max Glucose Level (mg/dL)",
       y = "Average Hospital Death Rate") +
  theme_minimal()

gg_potassium <- ggplot(potassium_death, aes(x = d1_potassium_max, y = avg_h
ospital_death)) +
  geom_line() +
  labs(title = "Relationship Between Potassium Levels and Hospital Death",
       x = "Max Potassium Level (mEq/L)",
       y = "Average Hospital Death Rate") +
  theme_minimal()

grid.arrange(gg_glucose,gg_potassium)
```

Ovde možemo videti da, što je njihov nivo veći, veća je i stopa smrtnosti, takođe smo na istraživanjem došli do zaključka da visok nivo kalijuma ukazuje na probleme sa bubrezima.

APACHE 4A može biti dobar pokazatelj performansi specijalizovanih medicinskih objekata. Hajde da pomoću njega uporedimo performanse svih odeljenja našeg dataseta.

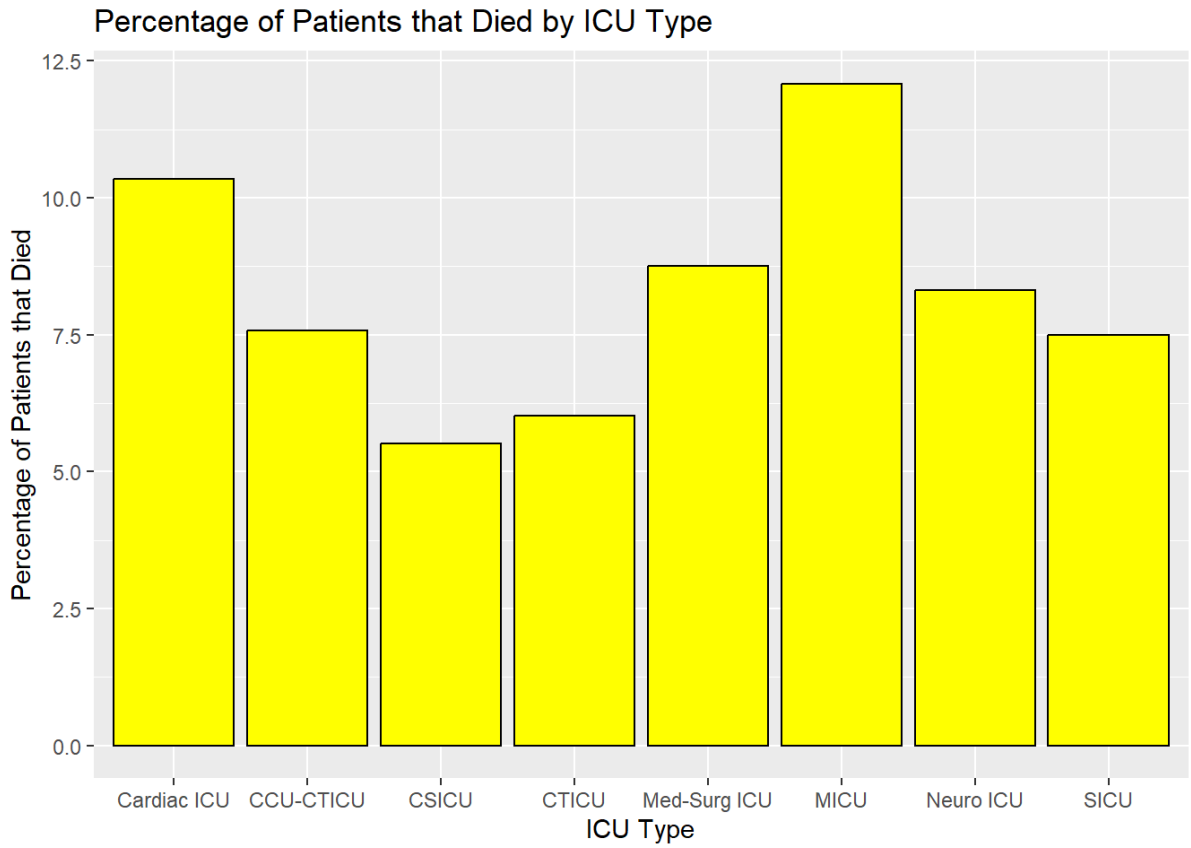
```
cleaned_dataset$death_prob <- ifelse(cleaned_dataset$hospital_death == 1, c
leaned_dataset$apache_4a_hospital_death_prob, cleaned_dataset$apache_4a_icu
_death_prob)


ggplot(cleaned_dataset, aes(x = icu_type, y = death_prob)) +
  geom_violin() +
  labs(title = "Distribution of APACHE IVa Death Probability by ICU Type",
x = "ICU Type", y = "APACHE IVa Death Probability")
```



Vidimo da su performanse relativno slične, i na zadovoljavajućem nivou(Većina pacijenata na bilo kom od odeljenja ima APACHE 4A verovatnoću smrti manju od 20-25%, i to nakon što smo ustanovili da model koji je korišćen za njeno računanje overshoot-uje) što nam govori da je rad svakog od odeljenja dobro regulisan.

Hajde sada da se nadovežemo, i proverimo realne performanse svih odeljenja. Doktori sa kog odeljenja imaju najduže pauze za kafu?

```
death_counts <- table(cleaned_dataset$icu_type, cleaned_dataset$hospital_death)

death_percentages <- death_counts[, "1"] / rowSums(death_counts) * 100


ggplot(data.frame(icu_type = names(death_percentages), death_percentage = death_percentages), aes(x = icu_type, y = death_percentage)) +

  geom_bar(stat = "identity", fill = "yellow", color = "black") +

  labs(title = "Percentage of Patients that Died by ICU Type", x = "ICU Type", y = "Percentage of Patients that Died")
```



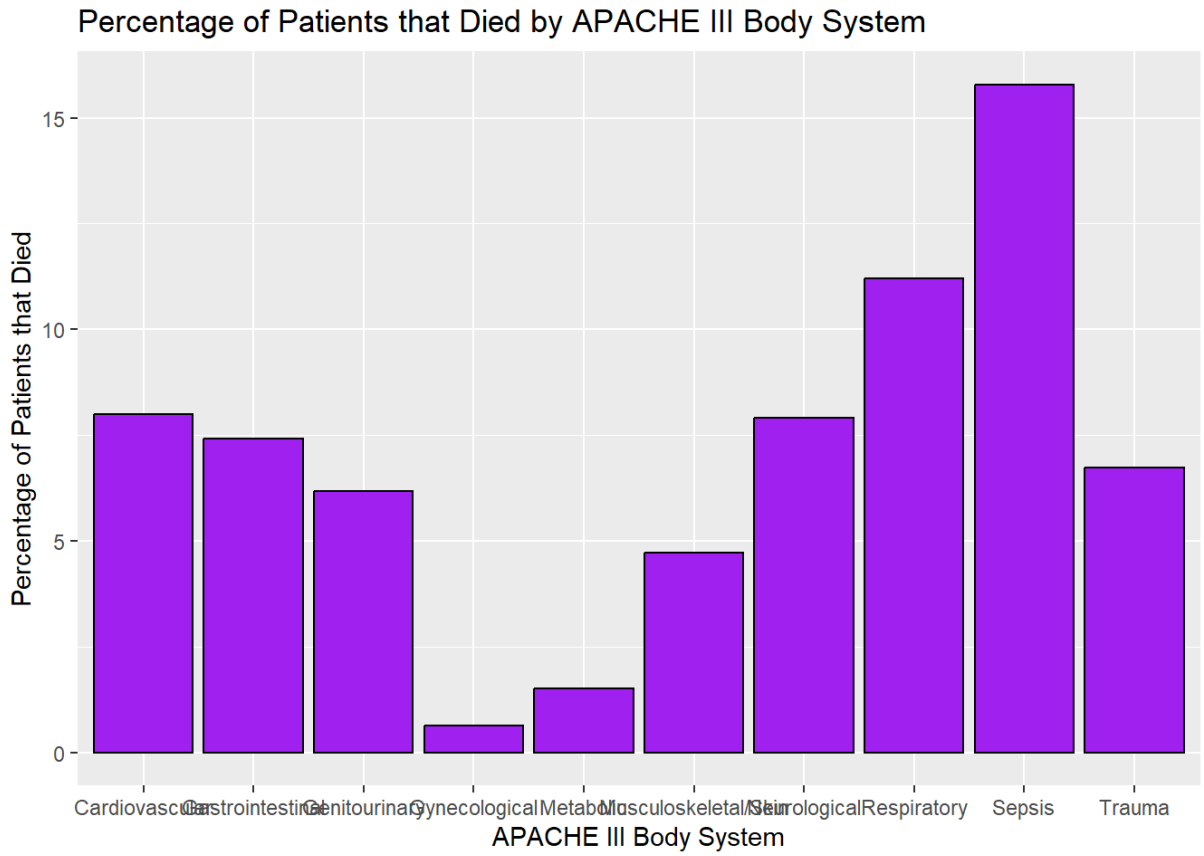
Vidimo da je stopa smrtnosti najviša za MICU, oko 12%. Da li doktori sa MIC odeljenja malo zabušavaju dok pijuckaju kaficu i dele tračeve? Ne. Priroda odeljenja je takva da su pacijenti primljeni na isto uglavnom ozbiljnijeg stanja nego na ostalim odeljenjima, pa je viša stopa smrtnosti i opravdana. Sa druge strane, doktori sa CSIC odeljenja sa vrlo niskom stopom smrtnosti od oko svega 5% zaslužuju jedan kraći s' mlekom.

Hajde sada da proverimo ozbiljnost svake od grupi bolesti.(APACHE 3 bodysystem. Korisitmo APACHE 3 umesto APACHE 2 jer je model tačniji)

```
death_counts_bs <- table(cleaned_dataset$apache_3j_bodysystem, cleaned_data
set$hospital_death)
death_percentages_bs <- death_counts_bs[, "1"] / rowSums(death_counts_bs) *
100
death_percentages_bs
```

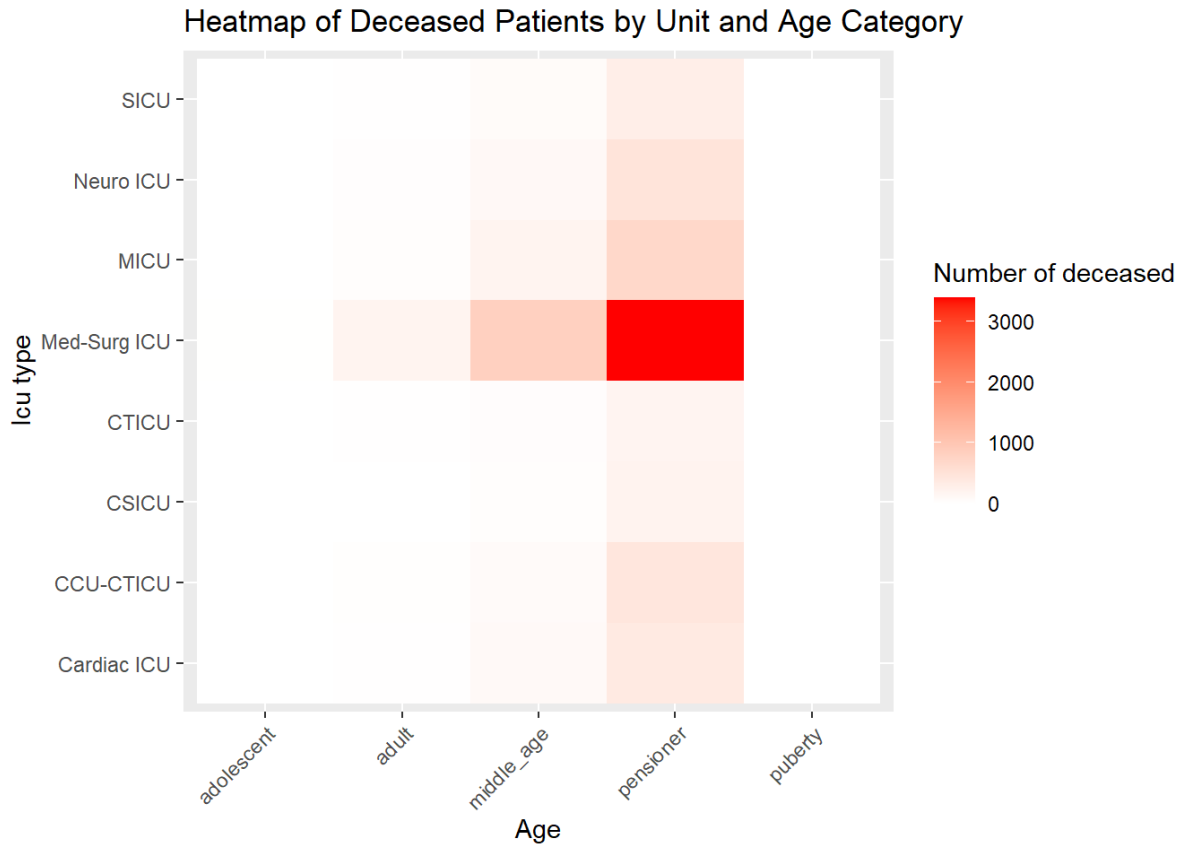
```
##       Cardiovascular     Gastrointestinal        Genitourinary
##            8.0002477            7.4230002            6.1694291
##         Gynecological           Metabolic Musculoskeletal/Skin
##            0.6389776            1.5163399            4.7169811
##          Neurological          Respiratory               Sepsis
##            7.9018157           11.2068223           15.7921635
##               Trauma
##            6.7412806
```

