Running Head: HOSPITAL READMISSION ANALYSIS

Hospital Readmission Prevalence: An analytics approach to reduce hospital readmissions

Anubha Adwani, Ankita Sharma, Ali Algarni, Imani Bansal

University of North Carolina Charlotte

Table of Contents

Abstract 3

1. Introduction..4
2. Data Gathering and Analysis...4
   - 2.1 Data Features...4
   - 2.2 Data Pre-processing..5
   - 2.3 Correlation Analysis 5
   - 2.4 Regression 7
   - 2.5 Chi Square Analysis 9
   - 2.6 Principal Component Analysis 11
   - 2.7 Pattern Mining...13
   - 2.8 Feature Selection...14
     - 2.8.1 Random Forest...14
     - 2.8.2 Boruta Package..15
   - 2.9 Classification 15
     - 2.9.1 Decision Tree...16
     - 2.9.2 Random Forest...17
     - 2.9.3 Neural Network..18
     - 2.9.4 K Nearest Neighbour...19
     - 2.9.5 Evaluation of classification models................. 19
   - 2.10 Clustering..20
     - 2.10.1 PAM..20
     - 2.10.2 Hierarchal Clustering.......................................22
3. Findings and Analysis...23
4. References...24
5. Appendices...25

Abstract

Hospital readmission is the most important contributor towards total medical care expenditure and is an emerging indicator towards quality of care. Research studies show that nearly 15%-20% of the people discharged from the hospital are readmitted within 30 days from the date of discharge. Some of these readmissions are voluntarily, but some are preventable. The process of admitting a person, readmitting is equally tough for both the patient and the hospital. Reducing hospital readmissions is a win for both the patient and the hospital. Diabetes like any other chronic disease is associated with increased risk of hospital readmissions

Risk factor includes previous hospitalizations, age, gender, previous diagnoses and other socio economic barriers.

Keywords: diabetes, readmission, hospital, diagnoses.

1. Introduction

A significant proportion of medical costs are contributed by small percentage of chronic diseases like diabetes, heart problems, knee problems, etc. These costs are large due to the repeated readmission of the patient for the same problem. Prevention of unplanned or emergency readmissions is therefore increasingly gaining attention. The purpose of this report is to analyze the predictors in the unplanned readmissions and to describe the role of diagnoses of diabetes and glycemic control.

This report is about the analysis of a large clinical database, to examine the pattern of diabetes care in patients with diabetes that were admitted to hospitals and to reduce their readmissions. We examined the use of HbA1c results, the previous diagnoses of the patients, their hospital admission history.

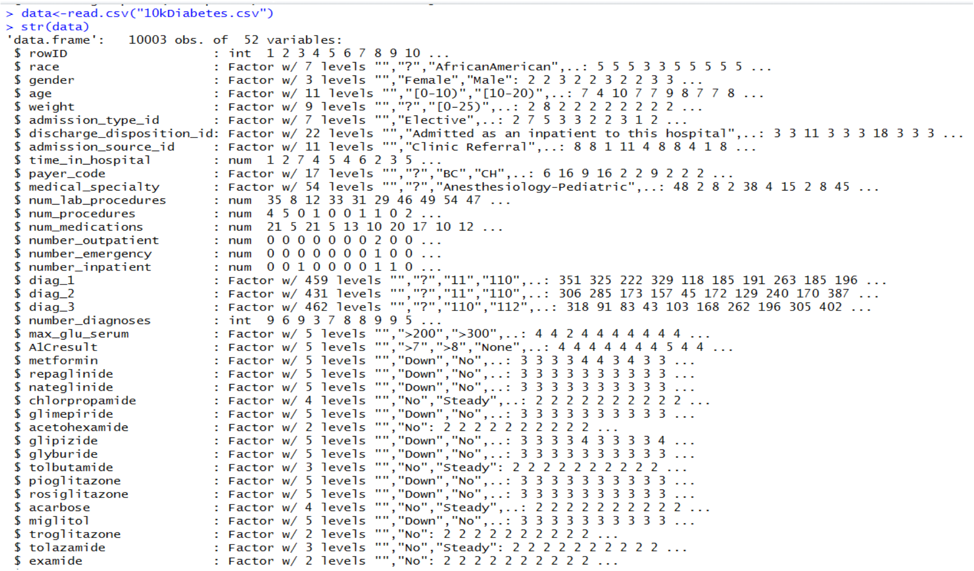
2. Data Gathering and Analysis Techniques

2.1 **Data features**

The dataset given is a clinical dataset that contains historical records of patients pertaining to admission and readmission due to their diabetes. The dataset contains valuable but heterogeneous and incomplete data in terms of missing values, inconsistent records and high dimensionality. The data contains records that includes hospital admission information (inpatient, outpatient or emergency), demographic patient information(age, gender, race, weight), diagnoses documented by ICD-9 codes, laboratory data, pharmacy data, medical caregiver data. The database has 10k observations for 52 variables. The dataset has categorical, numerical, logical data types.

Binary Data	Readmitted
Numerical Data	rowID, num_lab_procedures, num_medications, num_outpatients, num_emergency, num_inpatient, time_in_hospital
Categorical Data	Race, gender, age, weight, medical_speciality, diag_1, diag_2, diag_3, A1C result, max_glu_serum, insulin, change, troglitazone, examide, glyburide.metformin, glipizide,repaglinide.

Table 1: Data Attributes and type



```
> data<-read.csv("10kDiabetes.csv")
> str(data)
'data.frame':   10003 obs. of  52 variables:
 $ rowID                   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ race                    : Factor w/ 7 levels "","?","AfricanAmerican",..: 5 5 5 3 3 5 5 5 5 5 ...
 $ gender                  : Factor w/ 3 levels "","Female","Male": 2 2 3 2 2 3 2 2 3 3 ...
 $ age                     : Factor w/ 11 levels "","[0-10)","[10-20)",..: 7 4 10 7 7 9 8 7 7 8 ...
 $ weight                  : Factor w/ 9 levels "","?","[0-25)",..: 2 8 2 2 2 2 2 2 2 2 ...
 $ admission_type_id       : Factor w/ 7 levels "","Elective",..: 2 7 5 3 3 2 2 3 1 2 ...
 $ discharge_disposition_id: Factor w/ 22 levels "","Admitted as an inpatient to this hospital",..: 3 3 11 3 3 3 18 3 3 3 ...
 $ admission_source_id     : Factor w/ 11 levels "","Clinic Referral",..: 8 8 1 11 4 8 8 4 1 8 ...
 $ time_in_hospital        : num  1 2 7 4 5 4 6 2 3 5 ...
 $ payer_code              : Factor w/ 17 levels "","?","BC","CH",..: 6 16 9 16 2 2 9 2 2 2 ...
 $ medical_specialty       : Factor w/ 54 levels "","?","Anesthesiology-Pediatric",..: 48 2 8 2 38 4 15 2 8 45 ...
 $ num_lab_procedures      : num  35 8 12 33 31 29 46 49 54 47 ...
 $ num_procedures          : num  4 5 0 1 0 0 1 1 0 2 ...
 $ num_medications         : num  21 5 21 5 13 10 20 17 10 12 ...
 $ number_outpatient       : num  0 0 0 0 0 0 0 2 0 0 ...
 $ number_emergency        : num  0 0 0 0 0 0 0 1 0 0 ...
 $ number_inpatient        : num  0 0 1 0 0 0 0 1 1 0 ...
 $ diag_1                  : Factor w/ 459 levels "","?","11","110",..: 351 325 222 329 118 185 191 263 185 196 ...
 $ diag_2                  : Factor w/ 431 levels "","?","11","110",..: 306 285 173 157 45 172 129 240 170 387 ...
 $ diag_3                  : Factor w/ 462 levels "","?","110","112",..: 318 91 83 43 103 168 262 196 305 402 ...
 $ number_diagnoses        : int  9 6 9 3 7 8 8 9 9 5 ...
 $ max_glu_serum           : Factor w/ 5 levels "",">200",">300",..: 4 4 2 4 4 4 4 4 4 4 ...
 $ A1Cresult               : Factor w/ 5 levels "",">7",">8","None",..: 4 4 4 4 4 4 4 5 4 4 ...
 $ metformin               : Factor w/ 5 levels "","Down","No",..: 3 3 3 3 4 4 3 4 3 3 ...
 $ repaglinide             : Factor w/ 5 levels "","Down","No",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ nateglinide             : Factor w/ 5 levels "","Down","No",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ chlorpropamide          : Factor w/ 4 levels "","No","Steady",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ glimepiride             : Factor w/ 5 levels "","Down","No",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ acetohexamide           : Factor w/ 2 levels "","No": 2 2 2 2 2 2 2 2 2 2 ...
 $ glipizide               : Factor w/ 5 levels "","Down","No",..: 3 3 3 3 4 3 3 3 3 4 ...
 $ glyburide               : Factor w/ 5 levels "","Down","No",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ tolbutamide             : Factor w/ 3 levels "","No","Steady": 2 2 2 2 2 2 2 2 2 2 ...
 $ pioglitazone            : Factor w/ 5 levels "","Down","No",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ rosiglitazone           : Factor w/ 5 levels "","Down","No",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ acarbose                : Factor w/ 4 levels "","No","Steady",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ miglitol                : Factor w/ 5 levels "","Down","No",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ troglitazone            : Factor w/ 2 levels "","No": 2 2 2 2 2 2 2 2 2 2 ...
 $ tolazamide              : Factor w/ 3 levels "","No","Steady": 2 2 2 2 2 2 2 2 2 2 ...
 $ examide                 : Factor w/ 2 levels "","No": 2 2 2 2 2 2 2 2 2 2 ...
```

Picture 1: Structure of the initial dataset

2.2 Data Pre-processing

The dataset contains many incomplete, redundant, noisy information. The first step towards analysis is to have quality data. Data quality includes accuracy, completeness, consistency and interpretability. In this step, we examine the structure of the data and its quality. We found that there were several features in the data that could not be treated well as they had high percentage of missing data. These features were weight (97% missing data), payer code (40%) and medical specialty (47%). The weight feature was removed due to its highest percentage of missing data. Medical specialty was maintained as it is relevant to the hospital care-giver information. The missing values in the data were handled by kNN imputation. Imputation is a class of procedures that aims to fill the missing values with the estimated ones based on the information available. The kNN imputation replaces missing data with the corresponding value from the nearest neighbor.

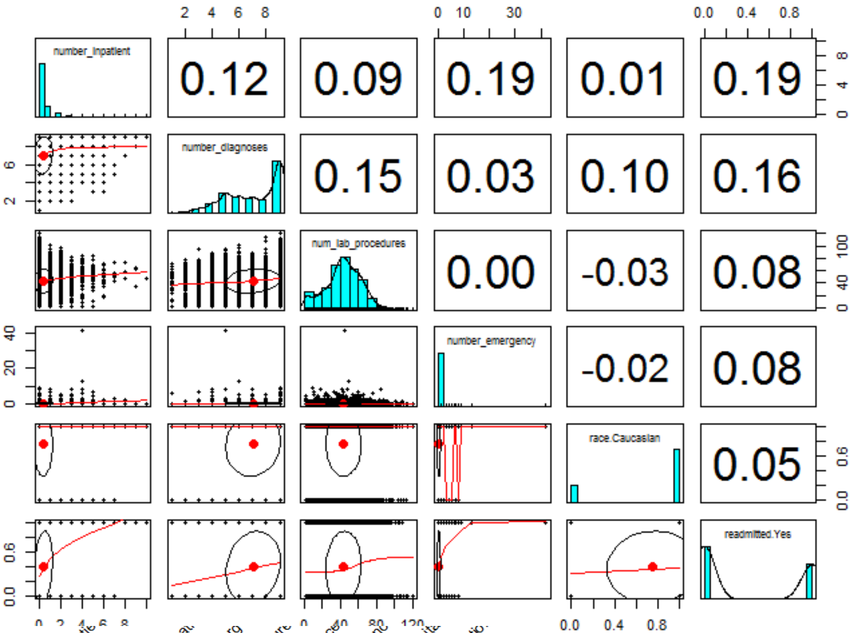
2.3 Correlation Analysis

Correlation was performed to measure the variability of variables. It is a way of measuring the extent to which two variables are related. Correlation coefficient represented the linear dependencies of two variables or sets of data. The value of correlation coefficient ranges from -1 to +1 with a value of ±.1 shows small effect, ±.3 shows medium effect and ±.5 shows large effect. Also, the p-values shows us the significance level of the correlation test performed. Value of p less than 0.05 shows the high significance of the test.

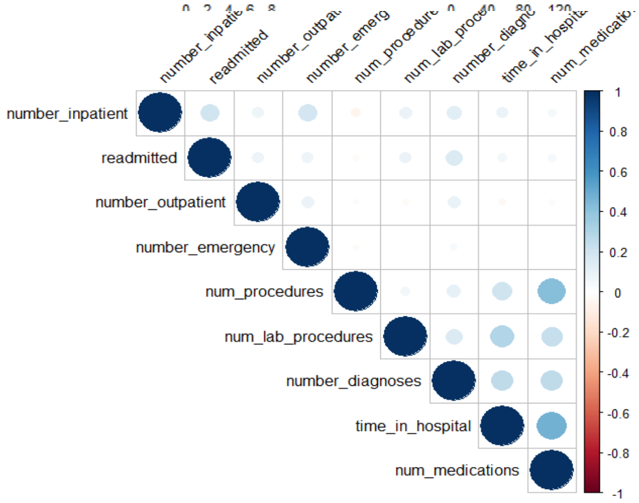
We have used two main functions cor() and rcorr() from Hmisc package to compute basic correlation coefficients. Pearson method is used for the correlation functions.

We have eliminated few of the variables that are not important for our analysis like payer code, index row, and description of the diagnoses. Cor () and rcorr () were performed on total 42 variables and found out the most significant ones to list down here.

