BA636 Data mining

# Prediction of readmission of a diabetic patient after treatment

Ravi Teja Reddy, Bandi

11012426

Table of Contents

[Prediction of readmission of a diabetic patient after treatment 1](#_Toc58448566)

[Abstract 3](#_Toc58448567)

[Data set 3](#_Toc58448568)

[Data exploration and cleaning 7](#_Toc58448569)

[Dimensionality and data reduction 10](#_Toc58448570)

[Data mining task 19](#_Toc58448571)

[Data mining techniques 19](#_Toc58448572)

[evaluation and results 23](#_Toc58448573)

[Conclusion 25](#_Toc58448574)

# **Abstract**

A large portion of hospital inpatient spending is because of hospital readmission rates. And Diabetes is one of the top ten leading causes of deaths, which is also considered most expensive disease in United states. Patients with diabetes are at high risk of getting readmitted than those patients who are not diabetic. There for understanding who/what kind of patients are readmitting and trying to minimize readmission rate will help patients to reduce medical costs. *The objective* of this study is to predict the likelihood of a diabetic patient being readmitted.

# **Data set**

After extensive search for data sets this data set has piqued my interest. The data was submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University and the de identified abstract data has been stored and collected from UCI Machine learning repository**1**.

The data set**2** contains instances nearly 100,000 and over 50 variables which are collected over a time of 10 years from 130 US hospitals. The data set contains information regarding

1. Inpatient hospital admission.
2. Diabetic encounter, diagnosis of which kind of diabetes was entered to the system.
3. Length of stay with min 1 day and max 14 days.
4. Laboratory tests performed during the encounter.
5. Medications administered during the encounter.

Based on above criteria the set contains attributes patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in previous year before hospitalization, etc.

The stakeholders in this study are Government, Insurance industry, Educational industry and patients.

1 [https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#](https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008)

2 <https://archive.ics.uci.edu/ml/machine-learning-databases/00296/dataset_diabetes.zip>

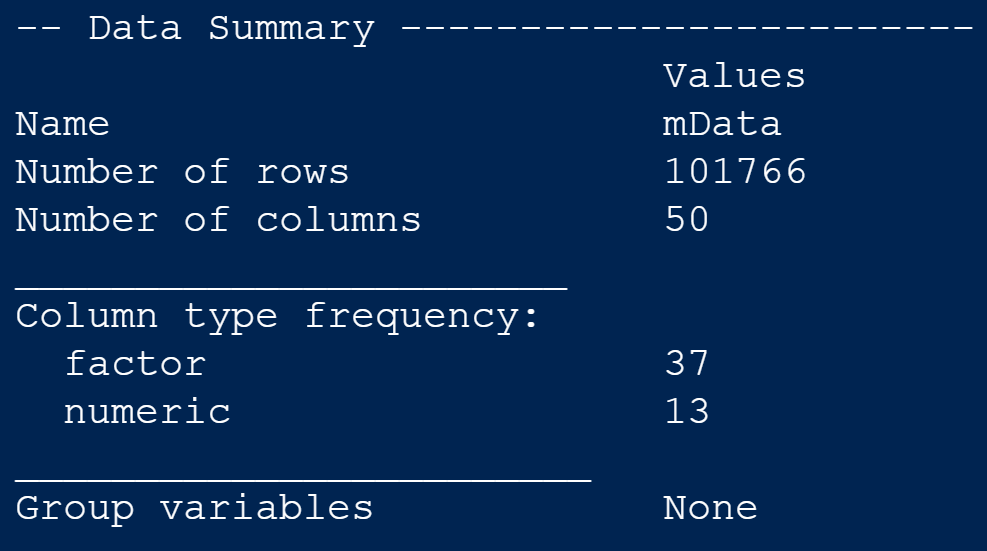
#### Data dictionary

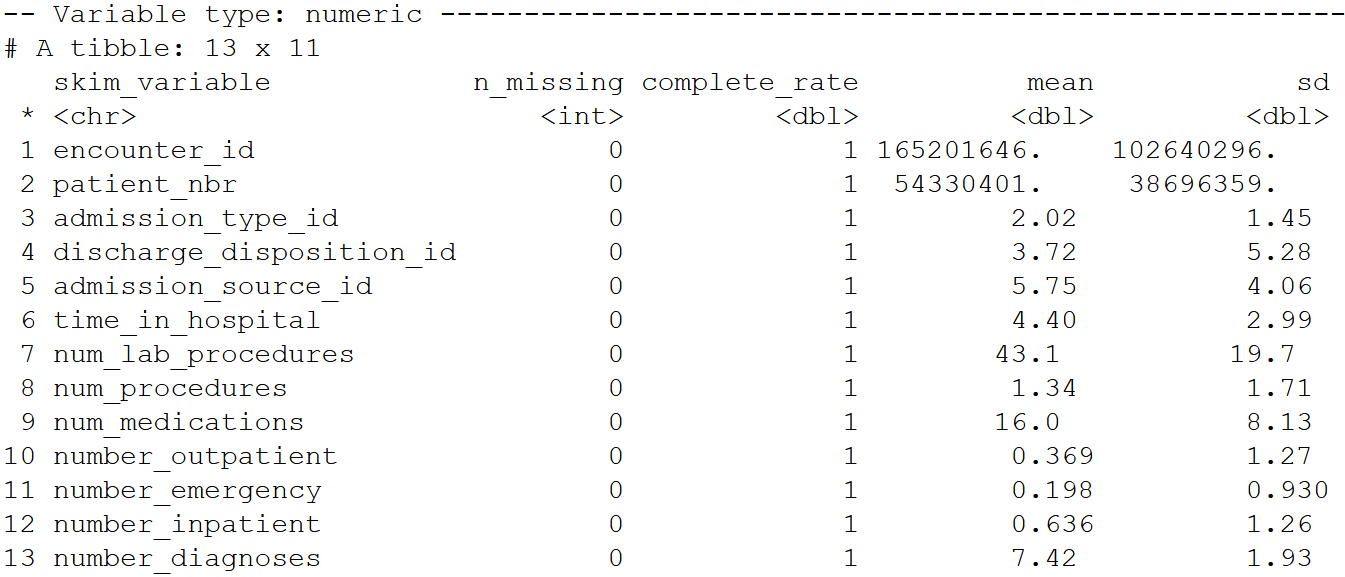
|  |  |
| --- | --- |
| **Variable** | **Description** |
| encounter\_id | Id given during visit of patient |
| patient\_nbr | Patient number |
| race | Race of the patient |
| Gender | Gender of patient |
| age | age of patient |
| weight | weight of patient |
| admission\_type\_id | 1=Emergency |
| 2=Urgent |
| 3=Elective |
| 4=Newborn |
| 5=Not Available |
| 6=NULL |
| 7=Trauma Center |
| 8=Not Mapped |
| discharge\_disposition\_id | 1=Discharged to home |
| 2=Discharged/transferred to another short-term hospital |
| 3=Discharged/transferred to SNF |
| 4=Discharged/transferred to ICF |
| 5=Discharged/transferred to another type of inpatient care institution |
| 6=Discharged/transferred to home with home health service |
| 7=Left AMA |
| 8=Discharged/transferred to home under care of Home IV provider |
| 9=Admitted as an inpatient to this hospital |
| 10=Neonate discharged to another hospital for neonatal aftercare |
| 11=Expired |
| 12=Still patient or expected to return for outpatient services |
| 13=Hospice / home |
| 14=Hospice / medical facility |
| 15=Discharged/transferred within this institution to Medicare approved swing bed |
| 16=Discharged/transferred/referred another institution for outpatient services |
| 17=Discharged/transferred/referred to this institution for outpatient services |
| 18=NULL |
| 19=Expired at home. Medicaid only, hospice. |
| 20=Expired in a medical facility. Medicaid only, hospice. |
| 21=Expired, place unknown. Medicaid only, hospice. |
| 22=Discharged/transferred to another rehab fac including rehab units of a hospital. |
| 23=Discharged/transferred to a long-term care hospital. |
| 24=Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare. |
| 25=Not Mapped |
| 26=Unknown/Invalid |
| 30=Discharged/transferred to another Type of Health Care Institution not Defined Elsewhere |
| 27=Discharged/transferred to a federal health care facility. |
| 28=Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital |
| 29=Discharged/transferred to a Critical Access Hospital (CAH). |
| admission\_source\_id | 1= Physician Referral |
| 2=Clinic Referral |
| 3=HMO Referral |
| 4=Transfer from a hospital |
| 5= Transfer from a Skilled Nursing Facility (SNF) |
| 6= Transfer from another health care facility |
| 7= Emergency Room |
| 8= Court/Law Enforcement |
| 9= Not Available |
| 10= Transfer from critical access hospital |
| 11=Normal Delivery |
| 12= Premature Delivery |
| 13= Sick Baby |
| 14= Extramural Birth |
| 15=Not Available |
| 17=NULL |
| 18= Transfer from Another Home Health Agency |
| 19=Readmission to Same Home Health Agency |
| 20= Not Mapped |
| 21=Unknown/Invalid |
| 22= Transfer from hospital input/same fac result in a sep claim |
| 23= Born inside this hospital |
| 24= Born outside this hospital |
| 25= Transfer from Ambulatory Surgery Center |
| 26=Transfer from Hospice |
| time\_in\_hospital | Time spent in hospital in months |
| payer\_code | Standardized payment code according to medical treatment received |
| medical\_specialty | Area/field of medicine |
| num\_lab\_procedures | lab procedures |
| num\_procedures | non lab procedures |
| num\_medications | number of medications |
| number\_emergency | admitted as an emergency |
| number\_outpatient | number of times admitted as an outpatient |
| number\_inpatient | number of times admitted as an in patient |
| diag\_1 | diagnosis 1 |
| diag\_2 | diagnosis 2 |
| diag\_3 | diagnosis 3 |
| number\_diagnoses | number of diagnoses done |
| max\_glu\_serum | Glucose serum test result |
| A1Cresult | Hb A1C or hemoglobin A1c (shows sugar level in blood) |
| metformin | one of the administered medicinal drugs |
| repaglinide | one of the administered medicinal drugs |
| nateglinide | one of the administered medicinal drugs |
| chlorpropamide | one of the administered medicinal drugs |
| glimepiride | one of the administered medicinal drugs |
| acetohexamide | one of the administered medicinal drugs |
| glipizide | one of the administered medicinal drugs |
| glyburide | one of the administered medicinal drugs |
| tolbutamide | one of the administered medicinal drugs |
| pioglitazone | one of the administered medicinal drugs |
| rosiglitazone | one of the administered medicinal drugs |
| acarbose | one of the administered medicinal drugs |
| miglitol | one of the administered medicinal drugs |
| troglitazone | one of the administered medicinal drugs |
| tolazamide | one of the administered medicinal drugs |
| examide | one of the administered medicinal drugs |
| citoglipton | one of the administered medicinal drugs |
| insulin | one of the administered medicinal drugs |
| glyburide-metformin | one of the administered medicinal drugs |
| glipizide-metformin | one of the administered medicinal drugs |
| glimepiride-pioglitazone | one of the administered medicinal drugs |
| metformin-rosiglitazone | one of the administered medicinal drugs |
| metformin-pioglitazone | one of the administered medicinal drugs |
| change | Change of medication |
| diabetesMed | Diabetes medications |
| readmitted | Readmission to hospital |

