PREDICTING DEATH OF HOSPITALIZED PATIENTS BY VARIABLES BASED ON PADUA-SCORE AND CHARLSON-INDEX

Vadim Litvinov | Shiran Ben-Meir

הקדמה

- הדאטה מכיל מידע אודות 18,890 חולים מאושפזים.
- עבור כל חולה מתועדים גיל החולה, מינו, תסמינים המאפיינים את החולה, מחלות בעבר, מחלות בהווה ועוד.
- הדאטה מתבסס על שני מודלים עיקריים:
  Charlson-Comobidity-Index – מודל המסוגל לחזות תמותה של חולה בצפי של שנה
  The Padua score- מודל המסוגל לחזות את מידת סיכון החולה לקבל אירוע Venous thromboemboli (VTE) (פקקת ורידים תסחיפית).
- בפרוייקט נרצה לשלב את שני המודלים ויחד עם שני משתנים נוספים - גיל ומין נרצה לחזות את תמותת החולים.

תיאור ה-Data לאחר התמקדות במדדים

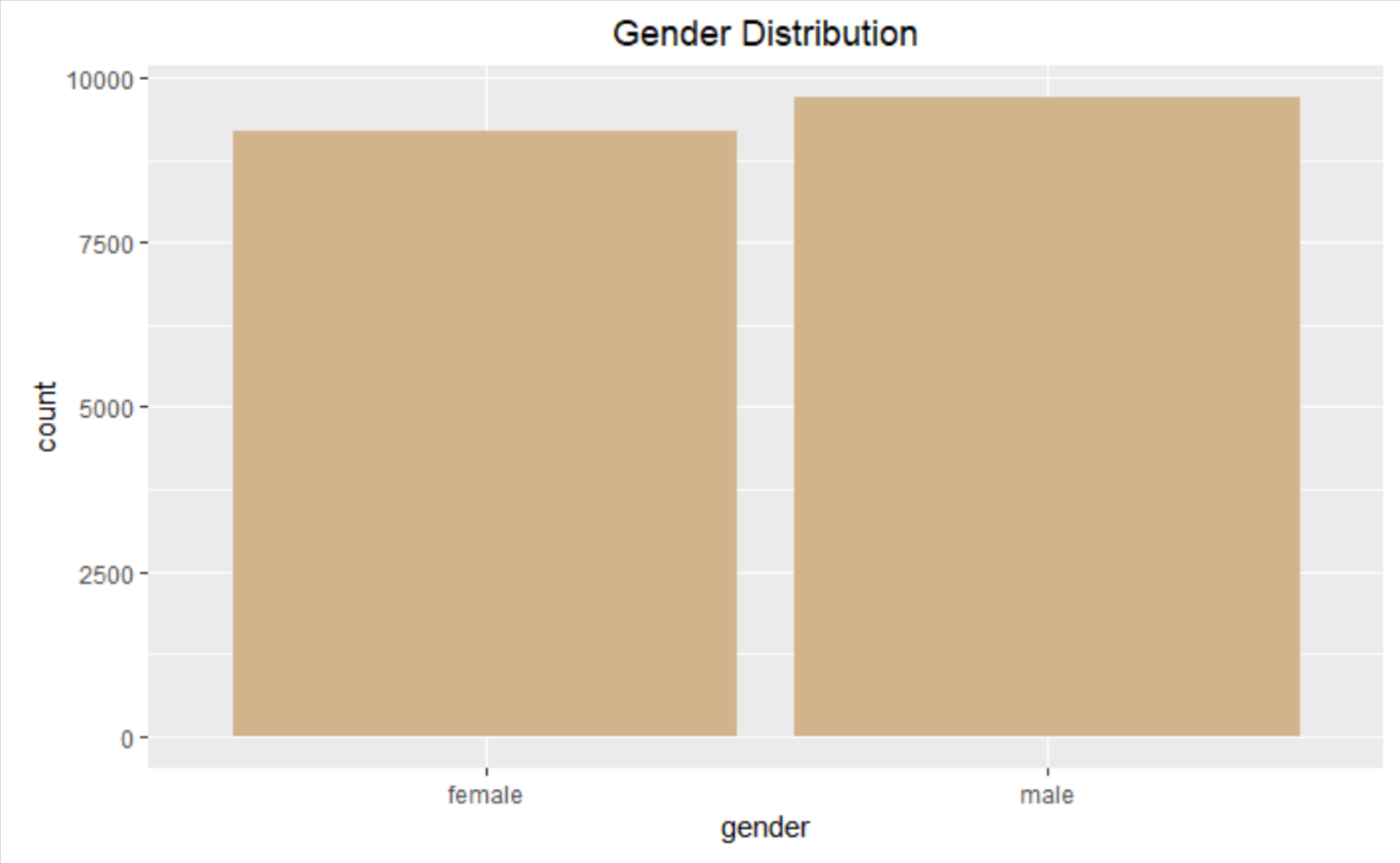
```
'data.frame':   18890 obs. of  34 variables:
 $ male.0.female.1                                  : chr  "1" "0" "0" "1" ...
 $ CPPS_KnownThrombophilia                          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CPPS_ActiveCancer                                : int  0 0 0 0 1 0 0 0 1 0 ...
 $ CPPS_PreviousVTE                                 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CPPS_ReducedMobility                             : int  0 0 1 0 0 0 0 0 0 1 ...
 $ CPPS_RecenTraumaSurgery                          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CPPS_Over70                                      : int  0 0 1 1 0 1 1 0 1 0 ...
 $ CPPS_HeartRespiratoryFailure                     : int  0 0 1 0 0 0 0 0 0 0 ...
 $ CPPS_MIorCVA                                     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CPPS_InfectionRheumatologicalDisorder            : int  0 0 1 0 0 0 0 0 0 1 ...
 $ CPPS_Obesity                                     : int  1 0 0 0 0 0 1 0 0 0 ...
 $ CPPS_OngoingHormonalTreatment                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.severe.renal.disease                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.Myocardial.Infarct                      : int  0 1 1 0 0 0 0 0 0 0 ...
 $ Charlson.Pulmonary.congestion                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.PVD                                     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.DB                                      : int  0 1 0 1 0 0 0 0 1 0 ...
 $ Charlson.DM...retinopathy...neuropathy...nephropathy: int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.Leukemia                                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.lymphoma                                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.lymphoma.or.leukemia                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.cerebrovascular.disease                 : int  0 0 0 0 0 0 1 0 0 0 ...
 $ Charlson.Hemiplegia                              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.Metastatic.solid.tumor                  : int  0 0 0 0 1 0 0 0 1 0 ...
 $ Charlson.Dementia                                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.chronic.pulmonary.diseas                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.Connective.tissue.disease               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.Peptic.ulcer.disease                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.AIDS                                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.Metastatic.solid.tumor.3                : int  0 0 0 0 1 0 0 0 1 0 ...
 $ Charlson.moderate.to.severe.renal.disease        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.liver.disease                           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ age                                              : num  49 47.4 82.1 80.1 68.8 ...
 $ death                                            : chr  "0" "0" "1" "0" ...
```