Variables	Readmitted(r values)	Readmitted(p values)
race	0.04	0
gender	-0.01	0.21
age	0.08	0
admission_type_id	0	0.81
discharge_disposition_id	0	0.81
admission_source_id	-0.09	0
time_in_hospital	0.05	0
num_lab_procedures	0.08	0
num_procedures	-0.03	0.01
num_medications	0.04	0
number_outpatient	0.07	0
number_emergency	0.08	0
number_inpatient	0.19	0
number_diagnoses	0.16	0
max_glu_serum	0	0.64
A1C result	0	0.64
metformin	-0.02	0.05
repaglinide	0.02	0.02
nateglinide	0.01	0.24
chlorpropamide	0.01	0.46
glimepiride	0.02	0.09
glipizide	0.01	0.16



This plot interprets that "number_inpatient" and "num_diagnoses" are positively correlated with "readmitted".



The value of r is most significant for number_inpatient and number_diagnoses which is 0.19 and 0.16. Also, he p-value is <0.05 which means these two variables are significantly correlated with the readmission and are important in predicting whether a patient will be readmitted or not.

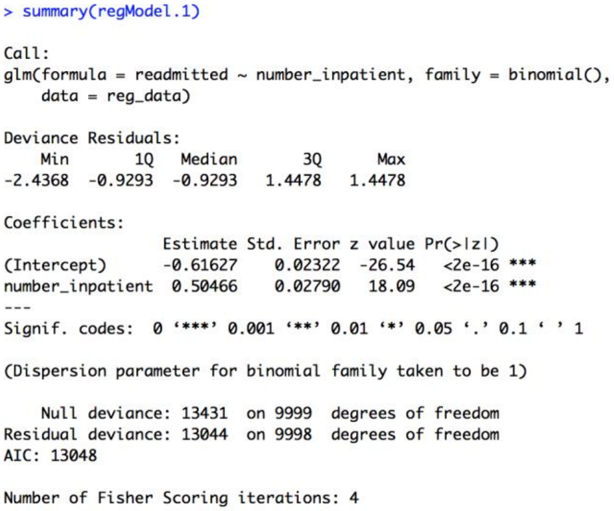
Corrplot: Using Method as Circle 1

2.4 Regression

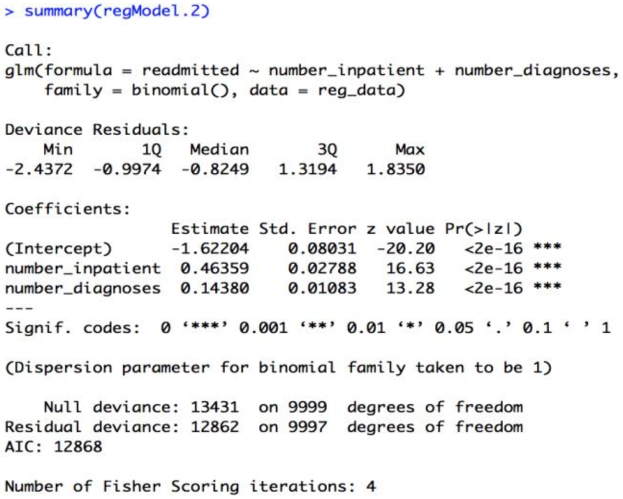
The next step in our hospital readmission analysis was regression which we have performed using logistic regression technique. Logistic regression is a method to predict an outcome variable that is categorical from predictor variables that are continuous and/or categorical.

In our analysis, the outcome variable is readmitted. Two models have been used to predict the value of outcome variable-first model consisting number_inpatient as predictor variable and second model with number_inpatient and number_diagnoses as predictor variables. These two variables have been chosen as predictors because they have been found to be strongly correlated with the readmitted variable because of the correlation analysis.

Below is the summary of the models:



```
> summary(regModel.1)

Call:
glm(formula = readmitted ~ number_inpatient, family = binomial(),
    data = reg_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4368  -0.9293  -0.9293   1.4478   1.4478

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.61627    0.02322  -26.54   <2e-16 ***
number_inpatient  0.50466    0.02790   18.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13431  on 9999  degrees of freedom
Residual deviance: 13044  on 9998  degrees of freedom
AIC: 13048

Number of Fisher Scoring iterations: 4
```



```
> summary(regModel.2)

Call:
glm(formula = readmitted ~ number_inpatient + number_diagnoses,
    family = binomial(), data = reg_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4372  -0.9974  -0.8249   1.3194   1.8350

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.62204    0.08031  -20.20   <2e-16 ***
number_inpatient  0.46359    0.02788   16.63   <2e-16 ***
number_diagnoses  0.14380    0.01083   13.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13431  on 9999  degrees of freedom
Residual deviance: 12862  on 9997  degrees of freedom
AIC: 12868

Number of Fisher Scoring iterations: 4
```

The odds ratio was found out for better understanding of the two models-

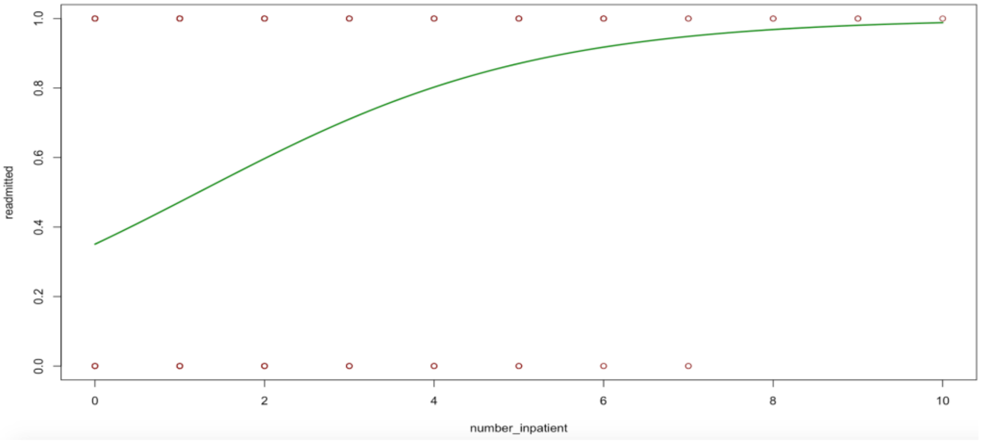
```
> exp(regModel.1$coefficients)
     (Intercept) number_inpatient
       0.5399532        1.6564174
> exp(regModel.2$coefficients)
     (Intercept) number_inpatient number_diagnoses
       0.1974953        1.5897764        1.1546509
```

Odds for number_inpatient in 1st model is 1.65 which is greater than 1 that means as the predictor increases, the odds of the outcome increasing also increases. For the second model, it is 1.59 for number_inpatient and 1.55 for number_diagnoses which means if the values of these two variables will increase the odds of getting readmitted for patient will also increase.

```
> modelChi <- regModel.1$null.deviance - regModel.1$deviance
> modelChi
[1] 387.3111
> chidf <- regModel.1$df.null - regModel.1$df.residual
> chidf
[1] 1
> chisq.prob <- 1 - pchisq(modelChi, chidf)
> chisq.prob
[1] 0
> modelChi2 <- regModel.2$null.deviance - regModel.2$deviance
> modelChi2
[1] 569.679
> chidf2 <- regModel.2$df.null - regModel.2$df.residual
> chidf2
[1] 2
> chisq.prob2 <- 1 - pchisq(modelChi2, chidf2)
> chisq.prob2
[1] 0
```

The deviance for the 1st model is more as compared to the second one thus it can be concluded that the first model is better fit for predicting the outcome.

The relationship between number_inpatient and readmitted can be visualized using the graph-



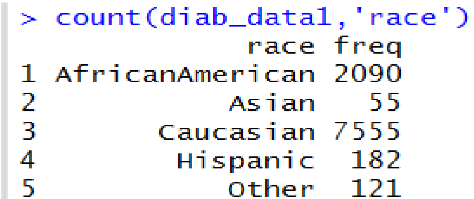
2.5 Chi Square Analysis

In case of categorical variables, we study their frequencies instead of the means as we do in numeric variables. This is because the mean of categorical variables does not convey any meaning, since they contain arbitrary values. To understand the relation between categorical variables, we implemented Pearson's chi square test of independence. This test is used to determine whether there is a significant association between two categorical variables or not. This test compares the observed frequencies in certain categories to the expected frequencies of the same category.

$$x^2 = (observed - expected)^2 / \, expected$$

Race and Readmission

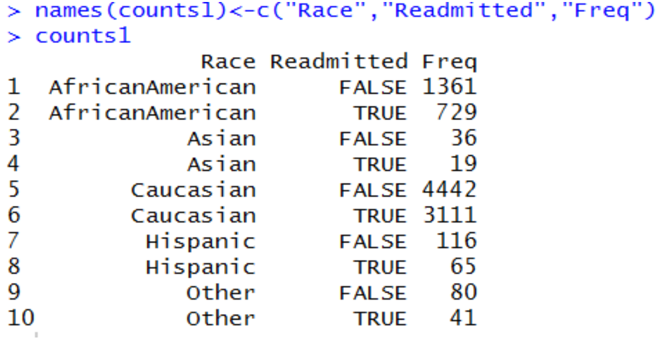
We first took "race" and "readmitted" as a categorical feature and ran pearson's chi square test for them.



```
> count(diab_data1,'race')
             race freq
1 AfricanAmerican 2090
2           Asian   55
3       Caucasian 7555
4        Hispanic  182
5           Other  121
```

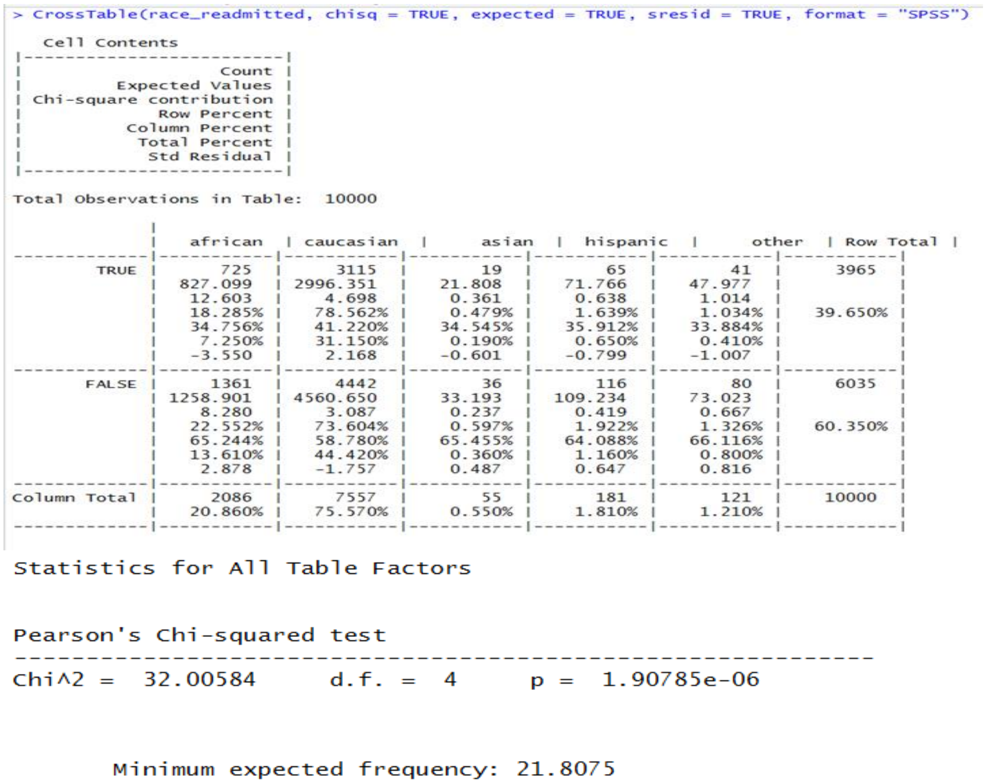
Step 1: Frequency of each race of patients in the dataset.

Step 2: Frequency of each race being readmitted or not:



```
> names(counts1)<-c("Race","Readmitted","Freq")
> counts1
              Race Readmitted Freq
1  AfricanAmerican      FALSE 1361
2  AfricanAmerican       TRUE  729
3            Asian      FALSE   36
4            Asian       TRUE   19
5        Caucasian      FALSE 4442
6        Caucasian       TRUE 3111
7         Hispanic      FALSE  116
8         Hispanic       TRUE   65
9            Other      FALSE   80
10           Other       TRUE   41
```

Step 3: Conduct the chi square test

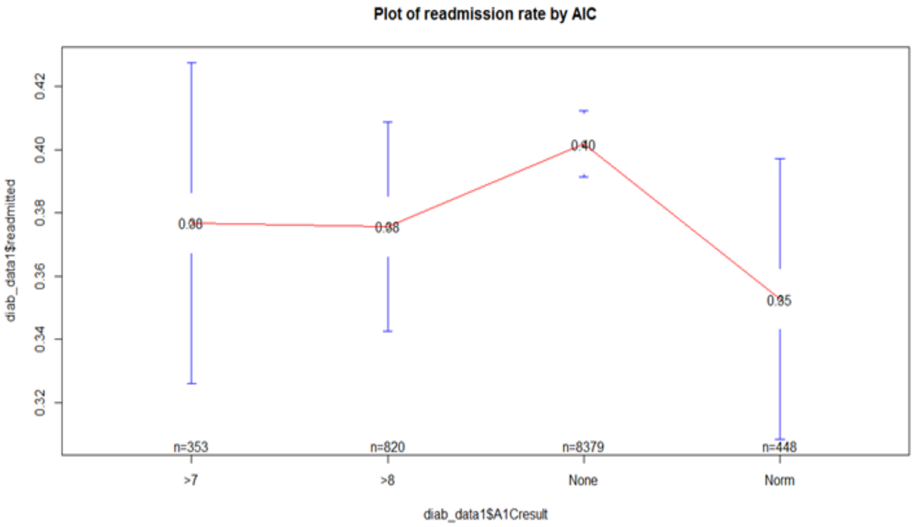


```
> CrossTable(race_readmitted, chisq = TRUE, expected = TRUE, sresid = TRUE, format = "SPSS")

   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  10000

             |
             |   african  | caucasian  |      asian  |  hispanic  |      other  | Row Total |
-------------|------------|------------|------------|------------|------------|------------|
        TRUE |       725  |      3115  |        19  |        65  |        41  |      3965  |
             |   827.099  |  2996.351  |    21.808  |    71.766  |    47.977  |            |
             |    12.603  |     4.698  |     0.361  |     0.638  |     1.014  |            |
             |   18.285%  |   78.562%  |    0.479%  |    1.639%  |    1.034%  |   39.650%  |
             |   34.756%  |   41.220%  |   34.545%  |   35.912%  |   33.884%  |            |
             |    7.250%  |   31.150%  |    0.190%  |    0.650%  |    0.410%  |            |
             |    -3.550  |     2.168  |    -0.601  |    -0.799  |    -1.007  |            |
-------------|------------|------------|------------|------------|------------|------------|
       FALSE |      1361  |      4442  |        36  |       116  |        80  |      6035  |
             |  1258.901  |  4560.650  |    33.193  |   109.234  |    73.023  |            |
             |     8.280  |     3.087  |     0.237  |     0.419  |     0.667  |            |
             |   22.552%  |   73.604%  |    0.597%  |    1.922%  |    1.326%  |   60.350%  |
             |   65.244%  |   58.780%  |   65.455%  |   64.088%  |   66.116%  |            |
             |   13.610%  |   44.420%  |    0.360%  |    1.160%  |    0.800%  |            |
             |     2.878  |    -1.757  |     0.487  |     0.647  |     0.816  |            |
-------------|------------|------------|------------|------------|------------|------------|
Column Total |      2086  |      7557  |        55  |       181  |       121  |     10000  |
             |   20.860%  |   75.570%  |    0.550%  |    1.810%  |    1.210%  |            |
-------------|------------|------------|------------|------------|------------|------------|

Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  32.00584     d.f. =  4     p =  1.90785e-06


       Minimum expected frequency: 21.8075
```

Step 4: Result Interpretation:

The chi square test is highly significant with chi^2(4) = 32.005 and p<0.05. This indicated that readmission rates vary along the race of the patient. Looking at the standard residuals, these results are non-significant for the Caucasian race, which show fairly even distribution for both readmitted (41%) and not admitted (59%). Within all other races, the standard residual varies much around 1.96. Like in the case of African race, expected readmission is higher than the observed frequency of readmitted patients.