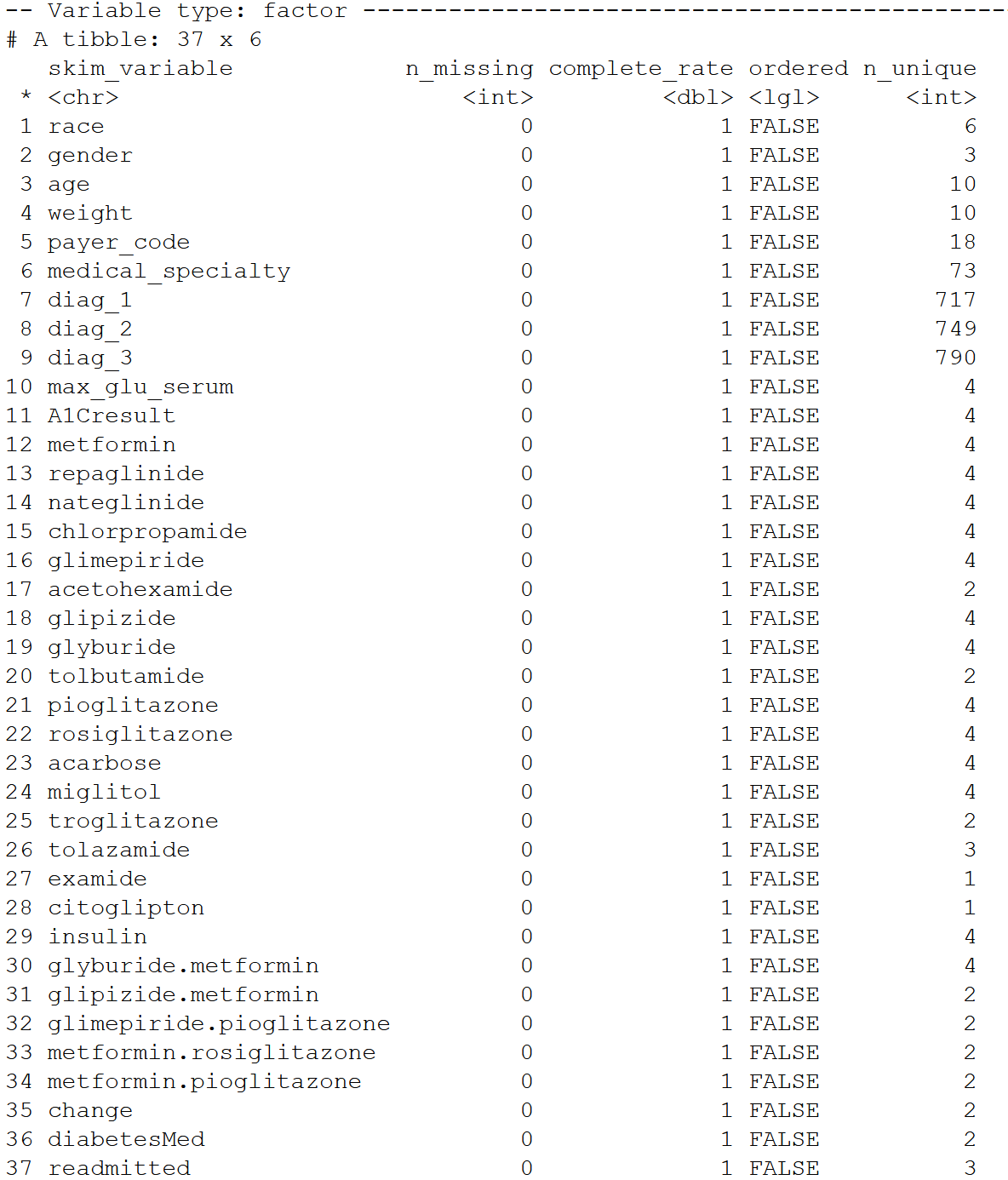
# Data exploration and cleaning

#### Data Dimension (variables and observations)

As mentioned earlier I have over one hundred thousand instances and 50 variables, out of which 13 are numeric (continuous) and 37 are factor variables (non-continuous).







Graphical user interface, application, table, Excel

Description automatically generated

#### preprocessing and challenges

1. Many variables have ‘?’ and ‘Unknown/Invalid’ in them as values, where there are missing or masked values for privacy purposes. So during cleaning process I have replaced them with NA.
2. Some variables which has only NA or monotonous or are not useful for our analysis such as encounter\_id, patient\_nbr, weight, payer\_code, medical speciality (diagnosis variables are similer to this) are dropped.
3. I have used heat maps to identify missings, once I remove all the variables not usefull for our analysis and fix all other variables values I can implement correlation matrix and other visualisations.
   1. Current challenges with cleaning are, I tried to replace all ‘?’ with NAs. But in three variables like race, diag\_1,diag\_2, diag\_3 there are still some ? left.
   2. Studying and understanding all the medicines in the variables is also a huge challenge as I need to understand how these medicines work patients. Once studying and understanding these will help in reduction of not necessary variables thus can implement suggested business questions.



# Dimensionality and data reduction

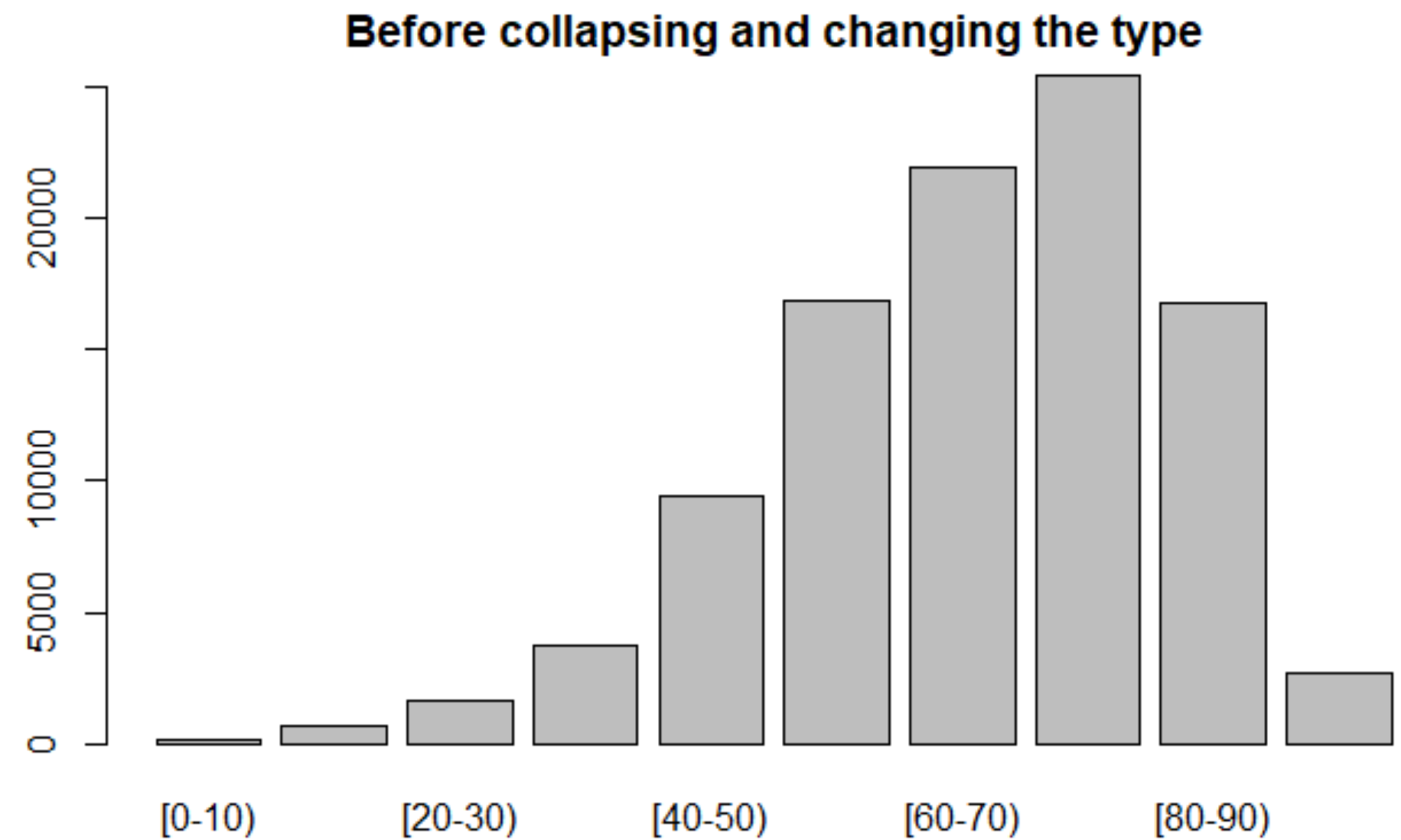
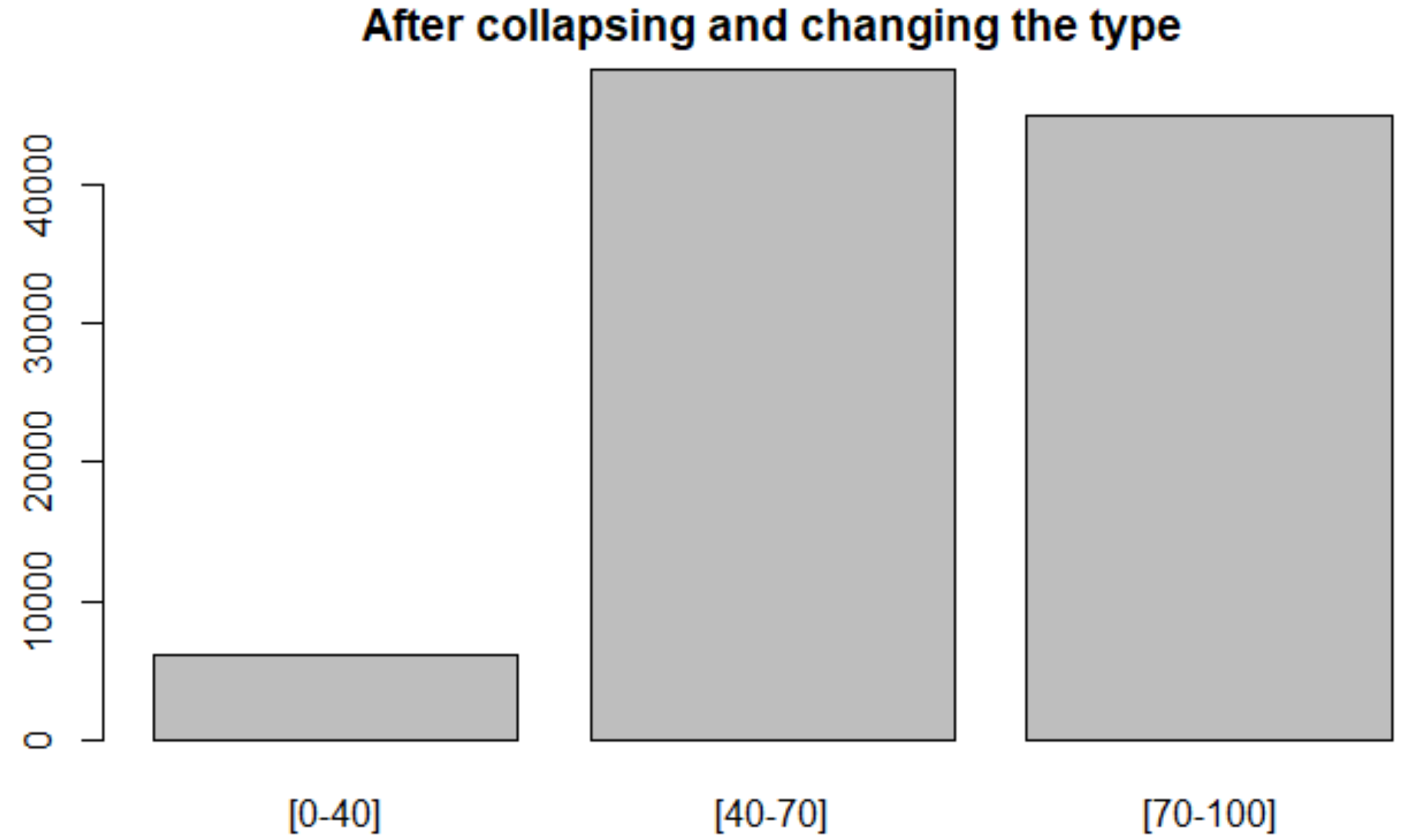
I will be clubbing some of the classes of categorical variables for convenience and better analysis, also will remove some outliers in numerical variables as it will help with accuracy of our model.