```
ggplot(data.frame(apache_3j_bodysystem = names(death_percentages_bs), death
_percentage = death_percentages_bs), aes(x = apache_3j_bodysystem, y = deat
h_percentages_bs)) +
  geom_bar(stat = "identity", fill = "purple", color = "black") +
  labs(title = "Percentage of Patients that Died by APACHE III Body System"
, x = "APACHE III Body System", y = "Percentage of Patients that Died")
```



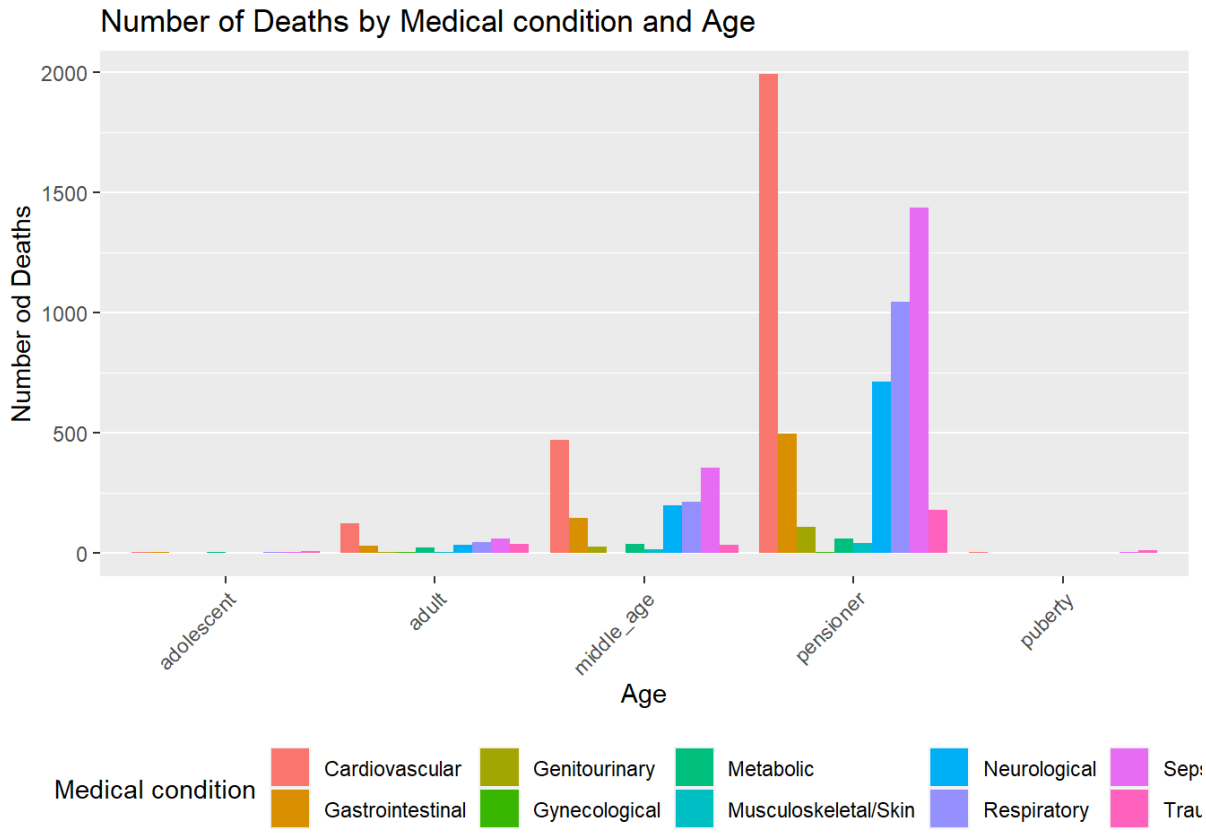
Vidimo da je sepsa najsmrtonosnija dijagnoza, što nam potvrđuje i prethodan zaključak.(Pacijenti oboleli od sepse se primaju na MIC odeljenje).

```
umrli_po_icu_age <- cleaned_dataset %>%
  group_by(icu_type, age) %>%
  summarise(ukupno_umrlih = sum(hospital_death))
heatmap_plot <- ggplot(umrli_po_icu_age, aes(x = age, y = icu_type, fill =
ukupno_umrlih)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "red") +
  labs(title = "Heatmap of Deceased Patients by Unit and Age Category",
       x = "Age",
       y = "Icu type",
       fill = "Number of deceased") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(heatmap_plot)
```



Primećujemo da je najviše umrlih imamo na Med-Surg ICU, specijalizovana jedinica unutar bolnice koja pruža intenzivnu medicinsku negu pacijentima koji su ozbiljno bolesni ili su nedavno prošli kroz hirurški zahvat. Potvrđujemo da najveći broj preminulih čine penizoneri zatim ljudi srednjih godina i odrasli.

```
data <- cleaned_dataset %>%
  group_by(apache_3j_bodysystem, age) %>%
  summarise(ukupno_umrlih = sum(hospital_death))
bar_plot <- ggplot(data, aes(x = age, y = ukupno_umrlih, fill = apache_3j_b
odysystem)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Number of Deaths by Medical condition and Age",
       x = "Age",
       y = "Number od Deaths",
       fill = "Medical condition") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),legend.position =
"bottom",
        panel.grid.major.x = element_blank())
bar_plot
```



Na ovom grafiku primećujemo da je u svakoj od starosnih grupa najviše preminulih imalo kardiovaskularne probleme. Kod penzionera, koji predstavljaju starosnu grupu sa najviše smrtnih ishoda, veliki broj preminulih je i od posledica sepse i respratornih problema.

Obrisaćemo varijable koje smo dodali u svrhu grafičkog prikaza.

```
cleaned_dataset <- subset(cleaned_dataset, select = -c(death_prob))
```

Selekcija

Sada cemo da predstavimo korelaciju izmedju feature-a kako bismo odredili koji od njih bi mogao da bude dobar prediktor. Za pocetak cemo predstaviti sledece kategorijske features (*Chi-squared test*).

```
anova_model <- lm(hospital_death ~ BMI_category + apache_2_bodysystem + apa
che_3j_bodysystem + icu_type + age + ethnicity + gender, data = cleaned_dat
aset)
anova_result <- anova(anova_model)
anova_result
```

```
## Analysis of Variance Table
##
## Response: hospital_death
##                          Df Sum Sq Mean Sq  F value    Pr(>F)
## BMI_category              3   21.3  7.0854  92.4862 < 2.2e-16 ***
```

```
## apache_2_bodysystem      8   87.7 10.9687 143.1759 < 2.2e-16 ***
## apache_3j_bodysystem     4   40.6 10.1538 132.5392 < 2.2e-16 ***
## icu_type                 7   11.2  1.5933  20.7974 < 2.2e-16 ***
## age                      4   46.1 11.5359 150.5796 < 2.2e-16 ***
## ethnicity                5    1.4  0.2715   3.5436  0.003323 **
## gender                   1    0.1  0.0519   0.6781  0.410251
## Residuals            91680 7023.6  0.0766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Zaključak:

1. BMI kategorija, apache_2_bodysystem, apache_3j_bodysystem, icu_type, age: za sve ove feature, p-vrednost je znatno manja od 0.05. To znači da postoji statistički značajna razlika u prosečnim vrednostima hospital_death i između različitih kategorija ovih feature-a.
2. ethnicity: iako p-vrednost za ovaj feature nije toliko mala kao za prethodne varijable, ona je ipak manja od 0.05, što ukazuje na statistički značajnu razliku u prosečnim vrednostima hospital_death među različitim područjima porekla (rase). Vrednost “**” nakon p-vrednosti označava da je razlika statistički značajna na nivou 0.01.
3. gender: za feature “gender”, p-vrednost je veća od 0.05, što znači da nema dovoljno dokaza da postoji statistički značajna razlika u prosečnim vrednostima hospital_death između polova.

```
numeric_subset <- cleaned_dataset[, sapply(cleaned_dataset, is.numeric)]
#Izračunavanje matrice korelacije
cor_matrix <- cor(numeric_subset, use = "complete.obs")
```

Za potrebe predikcije cemo zameniti mesta poslednjim dvema kolonama, kako bi nam *hospital_death* koju prediktujemo bila na poslednjem mestu.

```
cleaned_dataset <- cleaned_dataset %>%
  dplyr::select(-hospital_death) %>%
  bind_cols(hospital_death = cleaned_dataset$hospital_death)
#str(cleaned_dataset)
```

Sada ćemo sve vrednosti koje su tipa *string* da pretvorimo u numericke varijable, tj. da izvršimo faktorizaciju.

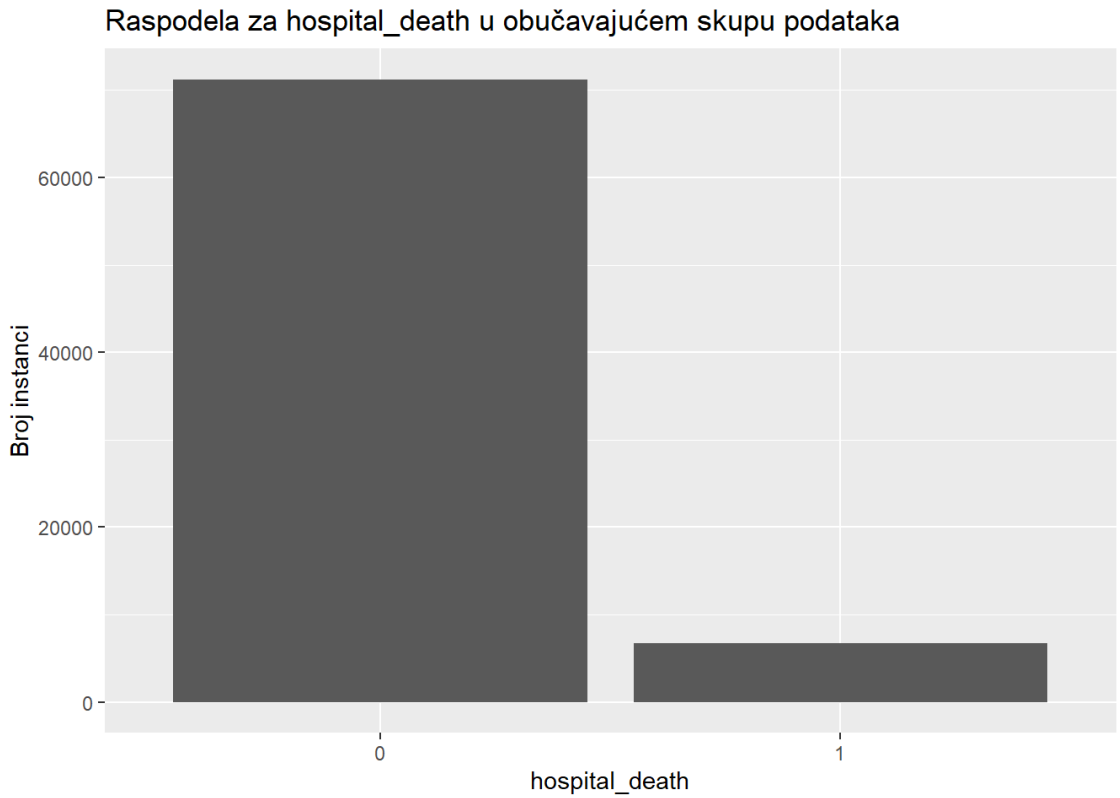
```
df <- cleaned_dataset
for (col in names(df)) {
  if (col != "hospital_death" && is.character(df[[col]])) {
    unique_vals <- unique(df[[col]])
    df[[col]] <- as.integer(factor(df[[col]]))}}
```

Modeli mašinskog učenja

Podelu smo izvršili tako da se 85% skupa koristi za treniranje, dok će se 15% koristiti za testiranje.