ניקוי ה-Data

```
[1] 34
'data.frame':   18890 obs. of  24 variables:
 $ male.0.female.1                                  : chr  "1" "0" "0" "1" ...
 $ CPPS_KnownThrombophilia                          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CPPS_ActiveCancer                                : int  0 0 0 0 1 0 0 0 1 0 ...
 $ CPPS_PreviousVTE                                 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CPPS_ReducedMobility                             : int  0 0 1 0 0 0 0 0 0 1 ...
 $ CPPS_HeartRespiratoryFailure                     : int  0 0 1 0 0 0 0 0 0 0 ...
 $ CPPS_InfectionRheumatologicalDisorder            : int  0 0 1 0 0 0 0 0 0 1 ...
 $ CPPS_Obesity                                     : int  1 0 0 0 0 0 1 0 0 0 ...
 $ CPPS_OngoingHormonalTreatment                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.Myocardial.Infarct                      : int  0 1 1 0 0 0 0 0 0 0 ...
 $ Charlson.Pulmonary.congestion                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.PVD                                     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.DB                                      : int  0 1 0 1 0 0 0 0 1 0 ...
 $ Charlson.DM...retinopathy...neuropathy...nephropathy: int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.cerebrovascular.disease                 : int  0 0 0 0 0 0 1 0 0 0 ...
 $ Charlson.Dementia                                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.chronic.pulmonary.diseas                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.Connective.tissue.disease               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.Peptic.ulcer.disease                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.AIDS                                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.moderate.to.severe.renal.disease        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Charlson.liver.disease                           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ age                                              : num  49 47.4 82.1 80.1 68.8 ...
 $ death                                            : chr  "0" "0" "1" "0" ...
```

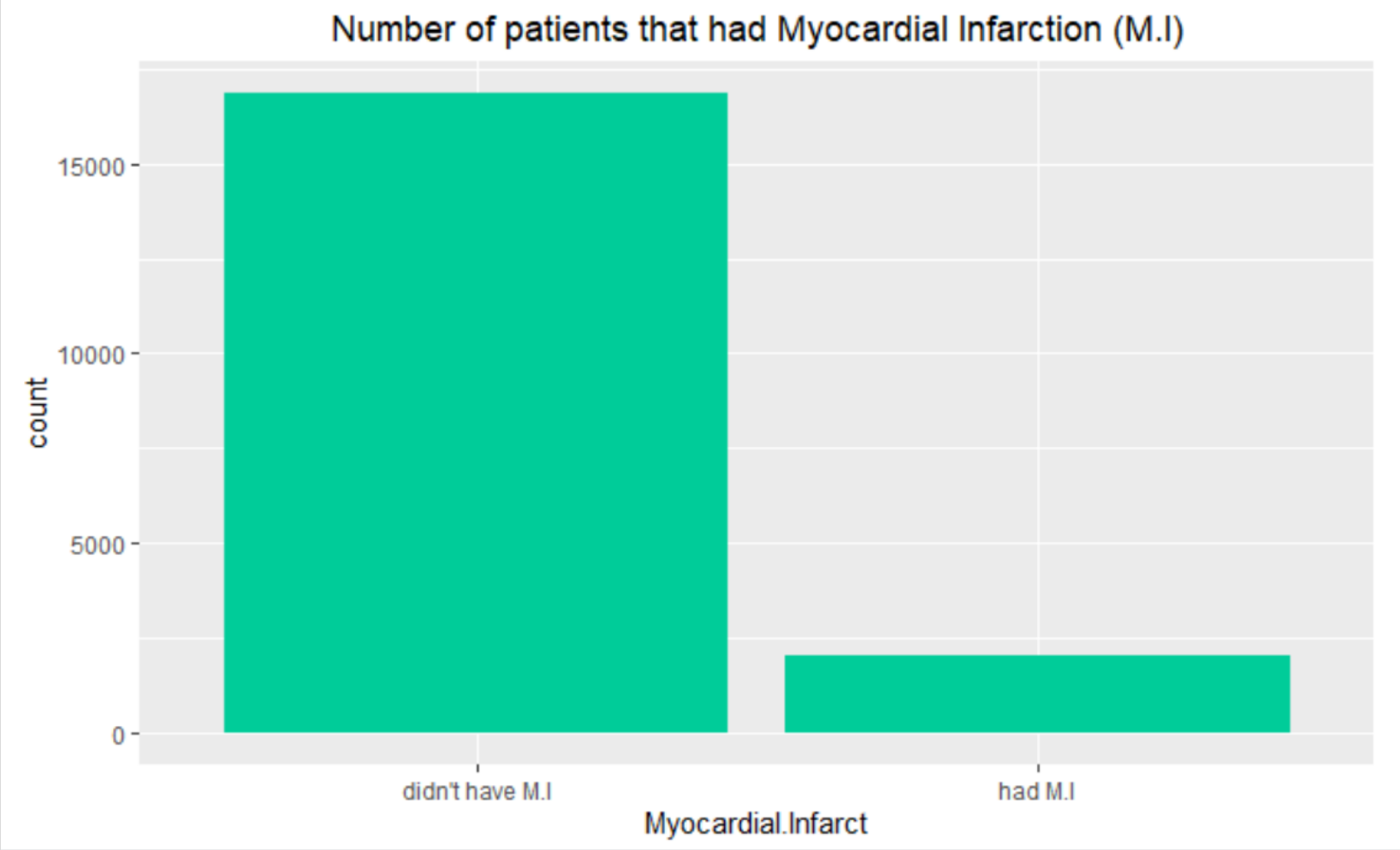
Gender variable

```
female   male
  9187   9697

female   male
   0.5    0.5
```