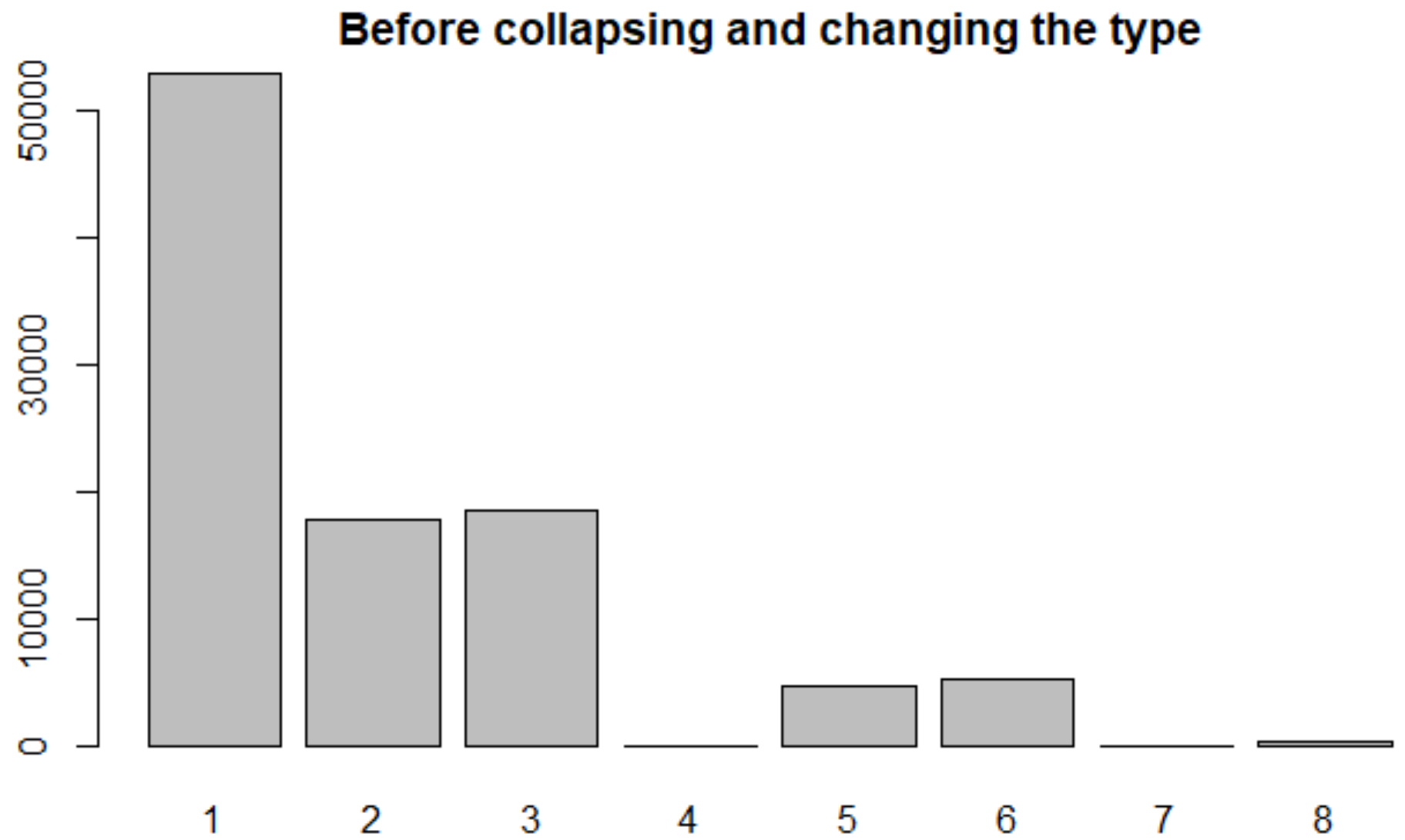
#### Category reduction

* I have taken the variable diag\_1 and plotted the medical codes according to CDC to understand what kind of problems patient is having and renamed it to primary diagnosis

Chart, bar chart

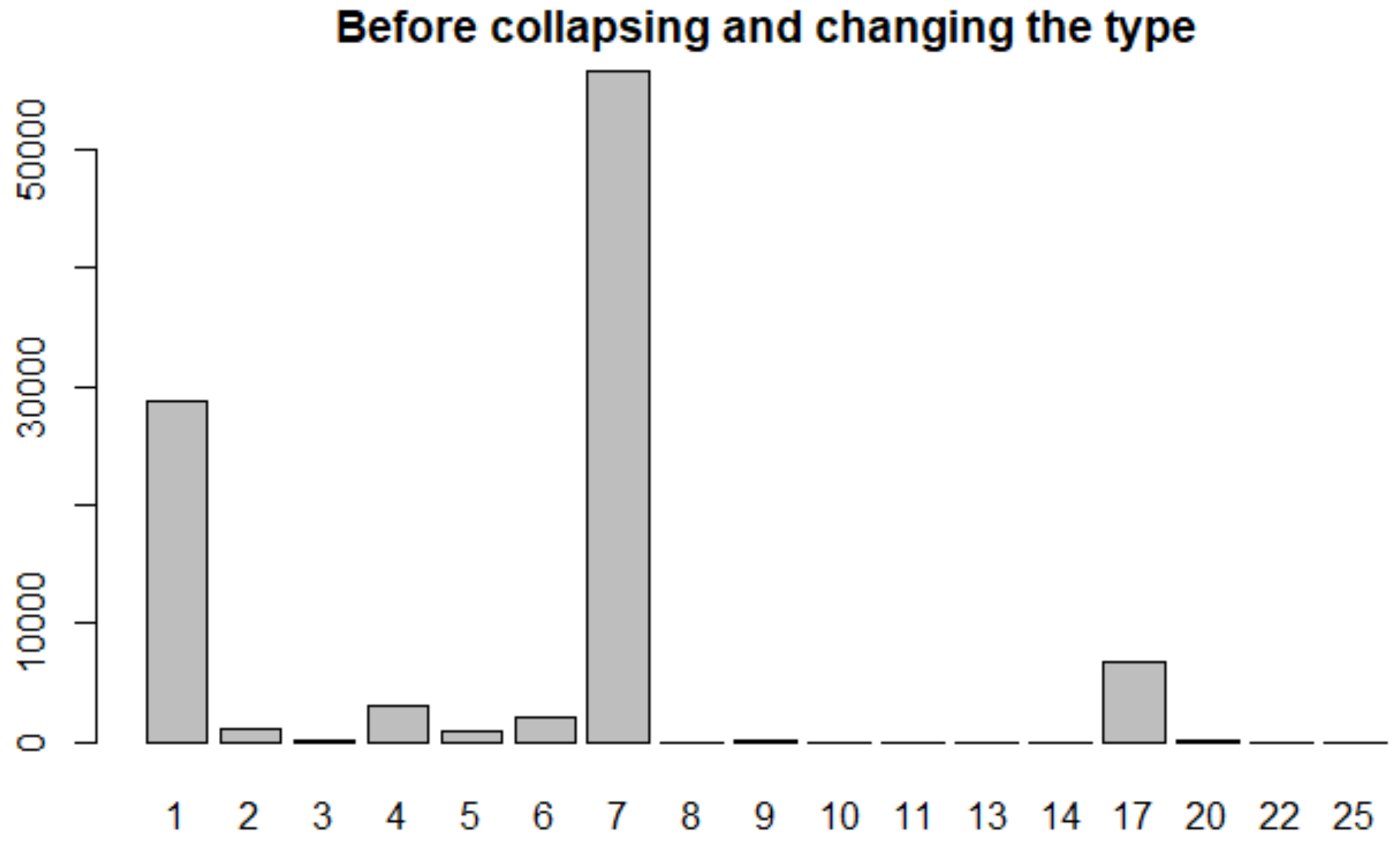
Description automatically generated

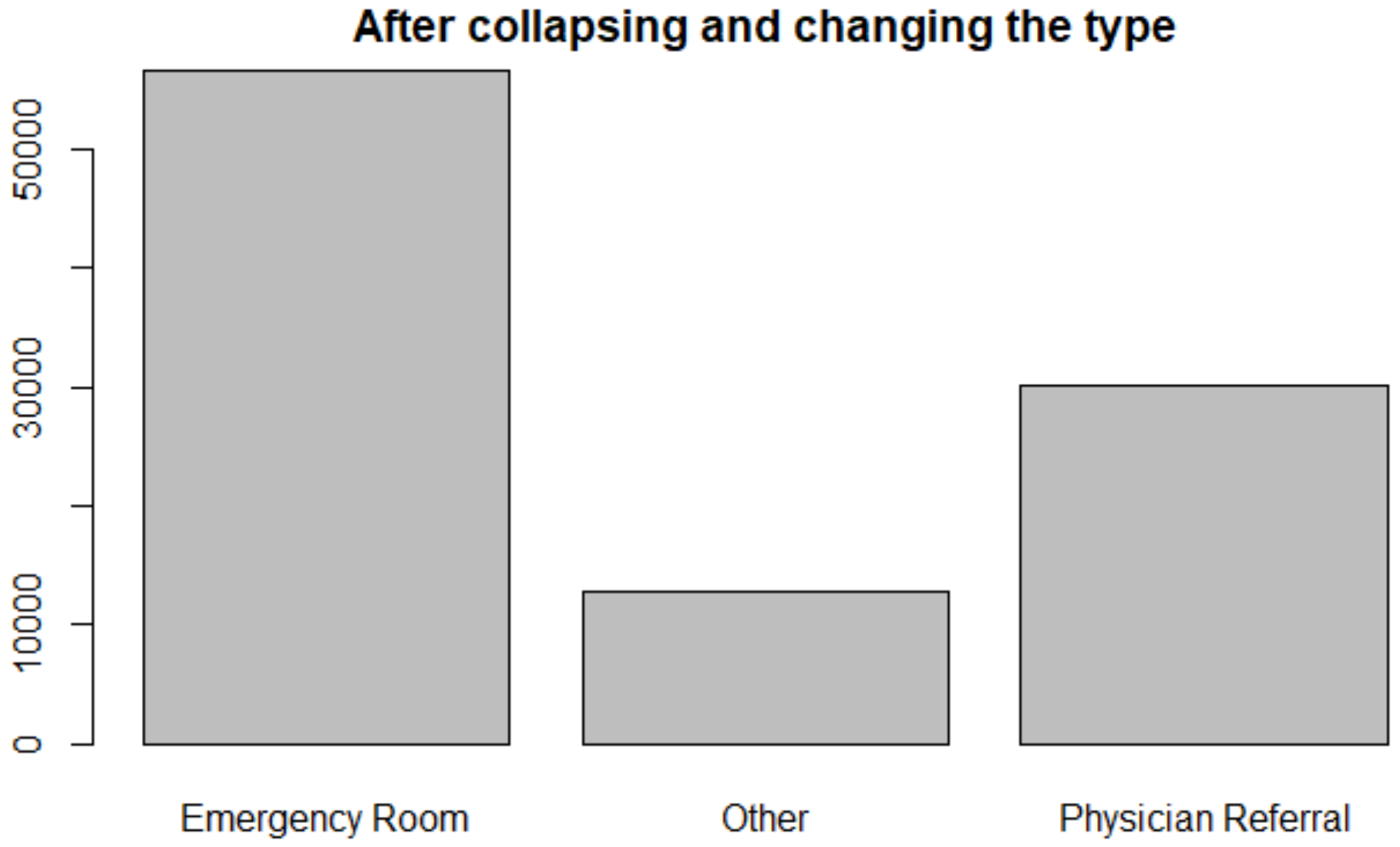
* Now I have dropped variables diag\_1, diag\_2, diag\_3 as we have stored mapped categories of necessary diagnosis classes
* I have taken age variable where I have different classes of age
* And I have collapsed 10 classes to 3 classes as the data is left skewed, the new classes are in the age variable are age group of 0-40, 40-70, 70-100
* In the variable admission type id we have 8 classes with their ID codes which we can refer from data dictionary, based on their similarities and necessesities I have clubbed 8 classes in to 4 and renamed variable and classes accordingly.