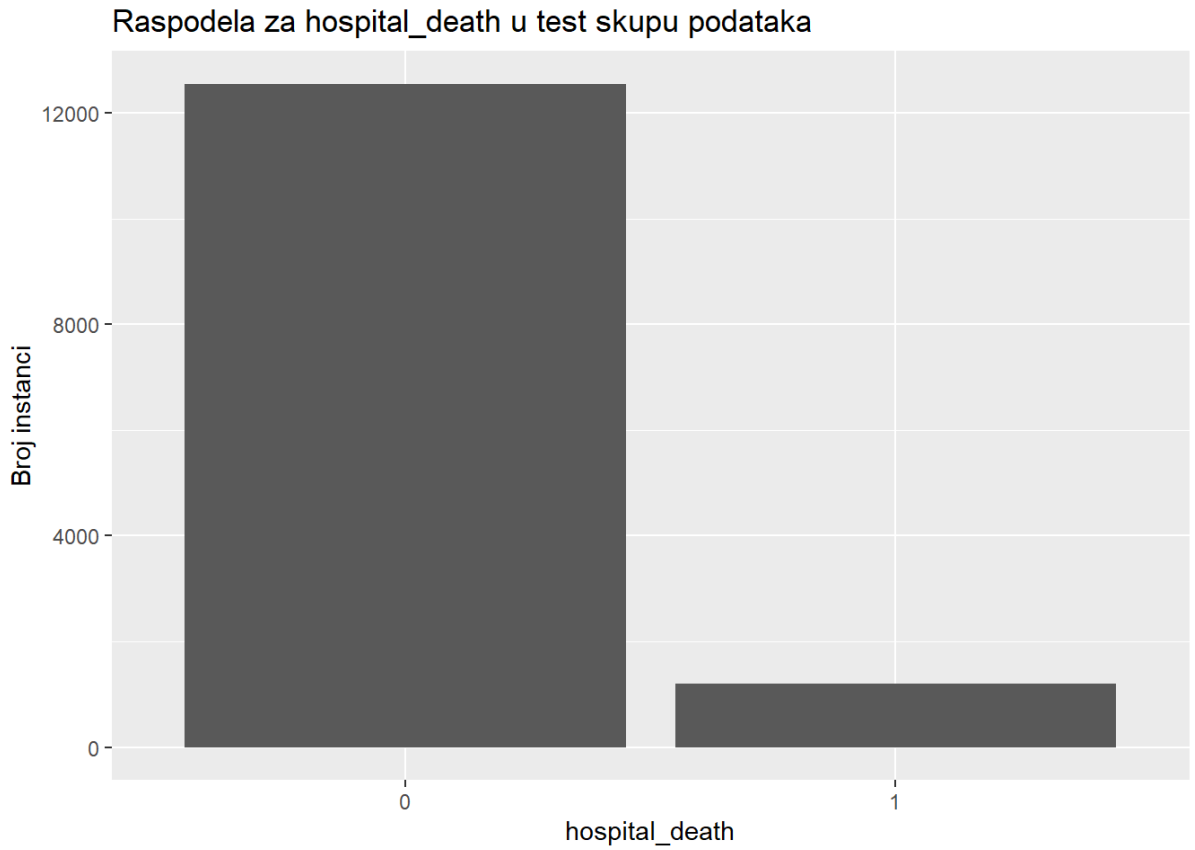
```
set.seed(123)
sample_size = floor(0.85*nrow(df))
train_index = sample(seq_len(nrow(df)), size = sample_size)
train = df[train_index,]
test = df[-train_index,]
```

```
train_plot <- ggplot(train, aes(x = factor(hospital_death))) +
  geom_bar() +
  labs(title = "Raspodela za hospital_death u obučavajućem skupu podataka",
       x = "hospital_death",
       y = "Broj instanci")
print(train_plot)
```




```
test_plot <- ggplot(test, aes(x = factor(hospital_death))) +
  geom_bar() +
  labs(title = "Raspodela za hospital_death u test skupu podataka",
       x = "hospital_death",
       y = "Broj instanci")
```

```
print(test_plot)
```




```
prop.table(table(train$hospital_death))
```

```
##
##          0          1
## 0.91388732 0.08611268
```

Dakle imamo otprilike 91% negativnih slučajeva i 9% pozitivnih što nam ukazuje na nebalansirane klase.

Ključna briga kod neravnoteženih(nebalansiranih) klasa je da modeli za mašinsko učenje mogu biti pristrasni prema većinskoj klasi i imati poteškoća u identifikaciji manjinske klase.

Resampling

Prvo ćemo korišćenjem decision tree algoritma da vidimo koliko nam loše ovo utiče na model.

```
library(ROSE)
```

```
proba <- rpart(hospital_death ~ ., data = train)
```

```
predikcija_proba <- predict(proba, newdata = test)
```

```
accuracy.meas(test$hospital_death,predikcija_proba[])
```

```
##
## Call:
## accuracy.meas(response = test$hospital_death, predicted = predikcija_pro
ba[])
##
## Examples are labelled as positive when predicted is greater than 0.5
##
## precision: 0.616
## recall: 0.224
## F: 0.164
```

Threshold vrednost je 0.5. Preciznos je 0.616, što znači da oko 61.6% pozitivnih predikcija vašeg modela su tačne, nije toliko dobro, odaziv je 0.224, što znači da je identifikovao samo 22.4% svih pozitivnih instanci, imamo dosta lažno negativnih vrednosti. Takođe F1-score koji je 0.164, što sugeriše da postoji prostor za poboljšanje ravnoteže između preciznosti i odziva.

Sada ćemo proveriti tačnost korišćenjem ROC krive. Ovo će nam dati jasnu sliku, koliko ovaj model vredi.

```
roc.curve(test$hospital_death, predikcija_proba[], plotit = F)
```

```
## Area under the curve (AUC): 0.810
```

AUC vrednost od 0.810 ukazuje na to da naš model ima dobru sposobnost razdvajanja klasa i bolje performanse od nasumičnog modela. To je pozitivan znak i sugeriše da model ima potencijal za donošenje korisnih predikcija. Dakle model nije loš, ali definitivno pre primene mašinskog učenja je potrebno da se podaci balansiraju.

```
xtabs(~hospital_death, data = train)
```

```
## hospital_death
##     0     1
## 71243  6713
```

Oversampling

```
data_balanced_over <- ovun.sample(hospital_death ~ ., data = train, method
= "over",N = 142486)$data
```

```
table(data_balanced_over$hospital_death)
```

```
##
##     0     1
## 71243 71243
```

Undersampling

```
data_balanced_under <- ovun.sample(hospital_death ~ ., data = train, method
= "under", N = 13426, seed = 1)$data
```

```
table(data_balanced_under$hospital_death)
```
```
##
##    0    1
## 6713 6713
```

Podaci su balansirani ali smo izgubili ključnu informaciju iz uzorka. Sada ćemo uraditi kombinaciju oversampling-a i undersampling-a.

```
dim(train)
```
```
## [1] 77956    66
```

```
data_balanced_both <- ovun.sample(hospital_death ~ ., data = train, method
= "both", p=0.5, N=77956, seed = 1)$data
```

```
table(data_balanced_both$hospital_death)
```
```
##
##     0     1
## 38853 39103
```

```
data.rose <- ROSE(hospital_death ~ ., data = train, seed = 1)$data
table(data.rose$hospital_death)
```
```
##
##     0     1
## 38853 39103
```

Sada treba da proverimo šta smo uradili.

```
tree.rose <- rpart(hospital_death ~ ., data = data.rose)
tree.over <- rpart(hospital_death ~ ., data = data_balanced_over)
tree.under <- rpart(hospital_death ~ ., data = data_balanced_under)
tree.both <- rpart(hospital_death ~ ., data = data_balanced_both)
```

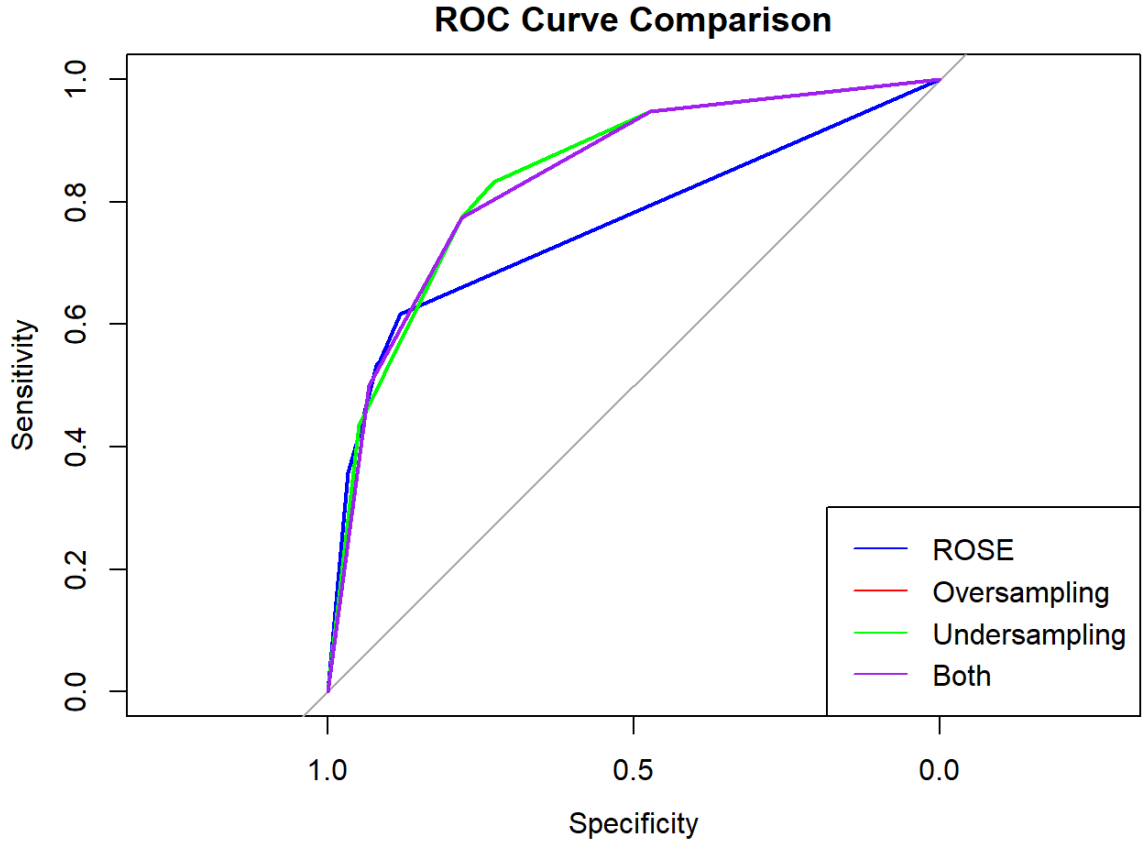
```
predict_rose <- predict(tree.rose, newdata = test)
predict_over <- predict(tree.over, newdata = test)
predict_under <- predict(tree.under, newdata = test)
predict_both <- predict(tree.both, newdata = test)
```