Similarly, we found that "medical_specialty" and "readmitted" are dependent with chi^2(4) = 128.16 and p<0.05. It means that hospital care giver's specialty played an important role in readmitting the patient. On the other hand, we found that "A1Cresult" and "readmitted" are independent with chi^2(3) = 6.61 and p>0.05.



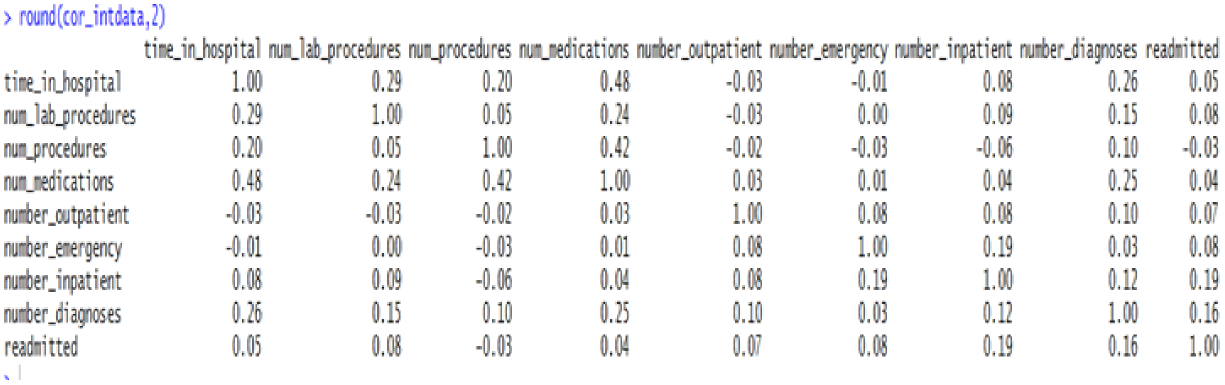
The graph shows that readmission rate is higher for patients for whom earlier A1C results were not considered or were not measured.

2.6 PCA

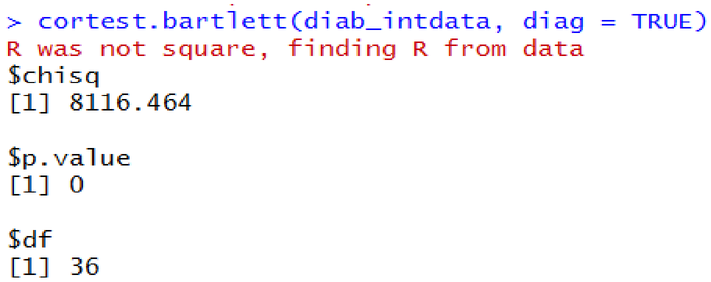
It is a multivariate technique that analyzes the data, in which observations are described by several inter correlated quantitative dependent variables. It is used to emphasize variation and bring out strong patterns in a dataset. The goal of doing PCA in our project was to extract the most important information from the data, compress the size of data by keeping only this important information and analyse the structure of observations and the variables.

Since this can be applied only to numeric variables, the attributes included in this were num_lab_procedures, num_procedures ,num_medications, number_outpatient, number_emergency, number_inpatient, time_in_hospital, number_inpatient, number_diagnoses and the response variable is readmitted.

Step1: Correlation matrix of the data



```
> round(cor_intdata,2)
                   time_in_hospital num_lab_procedures num_procedures num_medications number_outpatient number_emergency number_inpatient number_diagnoses readmitted
time_in_hospital               1.00               0.29           0.20            0.48             -0.03            -0.01             0.08             0.26       0.05
num_lab_procedures             0.29               1.00           0.05            0.24             -0.03             0.00             0.09             0.15       0.08
num_procedures                 0.20               0.05           1.00            0.42             -0.02            -0.03            -0.06             0.10      -0.03
num_medications                0.48               0.24           0.42            1.00              0.03             0.01             0.04             0.25       0.04
number_outpatient             -0.03              -0.03          -0.02            0.03              1.00             0.08             0.08             0.10       0.07
number_emergency              -0.01               0.00          -0.03            0.01              0.08             1.00             0.19             0.03       0.08
number_inpatient               0.08               0.09          -0.06            0.04              0.08             0.19             1.00             0.12       0.19
number_diagnoses               0.26               0.15           0.10            0.25              0.10             0.03             0.12             1.00       0.16
readmitted                     0.05               0.08          -0.03            0.04              0.07             0.08             0.19             0.16       1.00
```



```
> cortest.bartlett(diab_intdata, diag = TRUE)
R was not square, finding R from data
$chisq
[1] 8116.464

$p.value
[1] 0

$df
[1] 36
```

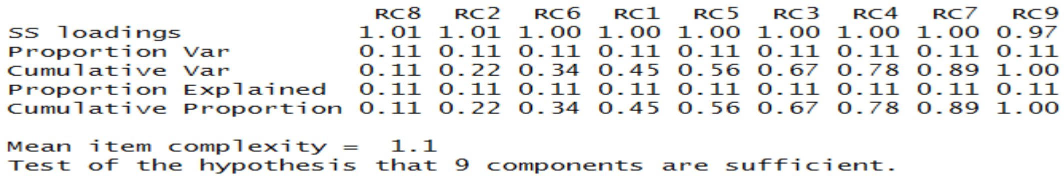
Step2: Cortest Bartlett's Test: Along with correlation matrix, we run this test using cortest.barlette() function from 'psych' package. We ran this test on the raw data.

The test results imply that the results are significant with chi^2(36) = 8116.46 (having significance value <0.05). This tells us that R-matrix is not an identity matrix. Therefore we proceed further with the PCA.

Step 3: PCA with 9 factors

We create the first principal component model by taking all the numeric variables in the data and then deciding which components to keep.

pc1<- principal(diab_intdata, nfactors = 9)



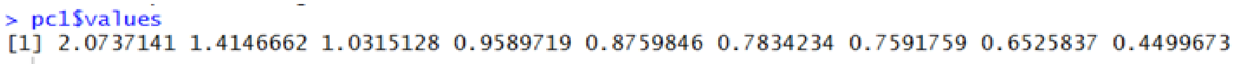
```
                          RC8  RC2  RC6  RC1  RC5  RC3  RC4  RC7  RC9
SS loadings              1.01 1.01 1.00 1.00 1.00 1.00 1.00 1.00 0.97
Proportion Var           0.11 0.11 0.11 0.11 0.11 0.11 0.11 0.11 0.11
Cumulative Var           0.11 0.22 0.34 0.45 0.56 0.67 0.78 0.89 1.00
Proportion Explained     0.11 0.11 0.11 0.11 0.11 0.11 0.11 0.11 0.11
Cumulative Proportion    0.11 0.22 0.34 0.45 0.56 0.67 0.78 0.89 1.00

Mean item complexity =  1.1
Test of the hypothesis that 9 components are sufficient.
```

The thing to look here is the eigenvalues. The eigenvalues associated with each factor represents the variance explained by that linear component.

The result also displays the proportion of variance explained by each of the variable. Like component 8 explains 1.01 units of variance out of possible 9, to proportion this is

1.01/9 = 0.11.

These variables associated with the model are called values. We can find them by:

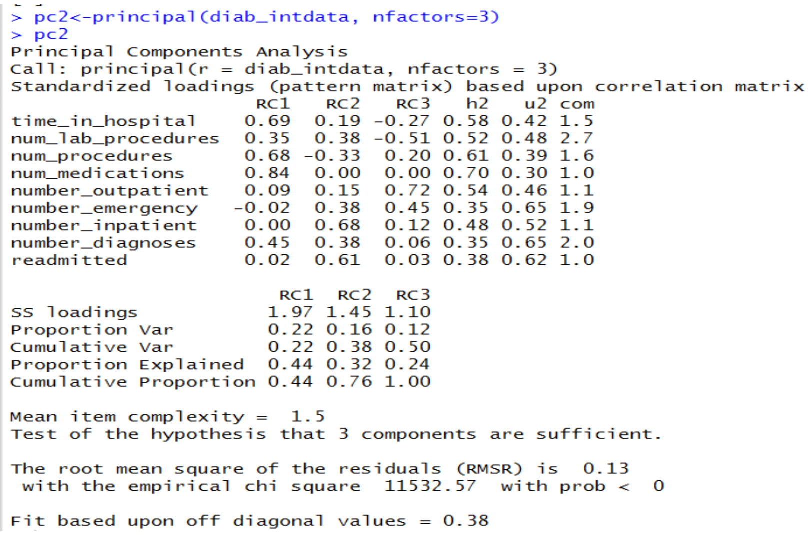


```
> pc1$values
[1] 2.0737141 1.4146662 1.0315128 0.9589719 0.8759846 0.7834234 0.7591759 0.6525837 0.4499673
```

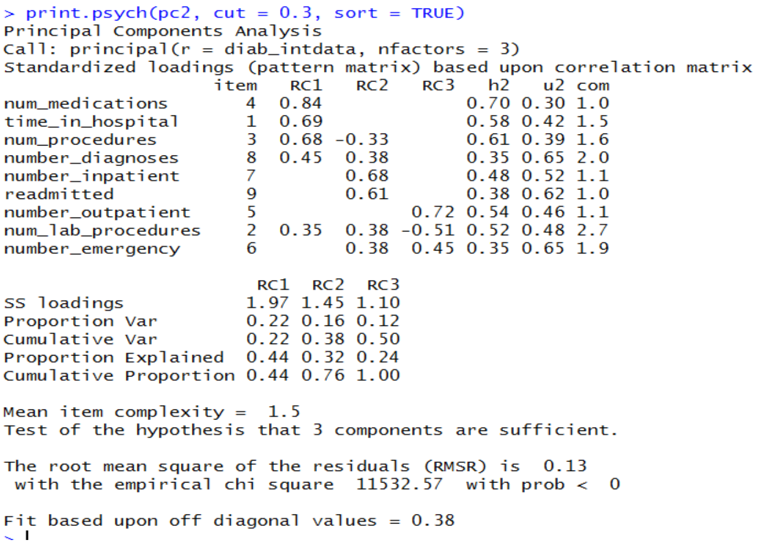
According to, Kaiser’s criterion, we retain only those components that have eigenvalue vectors greater than 1. In this case, we only retain 3 components.

Step 4: PCA with 3 factors

Now we do PCA again, only this time we have 3 factors instead of 9. The output shows the result of the second model.



```
> pc2<-principal(diab_intdata, nfactors=3)
> pc2
Principal Components Analysis
Call: principal(r = diab_intdata, nfactors = 3)
Standardized loadings (pattern matrix) based upon correlation matrix
                     RC1   RC2   RC3   h2   u2 com
time_in_hospital     0.69  0.19 -0.27 0.58 0.42 1.5
num_lab_procedures   0.35  0.38 -0.51 0.52 0.48 2.7
num_procedures       0.68 -0.33  0.20 0.61 0.39 1.6
num_medications      0.84  0.00  0.00 0.70 0.30 1.0
number_outpatient    0.09  0.15  0.72 0.54 0.46 1.1
number_emergency    -0.02  0.38  0.45 0.35 0.65 1.9
number_inpatient     0.00  0.68  0.12 0.48 0.52 1.1
number_diagnoses     0.45  0.38  0.06 0.35 0.65 2.0
readmitted           0.02  0.61  0.03 0.38 0.62 1.0

                       RC1  RC2  RC3
SS loadings           1.97 1.45 1.10
Proportion Var        0.22 0.16 0.12
Cumulative Var        0.22 0.38 0.50
Proportion Explained  0.44 0.32 0.24
Cumulative Proportion 0.44 0.76 1.00

Mean item complexity =  1.5
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.13
 with the empirical chi square  11532.57  with prob <  0

Fit based upon off diagonal values = 0.38
```



```
> print.psych(pc2, cut = 0.3, sort = TRUE)
Principal Components Analysis
Call: principal(r = diab_intdata, nfactors = 3)
Standardized loadings (pattern matrix) based upon correlation matrix
                   item  RC1   RC2   RC3   h2   u2 com
num_medications       4 0.84             0.70 0.30 1.0
time_in_hospital      1 0.69             0.58 0.42 1.5
num_procedures        3 0.68 -0.33       0.61 0.39 1.6
number_diagnoses      8 0.45  0.38       0.35 0.65 2.0
number_inpatient      7       0.68       0.48 0.52 1.1
readmitted            9       0.61       0.38 0.62 1.0
number_outpatient     5             0.72 0.54 0.46 1.1
num_lab_procedures    2 0.35  0.38 -0.51 0.52 0.48 2.7
number_emergency      6       0.38  0.45 0.35 0.65 1.9

                       RC1  RC2  RC3
SS loadings           1.97 1.45 1.10
Proportion Var        0.22 0.16 0.12
Cumulative Var        0.22 0.38 0.50
Proportion Explained  0.44 0.32 0.24
Cumulative Proportion 0.44 0.76 1.00

Mean item complexity =  1.5
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.13
 with the empirical chi square  11532.57  with prob <  0

Fit based upon off diagonal values = 0.38
```