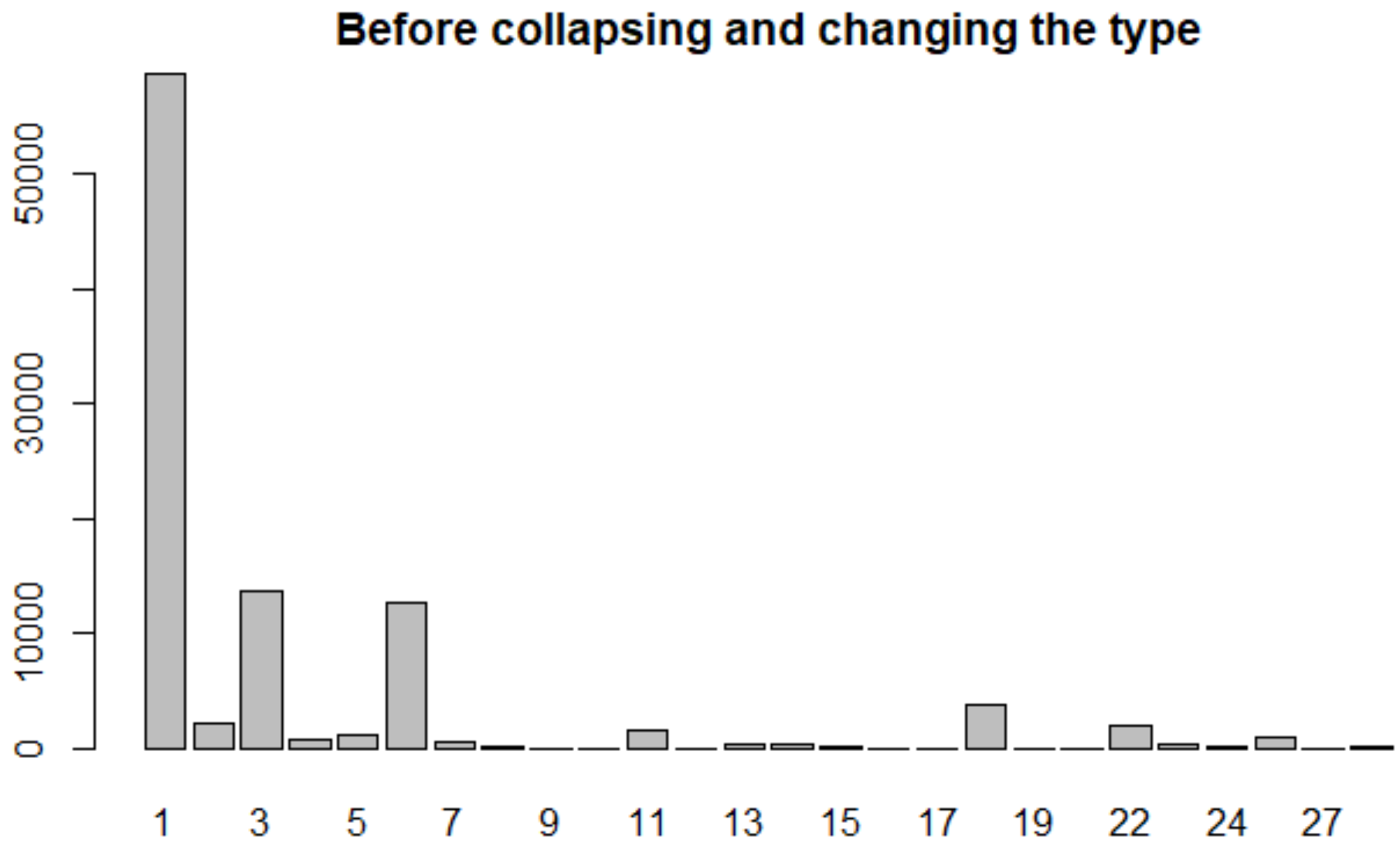


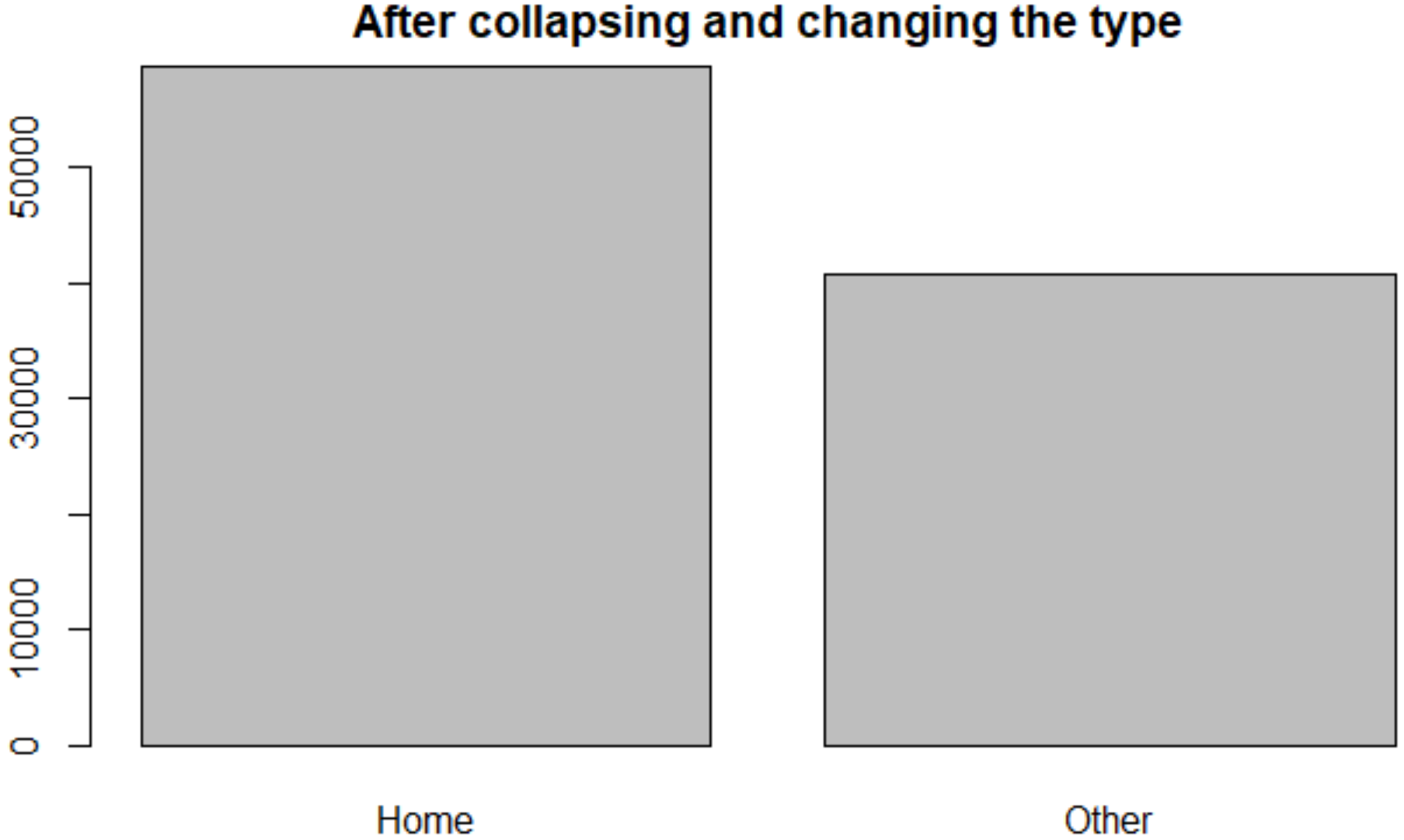
* Like above variable I have referred to data dictionary and collapsed the classes accordingly for admission source ID variable and renamed the classes and the variable to admission source. This variable will tell us from what source the patient has been admitted to the hospital.





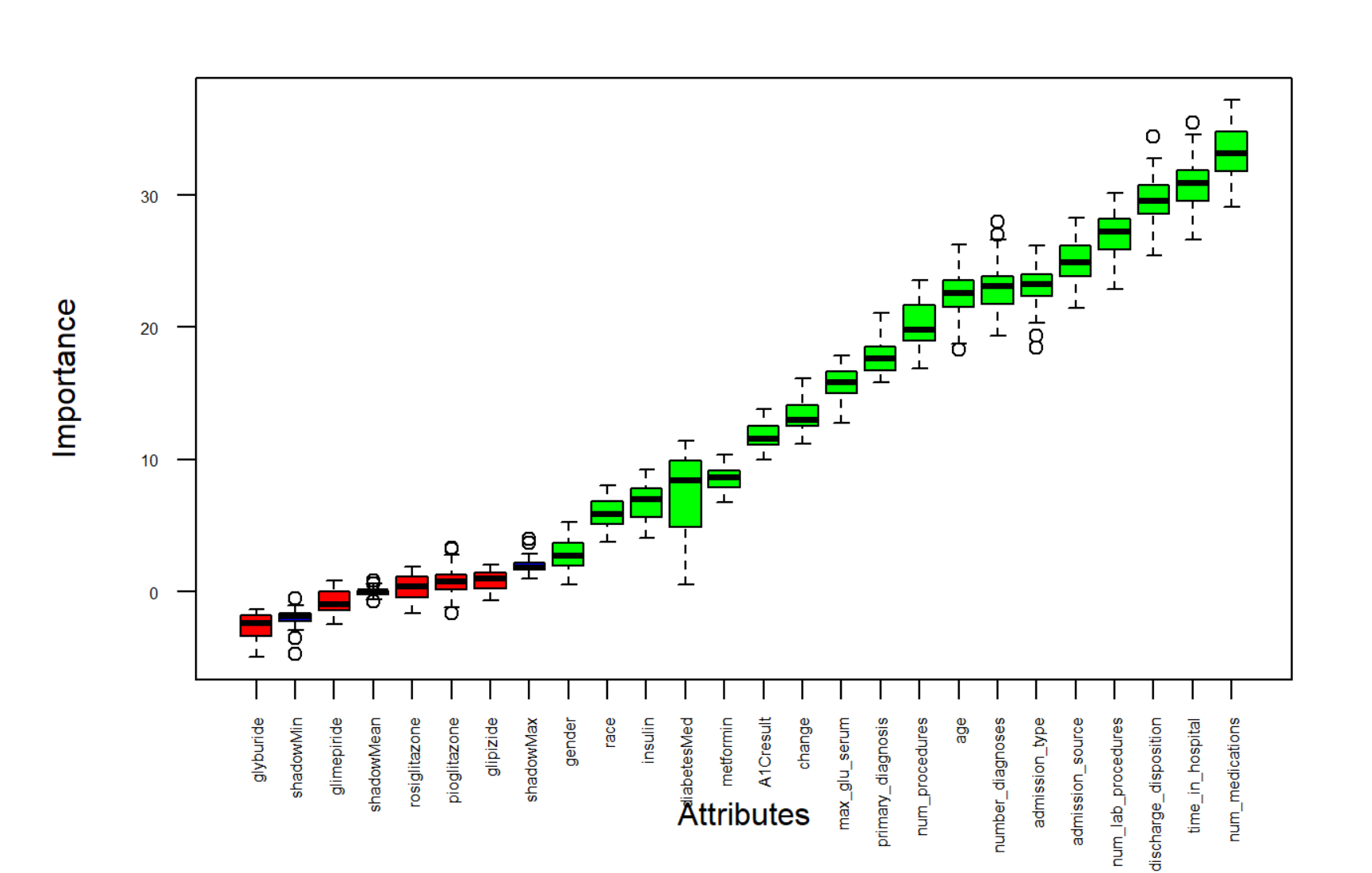
* Discharge disposition id tell us where the patient been to after discharging like home, hospice etc. I have collapsed 28 classes in to 2 and renamed them accordingly and renamed the variable.





#### Data and Dimension reduction

* Variables which has only NA or monotonous or are not useful for our analysis such as encounter\_id, patient\_nbr, weight, payer\_code, medical speciality (diagnosis variables are similer to this) are dropped.
* All the instances with NA values are dropped.
* Now I have used BORUTA function in R, “Boruta is an all relevant feature selection wrapper algorithm, capable of working with any classification method that output variable importance measure (VIM); by default, Boruta uses Random Forest. The method performs a top-down search for relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies, and progressively eliminating irrelevant features to stabilize that test”3. This helped me to get rid of 21 variables most of them are medicines which are used by very few patients compared to our data, so they are removed as they are not significant enough. If we see below Boruta plots box plots for all the variables we are going to involve in the model, blue boxes consists minimum, maximum z scores of shadow features we created, green ones are most significant, red ones are unimportant and yellow color code is given to the variable whose significance Boruta is not able to identify.