Pomoću ROC krive ćemo predstaviti naš rezultat.

```
roc_rose <- roc(test$hospital_death, predict_rose)
roc_over <- roc(test$hospital_death, predict_over)
roc_under <- roc(test$hospital_death, predict_under)
roc_both <- roc(test$hospital_death, predict_both)
```

```
plot(roc_rose, col = "blue", main = "ROC Curve Comparison")
```

```
lines(roc_over, col = "red")
lines(roc_under, col = "green")
lines(roc_both, col = "purple")
legend("bottomright", legend = c("ROSE", "Oversampling", "Undersampling", "
Both"),
       col = c("blue", "red", "green", "purple"), lty = 1)
```



Ne vidimo Oversampling liniju zato što joj je vrednost skoro ista kao za Undersampling.
ROSE (AUC): 0.761
Oversampling (AUC): 0.844
Undersampling (AUC): 0.843
Both (AUC): 0.839
Najbolji rezultat dobijamo oversampling metodom.

```
resampling_model <- ROSE.eval(hospital_death ~ ., data = train, learner = r
part, method.assess = "holdout", extr.pred = function(obj)obj[], seed = 1)
resampling_model
```

```
##
## Call:
## ROSE.eval(formula = hospital_death ~ ., data = train, learner = rpart,
##     extr.pred = function(obj) obj[], method.assess = "holdout",
##     seed = 1)
```

```
##
## Holdout estimate of auc: 0.755
```

```
X_train <- train[, -ncol(train)]
y_train <- train[, ncol(train)]
oversampled_data <- ROSE(hospital_death ~ ., data = train, seed = 1)$data
X_oversampled <- oversampled_data[, -ncol(oversampled_data)]
y_oversampled <- oversampled_data[, ncol(oversampled_data)]
```

F-regression

Sada ćemo da probamo da pronađemo fetures koji najviše utiču na naš model.

```
X_train <- X_oversampled
y_train <- y_oversampled
```

```
model <- lm(y_train ~ ., data = X_train)
f_regression <- summary(model)$fstatistic
p_values <- pf(f_regression[1], f_regression[2], f_regression[3], lower.tai
l = FALSE)
significant_features <- names(df)[-1][p_values < 0.05]
significant_features
```

```
##  [1] "bmi"                          "elective_surgery"
##  [3] "ethnicity"                    "gender"
##  [5] "height"                       "icu_type"
##  [7] "weight"                       "apache_2_diagnosis"
##  [9] "apache_3j_diagnosis"          "apache_post_operative"
## [11] "arf_apache"                   "gcs_eyes_apache"
## [13] "gcs_motor_apache"             "gcs_unable_apache"
## [15] "gcs_verbal_apache"            "heart_rate_apache"
## [17] "intubated_apache"             "map_apache"
## [19] "resprate_apache"              "temp_apache"
## [21] "ventilated_apache"            "d1_diasbp_max"
## [23] "d1_diasbp_min"                "d1_heartrate_max"
## [25] "d1_heartrate_min"             "d1_mbp_max"
## [27] "d1_mbp_min"                   "d1_resprate_max"
## [29] "d1_resprate_min"              "d1_spo2_max"
## [31] "d1_spo2_min"                  "d1_sysbp_max"
## [33] "d1_sysbp_min"                 "d1_temp_max"
## [35] "d1_temp_min"                  "h1_diasbp_max"
## [37] "h1_diasbp_min"                "h1_heartrate_max"
## [39] "h1_heartrate_min"             "h1_mbp_max"
## [41] "h1_mbp_min"                   "h1_resprate_max"
## [43] "h1_resprate_min"              "h1_spo2_max"
## [45] "h1_spo2_min"                  "h1_sysbp_max"
## [47] "h1_sysbp_min"                 "d1_glucose_max"
## [49] "d1_glucose_min"               "d1_potassium_max"
## [51] "d1_potassium_min"             "apache_4a_hospital_death_prob"
## [53] "apache_4a_icu_death_prob"     "aids"
## [55] "cirrhosis"                    "diabetes_mellitus"
## [57] "hepatic_failure"              "immunosuppression"
## [59] "leukemia"                     "lymphoma"
## [61] "solid_tumor_with_metastasis"  "apache_3j_bodysystem"
```

```
## [63] "apache_2_bodysystem"           "BMI_category"
## [65] "hospital_death"
```

Logistička regresija

Logistička regresija se koristi za modelovanje verovatnoće da se dogodi određeni događaj koji ima binarni izlaz (kod nas *hospital_death* ima izlaz 0 ili 1). Logistička regresija koristi logističku funkciju (sigmoidnu funkciju) kako bi transformisala linearnu kombinaciju prediktivnih feature-a u verovatnoću. Binomijalna raspodela se koristi za modeliranje slučajeva gde se događaji mogu podeliti u dve disktinte kategorije (obično uspeh i neuspeh) i interesuje nas koliko često se uspeh događa u nizu nezavisnih pokušaja, kod nas se odnosi na to da li je pacijent preziveo ili nije.

```
formula_str <- paste("y_train ~", paste(significant_features, collapse = "
+ "))
#cat("Formula:", formula_str, "\n")
glm1 <- glm( y_train ~ bmi + elective_surgery + ethnicity + gender + height
+ icu_type + weight + apache_2_diagnosis + apache_3j_diagnosis + apache_pos
t_operative + arf_apache + gcs_eyes_apache + gcs_motor_apache + gcs_unable_
apache + gcs_verbal_apache + heart_rate_apache + intubated_apache + map_apa
che + resprate_apache + temp_apache + ventilated_apache + d1_diasbp_max + d
1_diasbp_min + d1_heartrate_max + d1_heartrate_min + d1_mbp_max + d1_mbp_mi
n + d1_resprate_max + d1_resprate_min + d1_spo2_max + d1_spo2_min + d1_sysb
p_max + d1_sysbp_min + d1_temp_max + d1_temp_min + h1_diasbp_max + h1_diasb
p_min + h1_heartrate_max + h1_heartrate_min + h1_mbp_max + h1_mbp_min + h1_
resprate_max + h1_resprate_min + h1_spo2_max + h1_spo2_min + h1_sysbp_max +
h1_sysbp_min + d1_glucose_max + d1_glucose_min + d1_potassium_max + d1_pota
ssium_min + apache_4a_hospital_death_prob + apache_4a_icu_death_prob + aids
+ cirrhosis + diabetes_mellitus + hepatic_failure + immunosuppression + leu
kemia + lymphoma + solid_tumor_with_metastasis + apache_3j_bodysystem + apa
che_2_bodysystem + BMI_category , data.frame(X_train, y_train), family = "b
inomial")
summary(glm1)
```

```
##
## Call:
## glm(formula = y_train ~ bmi + elective_surgery + ethnicity +
##     gender + height + icu_type + weight + apache_2_diagnosis +
##     apache_3j_diagnosis + apache_post_operative + arf_apache +
##     gcs_eyes_apache + gcs_motor_apache + gcs_unable_apache +
##     gcs_verbal_apache + heart_rate_apache + intubated_apache +
##     map_apache + resprate_apache + temp_apache + ventilated_apache +
##     d1_diasbp_max + d1_diasbp_min + d1_heartrate_max + d1_heartrate_min +
##     d1_mbp_max + d1_mbp_min + d1_resprate_max + d1_resprate_min +
##     d1_spo2_max + d1_spo2_min + d1_sysbp_max + d1_sysbp_min +
##     d1_temp_max + d1_temp_min + h1_diasbp_max + h1_diasbp_min +
##     h1_heartrate_max + h1_heartrate_min + h1_mbp_max + h1_mbp_min +
##     h1_resprate_max + h1_resprate_min + h1_spo2_max + h1_spo2_min +
##     h1_sysbp_max + h1_sysbp_min + d1_glucose_max + d1_glucose_min +
##     d1_potassium_max + d1_potassium_min + apache_4a_hospital_death_prob +
##     apache_4a_icu_death_prob + aids + cirrhosis + diabetes_mellitus +
##     hepatic_failure + immunosuppression + leukemia + lymphoma +
##     solid_tumor_with_metastasis + apache_3j_bodysystem + apache_2_bodysystem +
##     BMI_category, family = "binomial", data = data.frame(X_train,
##     y_train))
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                  1.111e+01  6.033e-01  18.420  < 2e-16 ***
## bmi                         -4.463e-03  9.481e-04  -4.708 2.51e-06 ***
## elective_surgery            -2.271e-01  2.521e-02  -9.007  < 2e-16 ***
## ethnicity                   -7.706e-03  7.348e-03  -1.049 0.294302
## gender                       5.548e-02  1.507e-02   3.681 0.000232 ***
## height                      -1.401e-03  7.142e-04  -1.962 0.049748 *
## icu_type                     1.541e-02  4.259e-03   3.617 0.000298 ***
## weight                      -2.156e-03  3.302e-04  -6.530 6.59e-11 ***
## apache_2_diagnosis          -2.232e-06  8.716e-05  -0.026 0.979573
## apache_3j_diagnosis         -2.303e-04  1.999e-05 -11.525  < 2e-16 ***
## apache_post_operative       -1.156e-01  2.467e-02  -4.686 2.79e-06 ***
## arf_apache                   2.750e-01  3.821e-02   7.197 6.14e-13 ***
## gcs_eyes_apache             -7.363e-02  7.521e-03  -9.790  < 2e-16 ***
## gcs_motor_apache            -4.442e-02  5.172e-03  -8.589  < 2e-16 ***
## gcs_unable_apache            4.402e-01  5.698e-02   7.726 1.11e-14 ***
## gcs_verbal_apache           -6.872e-02  4.886e-03 -14.064  < 2e-16 ***
## heart_rate_apache            8.301e-04  2.556e-04   3.247 0.001164 **
## intubated_apache             6.813e-02  1.864e-02   3.654 0.000258 ***
## map_apache                   1.470e-04  1.704e-04   0.863 0.388385
## resprate_apache              4.157e-03  5.092e-04   8.163 3.26e-16 ***
## temp_apache                 -7.224e-02  7.424e-03  -9.731  < 2e-16 ***
## ventilated_apache            4.312e-01  1.615e-02  26.706  < 2e-16 ***
## d1_diasbp_max               -3.009e-04  4.088e-04  -0.736 0.461707
## d1_diasbp_min               -7.982e-03  6.342e-04 -12.585  < 2e-16 ***
## d1_heartrate_max             5.168e-03  3.644e-04  14.183  < 2e-16 ***
## d1_heartrate_min             7.874e-04  3.754e-04   2.098 0.035940 *
## d1_mbp_max                  -1.555e-03  4.383e-04  -3.548 0.000388 ***
## d1_mbp_min                  -7.845e-03  6.218e-04 -12.616  < 2e-16 ***
## d1_resprate_max              3.946e-03  6.883e-04   5.733 9.87e-09 ***
## d1_resprate_min              8.761e-03  1.302e-03   6.728 1.72e-11 ***
## d1_spo2_max                 -1.518e-02  3.714e-03  -4.087 4.36e-05 ***
## d1_spo2_min                 -1.154e-02  5.780e-04 -19.962  < 2e-16 ***
## d1_sysbp_max                 5.803e-04  3.098e-04   1.873 0.061088 .
## d1_sysbp_min                -4.464e-03  4.019e-04 -11.107  < 2e-16 ***
## d1_temp_max                 -1.660e-02  1.008e-02  -1.647 0.099594 .
## d1_temp_min                 -1.085e-01  8.387e-03 -12.939  < 2e-16 ***
## h1_diasbp_max               -9.956e-05  4.529e-04  -0.220 0.826010
## h1_diasbp_min               -2.322e-03  5.195e-04  -4.469 7.85e-06 ***
## h1_heartrate_max             9.977e-05  3.814e-04   0.262 0.793661
## h1_heartrate_min             1.611e-03  3.982e-04   4.045 5.24e-05 ***
## h1_mbp_max                  -3.236e-04  4.530e-04  -0.714 0.475124
## h1_mbp_min                  -3.151e-03  5.209e-04  -6.048 1.46e-09 ***
## h1_resprate_max              7.132e-03  9.931e-04   7.181 6.90e-13 ***
## h1_resprate_min              1.692e-02  1.196e-03  14.149  < 2e-16 ***
## h1_spo2_max                 -1.168e-02  2.113e-03  -5.526 3.27e-08 ***
## h1_spo2_min                 -5.468e-03  9.927e-04  -5.509 3.62e-08 ***
## h1_sysbp_max                -6.550e-04  3.061e-04  -2.140 0.032387 *
## h1_sysbp_min                -1.498e-03  3.383e-04  -4.427 9.54e-06 ***
## d1_glucose_max               3.959e-04  7.942e-05   4.985 6.19e-07 ***
## d1_glucose_min               1.344e-03  1.805e-04   7.450 9.35e-14 ***
## d1_potassium_max             1.027e-01  7.758e-03  13.234  < 2e-16 ***
## d1_potassium_min             8.003e-02  9.728e-03   8.227  < 2e-16 ***
## apache_4a_hospital_death_prob 1.885e+00  4.431e-02  42.540  < 2e-16 ***
## apache_4a_icu_death_prob      1.371e+00  5.168e-02  26.534  < 2e-16 ***
## aids                        -1.647e-01  2.337e-01  -0.705 0.480869
## cirrhosis                    2.646e-01  4.893e-02   5.409 6.35e-08 ***
## diabetes_mellitus           -7.667e-02  1.713e-02  -4.476 7.61e-06 ***
## hepatic_failure              2.499e-01  5.320e-02   4.696 2.65e-06 ***
## immunosuppression            2.325e-01  3.799e-02   6.121 9.32e-10 ***
## leukemia                     1.229e-01  7.160e-02   1.716 0.086118 .
## lymphoma                     3.832e-01  9.109e-02   4.207 2.58e-05 ***
```

```
## solid_tumor_with_metastasis     4.043e-01  4.138e-02   9.772  < 2e-16 ***
## apache_3j_bodysystem            2.348e-02  2.174e-03  10.801  < 2e-16 ***
## apache_2_bodysystem            -2.152e-02  2.757e-03  -7.806 5.88e-15 ***
## BMI_category                   -1.013e-04  7.632e-03  -0.013 0.989413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 108069  on 77955  degrees of freedom
## Residual deviance:  76838  on 77891  degrees of freedom
## AIC: 76968
##
## Number of Fisher Scoring iterations: 5
```