We can interpret these results using print.psych() function as follows:

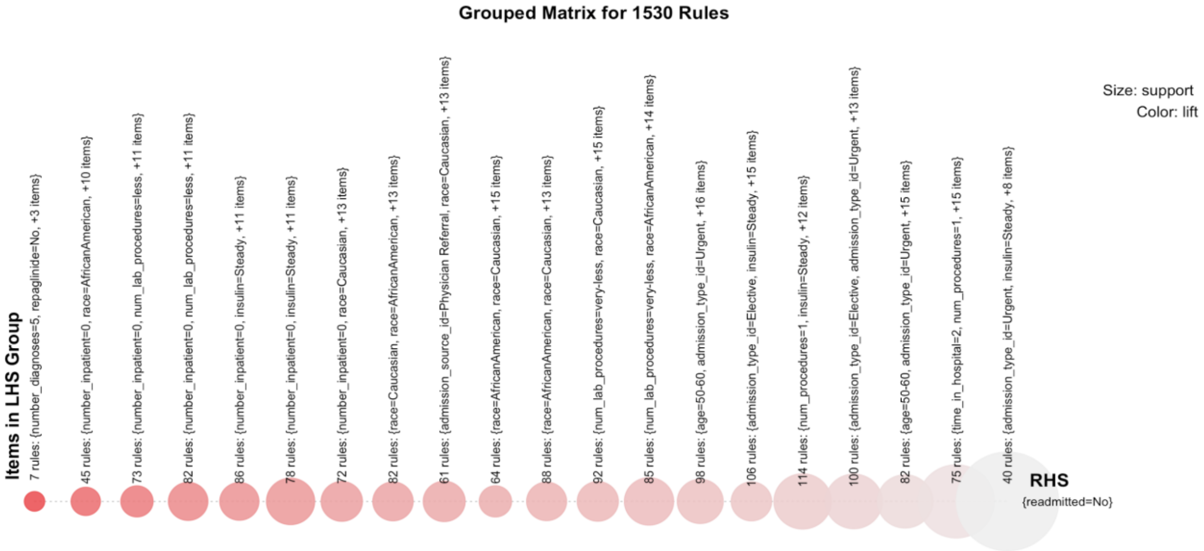
From these results we can interpret that num_medication have a higher loading on component 1 i.e num_diagnoses.So is num_procedures. Component 1 explains 44% of the total variance. We can interpret that if a person is diagnosed with multiple symptoms, he has more number of procedures and so the number of medications increases

2.7 Pattern Mining

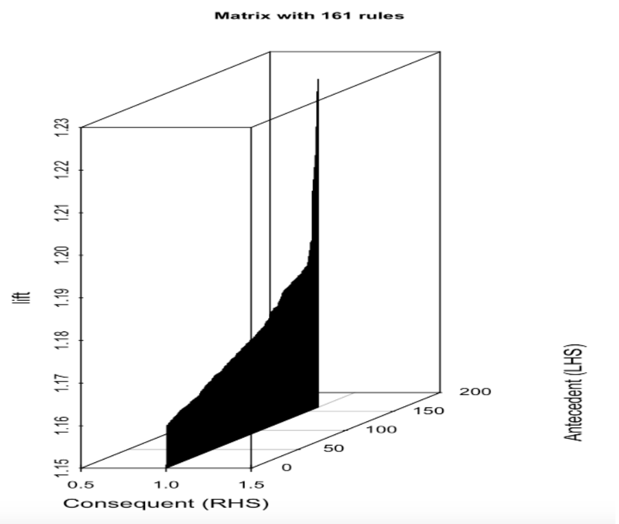
Pattern mining was done using apriori algorithm which is an influential algorithm for mining frequent item sets for Boolean association rules. This algorithm uses a bottom up approach where frequent subsets are extended one at a time.

Apriori was performed on 20 important variables that were extracted using feature selection method. Few of the variables that were numerical were first converted into factorial form such as num_lab_procedure and num_medications and were given a label per values like 0-25 for very less, 25-50 for moderate and so on. Apriori was applied on transformed data by taking support value of 0.1, confidence value of 0.6, minimum length of rules as 2 and readmitted as RHS. Rules were sorted and a subset matrix was formed to eliminate the redundant rules which was done by providing the values in the lower subset triangle to be zero. Redundant rules were eliminated and pruned rules were identified from the entire set of rules. There were total 1530 pruned rules identified.

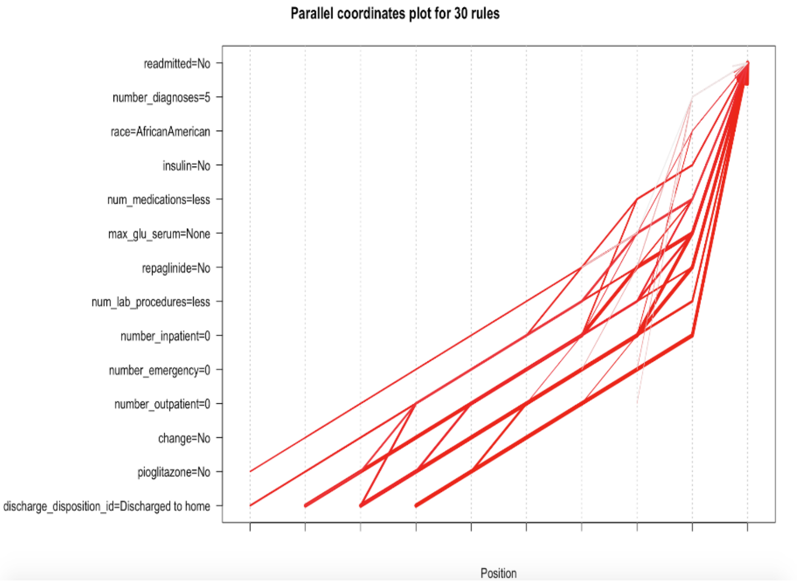
Below is the plot to visualize how these rules were categorized per lift and support values.



After the pruned rules, have been identified, a subset of rules was identified with confidence value greater than 0.7. This subset of rules contained total 161 rules which are important to predict frequent pattern that can lead to readmission values to be true or false.



3-D matrix graph to represent these rules with 'lift' on the y-axis and readmitted=No on the x-axis. Readmitted value is 'no' for almost all the rules because most of the rows have readmitted as false.



Parallel coordinate graph to visualize and understand the relationships.

<u>Interpretation-</u>

From the parallel coordinate graph of the 30 highest lift valued rules it can be interpreted that patients with less number of outpatient, less number of inpatient, less number of medications, no insulin taken, less number of emergency, who have not taken pioglitazone and repaglinide medicines, and whose race is african american are less likely to be readmitted.

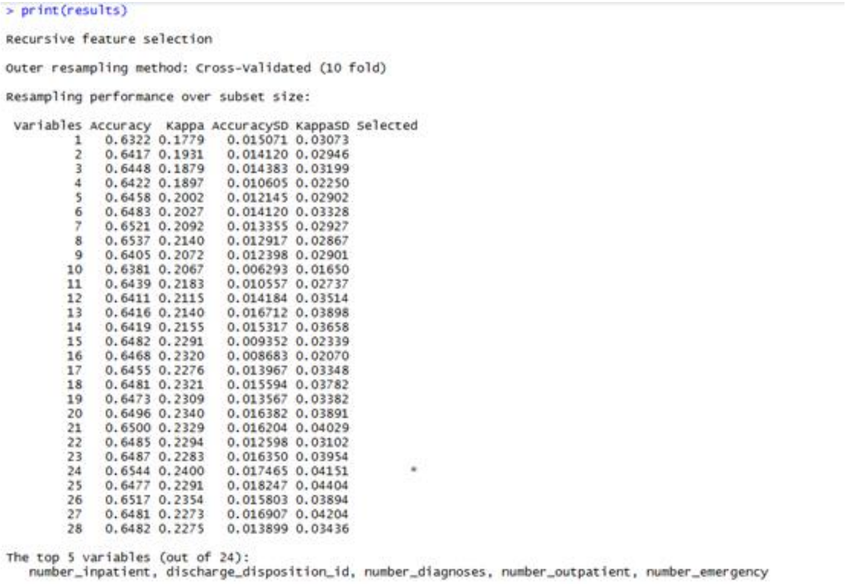
2.8 Feature Selection

To interpret the data in a more meaningful form, it is necessary to reduce the number of variables and remove attributes with less contribution in our data. Feature selection is important to subset important relevant attributes to build model and predict accurately. It helps to deal with curse of dimensionality, shortens the training time, improves results and avoids over fitting. We did feature selection using Random Forest and Boruta Package.

2.8.1 Random Forest

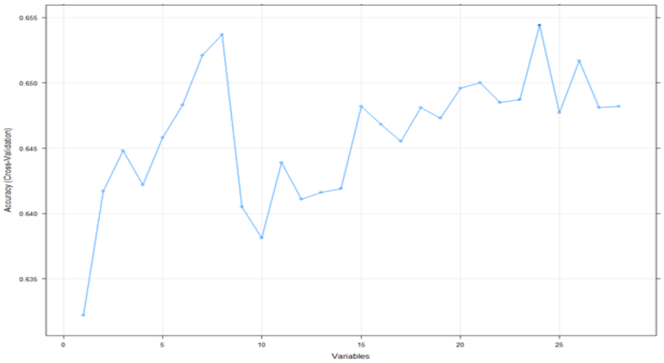
A popular automatic method for feature selection provided by the caret R package is called Recursive Feature Elimination or RFE. The algorithm is configured to explore all the possible subsets of the attribute. All the attributes are selected in this algorithm.

The output of rfe () is:



```
> print(results)

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         1   0.6322 0.1779   0.015071 0.03073
         2   0.6417 0.1931   0.014120 0.02946
         3   0.6448 0.1879   0.014383 0.03199
         4   0.6422 0.1897   0.010605 0.02250
         5   0.6458 0.2002   0.012145 0.02902
         6   0.6483 0.2027   0.014120 0.03328
         7   0.6521 0.2092   0.013355 0.02927
         8   0.6537 0.2140   0.012917 0.02867
         9   0.6405 0.2072   0.012398 0.02901
        10   0.6381 0.2067   0.006293 0.01650
        11   0.6439 0.2183   0.010557 0.02737
        12   0.6411 0.2115   0.014184 0.03514
        13   0.6416 0.2140   0.016712 0.03898
        14   0.6419 0.2155   0.015317 0.03658
        15   0.6482 0.2291   0.009352 0.02339
        16   0.6468 0.2320   0.008683 0.02070
        17   0.6455 0.2276   0.013967 0.03348
        18   0.6481 0.2321   0.015594 0.03782
        19   0.6473 0.2309   0.013567 0.03382
        20   0.6496 0.2340   0.016382 0.03891
        21   0.6500 0.2329   0.016204 0.04029
        22   0.6485 0.2294   0.012598 0.03102
        23   0.6487 0.2283   0.016350 0.03954
        24   0.6544 0.2400   0.017465 0.04151        *
        25   0.6477 0.2291   0.018247 0.04404
        26   0.6517 0.2354   0.015803 0.03894
        27   0.6481 0.2273   0.016907 0.04204
        28   0.6482 0.2275   0.013899 0.03436

The top 5 variables (out of 24):
   number_inpatient, discharge_disposition_id, number_diagnoses, number_outpatient, number_emergency
```

The output gives us top 5 important variables, number_inpatient, discharge_disposition_id, number_diagnoses, number_outpatient and number_emergency.

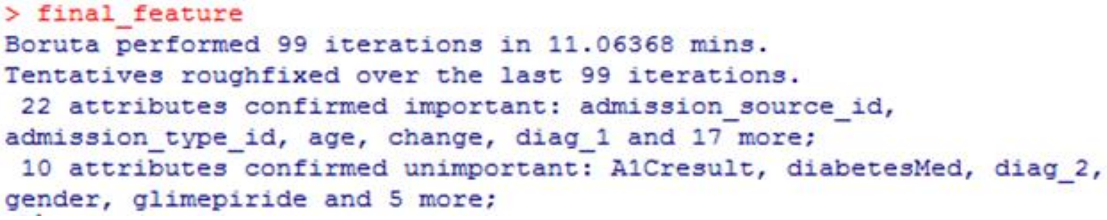


From this graph, we can interpret that 24 variables gives the most comparable result.

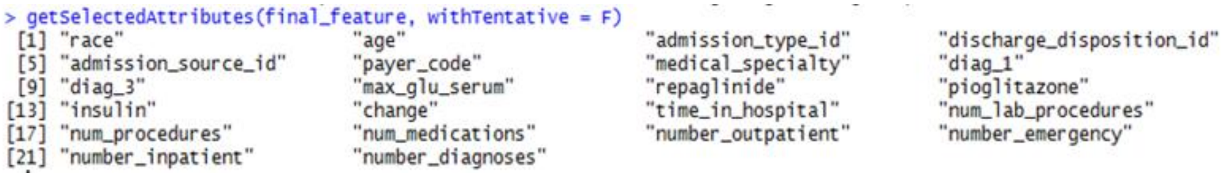
2.8.1 Boruta Package

The method performs top down search for relevant features by comparing original attributes importance with the importance achievable at random, progressively eliminating irrelevant features.