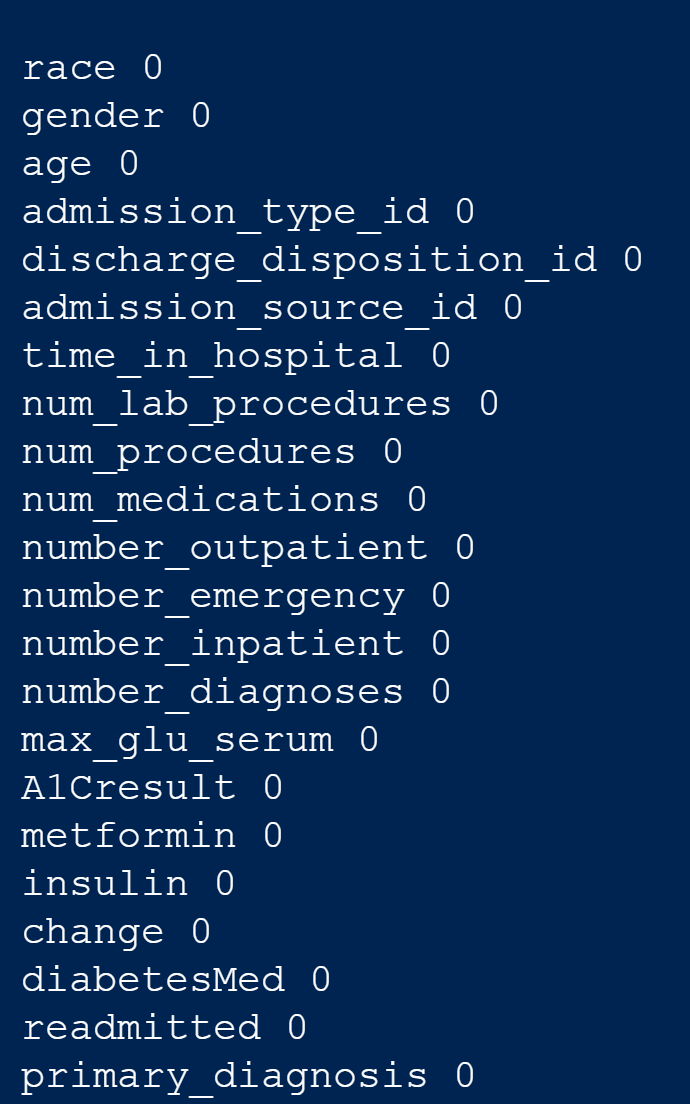


3 https://www.rdocumentation.org/packages/Boruta/versions/7.0.0/topics/Boruta

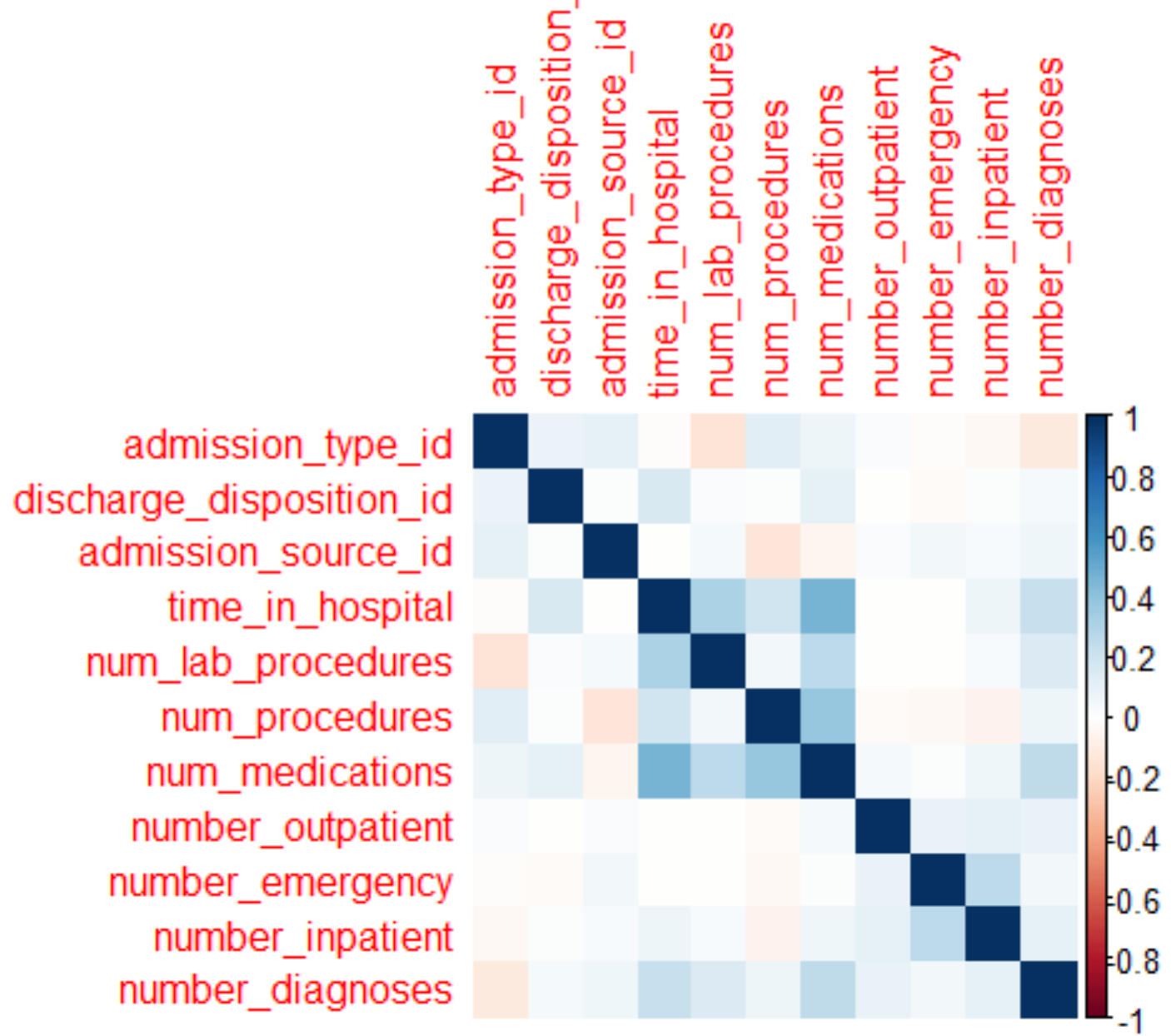
A screenshot of a computer

Description automatically generated

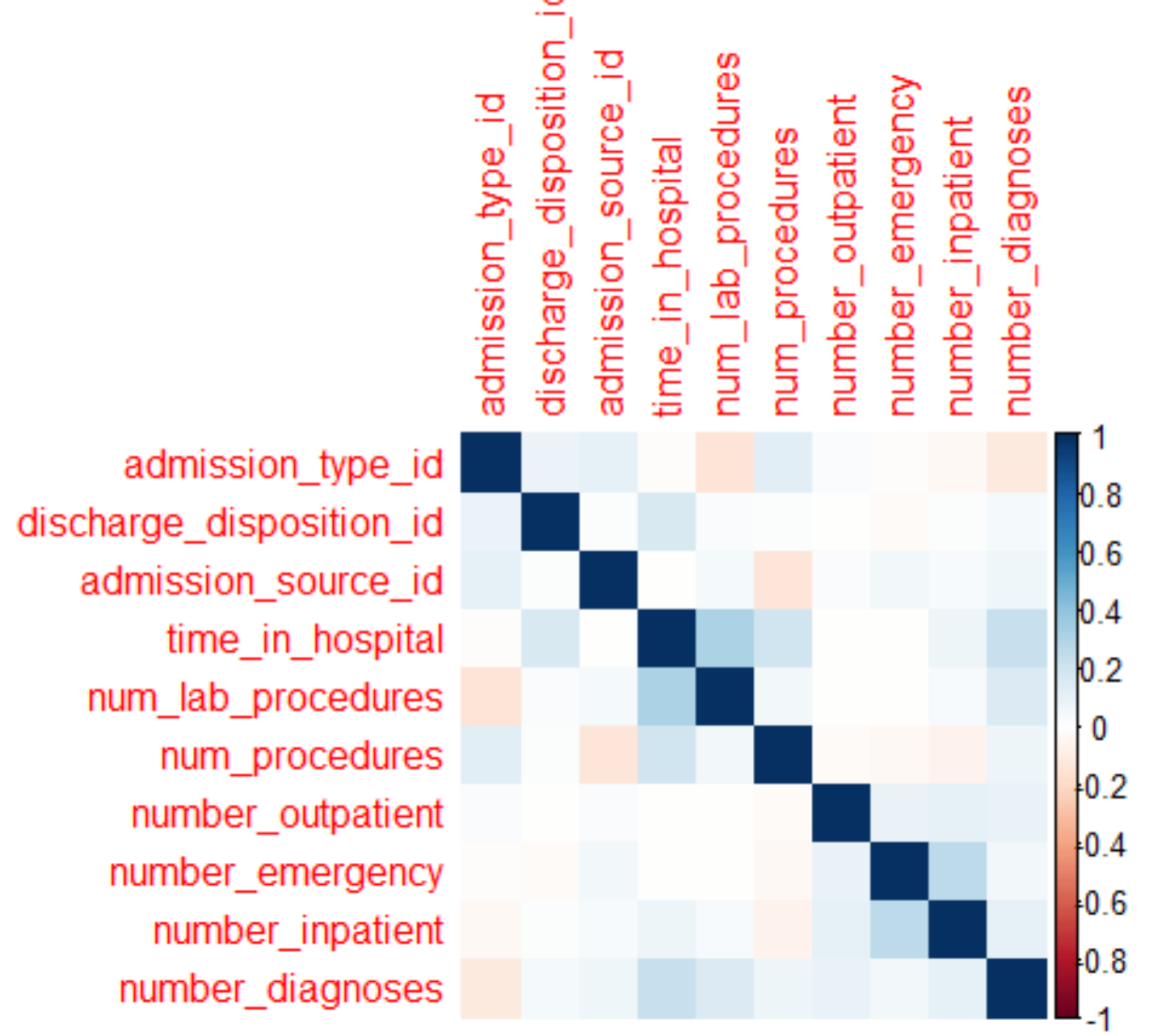
* Now we are left with the following variables



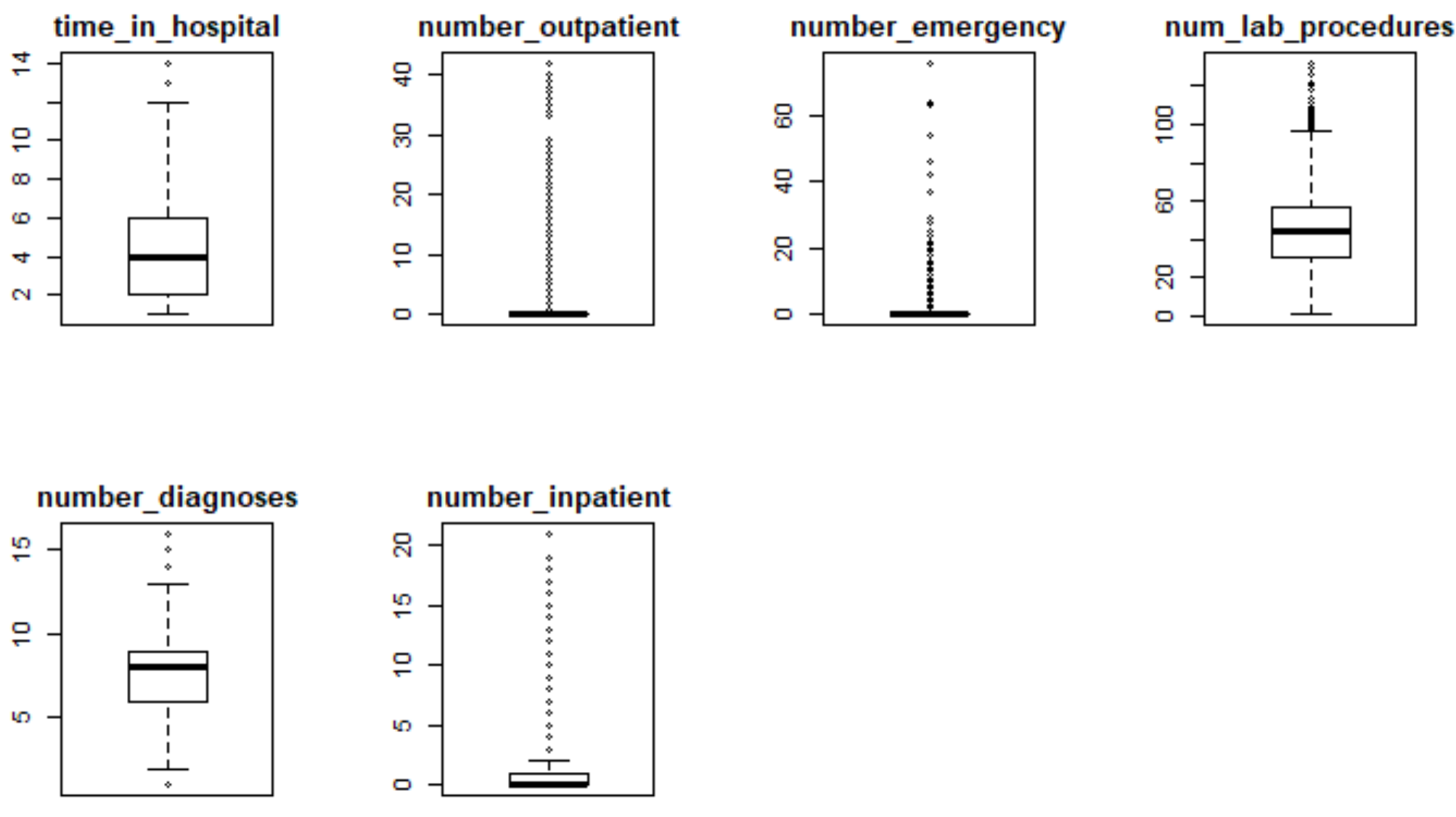
* I have implemented a correlation matrix to remove any variables with higher correlation, we can observe that number of medications and time spend in hospital are correlated which make sense.