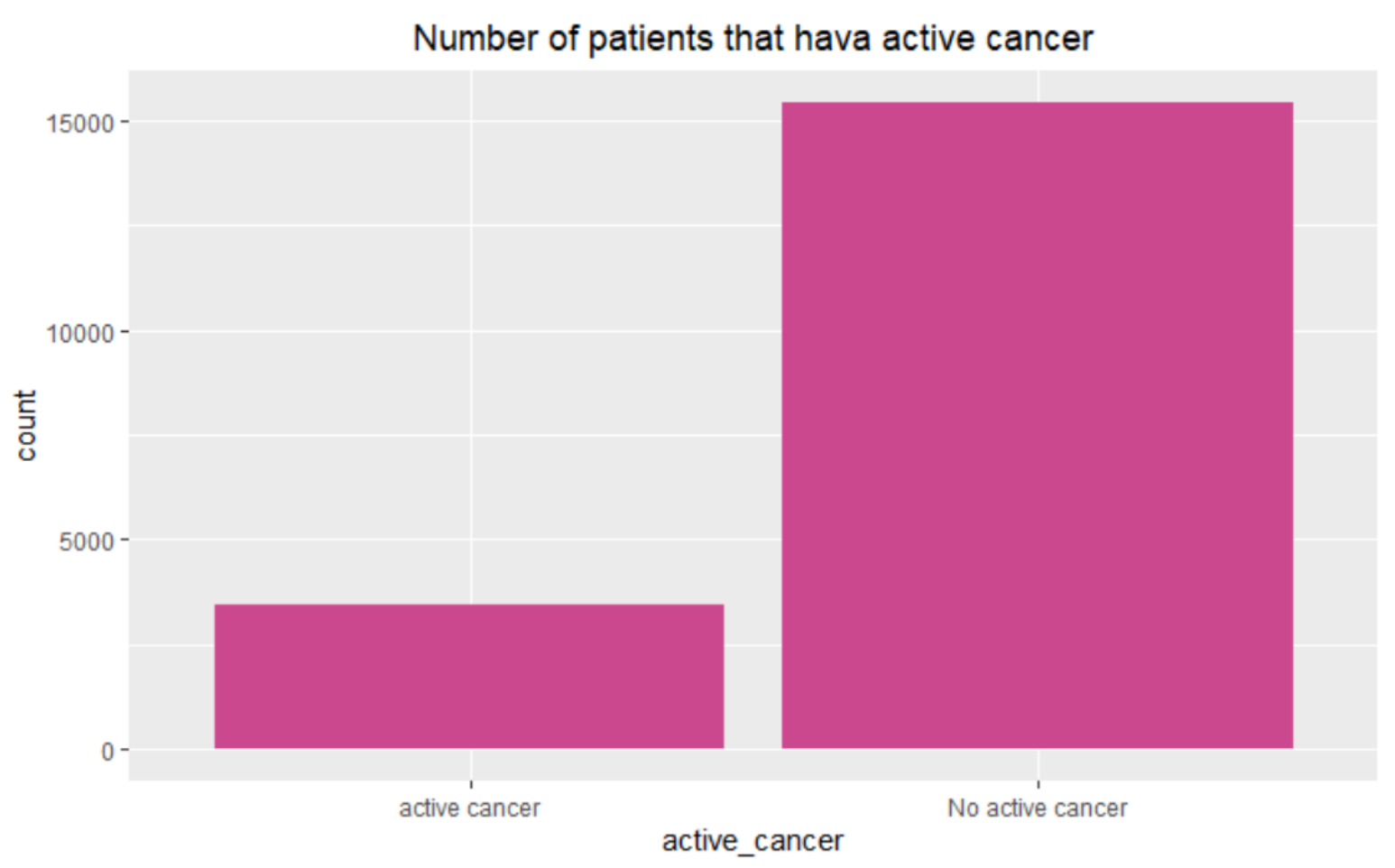
Myocardial Infarction variable

```
didn't have M.I        had M.I
          16855          2029

didn't have M.I        had M.I
            0.9            0.1
```