Početni model uključuje sve prediktore. Možemo da vidimo da imamo obeležja koja ne utiču na model (slabo utiču). AIC prvog modela je 76968. Cilj nam je da AIC bude što je moguće niža vrednost. Obratimo pažnju na sledeće: *elective_surgery*, *ethnicity*, *gender*, *height*, *apache_3j_diagnosis*, *apache_post_operative*, *arf_apache*, *gcs_eyes_apache*, *gcs_motor_apache*, *temp_apache*, *ventilated_apache*, *d1_diasbp_min*, *d1_heartrate_max*, *d1_mbp_min*, *d1_resprate_max*, *d1_resprate_min*, *d1_spo2_max*, *d1_spo2_min*, *d1_sysbp_min*, *d1_temp_min*, *h1_diasbp_min*, *h1_heartrate_min*, *h1_mbp_min*, *h1_resprate_max*, *h1_resprate_min*, *h1_spo2_max*, *h1_spo2_min*, *h1_sysbp_max*, *h1_sysbp_min*, *d1_glucose_max*, *d1_glucose_min*,*d1_potassium_max*, d*1_potassium_min*, *apache_4a_hospital_death_prob*, *apache_4a_icu_death_prob*, *cirrhosis*, *diabetes_mellitus*, *hepatic_failure*, *immunosuppression*, *leukemia*, *lymphoma*, *solid_tumor_with_metastasis*, *apache_3j_bodysystem*, *apache_2_bodysystem*, *BMI_category* imaju p-vrednosti manje od 0.05, što ukazuje na njihovu značajnost. Takođe pored p-vrednosti smo se bazirali na domenskom znanju prilikom izdvajanja feature-a.

```
glm2 <- glm(formula = y_train ~ elective_surgery + ethnicity + gender + hei
ght + apache_3j_diagnosis + apache_post_operative + arf_apache + gcs_eyes_a
pache + gcs_motor_apache + temp_apache + ventilated_apache + d1_diasbp_min
+ d1_heartrate_max + d1_mbp_min + d1_resprate_max + d1_resprate_min + d1_sp
o2_max + d1_spo2_min + d1_sysbp_min + d1_temp_min + h1_diasbp_min + h1_hear
trate_min + h1_mbp_min + h1_resprate_max + h1_resprate_min + h1_spo2_max +
h1_spo2_min + h1_sysbp_max + h1_sysbp_min + d1_glucose_max + d1_glucose_min
+ d1_potassium_max + d1_potassium_min + apache_4a_hospital_death_prob + apa
che_4a_icu_death_prob + cirrhosis + diabetes_mellitus + hepatic_failure + i
mmunosuppression + leukemia + lymphoma + solid_tumor_with_metastasis + apac
he_3j_bodysystem + apache_2_bodysystem + BMI_category, data.frame(X_train,
y_train), family = "binomial")
summary(glm2)
```

```
##
## Call:
## glm(formula = y_train ~ elective_surgery + ethnicity + gender +
##     height + apache_3j_diagnosis + apache_post_operative + arf_apache +
##     gcs_eyes_apache + gcs_motor_apache + temp_apache + ventilated_apache +
##     d1_diasbp_min + d1_heartrate_max + d1_mbp_min + d1_resprate_max +
##     d1_resprate_min + d1_spo2_max + d1_spo2_min + d1_sysbp_min +
##     d1_temp_min + h1_diasbp_min + h1_heartrate_min + h1_mbp_min +
##     h1_resprate_max + h1_resprate_min + h1_spo2_max + h1_spo2_min +
##     h1_sysbp_max + h1_sysbp_min + d1_glucose_max + d1_glucose_min +
##     d1_potassium_max + d1_potassium_min + apache_4a_hospital_death_prob +
##     apache_4a_icu_death_prob + cirrhosis + diabetes_mellitus +
```

```
##     hepatic_failure + immunosuppression + leukemia + lymphoma +
##     solid_tumor_with_metastasis + apache_3j_bodysystem + apache_2_bodysystem +
##     BMI_category, family = "binomial", data = data.frame(X_train,
##     y_train))
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    1.009e+01  5.194e-01  19.425  < 2e-16 ***
## elective_surgery              -2.214e-01  2.502e-02  -8.848  < 2e-16 ***
## ethnicity                     -3.132e-03  7.309e-03  -0.428 0.668293
## gender                         4.474e-02  1.494e-02   2.994 0.002752 **
## height                        -2.359e-03  6.926e-04  -3.406 0.000658 ***
## apache_3j_diagnosis           -2.183e-04  1.967e-05 -11.101  < 2e-16 ***
## apache_post_operative         -1.083e-01  2.444e-02  -4.431 9.36e-06 ***
## arf_apache                     2.726e-01  3.810e-02   7.155 8.37e-13 ***
## gcs_eyes_apache               -1.043e-01  7.169e-03 -14.547  < 2e-16 ***
## gcs_motor_apache              -5.484e-02  5.058e-03 -10.842  < 2e-16 ***
## temp_apache                   -7.342e-02  7.283e-03 -10.080  < 2e-16 ***
## ventilated_apache              4.754e-01  1.539e-02  30.892  < 2e-16 ***
## d1_diasbp_min                 -7.523e-03  6.276e-04 -11.987  < 2e-16 ***
## d1_heartrate_max               5.746e-03  3.283e-04  17.502  < 2e-16 ***
## d1_mbp_min                    -7.668e-03  6.174e-04 -12.420  < 2e-16 ***
## d1_resprate_max                5.188e-03  6.598e-04   7.864 3.73e-15 ***
## d1_resprate_min                1.090e-02  1.277e-03   8.534  < 2e-16 ***
## d1_spo2_max                   -1.300e-02  3.664e-03  -3.548 0.000389 ***
## d1_spo2_min                   -1.140e-02  5.733e-04 -19.881  < 2e-16 ***
## d1_sysbp_min                  -4.560e-03  3.977e-04 -11.465  < 2e-16 ***
## d1_temp_min                   -1.099e-01  8.343e-03 -13.177  < 2e-16 ***
## h1_diasbp_min                 -2.376e-03  5.076e-04  -4.681 2.86e-06 ***
## h1_heartrate_min               2.195e-03  3.563e-04   6.160 7.26e-10 ***
## h1_mbp_min                    -3.355e-03  5.146e-04  -6.520 7.03e-11 ***
## h1_resprate_max                7.519e-03  9.735e-04   7.723 1.13e-14 ***
## h1_resprate_min                1.805e-02  1.183e-03  15.254  < 2e-16 ***
## h1_spo2_max                   -1.078e-02  2.089e-03  -5.162 2.45e-07 ***
## h1_spo2_min                   -5.348e-03  9.845e-04  -5.433 5.55e-08 ***
## h1_sysbp_max                  -8.731e-04  2.662e-04  -3.280 0.001037 **
## h1_sysbp_min                  -1.597e-03  3.348e-04  -4.771 1.83e-06 ***
## d1_glucose_max                 4.100e-04  7.898e-05   5.191 2.09e-07 ***
## d1_glucose_min                 1.166e-03  1.792e-04   6.507 7.67e-11 ***
## d1_potassium_max               9.971e-02  7.708e-03  12.935  < 2e-16 ***
## d1_potassium_min               7.081e-02  9.655e-03   7.334 2.24e-13 ***
## apache_4a_hospital_death_prob  1.985e+00  4.389e-02  45.239  < 2e-16 ***
## apache_4a_icu_death_prob       1.454e+00  5.118e-02  28.402  < 2e-16 ***
## cirrhosis                      2.465e-01  4.885e-02   5.046 4.51e-07 ***
## diabetes_mellitus             -1.049e-01  1.695e-02  -6.187 6.13e-10 ***
## hepatic_failure                2.355e-01  5.297e-02   4.446 8.75e-06 ***
## immunosuppression              2.262e-01  3.784e-02   5.978 2.26e-09 ***
## leukemia                       1.025e-01  7.152e-02   1.434 0.151675
## lymphoma                       3.836e-01  9.064e-02   4.233 2.31e-05 ***
## solid_tumor_with_metastasis    4.048e-01  4.121e-02   9.822  < 2e-16 ***
## apache_3j_bodysystem           2.533e-02  2.129e-03  11.899  < 2e-16 ***
## apache_2_bodysystem           -1.980e-02  2.704e-03  -7.322 2.44e-13 ***
## BMI_category                  -2.306e-03  7.602e-03  -0.303 0.761641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 108069  on 77955  degrees of freedom
## Residual deviance:  77385  on 77910  degrees of freedom
## AIC: 77477
```

```
##
## Number of Fisher Scoring iterations: 5
```

Sada izbacivanjem feture-a koji imaju lošu p-vrednost dobijamo AIC (= 77477) koji je lošiji nego prilikom korišćenja svih feture-a u prvom modelu (AIC = 76968). Zadržaćemo se na našem prvom modelu.