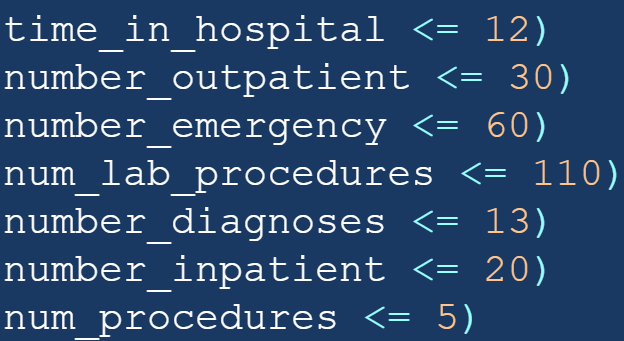
* So I have removed num\_medications variable and we can see the correlation matrix after removal of the variable



* Now in the data we have some outliers in numerical variables



* Initially I have removed all the outliers, after review as suggested by professor, I have now removed only some extreme outliers that could be disruption to data

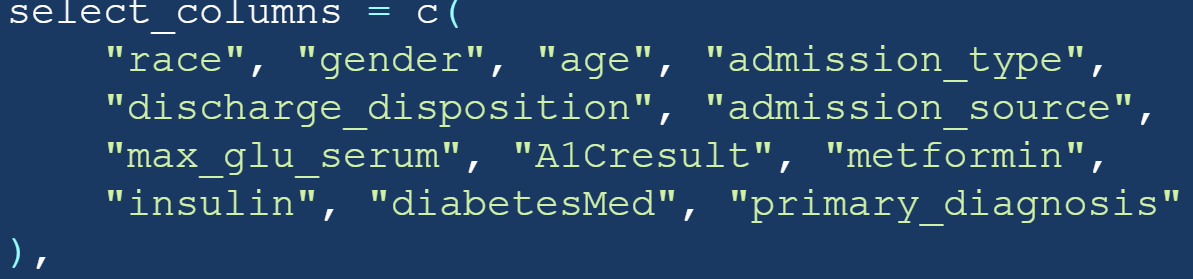




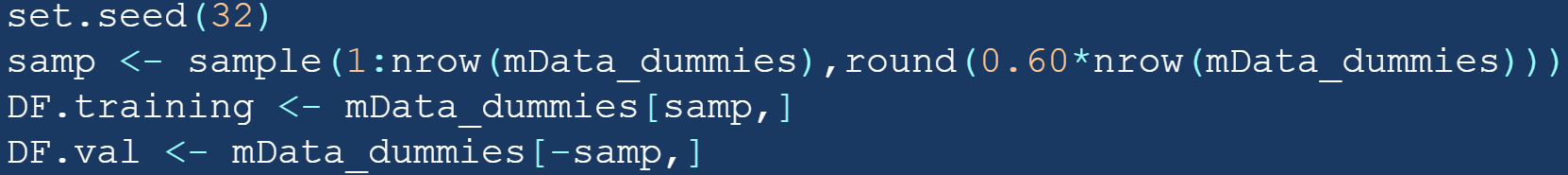
# Data mining task

As we are trying to find whether a patient is being readmitted and what factors could be reason, we will implement our data mining techniques on readmitted variable with other remaining variables as independent variables. Readmitted variables are a factor variable with values 1 and 0, readmitted and not readmitted. As we are trying to implement a model that is going to predict between two classes our job is now classification.

I have created dummies for remaining categorical variables, which got our variable count to 41.



And then I have split the 60 % data in to training and 40% to validation sets.



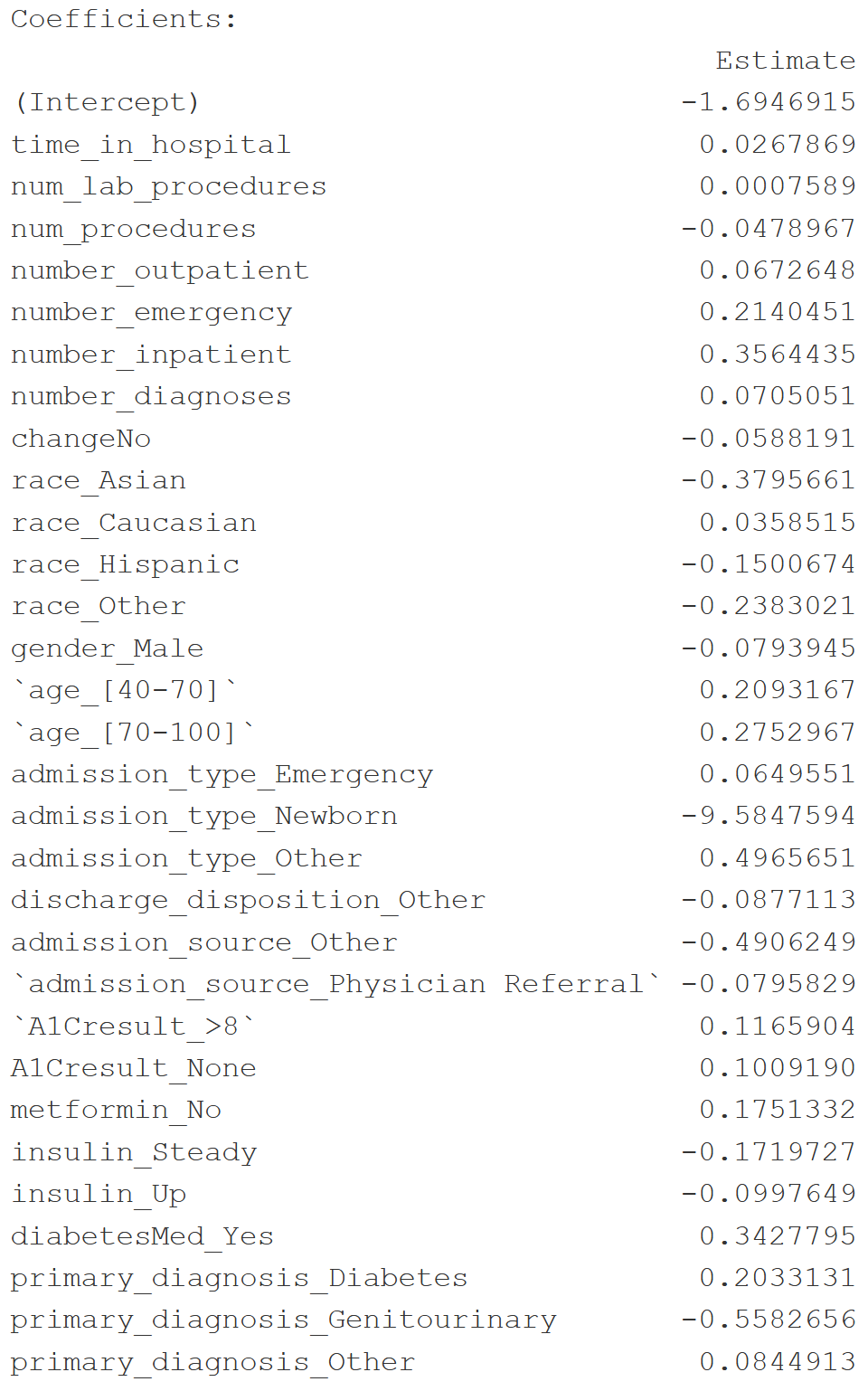
# Data mining techniques

After preprocessing the data and choosing the model, I am now going to run the model using three different techniques to compare the accuracy of the different techniques. And infer better results together.

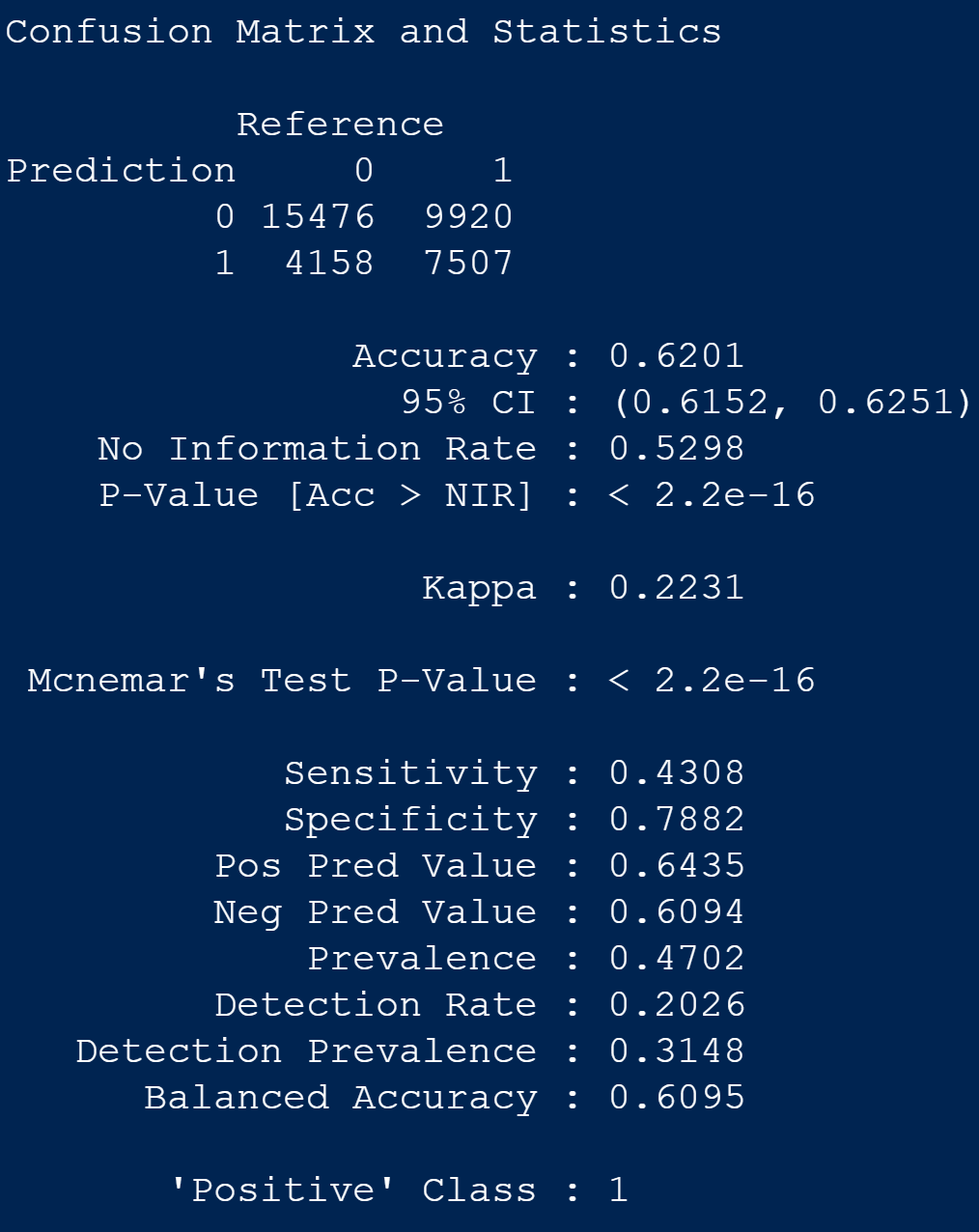
#### Step wise logistic regression

First, I have ran the stepwise logistic regression, where in each step variables are either added or removed to the model. When minimum possible AIC value is achieved our regression stops. We can look at the variables in the final step and their estimates.

The relation between input and output function is approximate better. Makes smoother decisions with well distributed estimations. Variables like higher age, inpatient are highly significant.



* This model suggests that Patients who are diagnosed diabetic, age above 40, being Outpatient, sugar level in hemoglobin more than 8 are more likely to readmitted
* People who are spending more time in hospital, inpatient, taking a greater number of medications, a greater number of procedures are less likely to readmit to hospital again.



#### Classification tree

In classification tree the parent node is readmitted as it is out dependent variable and we can observe all other ate leaves with each of them having a weight.