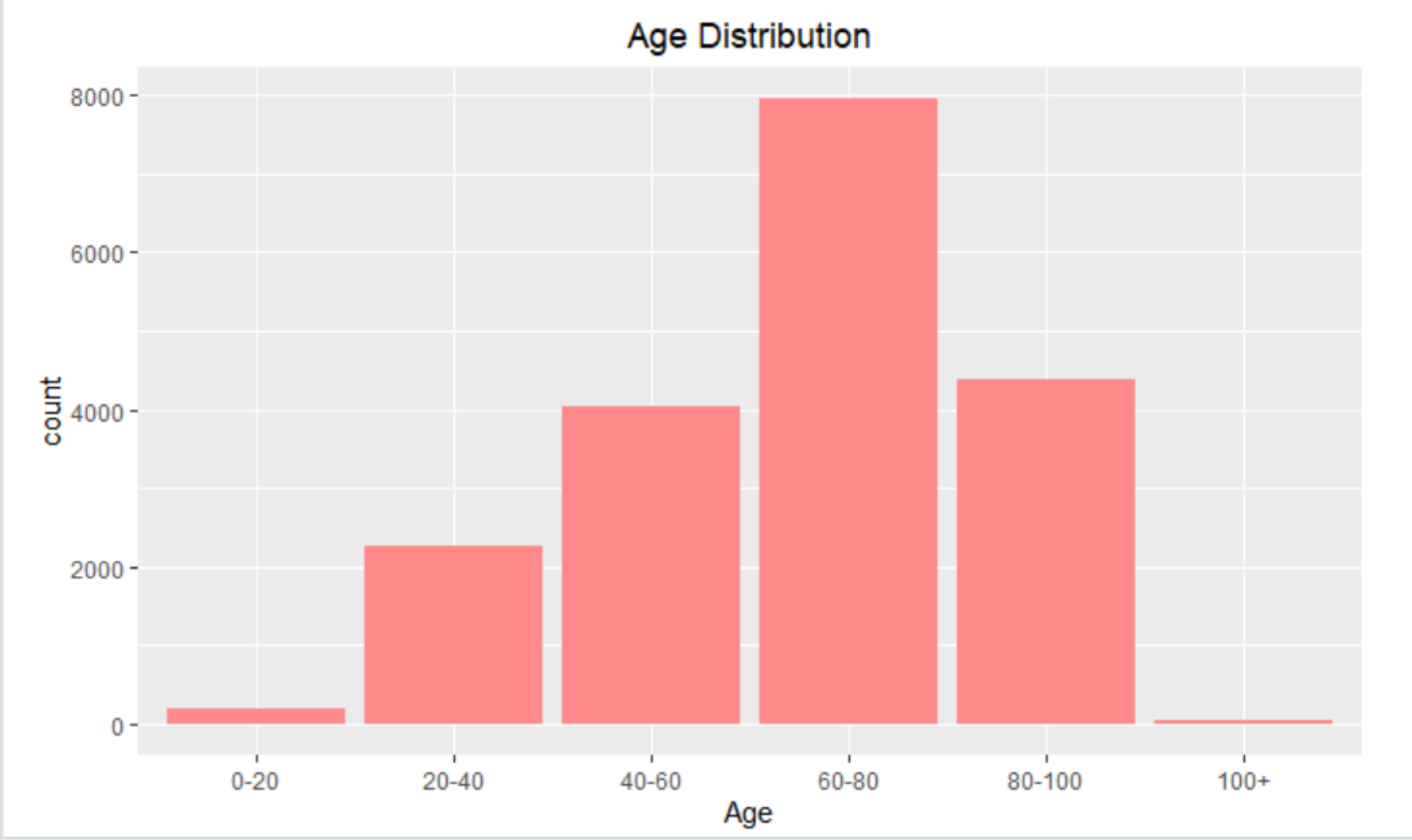


Active Cancer variable

```
don't have active cancer       have active cancer
                   15426                     3458

don't have active cancer       have active cancer
                     0.8                      0.2
```