The output of running boruta() are:



```
> final_feature
Boruta performed 99 iterations in 11.06368 mins.
Tentatives roughfixed over the last 99 iterations.
 22 attributes confirmed important: admission_source_id,
admission_type_id, age, change, diag_1 and 17 more;
 10 attributes confirmed unimportant: A1Cresult, diabetesMed, diag_2,
gender, glimepiride and 5 more;
```



```
> getSelectedAttributes(final_feature, withTentative = F)
 [1] "race"                "age"                 "admission_type_id"   "discharge_disposition_id"
 [5] "admission_source_id" "payer_code"          "medical_specialty"   "diag_1"
 [9] "diag_3"              "max_glu_serum"       "repaglinide"         "pioglitazone"
[13] "insulin"             "change"              "time_in_hospital"    "num_lab_procedures"
[17] "num_procedures"      "num_medications"     "number_outpatient"   "number_emergency"
[21] "number_inpatient"    "number_diagnoses"
```

2.9 Classification:

We built different classification models and compared between them. We have tried to find the most accurate model based on evaluation parameters such as precision, recall, sensitivity, specificity and F1 measure. Also, we planned to finds any similarities between patterns extracted from each model. We used the following algorithms:

1. Decision tree
2. Random Forest
3. Neural Network
4. KNN

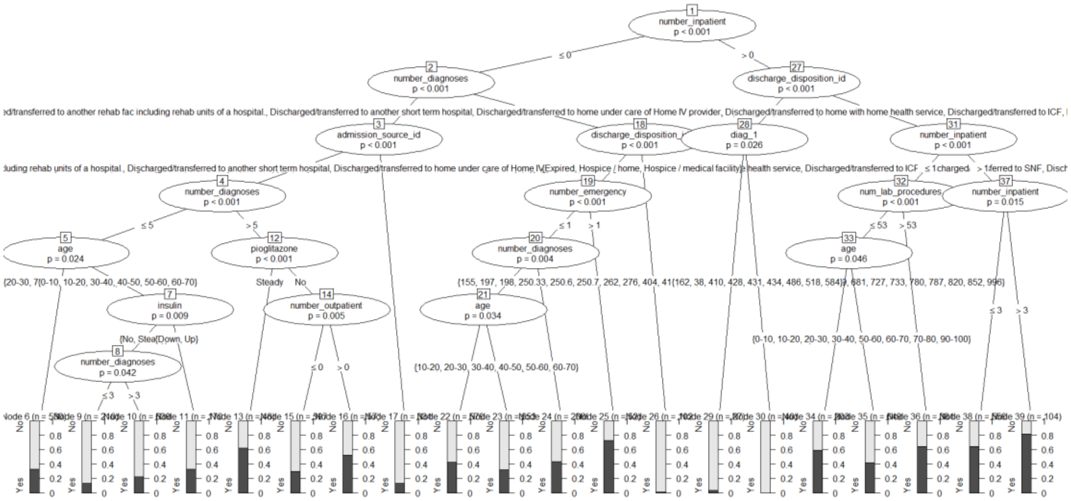
Data splitting used was 80% of data for training and 20% for testing. We also performed cross validation through KNN model, since it has built in function (traincontrol()) to do cross validation through building the model. For In the next section, we are going to explain each model in details.

2.9.1 Decision Tree:

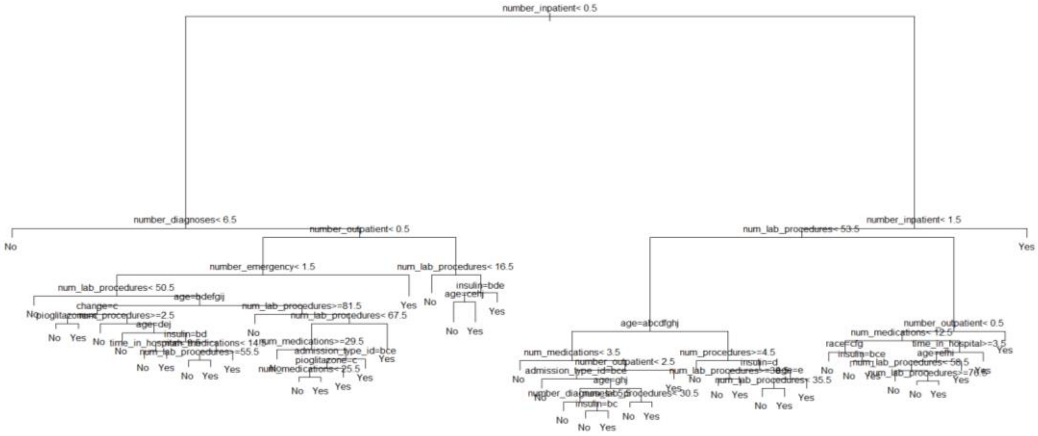
Decision tree is one of the most common and simple classification algorithm. Decision tree uses specific attributes to split data and finally reach the final decision. So, it would use variables that have significant contribution to identify the label. We used three decision tree models in R:

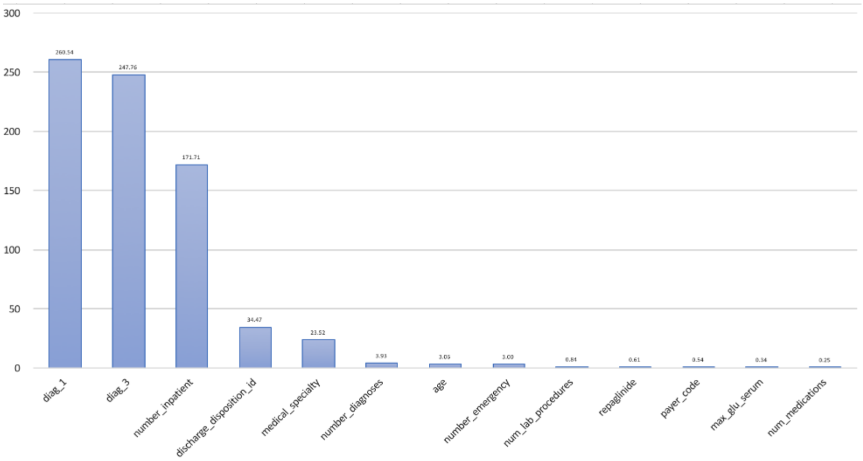
- Tree function under tree package
- Ctree function under party package
- Rpart function under rpart package.

The accuracy and other parameters are same, but trees built were different. We have the following trees : 1- ctree model 2- rpart model



1 number_inpatient p < 0.001
≤ 0
> 0
2 number_diagnoses p < 0.001
27 discharge_disposition_id p < 0.001
3 admission_source_id p < 0.001
18 discharge_disposition_i
28 diag_1 p = 0.026
31 number_inpatient p < 0.001
4 number_diagnoses p < 0.001
19 number_emergency p < 0.001
32 num_lab_procedures p < 0.001
37 number_inpatient p = 0.015
≤ 5
> 5
5 age p = 0.024
12 pioglitazone p < 0.001
20 number_diagnoses p = 0.004
33 age p = 0.046
Steady
No
7 insulin p = 0.009
14 number_outpatient p = 0.005
21 age p = 0.034
8 number_diagnoses p = 0.042
≤ 53
> 53
≤ 3
> 3
≤ 1
> 1
≤ 0
> 0
{No, Stea{Down, Up}
{10-20, 20-30, 30-40, 40-50, 50-60, 60-70}
{0-10, 10-20, 20-30, 30-40, 50-60, 60-70, 70-80, 90-100}
Node 39 (n = 104)
No
Yes
0.8
0.6
0.4
0.2
0

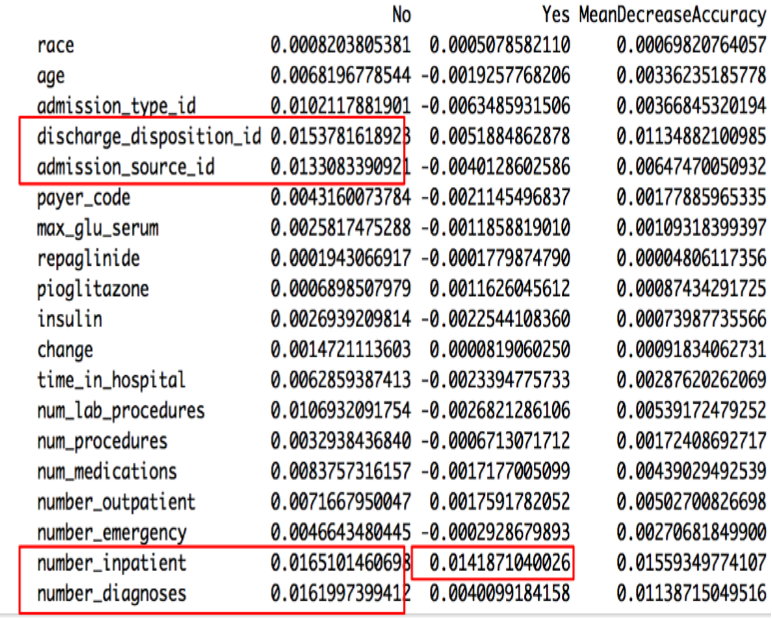




In addition, rpart model shows the importance of values used in the model. Here is a histogram that shows importance values of the attributes through building the tree.

2.9.2 Random Forest:

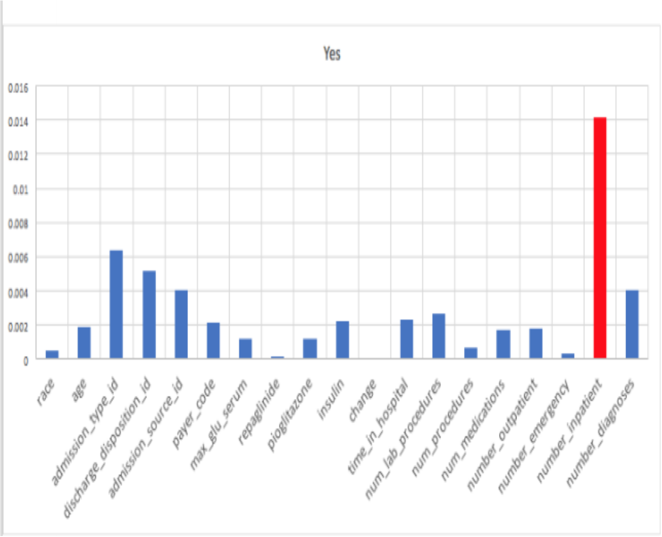
In Random Forest, we used randomForest package and function build our model. We used 500 tree to be generated. We have the following result:

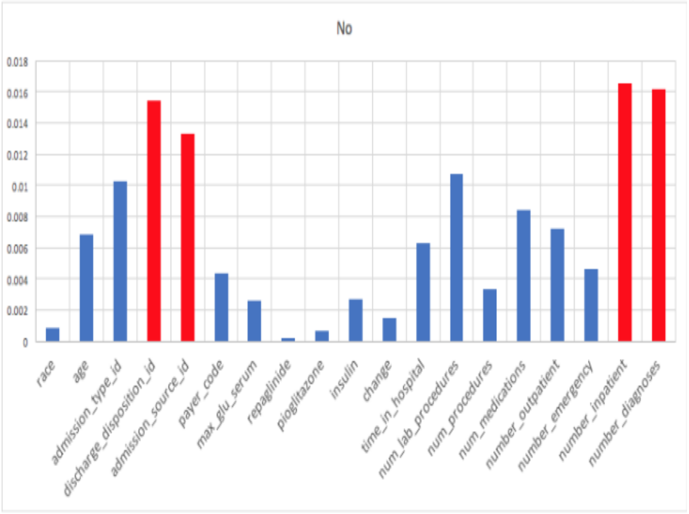


```
                                      No              Yes MeanDecreaseAccuracy
race                     0.0008203805381  0.0005078582110     0.00069820764057
age                      0.0068196778544 -0.0019257768206     0.00336235185778
admission_type_id        0.0102117881901 -0.0063485931506     0.00366845320194
discharge_disposition_id 0.0153781618923  0.0051884862878     0.01134882100985
admission_source_id      0.0133083390921 -0.0040128602586     0.00647470050932
payer_code               0.0043160073784 -0.0021145496837     0.00177885965335
max_glu_serum            0.0025817475288 -0.0011858819010     0.00109318399397
repaglinide              0.0001943066917 -0.0001779874790     0.00004806117356
pioglitazone             0.0006898507979  0.0011626045612     0.00087434291725
insulin                  0.0026939209814 -0.0022544108360     0.00073987735566
change                   0.0014721113603  0.0000819060250     0.00091834062731
time_in_hospital         0.0062859387413 -0.0023394775733     0.00287620262069
num_lab_procedures       0.0106932091754 -0.0026821286106     0.00539172479252
num_procedures           0.0032938436840 -0.0006713071712     0.00172408692717
num_medications          0.0083757316157 -0.0017177005099     0.00439029492539
number_outpatient        0.0071667950047  0.0017591782052     0.00502700826698
number_emergency         0.0046643480445 -0.0002928679893     0.00270681849900
number_inpatient         0.0165101460698  0.0141871040026     0.01559349774107
number_diagnoses         0.0161997399412  0.0040099184158     0.01138715049516
```

We found some attributes that became significant comparing to others for both “yes” and “no” situation. For example, number_inpatient represented high contribution to ‘yes’ classification by 0.014 while rest of variable were between 0 and 0.006.

The following graph presented the contribution of each variable in both ‘yes’ and ‘no’ classification.