Diagram

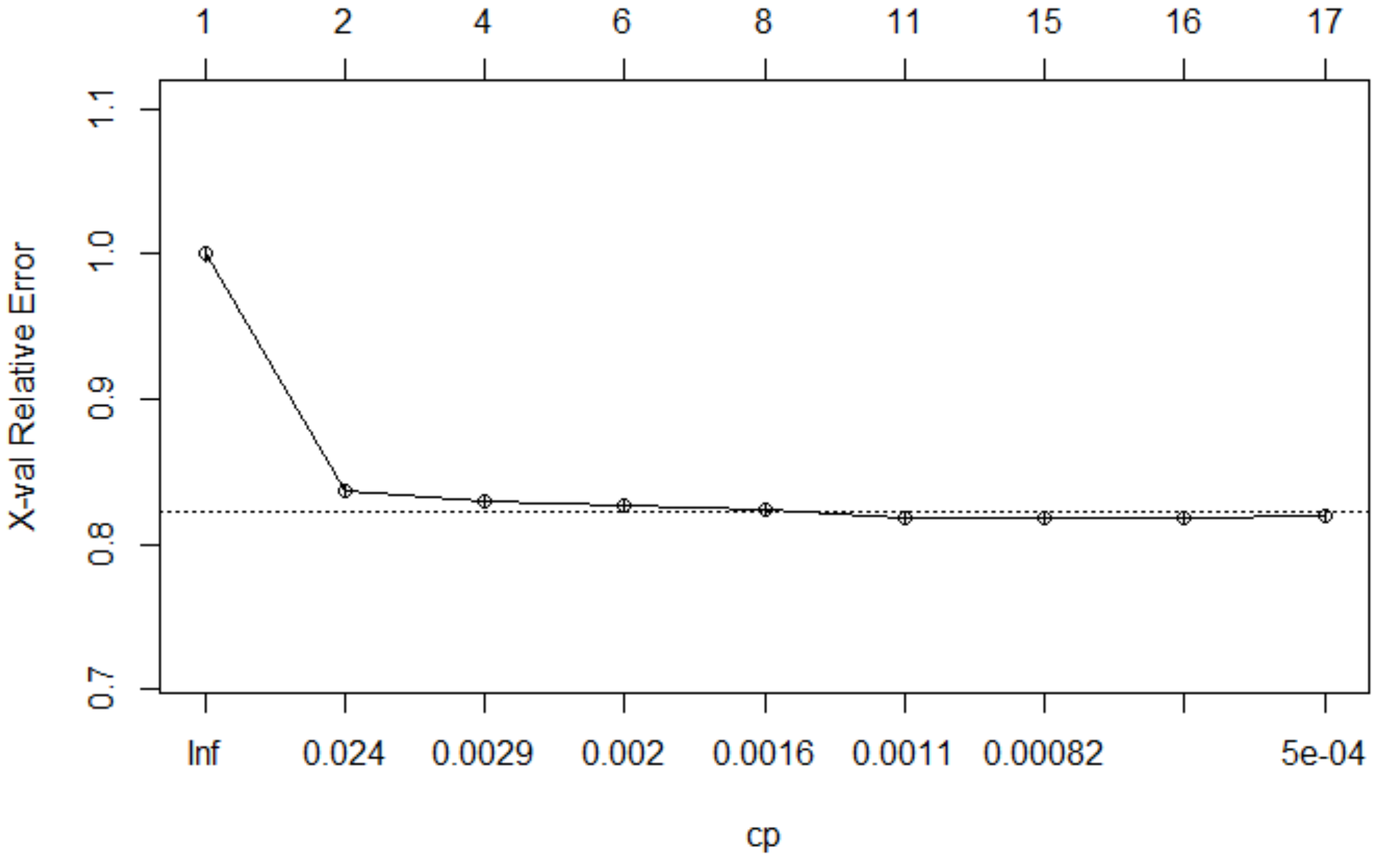
Description automatically generated

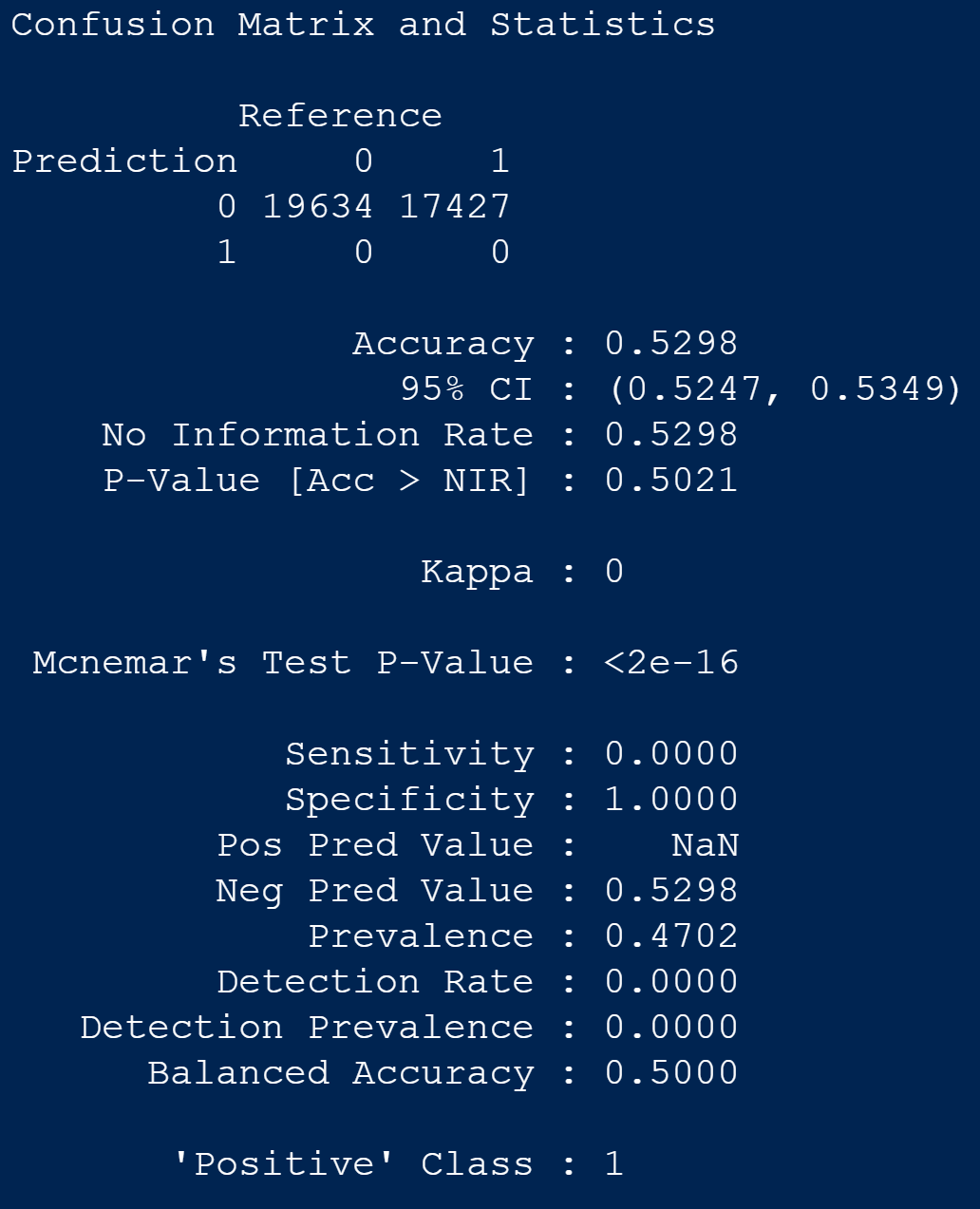
Some variables are undervalued. Some probability estimations are constant. Inpatient, admission source, admission type are significant variables here.

As we are implementing this technique on dummy variables, we can clearly have a look at what kind of patients are being readmitted and who are not. If we look at the left to the tree and inpatient who was diagnosed diabetic and joined hospital in an emergency will likely get readmitted 56% of times.

Or if we look at right side of the tree a person who is an outpatient is likely to get readmitted with chance of 15%. Similarly, we can interpret the tree and understand the who and through what causes are getting readmitted. Our classification tree accuracy is 53% appx.

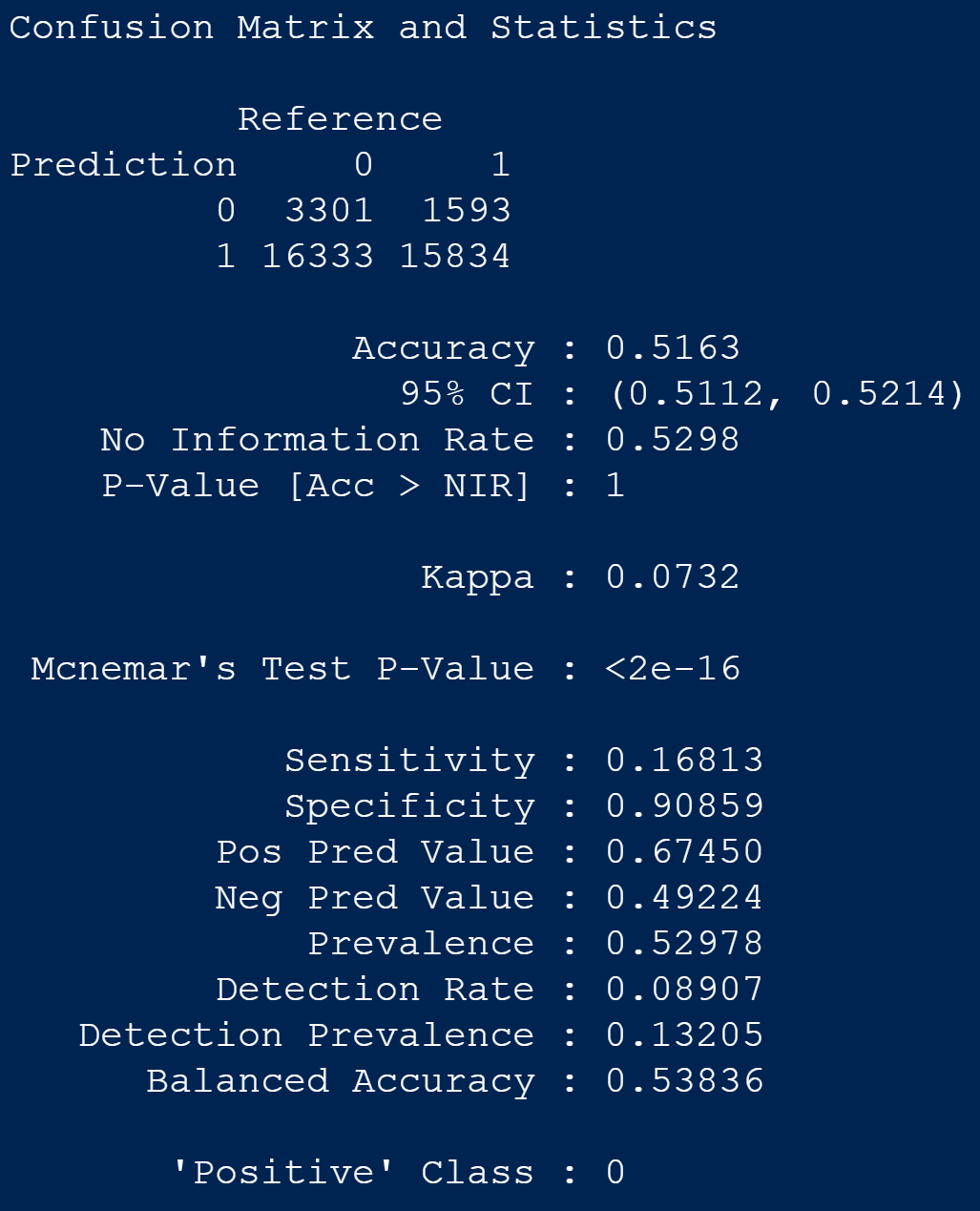
We can get more nodes to understand in the above tree by adjusting the complexity of the tree which will add more variables to tree. As we increase the complexity the error rate decreases as shown below.





#### naïve bayes

Like above techniques I have implemented Naïve Bayes where I have observed the accuracy of the model to be 51.6%.