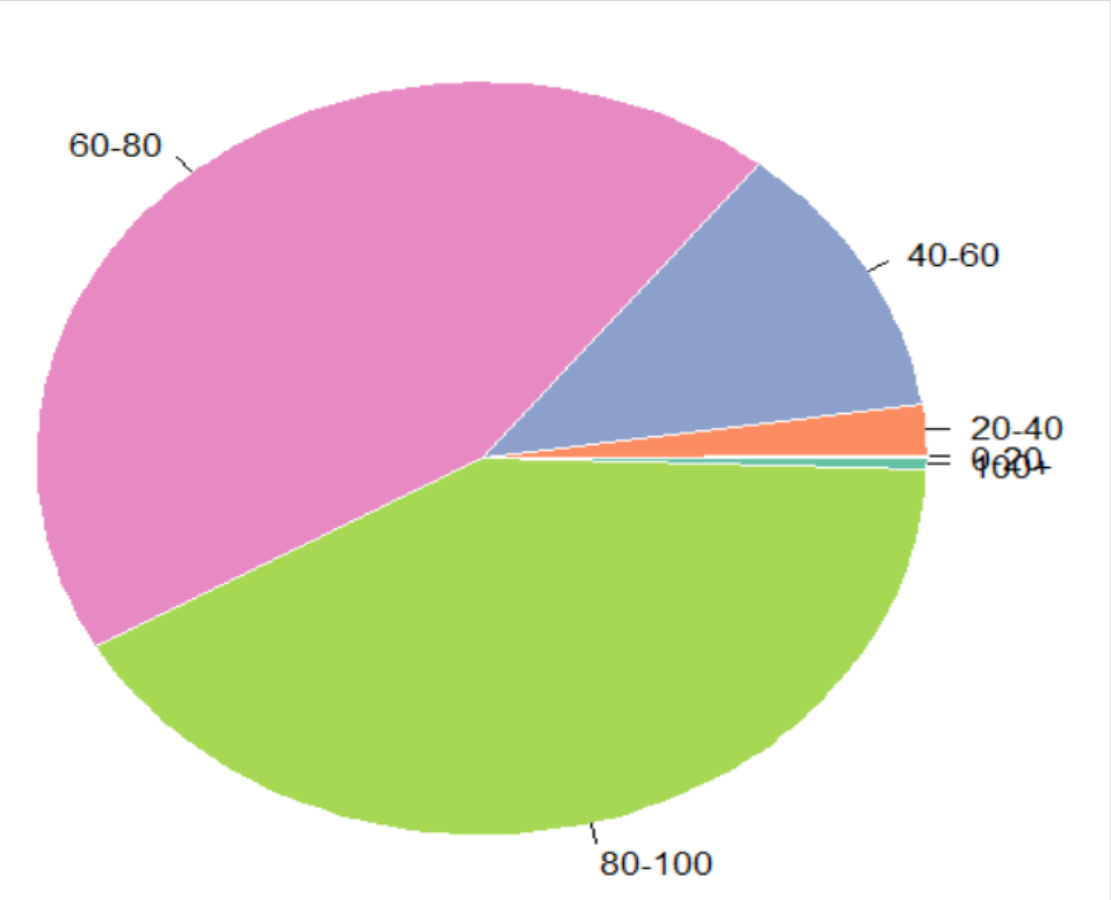
Age variable




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.02   53.43   67.17   64.65   79.27  116.25
```

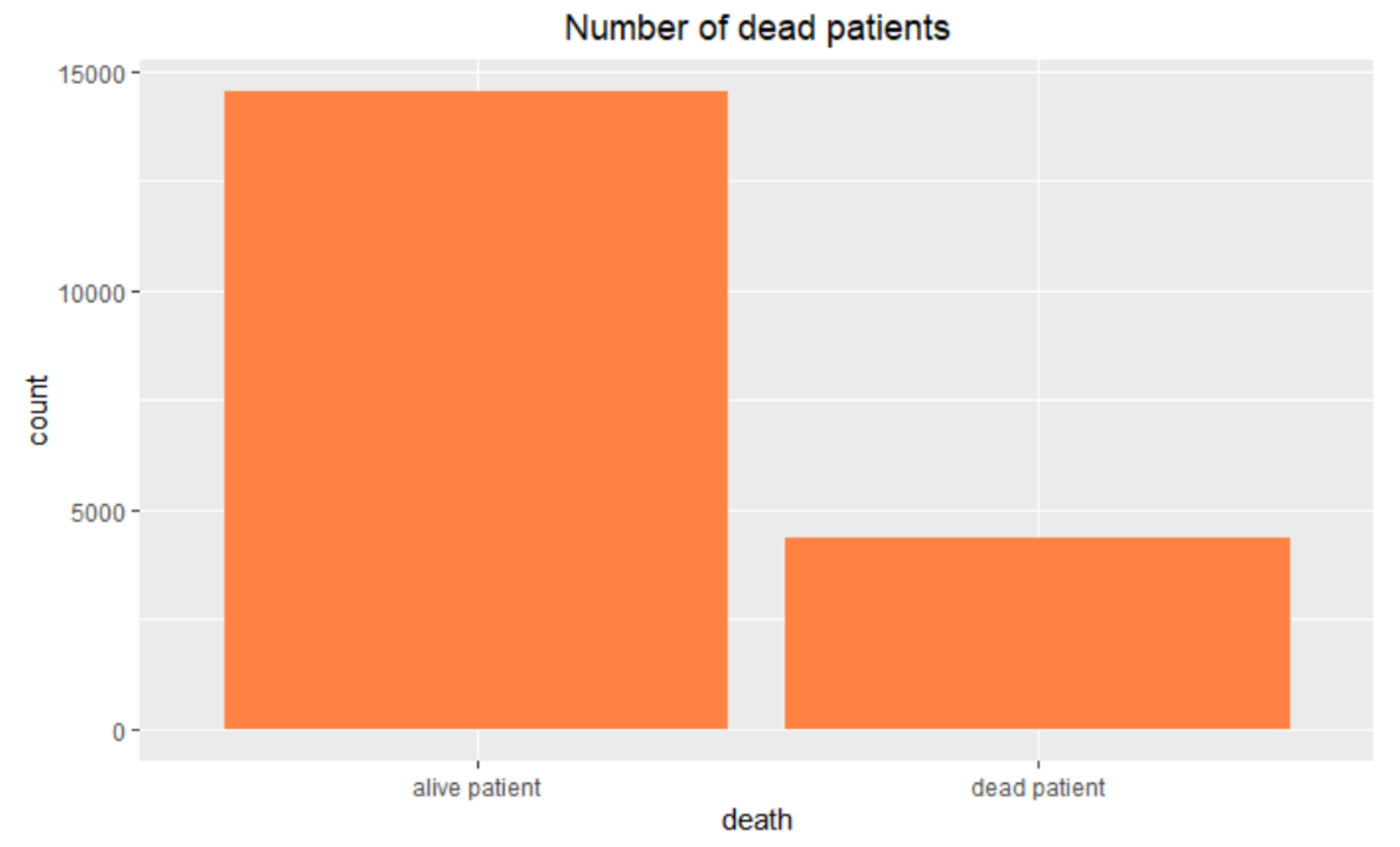
Age distribution by death

Age <fctr>	death <dbl>
0-20	6
20-40	93
40-60	523
60-80	1922
80-100	1801
100+	19

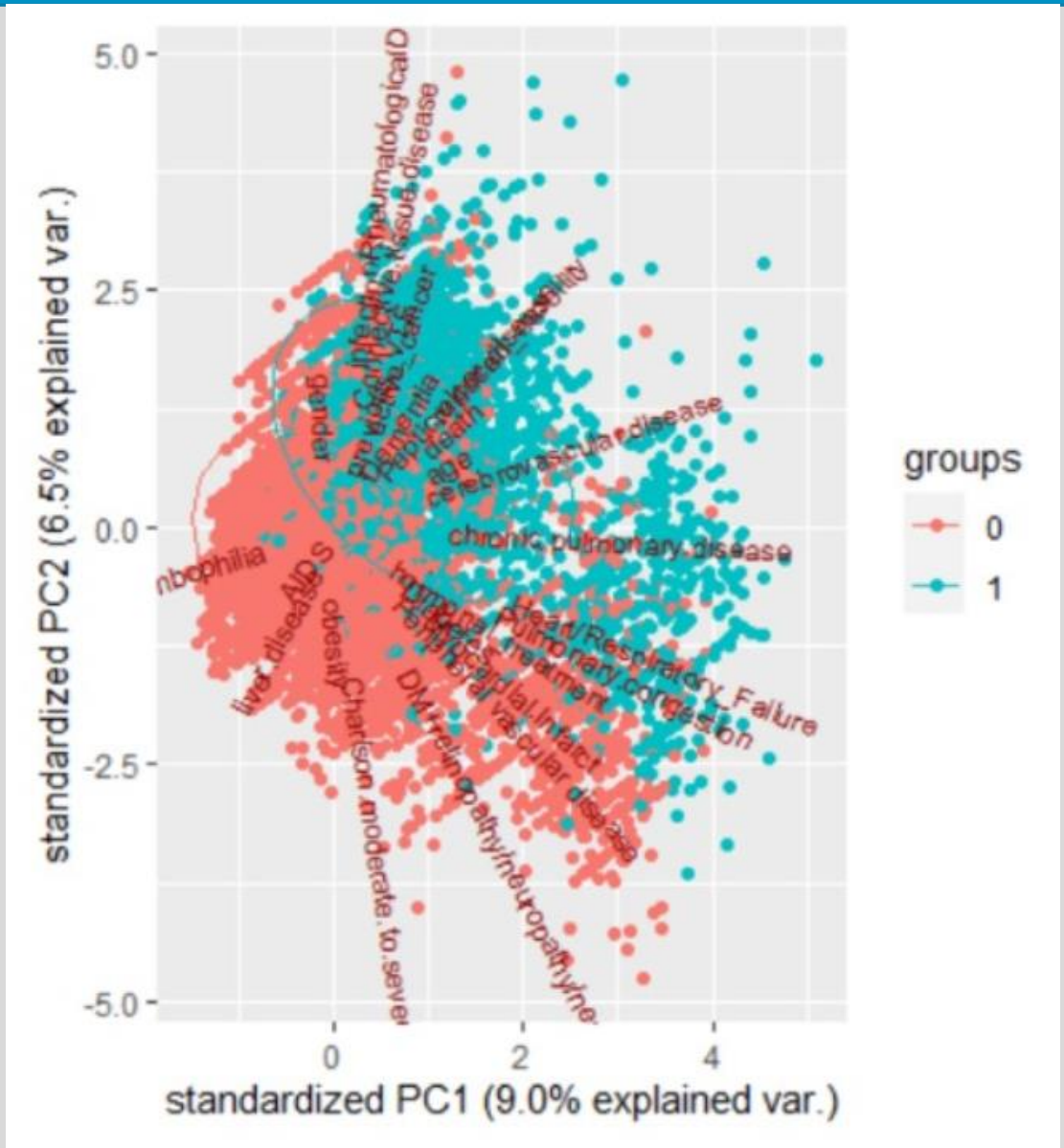


Death distribution