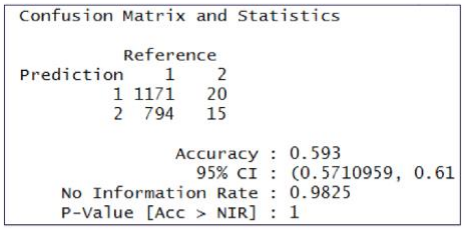
In addition, we used ***inTrees*** packages to extract and prune rules from *randomForest* model. Here is some rules extracted from our model.

len	fre	error	rules	predict
1	0.026	0.027	discharge_disposition_id %in% ● Admitted as an inpatient to this hospital', ● 'Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare' ● 'Discharged/transferred/referred another institution for outpatient services', ● 'Discharged/transferred/referred to this institution for outpatient services', ● 'Expired', ● Hospice / medical facility'	No
1	0.405	0.387	age = 20-30, 50-60, 60-70	No
3	0.001	0.429	age = 60-70 admission_type_id = Emergency discharge_disposition_id %in% ● Admitted as an inpatient to this hospital ● Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare ● Discharged/transferred/referred to this institution for outpatient services ● Hospice / home ● Hospice / medical facility	Yes

2.9.3 Neural Network

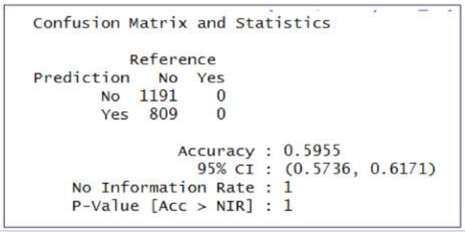
We used two models here just to compare between them. The first model uses multi-layer perceptron (MLP) algorithm in train function. Second model uses nnet function, and this model requires data to be scaled, centered and normalized. Here is the result of both model

nnet::nnet



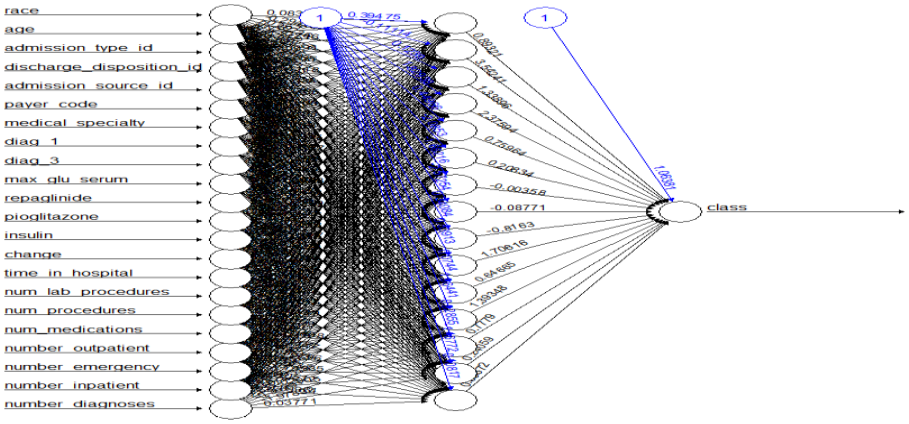
```
Confusion Matrix and Statistics

          Reference
Prediction    1    2
         1 1171   20
         2  794   15

               Accuracy : 0.593
                 95% CI : (0.5710959, 0.61
    No Information Rate : 0.9825
    P-Value [Acc > NIR] : 1
```

caret::train



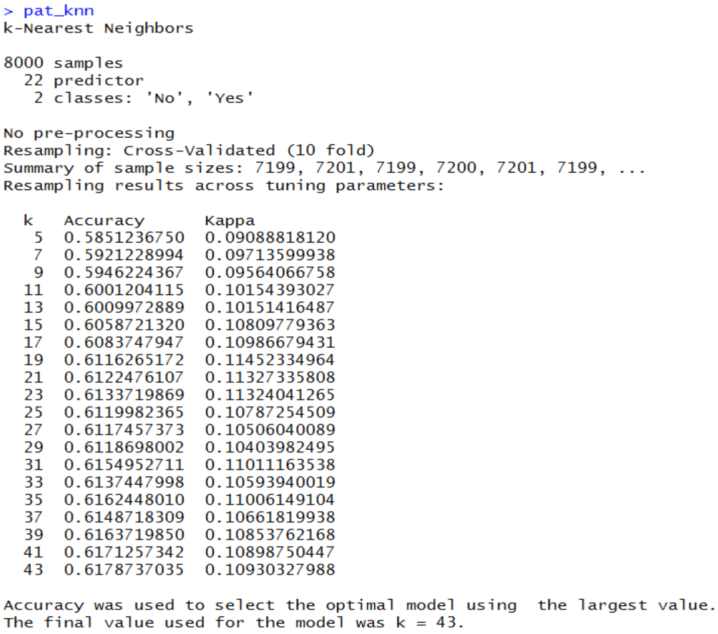
```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1191    0
       Yes  809    0

               Accuracy : 0.5955
                 95% CI : (0.5736, 0.6171)
    No Information Rate : 1
    P-Value [Acc > NIR] : 1
```

They have same accuracy, but MLP model faced some problem in calculating precision, recall and F1 since there is zero values for true negative(TN) and false negative(FN). We used *Neuralnet* package to do NN model and visualize it. Here is the plot diagram:



2.9.4 KNN

Under caret package, we used train () function with knn method. We also used cross validation in training our model. We used auto tuning to identify the best k value that has the highest accuracy. Here is the result of our model



```
> pat_knn
k-Nearest Neighbors

8000 samples
  22 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 7199, 7201, 7199, 7200, 7201, 7199, ...
Resampling results across tuning parameters:

  k   Accuracy      Kappa
   5  0.5851236750  0.09088818120
   7  0.5921228994  0.09713599938
   9  0.5946224367  0.09564066758
  11  0.6001204115  0.10154393027
  13  0.6009972889  0.10151416487
  15  0.6058721320  0.10809779363
  17  0.6083747947  0.10986679431
  19  0.6116265172  0.11452334964
  21  0.6122476107  0.11327335808
  23  0.6133719869  0.11324041265
  25  0.6119982365  0.10787254509
  27  0.6117457373  0.10506040089
  29  0.6118698002  0.10403982495
  31  0.6154952711  0.11011163538
  33  0.6137447998  0.10593940019
  35  0.6162448010  0.11006149104
  37  0.6148718309  0.10661819938
  39  0.6163719850  0.10853762168
  41  0.6171257342  0.10898750447
  43  0.6178737035  0.10930327988

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was k = 43.
```

The highest accuracy of our classification model achieved by using k= 43. We tuned our model by selecting length -20, so model tested 20 k values starting 5 and ended by 43. If we extended the length of tuning model, we may get better results than k = 43.

2.9.5 Evaluation our classification models:

Here is the table that comparing different classification algorithms:

Model	Accuracy	TP	TN	FP	FN	precision	recall	F-Mea s
Decision tree	0.638	213	1036	596	128	0.89	0.64	0.75
Random Forest	0.632	339	925	470	266	0.78	0.66	0.75
KNN	0.60	189	1016	175	620	0.62	.85	0.71
NN (train)	0.595	0	1191	809	0	NA	NA	NA
NN (nnet)	0.593	15	1171	794	20	0.98	0.60	0.74

Table shows that decision tree achieved the highest accuracy among other model while lowest accuracy went for NN models. NN is significant in achieving true negative (TN) among other models, but very low in true positive (TP). This is obvious in both precision and recall values. F-measure showed that both decision tree and random forest higher than others. However, we still have kind of bias here because decision tree didn't include all attributes (variables with levels less than 53), while other model included all data fields.

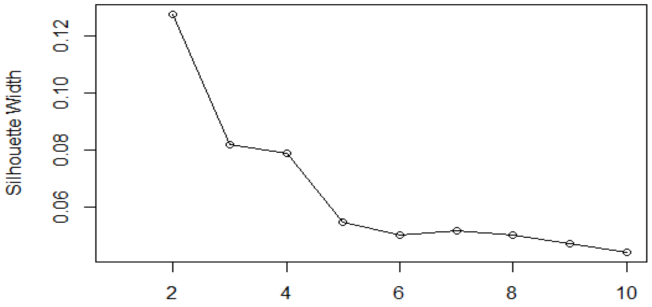
By seeing the results of models, we found some variables that become very important to assign accure label. Number_inpatient, for example, showed high importance in decision tree, used as the root of tree in both trees we built, and showed importance in classification using random forest model. Number_emerganct, discharge_disposition_id, diag_1, diag_3 represented importance in classification.

2.10 Clustering

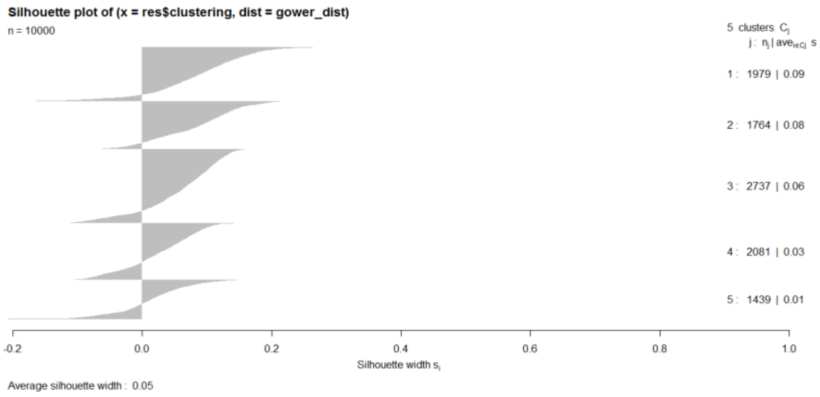
It is an explanatory data analysis technique used for identifying groups in the data set. Each group contains similar profile per the set criteria. The clusters have high intra similarity while low inter similarity. We implemented clustering to cluster patients based on their previous diagnoses. Hospital readmission is not only based on present symptoms but also on the medical history of the patient.

2.10.1 PAM

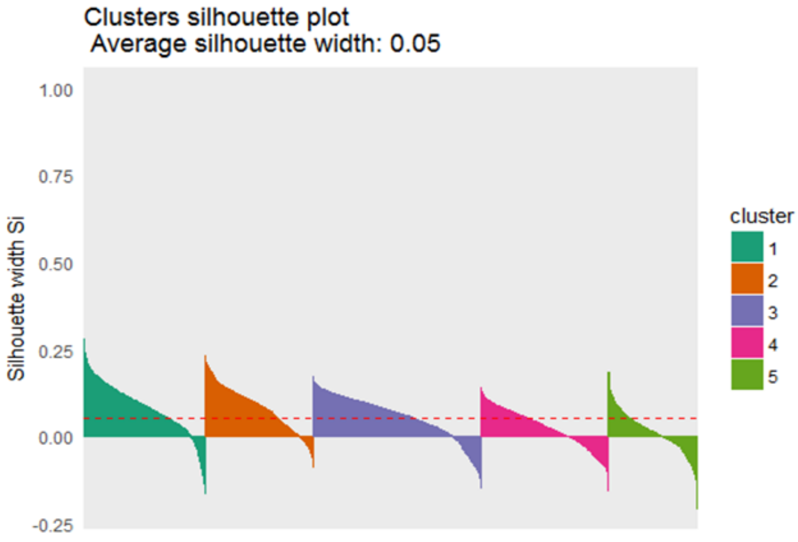
Since we had mixed data type, we implemented PAM(Partitioning Around Medoids) for clustering. This algorithm is based on k representative medoids among the dataset observations. The pam() function takes 2 inputs, dataset and the number of clusters(k). To find the optimal number of clusters we tried silhouette() function. This function is used to interpret and validate the consistency of the clusters.



From this silhouette graph, we can see that the line tends to straighten out at 5. The silhouette plot shows the cluster width. Larger the width, better the clusters. So we chose k =5 for our pam() function.

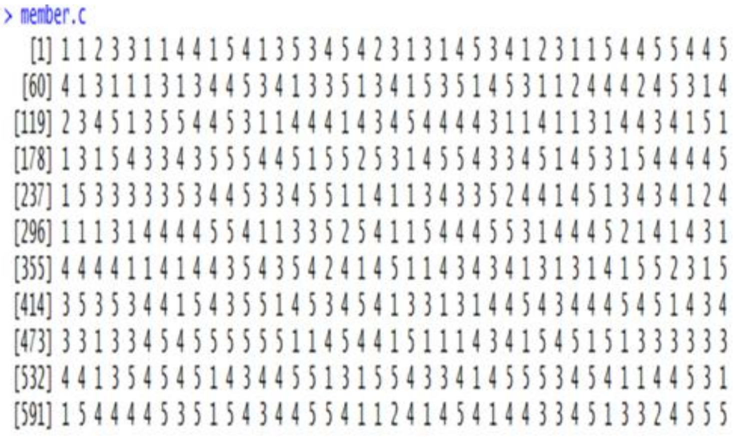


This plot shows 5 clusters, with total number of data items in each and the cluster width. Like cluster 1 has 1979 items with the width of 0.09.



This graph is more colored visualization of the silhouette plot using “factoextra” package. The lines below zero denotes that some items are wrongly clustered.

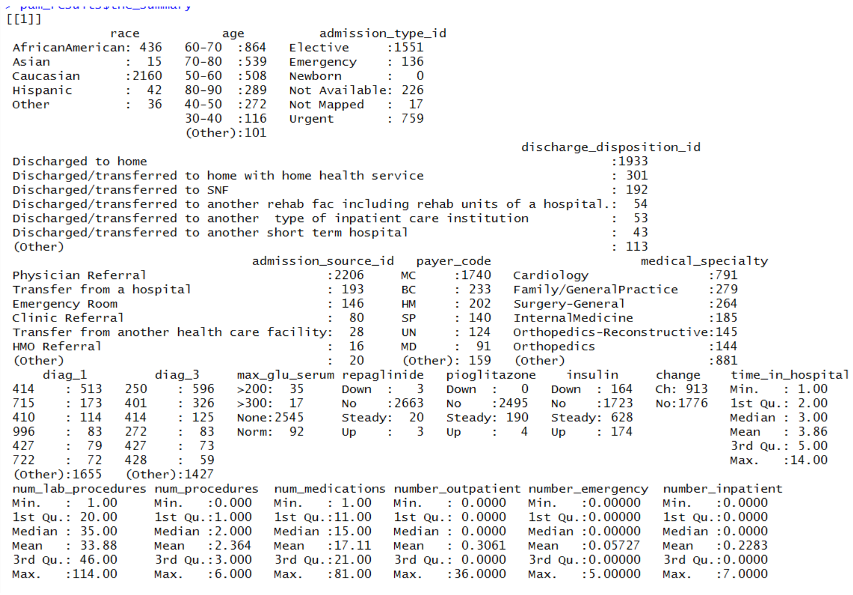
We can find the members of each cluster using member() function:



```
> member.c
  [1] 1 1 2 3 3 1 1 4 4 1 5 4 1 3 5 3 4 5 4 2 3 1 3 1 4 5 3 4 1 2 3 1 1 5 4 4 5 5 4 4 5
 [60] 4 1 3 1 1 1 3 1 3 4 4 5 3 4 1 3 3 5 1 3 4 1 5 3 5 1 4 5 3 1 1 2 4 4 4 2 4 5 3 1 4
[119] 2 3 4 5 1 3 5 5 4 4 5 3 1 1 4 4 4 1 4 3 4 5 4 4 4 4 3 1 1 4 1 1 3 1 4 4 3 4 1 5 1
[178] 1 3 1 5 4 3 3 4 3 5 5 5 4 4 5 1 5 5 2 5 3 1 4 5 5 4 3 3 4 5 1 4 5 3 1 5 4 4 4 4 5
[237] 1 5 3 3 3 3 3 5 3 4 4 5 3 3 4 5 5 1 1 4 1 1 3 4 3 3 5 2 4 4 1 4 5 1 3 4 3 4 1 2 4
[296] 1 1 1 3 1 4 4 4 4 5 5 4 1 1 3 3 5 2 5 4 1 1 5 4 4 4 5 5 3 1 4 4 4 5 2 1 4 1 4 3 1
[355] 4 4 4 4 1 1 4 1 4 4 3 5 4 3 5 4 2 4 1 4 5 1 1 4 3 4 3 4 1 3 1 3 1 4 1 5 5 2 3 1 5
[414] 3 5 3 5 3 4 4 1 5 4 3 5 5 1 4 5 3 4 5 4 1 3 3 1 3 1 4 4 5 4 3 4 4 4 5 4 5 1 4 3 4
[473] 3 3 1 3 3 4 5 4 5 5 5 5 5 5 1 1 4 5 4 4 1 5 1 1 1 4 3 4 1 5 4 5 1 5 1 3 3 3 3 3 3
[532] 4 4 1 3 5 4 5 4 5 1 4 3 4 4 5 5 1 3 1 5 5 4 3 3 4 1 4 5 5 5 3 4 5 4 1 1 4 4 5 3 1
[591] 1 5 4 4 4 4 5 3 5 1 5 4 3 4 4 5 5 4 1 1 2 4 1 4 5 4 1 4 4 3 3 4 5 1 3 3 2 4 5 5 5
```