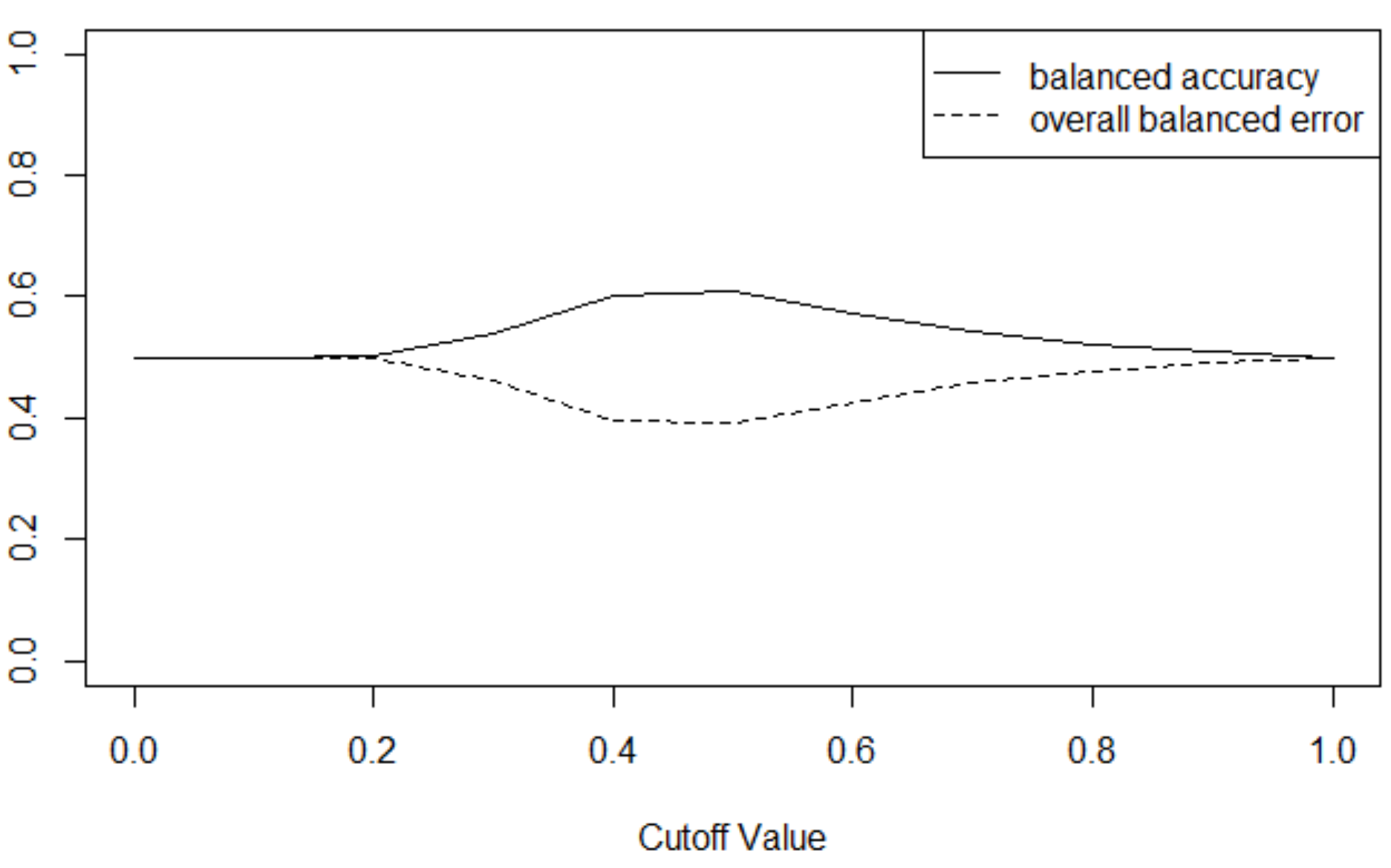


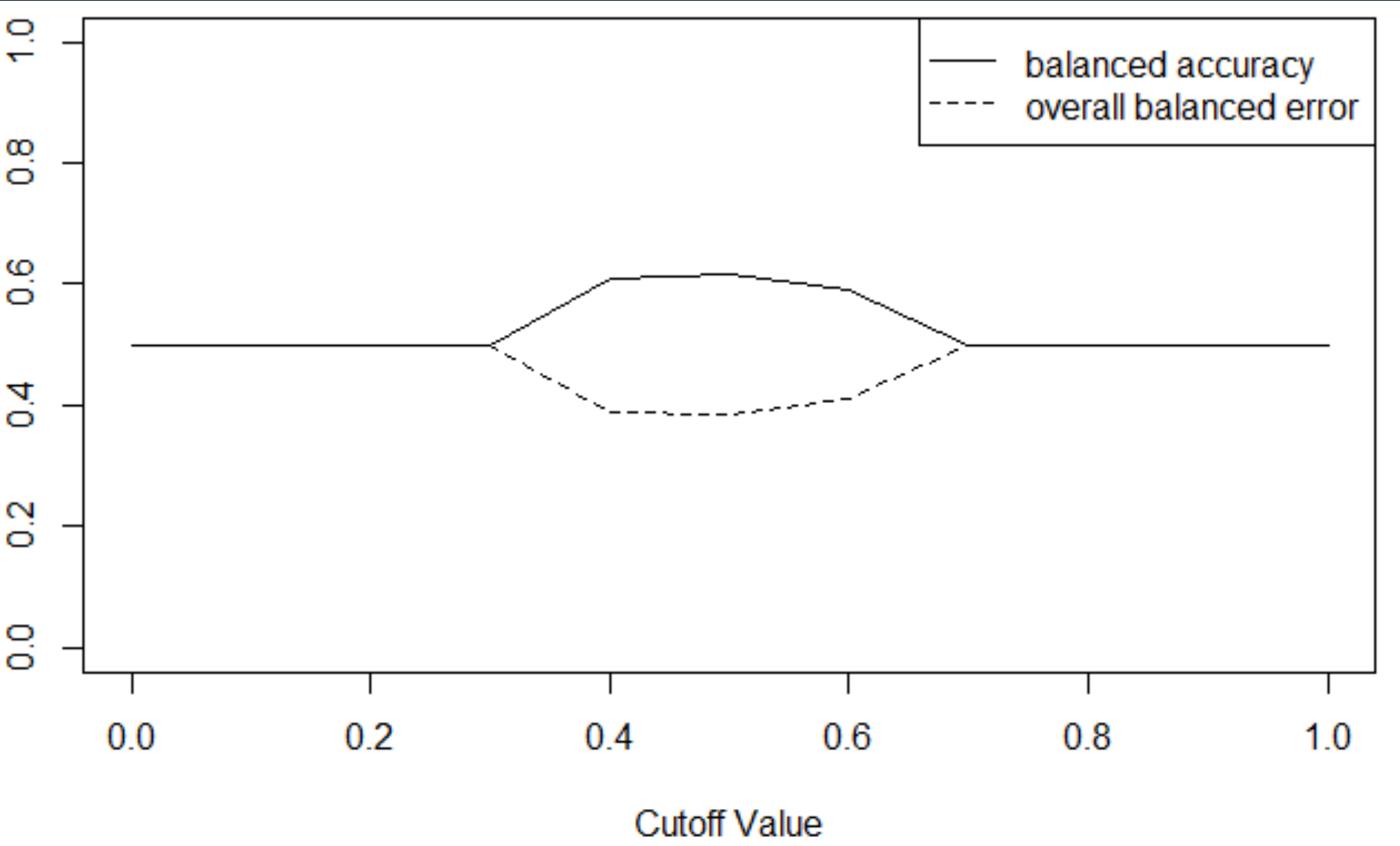
# evaluation and results

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Stepwise Logistic | 0.6201 |
| Classification tree | 0.5298 |
| Naïve Bayes | 0.5163 |

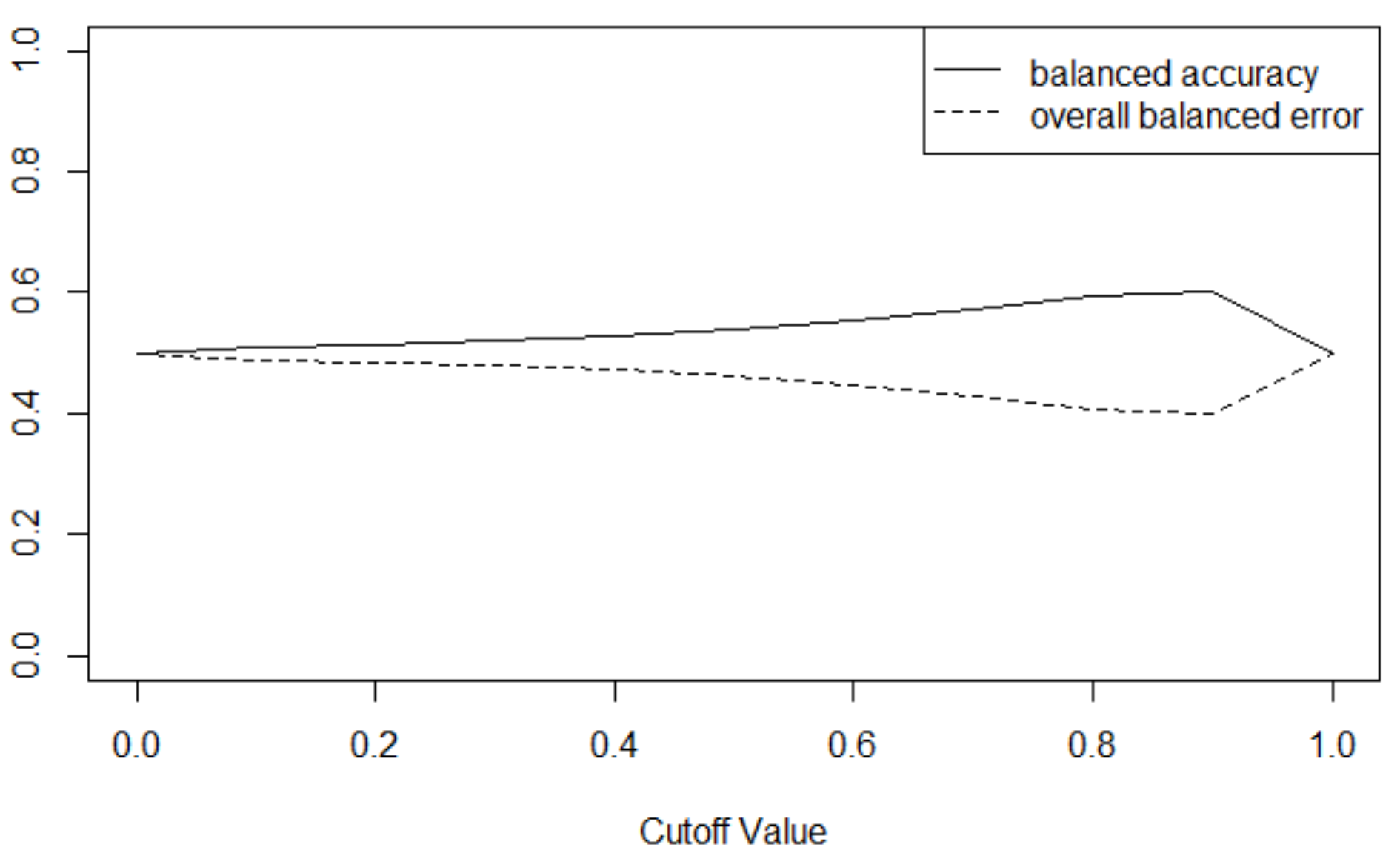
As we have done a classification task, I have considered balanced accuracy as my primary evaluation metric. Normal accuracy is sometimes misleading when the classes are not balanced. So, I am comparing balanced accuracy of all the three techniques.

The below I am comparing the performance of the model with the cutoff value, the step wise logistic regression model performed well with cutoff of 0.5. We can achieve balanced accuracy of above 60%.



And below we have balanced accuracy plot for classification tree, and as same as for logistic regression model we can achieve maximum balanced accuracy near 60% with cutoff value 0.5.

The below plot is balanced accuracy plot for Navies Bayes, and we are getting maximum accuracy as below 60% at cutoff value 0.9.



# Conclusion

My model is not for deployment instead I am making suggestions based on our model which would help a diabetic patient to stay healthy and avoid readmission into a hospital which will save their money.

All my model finds mostly similar patterns in readmitted and non-readmitted patients.

My suggestions are that Patients who are diagnosed diabetic, aged above 40, being Outpatient, sugar level in hemoglobin more than 8,admitted in emergency situation, spent less number of days in hospital for treatment, has low insulin level in body, used less medication *(in different combinations of these factors)* are more likely to get readmitted.

People who are spending more time in hospital, being inpatient, taking a greater number of medications, a greater number of procedures, with age less than 40, *(in different combinations of these factors)* are less likely to readmit to hospital again.

**Thank you**