```
alive patient  dead patient
        14520          4364

alive patient  dead patient
          0.8            0.2
```



PCA



- כאשר מצמצמים את מס' המשתנים ל-2 בעזרת PCA, ניתן לראות כי ישנה הפרדה סבירה.
- עקב תוצאות אלה ניתן להניח כי ניתן להפריד את הדוגמאות בעזרת אלגוריתמי למידה.

Feature Selection

gender	28.4906709
known_Thrombophilia	1.2685632
active_cancer	213.4689639
previous_VTE	14.2332025
reduced_mobility	264.0625796
Heart.Respiratory_Failure	63.0036293
infectionRheumatologicalDisorder	44.5754904
obesity	40.2110009
hormonal_treatment	30.1606354
Myocardial.Infarct	40.9274436
Pulmonary.congestion	36.3138864
Peripheral.vascular.disease	0.9771439
Diabetes	36.0714985
DM.retinopathy.neuropathy.nephropathy	11.9075847
cerebrovascular.disease	30.9965702
Dementia	11.0079568
chronic.pulmonary.disease	39.0778066
Connective.tissue.disease	14.6084725
Peptic.ulcer.disease	12.6173951
AIDS	1.1334451
moderate.severe.renal	6.9655295
liver.disease	30.0767165
age	488.7131463

- ניתן לראות כי המשתנים המשמעותיים ביותר הם: age, reduced mobility, active cancer.

Algorithms - KNN

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  5666

             | test_pred
 test_labels |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |      4324 |         0 |      4324 |
             |     1.000 |     0.000 |     0.763 |
             |     1.000 |     0.000 |           |
             |     0.763 |     0.000 |           |
-------------|-----------|-----------|-----------|
           1 |         2 |      1340 |      1342 |
             |     0.001 |     0.999 |     0.237 |
             |     0.000 |     1.000 |           |
             |     0.000 |     0.236 |           |
-------------|-----------|-----------|-----------|
Column Total |      4326 |      1340 |      5666 |
             |     0.764 |     0.236 |           |
-------------|-----------|-----------|-----------|

Setting levels: control = 0, case = 1
Setting direction: controls < cases
Area under the curve: 0.9993
```