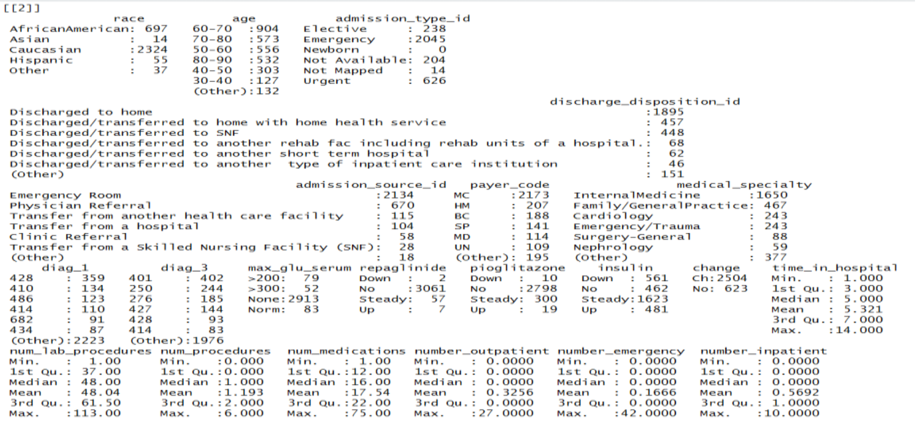
It shows that data item 1 belongs to cluster 1, item 8 belongs to cluster 4 and so on.

<u>Cluster 1:</u>



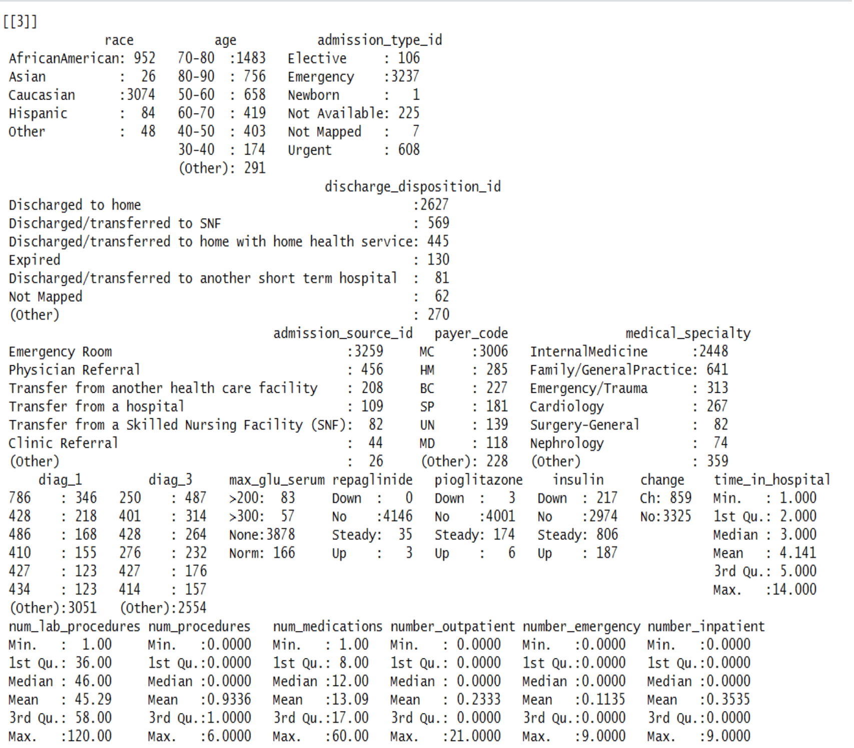
```
> pam_results$the_summary
[[1]]
              race            age        admission_type_id
 AfricanAmerican: 436   60-70   :864   Elective     :1551
 Asian          :  15   70-80   :539   Emergency    : 136
 Caucasian      :2160   50-60   :508   Newborn      :   0
 Hispanic       :  42   80-90   :289   Not Available: 226
 Other          :  36   40-50   :272   Not Mapped   :  17
                        30-40   :116   Urgent       : 759
                        (Other):101
                                                                     discharge_disposition_id
 Discharged to home                                                           :1933
 Discharged/transferred to home with home health service                      : 301
 Discharged/transferred to SNF                                                : 192
 Discharged/transferred to another rehab fac including rehab units of a hospital.:  54
 Discharged/transferred to another  type of inpatient care institution        :  53
 Discharged/transferred to another short term hospital                        :  43
 (Other)                                                                      : 113
                              admission_source_id   payer_code                   medical_specialty
 Physician Referral                      :2206      MC     :1740   Cardiology                :791
 Transfer from a hospital                : 193      BC     : 233   Family/GeneralPractice    :279
 Emergency Room                          : 146      HM     : 202   Surgery-General           :264
 Clinic Referral                         :  80      SP     : 140   InternalMedicine          :185
 Transfer from another health care facility:  28    UN     : 124   Orthopedics-Reconstructive:145
 HMO Referral                            :  16      MD     :  91   Orthopedics               :144
 (Other)                                 :  20      (Other): 159   (Other)                   :881
     diag_1         diag_3      max_glu_serum repaglinide  pioglitazone    insulin      change   time_in_hospital
 414     : 513   250     : 596   >200:  35     Down  :   3  Down  :   0   Down  : 164   Ch: 913  Min.   : 1.00
 715     : 173   401     : 326   >300:  17     No    :2663  No    :2495   No    :1723   No:1776  1st Qu.: 2.00
 410     : 114   414     : 125   None:2545     Steady:  20  Steady: 190   Steady: 628            Median : 3.00
 996     :  83   272     :  83   Norm:  92     Up    :   3  Up    :   4   Up    : 174            Mean   : 3.86
 427     :  79   427     :  73                                                                   3rd Qu.: 5.00
 722     :  72   428     :  59                                                                   Max.   :14.00
 (Other):1655   (Other):1427
 num_lab_procedures num_procedures  num_medications number_outpatient number_emergency  number_inpatient
 Min.   :  1.00     Min.   :0.000   Min.   : 1.00   Min.   : 0.0000   Min.   :0.00000   Min.   :0.0000
 1st Qu.: 20.00     1st Qu.:1.000   1st Qu.:11.00   1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.:0.0000
 Median : 35.00     Median :2.000   Median :15.00   Median : 0.0000   Median :0.00000   Median :0.0000
 Mean   : 33.88     Mean   :2.364   Mean   :17.11   Mean   : 0.3061   Mean   :0.05727   Mean   :0.2283
 3rd Qu.: 46.00     3rd Qu.:3.000   3rd Qu.:21.00   3rd Qu.: 0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
 Max.   :114.00     Max.   :6.000   Max.   :81.00   Max.   :36.0000   Max.   :5.00000   Max.   :7.0000
```

Cluster 2:



```
[[2]]
              race             age          admission_type_id
 AfricanAmerican: 697   60-70   :904   Elective        : 238
 Asian          :  14   70-80   :573   Emergency       :2045
 Caucasian      :2324   50-60   :556   Newborn         :   0
 Hispanic       :  55   80-90   :532   Not Available: 204
 Other          :  37   40-50   :303   Not Mapped      :  14
                        30-40   :127   Urgent          : 626
                        (Other):132
                                                              discharge_disposition_id
 Discharged to home                                                     :1895
 Discharged/transferred to home with home health service                : 457
 Discharged/transferred to SNF                                          : 448
 Discharged/transferred to another rehab fac including rehab units of a hospital.:  68
 Discharged/transferred to another short term hospital                  :  62
 Discharged/transferred to another  type of inpatient care institution  :  46
 (Other)                                                                : 151
                                      admission_source_id   payer_code                    medical_specialty
 Emergency Room                                   :2134     MC      :2173   InternalMedicine        :1650
 Physician Referral                               : 670     HM      : 207   Family/GeneralPractice: 467
 Transfer from another health care facility       : 115     BC      : 188   Cardiology              : 243
 Transfer from a hospital                         : 104     SP      : 141   Emergency/Trauma        : 243
 Clinic Referral                                  :  58     MD      : 114   Surgery-General         :  88
 Transfer from a Skilled Nursing Facility (SNF):  28     UN      : 109   Nephrology              :  59
 (Other)                                          :  18     (Other): 195   (Other)                 : 377
     diag_1          diag_3        max_glu_serum repaglinide   pioglitazone    insulin       change      time_in_hospital
 428     : 359   401     : 402   >200:  79     Down   :   2   Down   :  10   Down   : 561   Ch:2504   Min.   : 1.000
 410     : 134   250     : 244   >300:  52     No     :3061   No     :2798   No     : 462   No: 623   1st Qu.: 3.000
 486     : 123   276     : 185   None:2913     Steady:  57   Steady: 300   Steady:1623             Median : 5.000
 414     : 110   427     : 144   Norm:  83     Up     :   7   Up     :  19   Up     : 481             Mean   : 5.321
 682     :  91   428     :  93                                                                       3rd Qu.: 7.000
 434     :  87   414     :  83                                                                       Max.   :14.000
 (Other):2223   (Other):1976
 num_lab_procedures num_procedures  num_medications number_outpatient number_emergency  number_inpatient
 Min.   :  1.00     Min.   :0.000   Min.   : 1.00   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
 1st Qu.: 37.00     1st Qu.:0.000   1st Qu.:12.00   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
 Median : 48.00     Median :1.000   Median :16.00   Median : 0.0000   Median : 0.0000   Median : 0.0000
 Mean   : 48.04     Mean   :1.193   Mean   :17.54   Mean   : 0.3256   Mean   : 0.1666   Mean   : 0.5692
 3rd Qu.: 61.50     3rd Qu.:2.000   3rd Qu.:22.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 1.0000
 Max.   :113.00     Max.   :6.000   Max.   :75.00   Max.   :27.0000   Max.   :42.0000   Max.   :10.0000
```

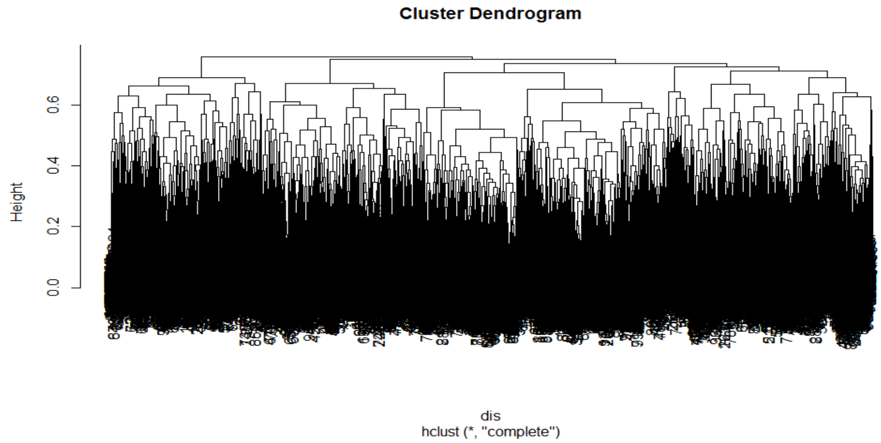
Cluster 3:



```
[[3]]
              race             age          admission_type_id
 AfricanAmerican: 952   70-80   :1483   Elective      : 106
 Asian          :  26   80-90   : 756   Emergency     :3237
 Caucasian      :3074   50-60   : 658   Newborn       :   1
 Hispanic       :  84   60-70   : 419   Not Available: 225
 Other          :  48   40-50   : 403   Not Mapped    :   7
                        30-40   : 174   Urgent        : 608
                        (Other): 291
                                               discharge_disposition_id
 Discharged to home                                        :2627
 Discharged/transferred to SNF                             : 569
 Discharged/transferred to home with home health service: 445
 Expired                                                   : 130
 Discharged/transferred to another short term hospital   :  81
 Not Mapped                                                :  62
 (Other)                                                   : 270
                                        admission_source_id   payer_code                 medical_specialty
 Emergency Room                                    :3259     MC     :3006   InternalMedicine      :2448
 Physician Referral                                : 456     HM     : 285   Family/GeneralPractice: 641
 Transfer from another health care facility        : 208     BC     : 227   Emergency/Trauma      : 313
 Transfer from a hospital                          : 109     SP     : 181   Cardiology            : 267
 Transfer from a Skilled Nursing Facility (SNF):  82     UN     : 139   Surgery-General       :  82
 Clinic Referral                                   :  44     MD     : 118   Nephrology            :  74
 (Other)                                           :  26     (Other): 228   (Other)               : 359
     diag_1         diag_3       max_glu_serum repaglinide   pioglitazone    insulin       change     time_in_hospital
 786     : 346   250     : 487   >200:  83    Down  :   0   Down  :   3   Down  : 217   Ch: 859   Min.   : 1.000
 428     : 218   401     : 314   >300:  57    No    :4146   No    :4001   No    :2974   No:3325   1st Qu.: 2.000
 486     : 168   428     : 264   None:3878    Steady:  35   Steady: 174   Steady: 806             Median : 3.000
 410     : 155   276     : 232   Norm: 166    Up    :   3   Up    :   6   Up    : 187             Mean   : 4.141
 427     : 123   427     : 176                                                                    3rd Qu.: 5.000
 434     : 123   414     : 157                                                                    Max.   :14.000
 (Other):3051   (Other):2554
 num_lab_procedures num_procedures  num_medications number_outpatient number_emergency number_inpatient
 Min.   :  1.00     Min.   :0.0000  Min.   : 1.00   Min.   : 0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.: 36.00     1st Qu.:0.0000  1st Qu.: 8.00   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median : 46.00     Median :0.0000  Median :12.00   Median : 0.0000   Median :0.0000   Median :0.0000
 Mean   : 45.29     Mean   :0.9336  Mean   :13.09   Mean   : 0.2333   Mean   :0.1135   Mean   :0.3535
 3rd Qu.: 58.00     3rd Qu.:1.0000  3rd Qu.:17.00   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :120.00     Max.   :6.0000  Max.   :60.00   Max.   :21.0000   Max.   :9.0000   Max.   :9.0000
```

From here we found that cluster 1 has the patients who had previous history of circulatory and connective tissue problems, were readmitted on physician's referral and having max_glu_serum not measured earlier.. Cluster 2 represents patients who have cardiac problems in the past and were admitted at emergency and having stable insulin. Cluster 3 represents patients having respiratory problems with the age of 70-80 and max_glu_serum>200.