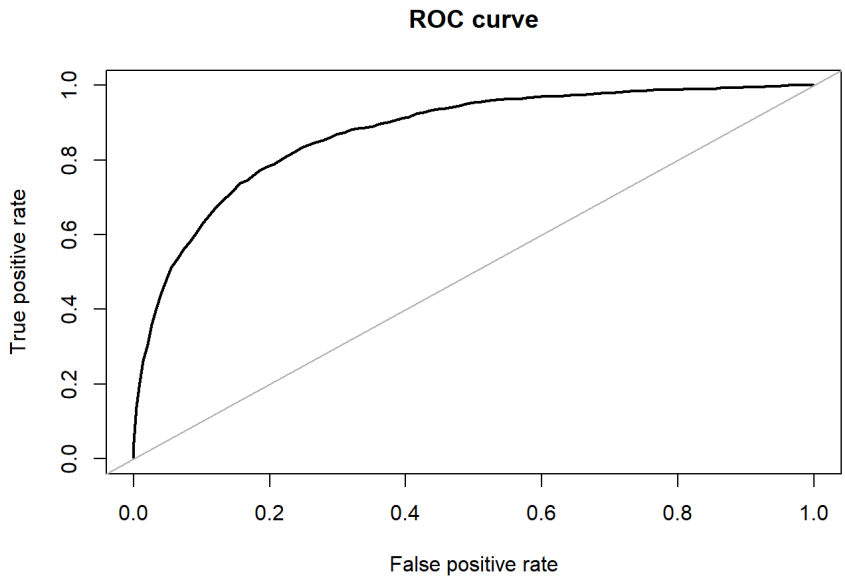
```
glm3 <- glm(formula = y_train ~ ., data.frame(X_train, y_train), family = "binomial")
summary(glm3)
```

```
##
## Call:
## glm(formula = y_train ~ ., family = "binomial", data = data.frame(X_train,
##     y_train))
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   8.835e+00  6.077e-01  14.539  < 2e-16 ***
## age                           3.591e-01  1.137e-02  31.576  < 2e-16 ***
## bmi                          -3.299e-03  9.562e-04  -3.449 0.000562 ***
## elective_surgery             -2.552e-01  2.550e-02 -10.010  < 2e-16 ***
## ethnicity                    -7.956e-03  7.428e-03  -1.071 0.284105
## gender                        5.006e-02  1.520e-02   3.294 0.000986 ***
## height                       -6.134e-04  7.201e-04  -0.852 0.394292
## icu_type                      1.684e-02  4.291e-03   3.925 8.67e-05 ***
## weight                       -1.805e-03  3.329e-04  -5.422 5.88e-08 ***
## apache_2_diagnosis           -3.819e-05  8.791e-05  -0.434 0.663975
## apache_3j_diagnosis          -2.128e-04  2.021e-05 -10.533  < 2e-16 ***
## apache_post_operative        -1.358e-01  2.493e-02  -5.448 5.11e-08 ***
## arf_apache                    2.933e-01  3.847e-02   7.624 2.45e-14 ***
## gcs_eyes_apache              -8.098e-02  7.574e-03 -10.692  < 2e-16 ***
## gcs_motor_apache             -5.293e-02  5.204e-03 -10.171  < 2e-16 ***
## gcs_unable_apache             4.759e-01  5.726e-02   8.312  < 2e-16 ***
## gcs_verbal_apache            -6.921e-02  4.920e-03 -14.065  < 2e-16 ***
## heart_rate_apache             1.017e-03  2.572e-04   3.952 7.74e-05 ***
## intubated_apache              7.701e-02  1.874e-02   4.110 3.95e-05 ***
## map_apache                    1.553e-04  1.713e-04   0.907 0.364649
## resprate_apache               4.321e-03  5.137e-04   8.412  < 2e-16 ***
## temp_apache                  -7.398e-02  7.448e-03  -9.934  < 2e-16 ***
## ventilated_apache             4.289e-01  1.626e-02  26.374  < 2e-16 ***
## d1_diasbp_max                -5.650e-05  4.116e-04  -0.137 0.890807
## d1_diasbp_min                -7.056e-03  6.396e-04 -11.033  < 2e-16 ***
## d1_heartrate_max              5.456e-03  3.669e-04  14.868  < 2e-16 ***
## d1_heartrate_min              9.386e-04  3.773e-04   2.487 0.012865 *
## d1_mbp_max                   -1.581e-03  4.415e-04  -3.580 0.000343 ***
## d1_mbp_min                   -7.445e-03  6.263e-04 -11.888  < 2e-16 ***
## d1_resprate_max               4.252e-03  6.930e-04   6.136 8.47e-10 ***
## d1_resprate_min               7.909e-03  1.311e-03   6.035 1.59e-09 ***
## d1_spo2_max                  -1.338e-02  3.711e-03  -3.605 0.000312 ***
## d1_spo2_min                  -1.133e-02  5.804e-04 -19.512  < 2e-16 ***
## d1_sysbp_max                  1.619e-04  3.121e-04   0.519 0.604006
## d1_sysbp_min                 -4.365e-03  4.044e-04 -10.793  < 2e-16 ***
## d1_temp_max                  -2.928e-03  1.015e-02  -0.288 0.773084
## d1_temp_min                  -1.113e-01  8.407e-03 -13.237  < 2e-16 ***
## h1_diasbp_max                 2.429e-04  4.562e-04   0.532 0.594393
## h1_diasbp_min                -1.667e-03  5.237e-04  -3.184 0.001455 **
## h1_heartrate_max              5.072e-04  3.844e-04   1.320 0.186998
## h1_heartrate_min              1.966e-03  4.010e-04   4.904 9.39e-07 ***
## h1_mbp_max                   -2.940e-04  4.562e-04  -0.644 0.519352
```

```
## h1_mbp_min                    -2.816e-03  5.247e-04  -5.366 8.07e-08 ***
## h1_resprate_max                7.309e-03  1.001e-03   7.305 2.77e-13 ***
## h1_resprate_min                1.633e-02  1.203e-03  13.580  < 2e-16 ***
## h1_spo2_max                   -1.018e-02  2.114e-03  -4.813 1.49e-06 ***
## h1_spo2_min                   -5.107e-03  9.981e-04  -5.116 3.12e-07 ***
## h1_sysbp_max                  -1.034e-03  3.083e-04  -3.353 0.000800 ***
## h1_sysbp_min                  -1.740e-03  3.405e-04  -5.110 3.23e-07 ***
## d1_glucose_max                 4.349e-04  8.010e-05   5.429 5.66e-08 ***
## d1_glucose_min                 1.110e-03  1.817e-04   6.106 1.02e-09 ***
## d1_potassium_max               1.062e-01  7.822e-03  13.578  < 2e-16 ***
## d1_potassium_min               7.125e-02  9.802e-03   7.268 3.64e-13 ***
## apache_4a_hospital_death_prob  1.763e+00  4.455e-02  39.572  < 2e-16 ***
## apache_4a_icu_death_prob       1.349e+00  5.181e-02  26.031  < 2e-16 ***
## aids                          -2.725e-02  2.366e-01  -0.115 0.908328
## cirrhosis                      3.219e-01  4.921e-02   6.542 6.06e-11 ***
## diabetes_mellitus             -1.012e-01  1.726e-02  -5.865 4.49e-09 ***
## hepatic_failure                2.860e-01  5.338e-02   5.359 8.38e-08 ***
## immunosuppression              2.274e-01  3.819e-02   5.955 2.60e-09 ***
## leukemia                       1.034e-01  7.179e-02   1.440 0.149797
## lymphoma                       3.699e-01  9.179e-02   4.030 5.57e-05 ***
## solid_tumor_with_metastasis    4.061e-01  4.147e-02   9.794  < 2e-16 ***
## apache_3j_bodysystem           2.460e-02  2.190e-03  11.232  < 2e-16 ***
## apache_2_bodysystem           -1.842e-02  2.778e-03  -6.630 3.37e-11 ***
## BMI_category                  -2.153e-03  7.691e-03  -0.280 0.779492
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 108069  on 77955  degrees of freedom
## Residual deviance:  75808  on 77890  degrees of freedom
## AIC: 75940
##
## Number of Fisher Scoring iterations: 5
```

Ostalo nam je još da isprobamo model koji uključuje sve feture i u ovom modelu dobijamo AIC (= 75940) koji je najbolji do sada. Zadržaćemo se na ovom modelu.

```
#AUC = 0.873
prediction_glm3 <- predict(glm3, test, type="response")
roc.curve(test$hospital_death, prediction_glm3[], plotit = T)
```




```
## Area under the curve (AUC): 0.873
```

Naš prvi model je veoma dobro prediktovao podatke (AUC = 0.873), ali hajde da to potvrdimo metrikama. Za optimalni threshold bismo mogli da uzmemo vrednost 0.7, ali ćemo za svaki slučaj to proveriti.

Accuracy = TP + TN / TP + TN + FP + FN
Precision = TP / TP + FP
Recall = TP / TP + FN
F1-score = 2 * (Precision * Recall) / (Precision + Recall)
Za početak ćemo odrediti *treshold* i *kappa-score*. Cohen's Kappa (kappa-score), je statistička mera koja se koristi za procenu stepena usklađenosti (concordance) između stvarnih i predviđenih klasa u binarnoj ili višeklasnoj klasifikaciji. Ova mera uzima u obzir slučajnu usklađenost i pruža bolju procenu performansi modela od same tačnosti kada se suočavate sa neuravnoteženim klasama ili slučajnim predviđanjima.

```
predicted_probabilities <- prediction_glm3
actual_classes <- test$hospital_death

threshold_grid <- seq(0.1, 0.9, by = 0.1)
best_kappa <- -Inf
optimal_threshold <- NULL

for (threshold in threshold_grid) {
  predicted_classes <- ifelse(predicted_probabilities >= threshold, 1, 0)
  kappa <- kappa2(data.frame(predicted = predicted_classes, actual = actual
_classes))$value
  if (kappa > best_kappa) {
    best_kappa <- kappa
```

```
    optimal_threshold <- threshold }}

print(paste("Optimalni threshold:", optimal_threshold))
## [1] "Optimalni threshold: 0.7"
print(paste("Najbolji Kappa-Score:", best_kappa))
## [1] "Najbolji Kappa-Score: 0.426849019373061"
```

Kappa-Score nam pokazuje da naš model OK usklađuje predviđene i stvarne klase. Threshold je 0.7, ista vrednost koju smo mi slobodnim odabirom na osnovu ROC krive odredili.

```
table(ifelse(prediction_glm3 > 0.7, 1, 0), test$hospital_death)
##
##         0     1
##   0 11709   551
##   1   846   651
```

```
conf_matrix_glm = confusionMatrix(table(ifelse(prediction_glm3 > 0.7, 1, 0)
, test$hospital_death))
```

Accuracy

```
#Accuracy = TP + TN / TP + TN + FP + FN => 0.90
accuracy <- conf_matrix_glm$overall["Accuracy"]
accuracy_str <- sprintf("Accuracy: %.2f", accuracy)
print(accuracy_str)
## [1] "Accuracy: 0.90"
```

Precision

```
#Precision = TP / TP + FP => 0.96
precision <- conf_matrix_glm$byClass["Pos Pred Value"]
precision_str <- sprintf("Precision: %.2f", precision)
print(precision_str)
## [1] "Precision: 0.96"
```

Recall

```
#Recall = TP / TP + FN => 0.93
print(paste(round(conf_matrix_glm$byClass["Sensitivity"], 2)))
## [1] "0.93"
```

F1-score

```
#F1 - score = 2 * (Precision * Recall) / (Precision + Recall) => 0.94
print(paste("F1-Score:", round(conf_matrix_glm$byClass["F1"], 2)))
```

```
## [1] "F1-Score: 0.94"
```

Decision tree

Stablo odlučivanja je moćan algoritam mašinskog učenja koji se koristi za klasifikaciju i regresiju. Ovaj algoritam ima široku primenu u analizi podataka i donošenju odluka.

Medicinska dijagnostika: Stablo odlučivanja se koristi u medicinskim istraživanjima i dijagnostici za donošenje odluka o dijagnozi na osnovu medicinskih simptoma i karakteristika.

type = 5: Generiše prikaz stabla sa podeocima i horizontalnim rasporedom. Ovo je često korisno za veća stabla kako bi se izbegao problem pretrpane vizualizacije.