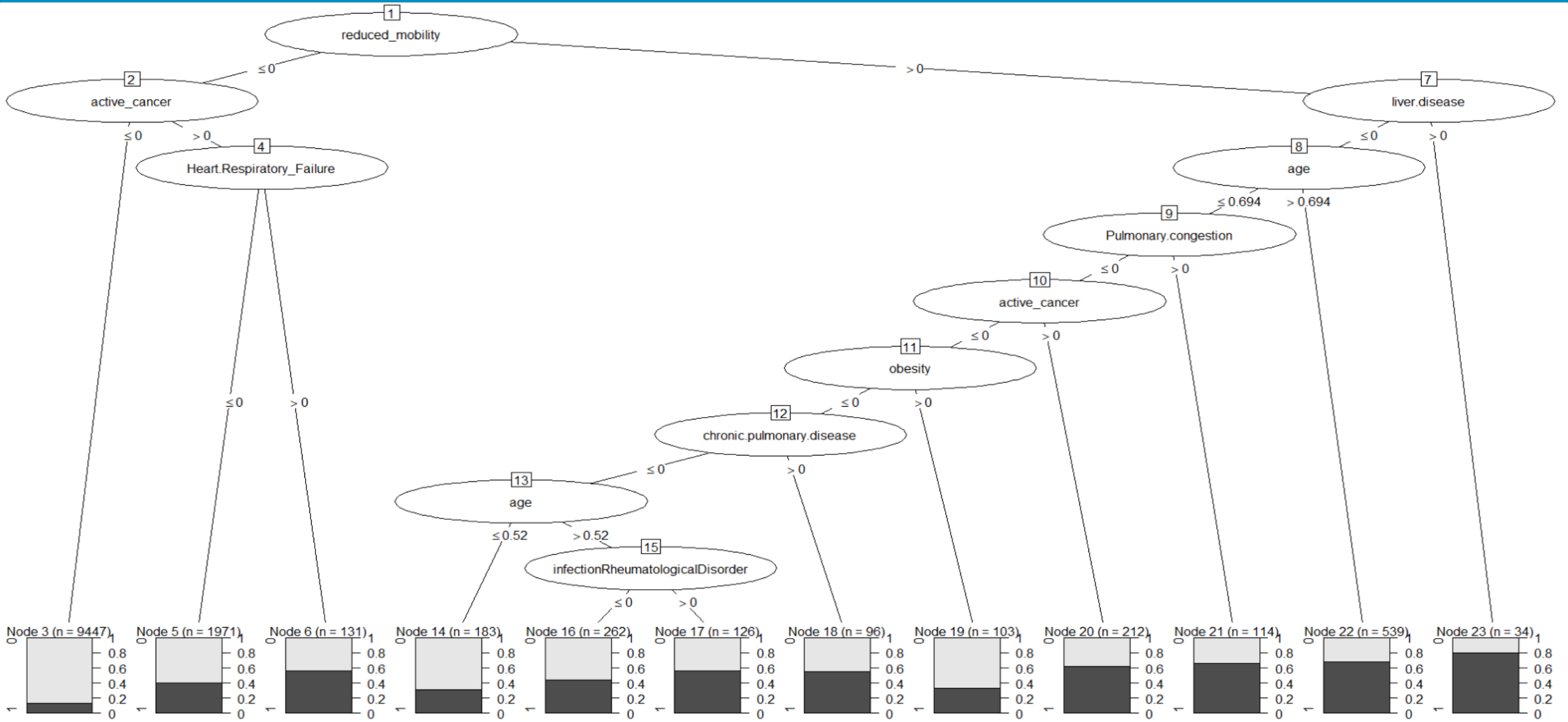
- בוצעו מספר ניסיונות הרצה עבור ערכי K שונים. ערך ה-K שנבחר (K=19) להרצת האלגוריתם הוא הערך בעל ה-accuracy הגבוה ביותר (99.96%)
- השטח מתחת לעקום ה-ROC שואף ל-1.
- קל לראות כי אלגוריתם ה-KNN מצליח לחזות באופן כמעט מושלם את מקרי המוות בממדים הנבחרים שפורטו קודם.

Algorithms – Decision Tree

```
Decision tree:

reduced_mobility <= 0:
:...active_cancer <= 0: 0 (9447/1223)
:   active_cancer > 0:
:   :...Heart.Respiratory_Failure <= 0: 0 (1971/794)
:       Heart.Respiratory_Failure > 0: 1 (131/57)
reduced_mobility > 0:
:...liver.disease > 0: 1 (34/7)
    liver.disease <= 0:
    :...age > 0.6938776: 1 (539/173)
        age <= 0.6938776:
        :...Pulmonary.congestion > 0: 1 (114/39)
            Pulmonary.congestion <= 0:
            :...active_cancer > 0: 1 (212/80)
                active_cancer <= 0:
                :...obesity > 0: 0 (103/34)
                    obesity <= 0:
                    :...chronic.pulmonary.disease > 0: 1 (96/43)
                        chronic.pulmonary.disease <= 0:
                        :...age <= 0.5204082: 0 (183/58)
                            age > 0.5204082:
                            :...infectionRheumatologicalDisorder <= 0: 0 (262/116)
                                infectionRheumatologicalDisorder > 0: 1 (126/56)
```

```
Evaluation on training data (13218 cases):

           Decision Tree
         ----------------
         Size      Errors

           12 2680(20.3%)   <<


          (a)   (b)    <-classified as
         ----  ----
         9741   455    (a): class 0
         2225   797    (b): class 1


       Attribute usage:

       100.00% reduced_mobility
        94.80% active_cancer
        15.90% Heart.Respiratory_Failure
        12.63% liver.disease
        12.37% age
         8.29% Pulmonary.congestion
         5.83% obesity
         5.05% chronic.pulmonary.disease
         2.94% infectionRheumatologicalDisorder


Time: 0.1 secs
```

Algorithms – Decision Tree



Algorithms – Decision Tree

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  5666

             | preticted death
actual death |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |      4117 |       240 |      4357 |
             |     0.727 |     0.042 |           |
-------------|-----------|-----------|-----------|
           1 |       942 |       367 |      1309 |
             |     0.166 |     0.065 |           |
-------------|-----------|-----------|-----------|
Column Total |      5059 |       607 |      5666 |
-------------|-----------|-----------|-----------|

Setting levels: control = 0, case = 1
Setting direction: controls < cases
Area under the curve: 0.6126
[1] 0.7913872
```

- ניתן לראות כי המשתנה שחלוקה על פיו נותן את ה-Information Gain הגבוה ביותר הוא reduced mobility.
  ולכן ניתן להניח שהוא בעל המתאם הגבוה ביותר עם מוות.
- ערך ה-accuracy הוא 0.791.
- ערך ה-AUC הוא 0.612.

Algorithms – Naive Bayes

```
       0    1
  0 3876  739
  1  525  526

               Accuracy : 0.7769
                 95% CI : (0.7658, 0.7877)
    No Information Rate : 0.7767
    P-Value [Acc > NIR] : 0.4948