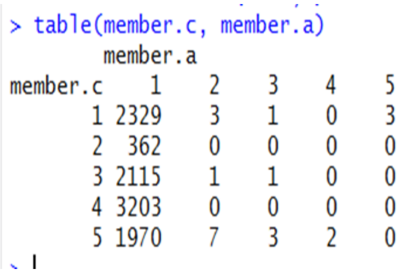
2.10.2 Hierarchical Clustering:



We also tried to implement hclust() from cluster package to plot dendrograms.

hclust() function was

implemented using complete linkage and average linkage. Complete linkage method uses maximum distance between the data clusters. While average linkage method uses the mean of the distance between the clusters. Members.c represents members of the clusters using complete linkage method, while members.a represents average linkage method.



```
> table(member.c, member.a)
        member.a
member.c    1    2    3    4    5
       1 2329    3    1    0    3
       2  362    0    0    0    0
       3 2115    1    1    0    0
       4 3203    0    0    0    0
       5 1970    7    3    2    0
```

This table shows that 2329 items belong to cluster 1, while 7 items were wrongly predicted for cluster 1.

3. Findings and Analysis

We have some findings through our analysis. We have focused on significant variables through our results in classification, which are: number_inpatient, number_diagnoses, ICD 9 diag_1, ICD 9 diag_3 , and Discharge disposition_id. By checking these important variables, we found some patterns. We found that if number_inpatient > 7.5, label will be ‘Yes’. We also find that if number_emergency more than 8.5, the class will be ‘yes’. However, we found many similarities between many variables by filtering both ‘Yes’ and ‘No’ observation.

We also examine ICD 9 codes frequency for both ‘Yes’ and ‘No’ observations. We found that code 428 is most frequent in Yes observation, while 414 for No observation. 428 is about heart attack and there are researches confirmed the relationship between diabetes and heart attack risk [5]. In our data, 428 has 8% of readmitted cases (342 out of 3965) and 4% of not readmitted cases (294 out of 6035). Diabetes, like other chronic medical condition, is associated with increased risk of hospital readmission. Efforts to reduce readmission should be multifactorial and encompass both general and specific diabetic measures. We found certain factors played important role in readmission analysis.

4. References:

- https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/
- http://machinelearningmastery.com/feature-selection-with-the-caret-r-package/
- http://www.sthda.com/english/wiki/partitioning-cluster-analysis-quick-start-guide-unsupervised-machine-learning
- http://www.sthda.com/english/wiki/cluster-analysis-in-r-unsupervised-machine-learning
- https://cran.r-project.org/web/packages/dendextend/vignettes/introduction.html
- http://www.joslin.org/info/diabetes_and_heart_disease_an_intimate_connection.html

5. Appendices:

R Code:

```
setwd("/Users/Imani/Desktop/UNCC/KDD/Final Project")
#converting all empty and ? to NA
diab_data<- read.csv("10kDiabetes.csv", header = TRUE, na.strings = c("","?","NA"))
head(diab_data)
str(diab_data)

#Part 1
#removing weight, since they are too sparse to process
diab_data[["weight"]]<-NULL



#Part 2
#CORR RCORR
diab_cordata <- subset(diab_data, select = c("time_in_hospital","num_lab_procedures","num_procedures","num_medications","number_outpatient","number_emergency","number_inpatient","number_diagnoses","readmitted"))
cor_data<-cor(diab_data$time_in_hospital,diab_data$num_lab_procedures,use="complete.obs",method="pearson")
cor_data

cor_cordata<-cor(diab_cordata,use="complete.obs",method="pearson")
cor_cordata

#rcorr
install.packages("Hmisc", dependencies = TRUE)
library(Hmisc)
tnlp<-rcorr(diab_data$time_in_hospital,diab_data$num_lab_procedures, type = "pearson")
tnlp

install.packages("plyr", dependencies = TRUE)
library('plyr')
count(diab_data,'race')
count(diab_data,'gender')

#chi-square analysis
counts1 <- ddply(diab_data,.(diab_data$race,diab_data$readmitted),nrow)
counts1
```

```
names(counts1)<-c("Race","Readmitted","Freq")

african<-c(725,1361)
caucasian<-c(3115,4442)
asian<-c(19,36)
hispanic<-c(65,116)
other<-c(41,80)

race_readmitted <- cbind(african,caucasian,asian,hispanic,other)
row.names(race_readmitted) = c("TRUE","FALSE")

install.packages("gmodels")
library(gmodels)
CrossTable(race_readmitted, chisq = TRUE, expected = TRUE, sresid = TRUE, format = "SPSS")

install.packages("mlogit")
library(mlogit)

model1 <- glm(diab_data$readmitted ~ diab_data$number_inpatient+diab_data$number_outpatient+diab_data$num_lab_procedures,data = diab_data, family = binomial())
model2 <- glm(diab_data$readmitted ~ diab_data$number_inpatient+diab_data$number_outpatient+diab_data$number_diagnoses,data = diab_data, family = binomial())
summary(model1)
summary(model2)

anova(model1,model2)

#FEATURE SELECTION
library(mlbench)

# feature selection using random forest
control2 <- rfeControl(functions=rfFuncs, method="cv", number=10)
# run the RFE algorithm
results <- rfe(patient[,1:32], patient[,33], sizes=c(1:32), rfeControl=control2)
# summarize the results
print(results)
# list the chosen features
```

```
predictors(results)
# plot the results
plot(results)


# feature selection using Boruta
install.packages("Boruta")
library(Boruta)
patient_boruta <- Boruta(readmitted~., data = patient, doTrace = 2)
final_feature<- TentativeRoughFix(patient_boruta)
getSelectedAttributes(final_feature, withTentative = F) # showes selected attributes

#CLASSIFICATION
#####Training data  ###########

pat_ind <- sample(1:nrow(patient),0.8*nrow(patient))
train <- new_pat[pat_ind,]
test <- new_pat[-pat_ind,]


#############Decision Tree  ################


# decision tree using party package
ctree <- party::ctree(class~.,train)
plot(tree)

#decision tree using rpart package
tree_rp <- rpart::rpart(class~.,train, method = "class", minsplit = 1, minbucket = 1, cp = 0.001)
#decision tree using tree package:
tree <- tree::tree(class~.,train)


############Random Forest  ##################

rf <- randomForest(class~., data = train[,c(1:6,10:23)], importance=TRUE, ntree=500)
library(inTrees)
# extract rules from rf model
rf_rules <- extractRules(RF2List(pat_rf),new_pat[,c(1:6,10:22)])
```

```
#############  KNN ##############

control <- trainControl(method="repeatedcv", number=10, repeats=3) # 10 fold cross validation
control_knn <- trainControl(method="cv", number=10) # 10 fold cross validation
clus<-makeCluster(spec=8,type="PSOCK")
registerDoParallel(clus)
pat_knn <- train(class ~ ., data = train, method = "knn", trControl=control_knn,tuneLength = 20)
stopCluster(clus)



################  Neural Network  ############
# using caret package
clus<-makeCluster(spec=8,type="PSOCK")
registerDoParallel(clus)
pat_nn <- train(class ~ ., data =
train,method="mlp",metric='Accuracy',tuneGrid=expand.grid(.size=1:15))
stopCluster(clus)

#another NN models
# scale
preprocessParams <- preProcess(tn[,1:22], method=c("scale"))
tn2 <- predict(preprocessParams, tn[,1:22])
tn2$class <- new_pat$readmitted

# normalize
preprocessParams2 <- preProcess(tn[,1:22], method=c("range"))
tn3 <- predict(preprocessParams2, tn[,1:22])
tn3$class <- tn$readmitted

tn_train <- tn3[pat_ind,]
tn_test <- tn3[-pat_ind,]

#build another NN package
library(neuralnet)
n <- names(tn_train)
f <- as.formula(paste('class ~', paste(n[!n %in% "class"], collapse = " + ")))
nnnn <- neuralnet(f,data=tn_train,hidden=15,linear.output=FALSE)
```

```
xnnx <- compute(nnnn,tn_test[1:22])
xnnx
plot(nnnn)

# another NN package
library(nnet)
ideal <- class.ind(tn$readmitted)
ANN = nnet(tn[pat_ind,1:22], ideal[pat_ind,], size=10)
nnp <- predict(ANN, tn[-pat_ind,1:22], type="class")
caret::confusionMatrix(tn_test$class,nnp)


###############  valdiating models   #################



#valdiate tree
val_tree <- predict(tree, test)
table(val == test[,23])

val_rp <- predict(tree_rp, test, type= "class")
table(val_rp == test[,23])

val_ctree <- predict(ctree, test)
table(val_ctree == test[,23])

val_knn2 <-  predict(pat_knn2, test)

# validate knn
  val_knn <-  predict(pat_knn, test)
table(val_knn == test[,23])

val_knn2 <- predict(pat_knn2, test)
table(val_knn2 == test[,23])

# validate random forest
val_rf <- predict(pat_rf, test)
table(val_rf == test[,23])

# validate NN
```

```
val_nn <- predict(pat_nn, test)
val_nnnn <- predict(nnnn, test2)
table(val_nn == test[,23])
```

```
#CLUSTERING
patient_data<- read.csv("patient.csv")
str(patient_data)
patient_data[["X"]]<-NULL
patient_data[["payer_code"]]<-NULL
library(cluster)
#scatter plot
plot(time_in_hospital~ diag_1, patient_data)
?daisy
dis<- daisy(patient_data, metric = "gower")
print(dis, digits = 3)
head(dis)

#complete linkage
hc.c<-hclust(dis)
plot(hc.c,hang = -1, cex = 0.6)

#average linkage
hc.a<- hclust(dis, method = "average")

#cluster membership
member.c<-cutree(hc.c, 5)
member.c

member.a<- cutree(hc.a,5)
table(member.c, member.a)

#cluster means
aggregate(patient_data, list(member.c), mean)

#silhoutte plot
library(cluster)
windows()
plot(silhouette(cutree(hc.c,5), dis))
```

```
table(member.a, patient_data$readmitted)


#dendextend
library(dendextend)
dend<-as.dendrogram(hc.c)
d1=color_branches(dend,k=5)
plot(d1)

plot(hc.c, type = "phylogram", show.tip.label = TRUE,
     edge.color = "red", edge.width = 1, edge.lty = 1,
     tip.color = "blue")


install.packages("devtools")
library(devtools)
devtools::install_github("kassambara/factoextra")
library(factoextra)
#clustering
clust_data<- patient_data
head(patient_num_data)
medians<- apply(clust_data,2,median)
mads<- apply(clust_data,2,mad)
clust_data = scale(clust_data,center=medians,scale=mads)

patient.dist<-dist(clust_data)
#hclust uses complete linkage
patient.clust<-hclust(patient.dist)
plot(patient.clust,labels=patient_data$readmitted,main='Default from hclust')

#dendextend package
install.packages("dendextend")
install.packages("colorspace")

library(dendextend)
library(colorspace)

k <- 4
cols <- rainbow_hcl(k)
```

```
dend <- as.dendrogram(patient.clust)
dend <- color_branches(dend, k = k)
plot(dend)
labels_dend <- labels(dend)
groups <- cutree(dend, k=4, order_clusters_as_data = FALSE)
dends <- list()
for(i in 1:k) {
  labels_to_keep <- labels_dend[i != groups]
  dends[[i]] <- prune(dend, labels_to_keep)
}
par(mfrow = c(2,2))
for(i in 1:k) {
  plot(dends[[i]],
       main = paste0("Tree number ", i))
}
groups.12 = cutree(patient.clust,12)
table(groups.12)

#Try clustering
library(dplyr)
library(cluster)
library(ggplot2)

?daisy

gower_dist<-daisy(patient_data, metric = "gower")
summary(gower_dist)
gower_mat <- as.matrix(gower_dist)
#find most similar data pairs
patient_data[which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]),arr.ind = TRUE)[1, ], ]

sil_width <- c(NA)
for(i in 2:10){

  pam_fit <- pam(gower_dist,
                 diss = TRUE,
                 k = i)

  sil_width[i] <- pam_fit$silinfo$avg.width
}

#Silhouette analysis measures how well an observation is clustered and
#it estimates the average distance between clusters. The silhouette plot displays a
#measure of how close each point in one cluster is to points in the neighboring clusters.
#plot sil width, higher the better

plot(1:10, sil_width,
     xlab = "Number of clusters",
     ylab = "Silhouette Width")
lines(1:10, sil_width)
```