1. Prvi model koristi varijable koje smo dobili kao najuticajnije modelom logističke regresije.

```
stablo1 = rpart(y_train ~ elective_surgery + weight + apache_3j_diagnosis +
apache_post_operative + arf_apache + gcs_eyes_apache + gcs_unable_apache +
heart_rate_apache + resprate_apache + ventilated_apache + d1_heartrate_max
+ d1_resprate_min + d1_spo2_min + h1_heartrate_max + h1_resprate_min + d1_g
lucose_min + d1_potassium_max + apache_4a_hospital_death_prob + apache_4a_i
cu_death_prob + diabetes_mellitus + hepatic_failure + immunosuppression + s
olid_tumor_with_metastasis + apache_3j_bodysystem + apache_2_bodysystem, da
ta.frame(X_train, y_train), method = "class")
#prp(stablo1, type = 5)
```

```
prediction_decision_tree = predict(stablo1, test, type="class")
table(prediction_decision_tree, test$hospital_death)
```

```
##
## prediction_decision_tree     0     1
##                        0 11139   468
##                        1  1416   734
```

```
confusion_matrix_dt1 = confusionMatrix(table(prediction_decision_tree, test
$hospital_death))
```

Accuracy

```
#Accuracy = TP + TN / TP + TN + FP + FN => 0.86
accuracy <- confusion_matrix_dt1$overall["Accuracy"]
accuracy_str <- sprintf("Accuracy: %.2f", accuracy)
print(accuracy_str)
```

```
## [1] "Accuracy: 0.86"
```

Precision

```
#Precision = TP / TP + FP => 0.96
precision <- confusion_matrix_dt1$byClass["Pos Pred Value"]
precision_str <- sprintf("Precision: %.2f", precision)
print(precision_str)
## [1] "Precision: 0.96"
```

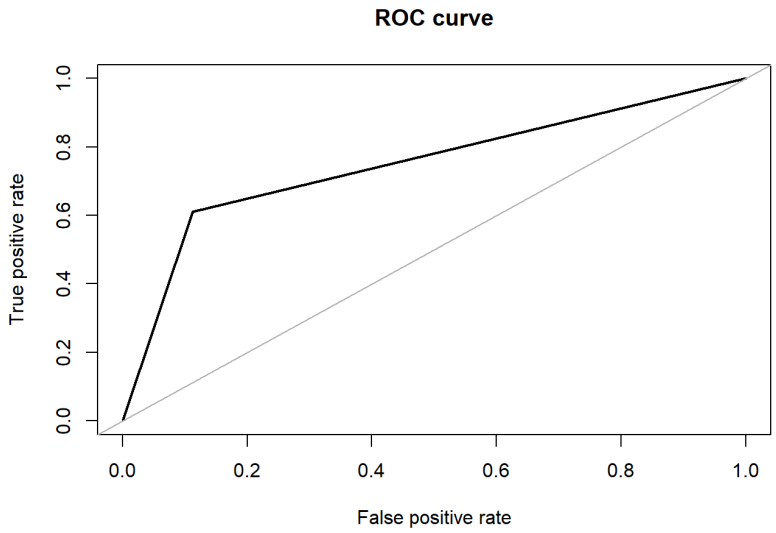
Recall

```
#Recall = TP / TP + FN => 0.89
print(paste(round(confusion_matrix_dt1$byClass["Sensitivity"], 2)))
## [1] "0.89"
```

F1-score

```
#F1 - score = 2 * (Precision * Recall) / (Precision + Recall) => 0.92
print(paste("F1-Score:", round(confusion_matrix_dt1$byClass["F1"], 2)))
## [1] "F1-Score: 0.92"
```

```
#AUC = 0.749
predict_dt1 <- predict(stablo1, test, type="class")
roc.curve(test$hospital_death, predict_dt1[], plotit = T)
```




```
## Area under the curve (AUC): 0.749
```

2. Koristimo varijable dobijene kao najrelevantnije(significant_fetures).

```
stablo2 <- rpart(y_train ~ bmi + elective_surgery + ethnicity + gender + he
ight + icu_type + weight + apache_2_diagnosis + apache_3j_diagnosis + apach
e_post_operative + arf_apache + gcs_eyes_apache + gcs_motor_apache + gcs_un
able_apache + gcs_verbal_apache + heart_rate_apache + intubated_apache + ma
p_apache + resprate_apache + temp_apache + ventilated_apache + d1_diasbp_ma
x + d1_diasbp_min + d1_heartrate_max + d1_heartrate_min + d1_mbp_max + d1_m
bp_min + d1_resprate_max + d1_resprate_min + d1_spo2_max + d1_spo2_min + d1
_sysbp_max + d1_sysbp_min + d1_temp_max + d1_temp_min + h1_diasbp_max + h1_
diasbp_min + h1_heartrate_max + h1_heartrate_min + h1_mbp_max + h1_mbp_min
+ h1_resprate_max + h1_resprate_min + h1_spo2_max + h1_spo2_min + h1_sysbp_
max + h1_sysbp_min + d1_glucose_max + d1_glucose_min + d1_potassium_max + d
1_potassium_min + apache_4a_hospital_death_prob + apache_4a_icu_death_prob
+ aids + cirrhosis + diabetes_mellitus + hepatic_failure + immunosuppressio
n + leukemia + lymphoma + solid_tumor_with_metastasis + apache_3j_bodysyste
m + apache_2_bodysystem + BMI_category, data.frame(X_train, y_train), metho
d = "class")

#prp(stablo2, type = 5)
```

```
prediction_decision_tree2 = predict(stablo2, test, type="class")
table(prediction_decision_tree2, test$hospital_death)
```

```
##
## prediction_decision_tree2     0     1
##                         0 11139   468
##                         1  1416   734
```

```
confusion_matrix_dt2 = confusionMatrix(table(prediction_decision_tree2, tes
t$hospital_death))
```

Accuracy

```
#Accuracy = TP + TN / TP + TN + FP + FN => 0.86
accuracy <- confusion_matrix_dt2$overall["Accuracy"]
accuracy_str <- sprintf("Accuracy: %.2f", accuracy)
print(accuracy_str)
```

```
## [1] "Accuracy: 0.86"
```

Precision

```
#Precision = TP / TP + FP => 0.96
precision <- confusion_matrix_dt2$byClass["Pos Pred Value"]
precision_str <- sprintf("Precision: %.2f", precision)
print(precision_str)
```

```
## [1] "Precision: 0.96"
```

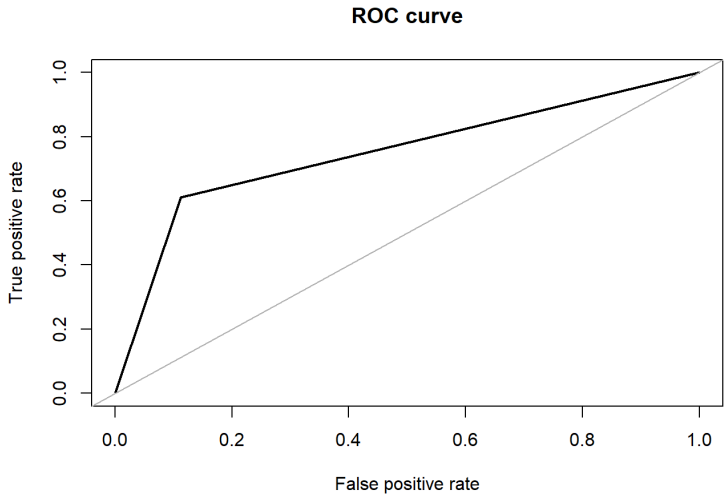
Recall

```
#Recall = TP / TP + FN => 0.89
print(paste(round(confusion_matrix_dt2$byClass["Sensitivity"], 2)))
## [1] "0.89"
```

F1-score

```
#F1 - score = 2 * (Precision * Recall) / (Precision + Recall) => 0.92
print(paste("F1-Score:", round(confusion_matrix_dt2$byClass["F1"], 2)))
## [1] "F1-Score: 0.92"
```

```
#AUC =  0.749
predict_dt2 <- predict(stablo2, test, type="class")
roc.curve(test$hospital_death, predict_dt2[], plotit = T)
```




```
## Area under the curve (AUC): 0.749
```

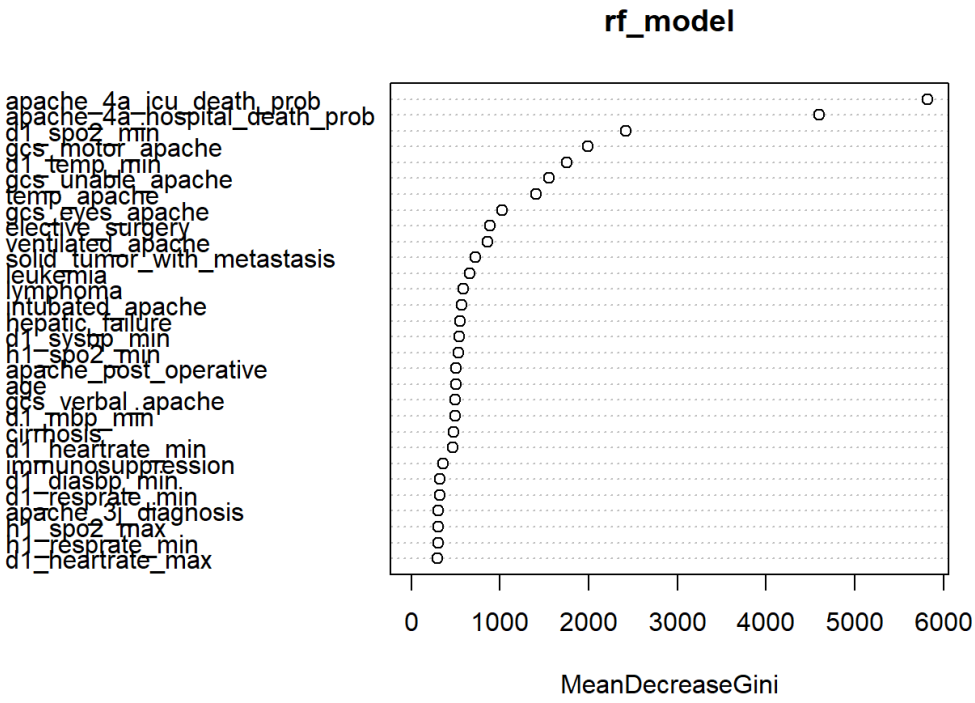
Primećujemo da nema značajne razlike između ova dva modela.

Random forest

Random Forest je moćan algoritam ansambla stabala odlučivanja. S obzirom na različite karakteristike pacijenata, kao i moguće interakcije među obeležjima, Random Forest može pružiti visoku tačnost i bolje upravljanje kompleksnošću.