                  Kappa : 0.3155

 Mcnemar's Test P-Value : 2.084e-09

            Sensitivity : 0.8807
            Specificity : 0.4158
         Pos Pred Value : 0.8399
         Neg Pred Value : 0.5005
             Prevalence : 0.7767
         Detection Rate : 0.6841
   Detection Prevalence : 0.8145
      Balanced Accuracy : 0.6483

       'Positive' Class : 0

Setting levels: control = 0, case = 1
Setting direction: controls < cases
Area under the curve: 0.6483
```

- ניתן לראות שערך ה-Accuracy הוא 0.776
- ערך ה-P-value גבוה (גדול מ-0.05) והדבר מעיד על פרדיקציה לא מדויקת.
- ערך ה-Kappa הוא נמוך יחסית ומעיד על סיכוי גבוה להגיע לחיזוי נכון באופן מקרי.
- ערך ה-AUC הוא 0.6483.

Algorithms – SVM

Confusion matrix:

```
   svm_pred
       0    1
  0 4089  312
  1  865  400
Setting levels: control = 0, case = 1
Setting direction: controls < cases
Area under the curve: 0.6227
[1] 0.7922697
```

- ערך ה-Accuracy הוא 0.7922
- ערך ה-AUC הוא 0.6227
- ה-Kernel שהשתמשנו בו הוא "linear"

Algorithms – k-means

cluster <int>	age <dbl>
1	0.5405652
2	0.4977485
3	0.4555810
4	0.4698856

cluster <int>	Dementia <dbl>
1	0.007270694
2	0.001290323
3	0.003586640
4	0.005676443

cluster <int>	death <dbl>
1	0.2639821
2	0.2464516
3	0.2354853
4	0.2162454

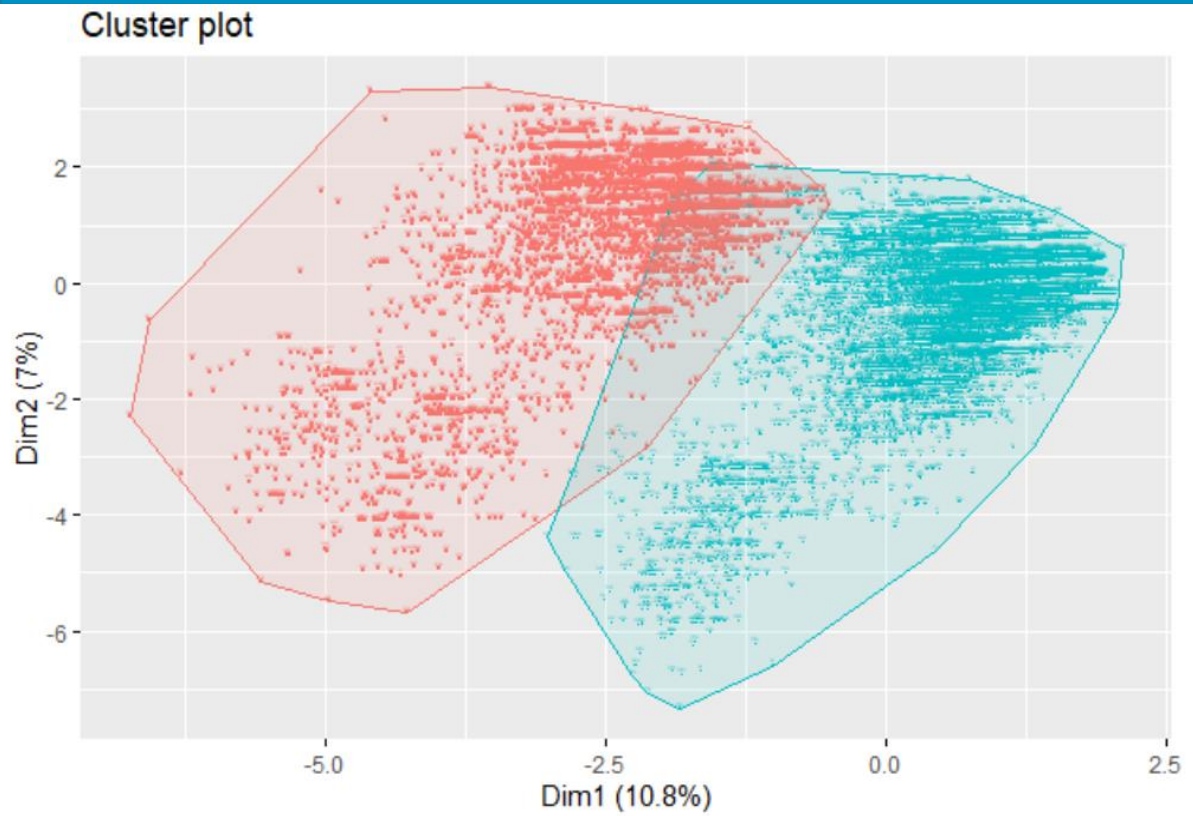
cluster <int>	active_cancer <dbl>
1	0.1901566
2	0.1445161
3	0.1775387
4	0.1921881

Algorithms – k-means

cluster <int>	obesity <dbl>
1	0.2829978
2	0.1858065
3	0.1912127
4	0.1681308

cluster <int>	Diabetes <dbl>
1	1.0000000
2	0.2296774
3	0.2158709
4	0.0000000

Algorithms – k-means



K=2

K=4

- בוצעו הרצות עבור מספר ערכי K שונים בטווח של 2 עד 6 (באזור שורש מס' המשתנים-24) וה-K עבורו ההפרדה הייתה הטובה ביותר הוא K=2.

סיכום

AUC	Accuracy	אלגוריתם
0.9993	0.9996	KNN
0.612	0.791	Decision Tree
0.6483	0.776	Naïve Bayes
0.6227	0.7299	SVM