1. Prvi model ćemo primeniti nad svim varijablama.

```
y_train <- factor(y_train)
rf_model <- randomForest(y_train ~ ., data = data.frame(X_train, y_train),
ntree = 100)
varImpPlot(rf_model)
```




```
predictions_rf1 <- predict(rf_model, newdata = test)
table(predictions_rf1, test$hospital_death)
```

```
##
## predictions_rf1     0     1
##               0 12127   709
##               1   428   493
```

```
confusion_matrix_rf1 = confusionMatrix(table(predictions_rf1, test$hospital
_death))
```

Accuracy

```
#Accuracy = TP + TN / TP + TN + FP + FN => 0.92
accuracy <- confusion_matrix_rf1$overall["Accuracy"]
accuracy_str <- sprintf("Accuracy: %.2f", accuracy)
print(accuracy_str)
```

```
## [1] "Accuracy: 0.92"
```

Precision

```
#Precision = TP / TP + FP => 0.94
precision <- confusion_matrix_rf1$byClass["Pos Pred Value"]
precision_str <- sprintf("Precision: %.2f", precision)
print(precision_str)
## [1] "Precision: 0.94"
```

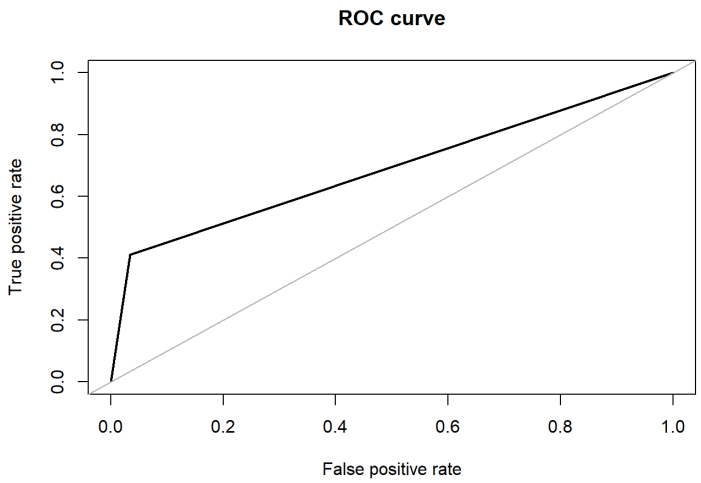
Recall

```
#Recall = TP / TP + FN => 0.97
print(paste(round(confusion_matrix_rf1$byClass["Sensitivity"], 2)))
## [1] "0.97"
```

F1-score

```
#F1 - score = 2 * (Precision * Recall) / (Precision + Recall) => 0.96
print(paste("F1-Score:", round(confusion_matrix_rf1$byClass["F1"], 2)))
## [1] "F1-Score: 0.96"
```

```
#AUC = 0.688
predict_rf1 <- predict(rf_model, newdata = test)
roc.curve(test$hospital_death, predict_rf1[], plotit = T)
```




```
## Area under the curve (AUC): 0.688
```

2. Sada ćemo ovaj algoritam primeniti na feature koje smo dobili kao najrelevantnije (significant_fetures).

```
rf_model2 <- randomForest(y_train ~ bmi + elective_surgery + ethnicity + ge
nder + height + icu_type + weight + apache_2_diagnosis + apache_3j_diagnosi
s + apache_post_operative + arf_apache + gcs_eyes_apache + gcs_motor_apache
+ gcs_unable_apache + gcs_verbal_apache + heart_rate_apache + intubated_apa
che + map_apache + resprate_apache + temp_apache + ventilated_apache + d1_d
iasbp_max + d1_diasbp_min + d1_heartrate_max + d1_heartrate_min + d1_mbp_ma
x + d1_mbp_min + d1_resprate_max + d1_resprate_min + d1_spo2_max + d1_spo2_
min + d1_sysbp_max + d1_sysbp_min + d1_temp_max + d1_temp_min + h1_diasbp_m
ax + h1_diasbp_min + h1_heartrate_max + h1_heartrate_min + h1_mbp_max + h1_
mbp_min + h1_resprate_max + h1_resprate_min + h1_spo2_max + h1_spo2_min + h
1_sysbp_max + h1_sysbp_min + d1_glucose_max + d1_glucose_min + d1_potassium
_max + d1_potassium_min + apache_4a_hospital_death_prob + apache_4a_icu_dea
th_prob + aids + cirrhosis + diabetes_mellitus + hepatic_failure + immunosu
ppression + leukemia + lymphoma + solid_tumor_with_metastasis + apache_3j_b
odysystem + apache_2_bodysystem + BMI_category, data = data.frame(X_train,
y_train), ntree = 100)

predictions_rf2 <- predict(rf_model2, newdata = test)
```

```
table(predictions_rf2, test$hospital_death)
```

```
##
## predictions_rf2     0     1
##               0 12129   713
##               1   426   489
```

```
confusion_matrix_rf2 = confusionMatrix(table(predictions_rf2, test$hospital
_death))
```

Accuracy

```
#Accuracy = TP + TN / TP + TN + FP + FN => 0.92
accuracy <- confusion_matrix_rf2$overall["Accuracy"]
accuracy_str <- sprintf("Accuracy: %.2f", accuracy)
print(accuracy_str)
```

```
## [1] "Accuracy: 0.92"
```

Precision

```
#Precision = TP / TP + FP => 0.94
precision <- confusion_matrix_rf2$byClass["Pos Pred Value"]
precision_str <- sprintf("Precision: %.2f", precision)
print(precision_str)
```

```
## [1] "Precision: 0.94"
```

Recall

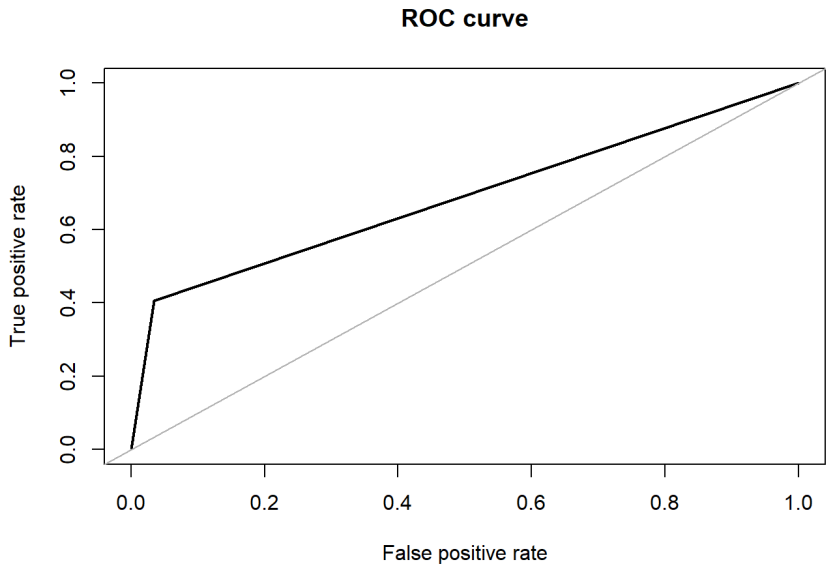
```
#Recall = TP / TP + FN => 0.97
print(paste(round(confusion_matrix_rf2$byClass["Sensitivity"], 2)))
```

```
## [1] "0.97"
```

F1-score

```
#F1 - score = 2 * (Precision * Recall) / (Precision + Recall) => 0.96
print(paste("F1-Score:", round(confusion_matrix_rf2$byClass["F1"], 2)))
## [1] "F1-Score: 0.96"
```

```
#AUC = 0.686
predict_rf2 <- predict(rf_model2, newdata = test)
roc.curve(test$hospital_death, predict_rf2[], plotit = T)
```




```
## Area under the curve (AUC): 0.686
```

Primećujemo da ovaj model nema značajnija poboljšanja.

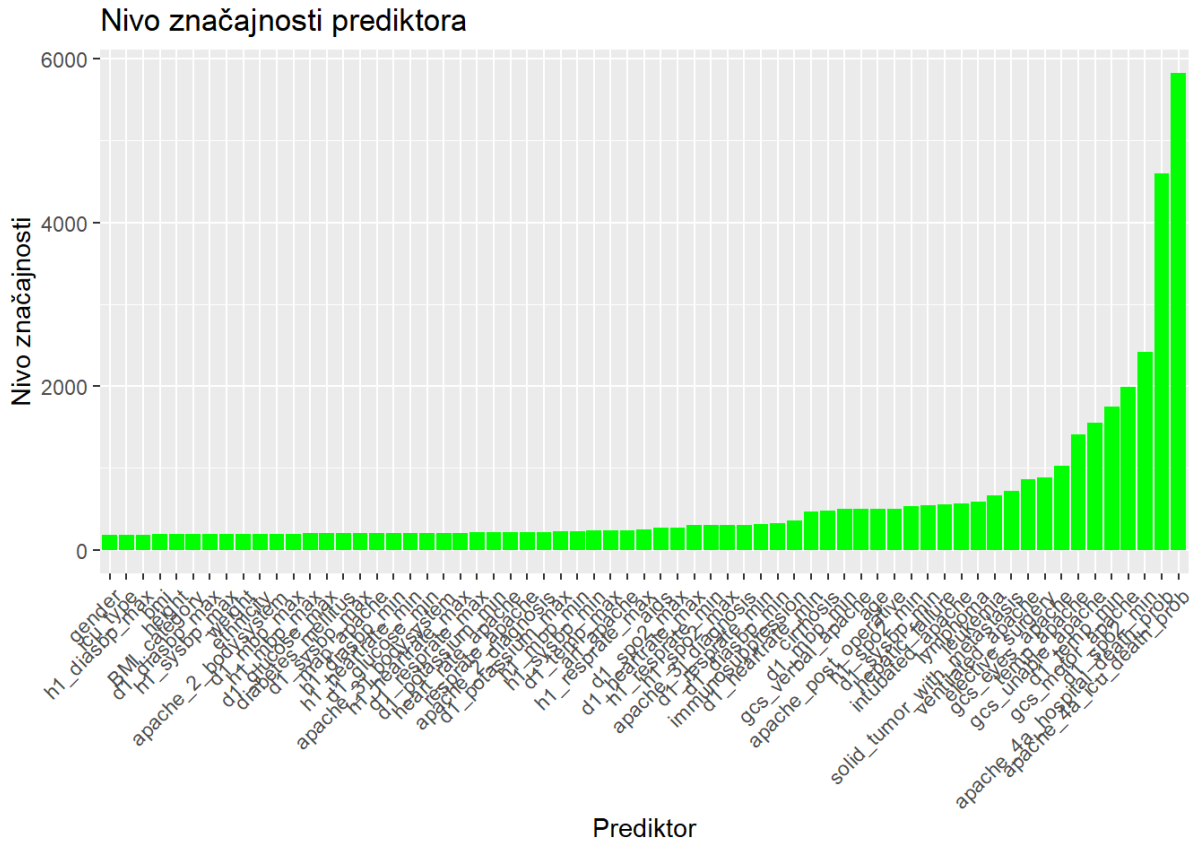
Hajde za kraj da vidimo koliki ucinak imaju prediktori, tj. koliki im je nivo znacajnosti na osnovu RANDOM FOREST algoritma.

```
feature_weight = data.frame( Feature = row.names(importance(rf_model)), Mea
nDecreaseGini = importance(rf_model))
```

Sada ćemo to predstaviti grafički.

```
gg_feature_weight <- ggplot(feature_weight, aes(x = reorder(Feature, MeanDe
creaseGini), y = MeanDecreaseGini)) +
 geom_bar(stat = "identity", fill = "green") +
```

```
 theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
 labs(
 title = "Nivo značajnosti prediktora",
 x = "Prediktor",
 y = "Nivo značajnosti"
 )
gg_feature_weight
```



Primećujemo da *apache_a4_hospital_death_prob* i *apache_a4_icu_death_prob* imaju najveću značajnost korišćenjem algoritma random forest. Pored njih su tu i *d1_spo2_min*, *gcs_motor_apache*.

Zaključak

Za oversampling smo koristili još jedan pristup:

- oversampled_data_1 <- ovun.sample(hospital_death ~ ., data = data, method = “over”, seed = 1)$data

Kako ne bismo ponovo pokretali sve ponovo, ispaćemo rezultat koji smo dobili ovim pristupom i uporedićemo ga sa metodom oversampled_data.

	oversampled_data_1		
	Logistička regresija	Stablo odlučivanja	Slučajna šuma
AUC	0.878	0.777	0.693
Matrica konfuzije	0: 11513, 479 1: 1042, 723	0: 9798, 272 1: 2757, 930	0: 12328, 834 1: 227, 368
Tačnost	0.89	0.78	0.92
Preciznost	0.96	0.97	0.94
Odaziv (Recall)	0.93	0.78	0.98
F1-score	0.94	0.87	0.96

	oversampled_data		
	Logistička regresija	Stablo odlučivanja	Slučajna šuma
AUC	0.873	0.749	0.688
Matrica konfuzije	0: 11709, 551 1: 846, 651	0: 11139, 468 1: 1416, 734	0: 12127, 709 1: 428, 493
Tačnost	0.90	0.86	0.91
Preciznost	0.96	0.96	0.94
Odaziv (Recall)	0.93	0.82	0.97
F1-score	0.94	0.92	0.96

Literatura

[1] Microsoft Teams, kanal Uvod u nauku o podacima
[2] Resampling the Data
[3] Random Forest
[4] Decision tree
[5] Dataset